Behaviour & Information Technology. Vol. 28, No. 2, March-April 2009, 165-181.
Preprint version

Scrutinizing usability evaluation: Does thinking aloud affect behaviour and mental workload?

MORTEN HERTZUM[* a], KRISTIN D. HANSEN[a b], and HANS H. K. ANDERSEN[b c]

[a] Computer Science, Roskilde University, Roskilde, Denmark

[b] Systems Analysis Department, Risø National Laboratory, Roskilde, Denmark

[c] Now at: National Research Centre for the Working Environment, Copenhagen, Denmark

Abstract. Thinking aloud is widely used for usability evaluation. The validity of the method is, however, debatable because it is generally used in a relaxed way that conflicts with the prescriptions of the classic model for obtaining valid verbalizations of thought processes. This study investigates whether participants that think aloud in the classic or relaxed way behave differently compared to performing in silence. Results indicate that whereas classic thinking aloud has little or no effect on behaviour apart from prolonging tasks, relaxed thinking aloud affects behaviour in multiple ways. During relaxed thinking aloud participants took longer to solve tasks, spent a larger part of tasks on general distributed visual behaviour, issued more commands to navigate both within and between the pages of the web sites used in the experiment, and experienced higher mental workload. Implications for usability evaluation are discussed.

Keywords: thinking aloud; verbalization; usability evaluation methods; user testing; validity

1 Introduction

Evaluation is an important and complex element of systems development, and effective and valid evaluation methods are, consequently, in high regard. The thinking-aloud method has become popular in practical usability evaluation as well as in usability research and is by many considered the single most valuable usability evaluation method (Nielsen 1993, Dumas 2003). However, the possible effects of thinking aloud on the behaviour of participants in usability evaluations have only been examined in a few studies (e.g. Van den Haak *et al.* 2003, Krahmer and Ummelen 2004). Instead, claims to validity have been adopted from studies in cognitive psychology, particularly Ericsson and Simon's (1980, 1993) work on verbal reports. Descriptions of the thinking-aloud method for usability evaluation differ, however, in important respects from thinking aloud as defined by Ericsson and Simon (1993), particularly regarding instructions and reminders to think aloud, and these differences are likely exacerbated in practical use of the method (Boren and Ramey 2000, Nørgaard and Hornbæk 2006). Hence, the validity of the thinking-aloud method for usability evaluation is presently debatable.

This study aims to investigate whether thinking aloud causes participants in usability evaluations to behave differently and experience a different level of mental workload compared to performing in silence. To address the variation in – and uncertainty about – what test participants are more specifically asked to do when they are asked to think out loud we distinguish between classic thinking aloud and relaxed thinking aloud. Classic thinking aloud complies with the prescriptions of Ericsson and Simon (1993). Relaxed thinking aloud complies with typical descriptions of thinking aloud in the context of usability evaluation (e.g. Nielsen 1993, Dumas and Redish 1999). We conduct an experiment involving both types of thinking aloud and assess their validity by measures comprising participants' task completion times, eye movements, interaction with the system, mental workload, and the correctness of their task solutions. Some authors have suggested that participants initially complete tasks in silence and then think aloud

retrospectively, often while viewing a video recording of their activities solving the tasks (e.g. Van den Haak *et al.* 2003, 2004, Guan *et al.* 2006). We restrict our study to concurrent thinking aloud; that is, participants think aloud while solving tasks.

Usability evaluation has become widely practiced, not least through the uptake of lightweight, or discount, methods (e.g. Monk *et al.* 1993, Nielsen 1993). These methods promise to require little time, few skills, hardly any facilities, and yet to yield good though not perfect results. The entire approach stands in stark contrast to the rigour of Ericsson and Simon's (1993) description of thinking aloud and to their assessment of the consequences of failing to think aloud in the proper way. In the next section we summarize Ericsson and Simon's (1993) framework for understanding verbalization of thought processes and review previous work on the possible effects of such verbalization on behaviour. Section 3 describes the method of our experiment, and in Section 4 we present the results. In Section 5 we discuss the results and their implications for usability evaluation. In conclusion, Section 6 asserts that whereas classic thinking aloud has little or no effect on behaviour apart from prolonging tasks, relaxed thinking aloud – that is, the type of thinking aloud typically used in usability evaluation – affects behaviour in multiple ways.

2 Thinking aloud

Thinking aloud was introduced as a method for usability evaluation by Lewis (1982) but within psychology the history of the method dates back at least a hundred years (Ericsson and Simon 1993).

2.1 Verbalization of thought processes

Prior to Ericsson and Simon (1980), people's verbal reports of their thought processes while performing a task were generally treated alike irrespective of the type of verbalizations people produced. For example, Nisbett and Wilson (1977: 246) concluded that 'in summary, it would appear that people may have little ability to report accurately on their cognitive processes' although their study concerned *retrospective* verbalizations by people that were often asked to provide *reasons* for their behaviour. Ericsson and Simon (1980, 1993) distinguished three levels of decreasingly valid verbalizations, each characterized by the amount of interference caused by the additional processing involved in producing the verbalizations:

Level 1 verbalization is the vocalization of thoughts and information that are already in a person's present focus of attention in verbal form. No intermediate processes are needed to report these thoughts and people need expend no special effort to communicate them. For example, when people report sequences of numbers while mentally solving math problems they are producing level 1 verbalizations because the reported numbers – that is, the intermediate results of the calculations – are directly available in the form needed to report them.

Level 2 verbalization is the explication of information that is presently in a person's focus of attention but must be recoded into verbal form before it can be reported. The explication or recoding involves additional processing but does not bring new information into the person's focus of attention. For example, images and abstract concepts must be transformed into words before they can be reported but as long this transformation is the only additional processing that is performed such verbalization is at level 2.

Level 3 verbalization introduces mental processing that influences a person's focus of attention in ways beyond those occasioned by task performance. The influence on the person's focus of attention consists of requiring that present thoughts and information attended to at the moment are linked to earlier thoughts and information attended to previously. People, for example, produce level 3 verbalizations when they are asked to provide explanations of their thoughts and behaviour or to retrieve information from memory.

According to Ericsson and Simon (1993) verbalization at levels 1 and 2 are valid data about task performance, whereas level 3 verbalizations are not. This restricts valid verbalization to the currently heeded information; that is, the contents of short-term memory. Ericsson and Simon (1993: 222-223) explicitly state that their conception of thought as the heeded information in short-term memory captures only part of what others have termed thought and that it is, for example, narrower than James' (1890) concept of the stream of thought. James (1890) contends that the focus of a person's thoughts is at every instance influenced by a fringe of dimly perceived relations and objects. The fringe is an integral part of the stream of thought and determines people's affinity, feelings, or mood toward their present focus of attention, but the fringe is by definition not heeded information.

In this study we use the terms thinking aloud and verbalization interchangeably. This is contrary to Ericsson and Simon (1993), who reserve the term thinking aloud for level 2 verbalization, but in accordance with how 'thinking aloud' is used in usability evaluation. Our focus in this study is on thinking aloud at all three levels of verbalization.

2.2 Instructions for thinking aloud during usability evaluations

Instructions for thinking aloud during usability evaluations provide ample evidence of an interest in the user's thought processes that goes beyond the presently heeded information and also includes explanations and preferences. Though instructions vary, there is a widespread lack of differentiation between verbalization at different levels. For example, Dumas and Redish (1999: 281) provide the following suggestions for prompts to remind people to think out loud:

- 'Mary, could you tell us what you are thinking now?'
- 'John, could you tell us why you pressed the enter key?'
- 'Abigail, what are you looking for in the index?'
- 'Jim, we couldn't hear what you said just then.'

These suggestions for prompts follow immediately after quoting Ericsson and Simon's (1993) instructions for learning participants to think aloud at levels 1 and 2. Yet, the suggestions, which are representative of many instructions for thinking aloud during usability evaluation (e.g. Monk *et al.* 1993, Nielsen 1993), request explanations of behaviour, elicit retrospection, and thus go beyond the heeded information in short-term memory. Boren and Ramey (2000) note that the suggestions are also long compared to the recommended 'keep talking', establish personal contact (e.g. between 'Mary' and 'us'), redirect the participant's attention (e.g. to 'the enter key'), and require a pause in the task flow. The suggested prompts may indicate that the objectives of usability evaluation call for a less restrictive concept of thought to capture the issues of relevance to the usability practitioner. However, the prompts are inconsistent with the dominant theoretical model of verbalization and may thus mislead the practitioner by yielding invalid data.

2.3 Effects of thinking aloud

Ericsson and Simon (1993) review several hundred studies and conclude that the findings are consistent with their three-level model of verbalization. A more cautious conclusion is made by Russo *et al.* (1989) who find that the influence of verbalization on task performance depends strongly on the nature of the task.

Independently of verbalization, data about what information is heeded can be obtained from recordings of eye fixations (e.g. Just and Carpenter 1980, Rayner 1998). This has been exploited by Winikoff (1967, see Ericsson and Simon 1993) and Rhenius and Deffner (1990) to investigate the validity of verbal reports at levels 1 and 2. In both studies eye fixations and verbalizations correspond very well. Rhenius and Deffner (1990) report that 87-98% of the verbalizations of an object could be mapped to fixations. They also report longer task completion times for tasks performed while thinking aloud compared to tasks performed in silence, consistent with the requirement for additional processing to produce level 2 verbalizations. There were no differences in the number of correctly solved tasks. An analysis of task-solution strategies was less conclusive in that there were substantial mean differences in preferred strategy but due to large variances only very few of these differences were significant. However, for thinking aloud the preferred strategy differed significantly between early and late instances of a task, indicating that thinking aloud led to delayed discovery and adoption of the optimal strategy.

With respect to level 2 verbalization, a series of studies has found that verbally describing a non-verbal stimulus such as a face or shape can impair subsequent attempts at identification of the stimulus (Schooler and Engstler-Schooler 1990, Meissner and Brigham 2001). This phenomenon, known as verbal overshadowing, was originally encountered in studies of eye-witness reports. After watching a video of a robbery, participants that spent five minutes writing a detailed description of the robber's face were less successful at identifying the robber from a photo array than control-group participants (Schooler and Engstler-Schooler 1990). Similar disruptive effects of verbalization have also been reported for other activities that rely on non-verbal thinking, for example visual problem solving (DeShon *et al.* 1995). It

appears that verbalization produces a processing shift that gives preference to verbal thinking and, temporarily, dampens people's capacity for non-verbal thinking (Schooler 2002). Thus, verbal overshadowing challenges the assumption that level 2 verbalization can be produced without causing changes to people's thought processes.

With respect to level 3 verbalization, it has repeatedly been found that retrospective verbalization yields substantial forgetting and fabrication of information (Nisbett and Wilson 1977, Russo *et al.* 1989). In addition, a number of studies have examined how concurrent verbal reflection affects people's task performance and found that such level 3 verbalization improves understanding. Chi *et al.* (1989) found that people who spontaneously produced verbal explanations to themselves of a text while reading it were subsequently more successful at answering questions about their understanding of the text. Subsequently, Chi *et al.* (1994) extended this finding by demonstrating that improved understanding could be achieved by instructing people to self-explain as they read a text. Furthermore, the positive effect of these requested self-explanations was more pronounced for difficult tasks than for simple tasks. It appears that generating explanations to oneself is both a constructive activity in which people formulate candidate understandings of the topic and an integration activity in which these candidate understandings are confronted with and incorporated into existing knowledge. However, for tasks intended to elicit opinions, rather than generate understanding, the result of encouraging people to reflect on the reasons for their opinions may be a reduction in the quality of their preferences and decisions (Wilson and Schooler 1991). Consistent with Ericsson and Simon (1993), these studies of level 3 verbalization find that requesting people to verbalize reasons and reflections influences their task performance.

2.4 Studies of thinking aloud during usability evaluations

Below we summarize the main findings of four studies that have investigated thinking aloud in the context of usability evaluation. The studies, which span a decade of research on usability evaluation, all compare thinking aloud with performing in silence, and two of the studies also investigate how effective thinking aloud is at making evaluators aware of usability problems. Three of the four studies concern level 3 verbalization.

Held and Biers (1992) had participants with and without computer experience learn and use a text-processing system. Participants either performed in silence or were prompted for explanations whenever they seemed to be experiencing difficulties; that is, level 3 verbalization. For the experienced participants it was found that verbalization led to a less favourable subjective impression of the system and to less confidence in being able to use it, compared to the non-verbalizing participants. For the novice participants there were no differences between the verbalizing and non-verbalizing conditions.

Wright and Converse (1992) had participants solve file-management tasks either in silence or while providing explanations for their actions; that is, level 3 verbalization. If participants in the verbalization condition were silent for more than 30 seconds or issued a command without giving an explanation they were prompted for their thoughts and reasons. Participants in the verbalization condition committed fewer errors and consumed less task time than participants in the silent condition. No differences were found for perceived ease of use and mental workload, as measured by the task load index (TLX; Hart and Staveland 1988).

Van den Haak *et al.* (2003) had users solve tasks with an online library catalogue and verbalize their thoughts either concurrently or retrospectively. In the concurrent condition users were instructed to verbalize at levels 1 and 2. Concurrent thinking aloud had a negative effect on task performance in that users failed to correctly complete more tasks while verbalizing than while performing in silence and verbalizing retrospectively. While the concurrent and retrospective conditions led to the detection of comparable sets of usability problems, there seemed to be an interesting difference in the evidence that made evaluators aware of the problems. In retrospective thinking aloud most problems were detected by means of users' verbalizations, while in concurrent thinking aloud most problems were detected by means of evaluators' observations of user behaviour. This finding, confirmed in a subsequent study (Van den Haak *et al.* 2004), suggests that concurrent verbalization may add little value to usability evaluation.

Lesaigle and Biers (2000) had users solve tasks with a library-search system while verbalizing their actions, expectations about the system's response, and problems using it; that is, level 3 verbalization. Subsequently, evaluators performed a usability evaluation based on video recordings of either (1) the user's screen with no audio, (2) the user's screen with verbalizations, or (3) the user's screen with

verbalizations combined with a video feed of the user's face. Evaluators in the three viewing conditions detected the same number of problems. However, evaluators' severity assessments differed significantly as a function of the viewing condition. Increasing the amount of information available to evaluators made them rate problems as increasingly severe.

3 Experimental method

To empirically investigate whether and how thinking aloud at either levels 1 and 2 or level 3 influences people's behaviour we conducted an experiment about people's behaviour and mental workload when they were thinking out loud while performing tasks and when they performed in silence.

3.1 Participants

Eight participants (three female, five male) took part in the experiment. Participants' age ranged from 23 to 33 years with an average of 28.5 years. All participants were experienced computer users and indicated that they used computers daily. Participants knew of the television-channel web sites used for half of the experimental tasks but indicated that they had had no preconceived knowledge about how to solve the tasks. No participant had prior knowledge of the bookstore web sites used for the other half of the tasks. Furthermore, none of the participants had knowledge or experience of verbalization at levels 1 and 2, but three had previously tried thinking aloud in an informal manner (i.e. involving verbalization at levels 1 to 3). Finally, all participants had normal or corrected-to-normal vision and none used hard contact lenses or multi-focal glasses.

3.2 Apparatus

Participants' eye movements were recorded with a head-mounted eye tracker from SMI, sampling at 50 Hz. Acceptable calibration required the participants' eye gaze to be within 2.5 mm of the five calibration points.

3.3 Thinking-aloud conditions

The experiment consisted of two sessions, each involving two thinking-aloud conditions. The two conditions in the first session were:

Classic thinking aloud, in which participants performed the tasks while thinking out loud and the experimenter, when needed, reminded participants to 'keep talking'. This condition corresponds to how thinking aloud is defined by Ericsson and Simon (1993) as consisting of verbalization at levels 1 and 2.

Silent, in which participants performed the tasks without verbalizing their thoughts. Participants were simply instructed to solve the tasks and report their answer to the experimenter upon completion. This condition is similar to how people work when they are not enrolled in usability evaluations or other tests.

The two thinking-aloud conditions in the second session were:

Relaxed thinking aloud, in which participants performed the tasks while thinking out loud and the experimenter intervened with questions asking participants for explanations and comments. This condition includes level 3 verbalization and corresponds to how thinking aloud is commonly employed in the context of usability evaluation.

Silent, the same as in the first session.

3.4 Tasks

The experimental tasks involved looking for information on four web sites – two web sites for Danish television channels and two for online bookstores. These web sites were selected as examples of information-rich, state-of-the-art web sites intended for broad audiences. Each task was paired with another, near-identical task (see Appendix A for a list of the full set of tasks). The two tasks in a pair were performed on the web sites for either the two television channels or the two online bookstores. That is, the tasks in a pair were performed on similar but different web sites, thus reducing any learning effects. To further even out effects of learning, the order in which participants solved the tasks in a pair was

counterbalanced across participants, see Section 3.6. Separate pairs of tasks were used for the first and second session.

We used one of the dimensions suggested by Spool *et al.* (1999) to divide the tasks into two types:

Fact, in which participants gathered information that was explicitly available on the web sites. Fact tasks concerned book titles, television listings, and today's weather. For example, 'Which city has the highest temperature today – Copenhagen or Aarhus?'

Assessment, in which participants gathered information and based on this information formed an opinion. The assessment tasks asked participants to assess their interest in books and the importance of news stories. For example, 'What is the biggest domestic news story on the front page?'

3.5 Procedure

Upon arriving at the lab, participants were introduced to the experiment and asked questions about their background. Then, participants were instructed about how to think aloud at levels 1 and 2 and practiced thinking aloud on four training tasks: (1) What is the result of multiplying 11 × 11? (2) Think of a friend. How many windows are there in your friend's house or flat? (3) Name 20 animals. (4) Take the pen on the table. Take it apart and put it back together, while thinking aloud. The thinking-aloud instructions were copied from Ericsson and Simon (1993: 377-379) and the three first training tasks were near identical to their training tasks. The last training task was added to provide participants with additional practice in verbalizing at levels 1 and 2. After practicing thinking aloud, participants were introduced to the task load index (TLX; Hart and Staveland 1988) and explained the definitions of its six rating scales. The preparations for the experimental tasks were completed by setting up and calibrating the eye tracker so that it accurately captured the participant's line of gaze. To maintain high accuracy, the eye tracker was periodically recalibrated during the two experimental sessions.

In the first session participants performed tasks in the classic thinking aloud and silent conditions. Tasks were available in a printed booklet, and the experimenter kept silent except when participants stopped talking for more than about 30 seconds during thinking-aloud tasks. When this happened the experimenter reminded participants to 'keep talking'. After completing the first session participants were allowed a break before they commenced on the second session. In the second session participants performed tasks in the relaxed thinking aloud and silent conditions. Tasks were again available in a printed booklet, but during thinking-aloud tasks the experimenter asked questions as participants progressed through the tasks. Examples of the questions include: 'What are you thinking about?', 'What are you trying to achieve?', and 'Did you find this easy or difficult?' After the second session participants were debriefed and invited to comment on the experiment. The experiment lasted about three hours per participant.

3.6 Design

Both sessions of the experiment employed a within-subjects design with the factors thinking-aloud condition (two levels that differed between sessions) and task type (fact, assessment). In *the first session* the thinking-aloud conditions were classic and silent. The session was divided into two blocks. Half of the participants performed tasks while thinking out loud during the first block and performed silently during the second block. The other half of the participants performed silently first, then performed tasks while thinking aloud. In the first block a participant solved four tasks, each task drawn from a different pair of tasks. The order of the four pairs of tasks was the same for all participants and mixed fact and assessment tasks; the task selected from each pair was counterbalanced. In the second block the participant went through the same sequence of pairs of tasks and for each pair solved the task not yet tried. In *the second session* the thinking-aloud conditions were relaxed and silent. Otherwise the design of the second session was identical to that of the first. We did not counterbalance the order of classic thinking aloud (first session) and relaxed thinking aloud (second session) because we were not making a direct comparison between classic and relaxed thinking aloud. Across the two sessions each of the eight participants performed $2 \times 2 \times 4 = 16$ tasks.

3.7 Dependent measures

Participants' behaviour was measured by their solution correctness, task completion times, eye movements, hand movements, and assessments of mental workload.

Correctness of task solutions was determined for fact tasks only and measured by comparing participants' answers with the correct answer. The tasks were rated as either correctly solved or incorrectly solved.

Task completion time was measured from participants had read the task and first saw the related web site to participants announced their answer to the task.

Eye movements were recorded by the eye tracker. Fixations were identified using a dispersion-based algorithm with a minimum fixation duration of 100 ms and a deviation threshold of 0.5 degrees of visual angle. These parameter settings correspond to typical values reported by Salvucci and Goldberg (2000), who also report that the dispersion-based algorithm has very good accuracy and robustness. At a viewing distance of 60-65 cm the deviation threshold is equivalent to a fixation area with a diameter of about 11 mm on the screen participants used for solving the tasks.

In addition to fixations and saccades, which were determined by the dispersion-based algorithm, we coded the eye-movement data manually with respect to focused and distributed visual behaviour. This coding was made on the basis of video recordings of the sessions with a superimposed crosshair indicating participants' point of gaze. *Focused visual behaviour* was defined as the behaviour occurring whenever participants examined a restricted area of the screen; that is, when their gaze dwelled on screen elements in close vicinity of each other. We distinguished three types of mutually exclusive focused visual behaviour. (1) *General*: participants visually examined screen elements of different types within a restricted area of the screen. The screen elements might, for example, be a combination of pictures, text, and input fields. (2) *Text*: participants read text. This visual behaviour was easily identifiable by a series of horizontal movements of the eyes from left to right. (3) *Illustrations*: participants' gaze dwelled on a picture, graph, or other illustration.

In contrast to focused visual behaviour, we defined *distributed visual behaviour* as the behaviour occurring whenever participants searched or scanned across the screen; that is, when their gaze jumped among screen elements that were not in close vicinity of each other. We distinguished three types of mutually exclusive distributed visual behaviour. (1) *General*: participants' gaze jumped among screen elements distributed over a large area of the screen. Further, the screen elements were of different types, such as text in combination with illustrations and their captions. (2) *Text*: participants jumped from text block to text block without reading the text or they scanned a list of textual headings, typically links. Scanning of headings is on the border between distributed and focused visual behaviour but has in this study been considered a distributed visual behaviour. (3) *Illustrations*: participants' gaze jumped successively from one illustration to another without dwelling on any of them.

The coding of focused and distributed visual behaviour was performed by the second author. To assess the reliability of the coding 15 of the 128 tasks were coded twice with at least a week between the first and second coding. The average Kappa value of the agreement between the two codings was 0.80 (range: 0.66-0.93). According to Landis and Koch (1977) this represents 'substantial' agreement.

Hand movements were coded manually on the basis of the video recordings of the sessions. We distinguished three types of hand behaviour. (1) *Mouse clicks*: participants clicked a menu item, link, or another screen element that took participants to another web page. That is, there was a one-to-one correspondence between mouse clicks and page shifts. (2) *Scrolling instances*: participants scrolled the current web page up or down. (3) *Writing instances*: participants entered text, typically search queries, by typing on the keyboard. To assess the reliability of the hand-movement coding 15 of the 128 tasks were coded twice with at least a week between the first and second coding, both codings performed by the second author. The agreement between the two codings showed an average Kappa value of 0.85 (range: 0.64-1.00), corresponding to 'almost perfect' agreement (Landis and Koch 1977).

Perceived mental workload was measured by TLX (Hart and Staveland 1988). TLX consists of six subscales (mental demand, physical demand, temporal demand, effort, performance, and frustration), which are rated on a scale from low (0) to high (100) in increments of five, except for performance where the anchors are good (0) and bad (100). Participants rated the six TLX subscales immediately after completing each task. We left out the weighting procedure for combining the six TLX rating scales into a

single measure of mental workload and, instead, report participants' answers to the six subscales. This is done to increase the diagnostic information acquired from the workload measurements and because the weighting procedure has been discouraged (Nygren 1991, Hendy *et al.* 1993).

4 Results

Below we use repeated-measures analyses of variance (ANOVA) to analyse the obtained data. For one task no timing and behavioural data were recorded due to problems with the video equipment, leaving 127 tasks. All 128 tasks are included in the analysis of participants' ratings of mental workload.

4.1 Correctness of task solutions

The correctness of participants' solutions of the tasks was determined for fact tasks only.

For the session comparing classic thinking aloud with performing in silence, a total of four solutions of fact tasks were incorrect, $N = 31$. One of 15 solutions was incorrect in the thinking-aloud condition and 3 of 16 solutions were incorrect in the silent condition. We found no significant difference between conditions, $F(1, 7) = 1.00$, $p = 0.4$.

For the session comparing relaxed thinking aloud with performing in silence, only one fact task was solved incorrectly, $N = 32$. All 16 solutions were correct in the thinking-aloud condition and 15 of 16 solutions were correct in the silent condition. Again, we found no significant difference between conditions, $F(1, 7) = 1.00$, $p = 0.4$.

4.2 Task completion time

The criterion for determining when a fact task had been completed was straightforward. For assessment tasks, participants individually determined when they had gathered the information necessary for them to make the assessment. In addition to analysing task completion times for all tasks combined we, therefore, also analyse task completion times for fact and assessment tasks separately.

For the session comparing classic thinking aloud with performing in silence we found a significant difference between conditions, $F(1, 7) = 6.80$, $p < 0.05$, indicating that tasks took longer when participants were thinking aloud than when they performed in silence. Looking at fact and assessment tasks separately, we found no difference between conditions for fact tasks, $F(1, 7) = 1.96$, $p = 0.2$, but a significant difference for assessment tasks, $F(1, 7) = 6.48$, $p < 0.05$, see Table 1. We also found a significant effect of task type on task completion time, $F(1, 7) = 68.26$, $p < 0.001$, indicating that fact tasks took less time than assessment tasks.

For the session comparing relaxed thinking aloud with performing in silence we found a significant difference between conditions, $F(1, 7) = 59.28$, $p < 0.001$, indicating that tasks took longer when participants were thinking aloud, compared to performing in silence. Looking at fact and assessment tasks separately, we found a marginally significant difference between conditions for fact tasks, $F(1, 7) = 4.26$, $p = 0.08$, and a significant difference for assessment tasks, $F(1, 7) = 24.60$, $p < 0.01$, see Table 2. There was also a marginal effect of task type on task completion time, $F(1, 7) = 4.90$, $p = 0.06$, suggesting that fact tasks may have taken less time to complete than assessment tasks.

4.3 Eye movements: fixations and saccades

For both sessions, the number of fixations in a task was strongly correlated with task completion time (session 1: $r = 0.98$, $p < 0.001$; session 2: $r = 0.99$, $p < 0.001$). We, therefore, report the rate of fixation (i.e. the number of fixations per second) rather than the number of fixations. This way, the reported eye-movement measures are independent of task completion time.

Table 3 shows the results for the session comparing classic thinking aloud with performing in silence. We found no significant differences between conditions for fixation rate, $F(1, 7) = 3.10$, $p = 0.1$, fixation duration, $F(1, 7) = 3.26$, $p = 0.1$, saccade duration, $F(1, 7) = 1.82$, $p = 0.2$, and saccade length, $F(1, 7) = 2.19$, $p = 0.2$. We also calculated how far into a task participants had performed half of the fixations performed during that task. Participants spent a significantly larger part of their tasks performing the first half of their fixations while thinking aloud compared to performing in silence, $F(1, 7) = 15.45$, $p < 0.01$.

There were also several significant effects of task type. Assessment tasks had lower fixation rate, shorter fixation duration, and longer saccade duration than fact tasks, Fs(1, 7) = 19.99, 5.99, and 18.43, respectively (all ps < 0.05). We found no interactions between thinking-aloud condition and task type.

Table 4 shows the results for the session comparing relaxed thinking aloud with performing in silence. We found no significant differences between conditions (half of fixations: $F(1, 7) = 0.14$, $p = 0.7$; fixation rate: $F(1, 7) = 3.41$, $p = 0.1$; fixation duration: $F(1, 7) = 0.01$, $p = 0.9$; saccade duration: $F(1, 7) = 2.57$, $p = 0.2$; saccade length: $F(1, 7) = 0.91$, $p = 0.4$). For the part of tasks containing the first half of the fixations there was, however, a significant interaction between thinking-aloud condition and task type, $F(1, 7) = 11.25$, $p < 0.05$. That is, the first half of participants' fixations in the relaxed condition were performed during a smaller part of the tasks for assessment tasks compared to solving tasks in silence (relaxed: $M = 51\%$, silent: $M = 53\%$) but during a larger part of the tasks for fact tasks (relaxed: $M = 53\%$, silent: $M = 50\%$). For the four other eye-movement measures we found no interactions between thinking-aloud condition and task type. With respect to task type, assessment tasks had higher fixation rate and shorter saccade duration than fact tasks, Fs(1, 7) = 43.12 and 17.97, respectively (both ps < 0.01).

Looking at fact and assessment tasks separately, we found two significant differences for the assessment tasks: The duration of fixations was shorter during classic thinking aloud (M = 266ms, SD = 28) compared to performing in silence (M = 284ms, SD = 33), and the duration of saccades was shorter during relaxed thinking aloud (M = 135ms, SD = 61) compared to performing in silence (M = 138ms, SD = 54), Fs(1, 7) = 7.02 and 6.23, respectively (both ps < 0.05).

4.4 Eye movements: focused and distributed visual behaviour

Focused and distributed visual behaviour were manually scored for text, illustrations, and in general, see Section 3.7. To avoid that the differences in task completion times confound the analysis of participants' visual behaviour, we measure the extent of the different visual behaviours as percentages of task completion time.

Table 5 shows the results for the session comparing classic thinking aloud with performing in silence. We found no significant differences between conditions for the measures of focused visual behaviour (general: $F(1, 7) = 3.39$, $p = 0.1$; text: $F(1, 7) = 0.12$, $p = 0.7$; illustrations: $F(1, 7) = 0.07$, $p = 0.8$), nor for the measures of distributed visual behaviour (general: $F(1, 7) = 2.79$, $p = 0.1$; text: $F(1, 7) = 1.16$, $p = 0.3$; illustrations: $F(1, 7) = 0.03$, $p = 0.9$). There were, however, significant interactions between thinking-aloud condition and task type for general focused visual behaviour, $F(1, 7) = 10.11$, $p < 0.05$, and general distributed visual behaviour, $F(1, 7) = 8.11$, $p < 0.05$. That is, the part of tasks spent on general focused visual behaviour in the classic condition was larger for fact tasks compared to solving tasks in silence (classic: $M = 12\%$, silent: $M = 6\%$) but identical for assessment tasks (classic: $M = 25\%$, silent: $M = 25\%$). And, the part of tasks spent on general distributed visual behaviour in the classic condition was larger for fact tasks compared to solving tasks in silence (classic: $M = 15\%$, silent: $M = 10\%$) but virtually identical for assessment tasks (classic: $M = 8\%$, silent: $M = 9\%$).

Table 6 shows the results for the session comparing relaxed thinking aloud with performing in silence. We found a significant difference between conditions for general distributed visual behaviour, $F(1, 7) = 14.44$, $p < 0.01$, indicating that participants spent a larger part of their tasks on general distributed visual behaviour during thinking aloud compared to performing in silence. There were no significant differences between conditions for the other measures (focused general: $F(1, 7) = 0.95$, $p = 0.4$; focused text: $F(1, 7) = 2.46$, $p = 0.2$; focused illustrations: $F(1, 7) = 0.61$, $p = 0.5$; distributed text: $F(1, 7) = 0.13$, $p = 0.7$; distributed illustrations: $F(1, 7) = 0.35$, $p = 0.6$). Also, there were no interactions between thinking-aloud condition and task type.

4.5 Hand movements

Hand movements were scored manually with respect to three measures of participants' interaction with the system, see Section 3.7. We measure the hand movements as instances per task. This way, the analysis will reveal whether the differences in task completion times merely indicate that participants were slowed down or whether they behaved differently.

Table 7 shows the results for the session comparing classic thinking aloud with performing in silence. We found marginally significant differences between conditions for mouse clicks, $F(1, 7) = 5.21$, $p =$

0.06, and writing instances, $F(1, 7) = 5.02$, $p = 0.06$, suggesting that participants may make more mouse clicks (corresponding to page shifts) and writing instances when thinking aloud compared to when they perform in silence. There was no significant difference between thinking aloud and performing in silence for scrolling instances, $F(1, 7) = 2.07$, $p = 0.2$. Also, we found no interactions between thinking-aloud condition and task type.

Table 8 shows the results for the session comparing relaxed thinking aloud with performing in silence. We found significant differences between conditions for mouse clicks, $F(1, 7) = 8.66$, $p < 0.05$, and scrolling instances, $F(1, 7) = 12.24$, $p < 0.01$, indicating that participants made more mouse clicks and scrolling instances when thinking aloud compared to when they performed in silence. The effect size was particularly large for scrolling instances, a 143% increase. There was no significant difference between thinking aloud and performing in silence for writing instances, $F(1, 7) = 0.65$, $p = 0.4$. Also, we found no interactions between thinking-aloud condition and task type.

4.6 Mental workload

For the session comparing classic thinking aloud with performing in silence, an overall multivariate analysis of the six TLX subscales showed a marginally significant difference between conditions, $F(1, 7) = 4.69$, $p = 0.07$. Table 9 summarizes the results of individual analyses of variance for the TLX subscales. We found a significant difference between conditions for mental demand, $F(1, 7) = 10.91$, $p < 0.05$, and a marginally significant difference for physical demand, $F(1, 7) = 5.44$, $p = 0.05$, both indicating that classic thinking aloud was perceived as more demanding than performing in silence. We found no significant differences between the classic and silent conditions for temporal demand, effort, performance, and frustration, Fs(1, 7) = 0.74, 2.21, 0.04, and 2.03, respectively (all ps > .1). Also, none of the interactions between thinking-aloud condition and task type were significant.

For the session comparing relaxed thinking aloud with performing in silence, an overall multivariate analysis of the six TLX subscales showed a significant difference between conditions, $F(1, 7) = 15.13$, $p < 0.01$. Table 10 summarizes the results of individual analyses of variance for the TLX subscales. We found significant differences between conditions for five of the six subscales (mental demand: $F(1, 7) = 19.55$, $p < 0.01$; temporal demand: $F(1, 7) = 8.13$, $p < 0.05$; effort: $F(1, 7) = 7.39$, $p < 0.05$; performance: $F(1, 7) = 11.42$, $p < 0.05$; frustration: $F(1, 7) = 5.65$, $p < 0.05$). For the last subscale, physical demand, we found a marginally significant difference between the relaxed and silent conditions, $F(1, 7) = 3.79$, $p = 0.09$. On all subscales relaxed thinking aloud was perceived as more demanding than performing in silence.

For performance we also found a significant interaction between thinking-aloud condition and task type, $F(1, 7) = 10.15$, $p < 0.05$. That is, performance in the relaxed condition was rated worse for assessment tasks compared to solving tasks in silence (relaxed 55% worse than silent) and essentially the same for fact tasks (relaxed 4% better than silent). For the five other subscales we found no interaction between thinking-aloud condition and task type.

4.7 Correlations between measures

The measures analysed in the preceding sections may correlate. We analysed correlations between the measures for which significant differences were found in Sections 4.1-4.6. In the correlation analysis, each data point was a participant's solution of a task solved while thinking aloud. For mental workload a single combined value was obtained for each task by taking the mean of the six TLX subscales, as recommended by Nygren (1991) and Hendy *et al.* (1993).

Table 11 summarizes the correlation analysis for classic thinking aloud. We found moderate to strong, significant correlations between all pairs of five of the measures: task completion time, general focused visual behaviour, general distributed visual behaviour, mouse clicks, and scrolling instances. While it is unsurprising that the number of mouse clicks and scrolling instances co-varied with task completion time, the correlations between task completion time and the percentages of time spent on general focused and general distributed visual behaviours suggest that these behaviours might become more salient as tasks become longer. In addition, mental workload had a weak, significant correlation with general focused visual behaviour and a moderate, significant correlation with scrolling instances, suggesting that $r^2 = 13\%$ and 21%, respectively, of the variation in mental workload can be predicted from these two measures.

Table 12 summarizes the correlation analysis for relaxed thinking aloud. We found moderate to very strong, significant correlations between all but one of the pairs of the same five measures as for classic thinking aloud: task completion time, general focused visual behaviour, general distributed visual behaviour, mouse clicks, and scrolling instances. In addition, mental workload was moderately and significantly correlated with task completion time and scrolling instances, suggesting that r^2 = 19% and 24%, respectively, of the variation in mental workload can be predicted from these two measures. Across all seven measures, the correlations for classic and relaxed thinking aloud were highly similar.

5 Discussion

Relaxed thinking aloud affected participants' behaviour in multiple ways. For classic thinking aloud most effects were merely suggestive. Below we discuss the effects of thinking aloud, their implications for usability evaluation, and the limitations of our experiment.

5.1 Effects of thinking aloud

Any effect of thinking aloud questions whether measurements made during thinking aloud can be taken as evidence of how people perform when they are not thinking aloud. Such threats to the validity of the method may ultimately reduce its applicability to comparisons between a situation in which participants have been thinking aloud and another situation in which participants have also been thinking aloud. We investigated the presence of effects with respect to correctness of task solutions, task completion times, eye movements, hand movements, and mental workload.

The *correctness of task solutions* was not affected by whether participants were thinking aloud or performing in silence. This was the case for both classic and relaxed thinking aloud. For classic thinking aloud, this result accords with Ericsson and Simon (1993) but discords with Van den Haak *et al.* (2003) who found that verbalization at levels 1 and 2 led to more tasks being solved incorrectly. For relaxed thinking aloud, previous studies (Chi *et al.* 1994, Wright and Converse 1992) indicate that verbalization leads to fewer errors, compared to performing in silence. In addition, Krahmer and Ummelen (2004) find that a variant of relaxed thinking aloud consistent with Boren and Ramey's (2000) recommendations leads to more tasks being solved correctly compared to classic thinking aloud. One reason for the disagreements between our results and those of others may be the low number of incorrect solutions to any of the fact tasks, suggesting that the fact tasks were fairly simple or participants very thorough. In Chi *et al.* (1994), Krahmer and Ummelen (2004), Van den Haak *et al.* (2003), and Wright and Converse (1992) the error rates were considerably higher.

Task completion times were longer during thinking aloud than when participants performed in silence. This difference was present for classic thinking aloud as well as for relaxed thinking aloud, and it was mainly due to participants' performance on assessment tasks. The extra time required during thinking aloud accords with previous studies (Rhenius and Deffner 1990, Ericsson and Simon 1993) and indicates that verbalization is a slower process than thinking. Further, fact tasks took less time to complete than assessment tasks. Reasons for this may include that assessment tasks had lower a priori determinability, a larger number of alternative paths of task performance, and potentially conflicting dependencies among the characteristics of candidate answers. Such characteristics, which may make the task completion times for assessment tasks more susceptible to individual variation, are often seen as indicators of higher task complexity (Wood 1986, Campbell 1988).

Contrary to our results, Wright and Converse (1992) found that participants verbalizing at level 3 solved tasks faster than participants performing in silence. They attribute this to a better understanding of the tasks, achieved by prompting participants for explanations for all their behaviours in the verbalization condition. One possible reason for the disagreement between our results and those of Wright and Converse (1992) may be that during relaxed thinking aloud we did not prompt participants for explanations for everything they did in the periods during which they verbalized by themselves, even when these verbalizations were mainly at levels 1 and 2. Consequently, participants may not have achieved the same improvement in their understanding of the tasks. We believe, however, that our level of prompting for explanations is more representative of how thinking aloud is typically done in practical usability evaluations. A related explanation is provided by Berry and Broadbent (1990), who found that the effect of level 3 verbalization on task completion time depended on whether participants had prior knowledge of an efficient strategy for solving the tasks. Participants that had been instructed about how to

solve tasks efficiently were 5% faster when verbalizing while solving tasks compared to performing in silence. Conversely, participants that had not received the instruction were 8% slower when verbalizing while solving tasks.

Participants' *eye movements* differed in some respects but did not provide evidence of a trend indicating that marginal effects for classic thinking aloud become significant for relaxed thinking aloud. Previous studies (e.g. Rhenius and Deffner 1990) have also yielded somewhat fragmented results with respect to the effects of thinking aloud on eye movements. During classic thinking aloud a larger part of task completion times elapsed before participants had performed half of the fixations they performed during a task. For assessment tasks this effect co-occurred with fixations of shorter duration. Combined these two effects suggest that mental activity was shifted slightly from the start toward the end of tasks. A similar effect was not found for relaxed thinking aloud. Rather, the part of task completion times that elapsed before participants had performed half of their fixations interacted with task type.

During relaxed thinking aloud saccades were of shorter duration for assessment tasks, the more complex type of tasks. This is often seen as an indication of decreased visual search (Goldberg and Kotval 1999). However, in this condition participants also spent a larger part of task completion times in general distributed visual behaviour. This seems to indicate that participants to a larger extent needed to fixate briefly on various screen elements to assess their relevance and contribution to the tasks. Distributed visual behaviour is akin to visual search but at a level of aggregation where brief fixations intersperse an activity primarily characterized by frequent saccades between screen elements that are spatially far apart and typically also distinct in contents. One reason for the increase in general distributed visual behaviour during relaxed thinking aloud could be that verbalizing at level 3 disrupted participants' mental activities and made it more difficult to maintain a focus, necessitating more distributed visual behaviour to regain a focus. Another reason could be that relaxed thinking aloud made participants in doubt about their approach to solving tasks or aware of other ways of solving them, leading to more distributed visual exploration of the screen.

Participants' *hand movements* revealed considerable differences in how participants interacted with the system while solving the tasks. These differences were marginal for classic thinking aloud and significant for relaxed thinking aloud, consistent with an interpretation that verbalization at levels 1 and 2 may at most have a limited effect on participants' operation of the system, whereas verbalization at level 3 has a distinct effect on how the system is operated. Participants made marginally more mouse clicks and writing instances during classic thinking aloud compared to the silent condition. This suggests that participants may be paying more attention to obtaining information from web pages other than the current one, because each mouse click corresponds to a shift to another web page and most writing instances consist of typing a query in a search field. During relaxed thinking aloud participants made more mouse clicks and scrolling instances compared to the silent condition. This indicates that participants made more efforts to obtain information from other web pages by making more shifts between web pages, and they also explored the current web page more comprehensively by scrolling more. Hence, the increase in task completion times must be interpreted differently for classic and relaxed thinking aloud. During classic thinking aloud the increase in task completion times was primarily a slowdown in participants' performance but during relaxed thinking aloud participants performed the tasks in a different way.

By issuing more operations to navigate both within and between web pages participants appear to come by more information and possibly explore more solution paths. It should, however, be noted that participants were not necessarily more systematic. An alternative explanation may be that verbalizing, particularly at level 3, disrupted participants' mental processes and forced them to redo some interactions with the system. The absence of differences in the number of correctly solved tasks lends some support to this alternative explanation. Except for the marginally significant effects for classic thinking aloud, our results about participants' hand movements are consistent with Ericsson and Simon (1993).

Mental workload was rated marginally higher for classic thinking aloud, compared to performing in silence. For relaxed thinking aloud participants rated mental workload higher than for performing in silence. This overall picture was repeated for the individual TLX subscales. Verbalization at levels 1 and 2 added to one of the six subscales of mental workload, whereas verbalization at level 3 added to all but one of the subscales. The results for mental workload, a subjective measure, are consistent with the performance measures. Specifically, mental workload correlated with number of scrolling instances and, for relaxed thinking aloud, task completion time. Effect sizes tend to be larger for mental workload than

for the performance measures, suggesting that participants may moderate performance differences by putting in extra mental effort while thinking aloud.

The results for verbalization at levels 1 and 2 are in slight disagreement with Ericsson and Simon (1993), who maintain that such verbalizations do not affect participants' mental processes, except by prolonging them. Our experiment yields increased ratings on the mental-demand subscale, suggesting that verbalization may consume some mental resources and thus leave fewer resources for performing the task. The additional mental workload entailed in producing verbalizations at level 3 is consistent with Ericsson and Simon (1993). According to Ericsson and Simon (1993) these verbalizations constitute a separate mental activity introduced by the requirement to produce explanations, an activity that is neither present when performing in silence nor when participants verbalize at levels 1 and 2 only. Our results discord, however, with Wright and Converse (1992), who found no difference in mental workload, which they also measured by TLX, when verbalization at level 3 was compared with performing in silence. One candidate explanation for this disagreement could be the difficulty of the tasks. This explanation is, however, not supported by the data as Wright and Converse (1992) used four tasks of increasing difficulty but found no interaction between verbalization condition and task, and we likewise found no interaction between thinking-aloud condition and task type, except for the performance subscale. Another reason could be that the between-subjects design of Wright and Converse's (1992) experiment made their experiment less sensitive to differences between conditions in participants' perception of mental workload, compared to our within-subjects design.

5.2 Implications for usability evaluation

Dumas and Redish (1999), Nielsen (1993), and multiple other texts on usability evaluation use the term thinking aloud with reference to Ericsson and Simon (1993) but without consistently distinguishing between verbalization at levels 1 and 2 and verbalization at level 3. Further, studies of what usability evaluators do in practice suggest that relaxed thinking aloud is common and that the additional comments obtained in this way are valued by usability evaluators (Boren and Ramey 2000, Nørgaard and Hornbæk 2006). Failure to distinguish between classic and relaxed thinking aloud is, however, misleading because relaxed thinking aloud has much more profound effects on participants' behaviour than classic thinking aloud. Three implications – applicable in different situations – suggest themselves:

Complying with Ericsson and Simon's (1993) description of thinking aloud. Classic thinking aloud seems to exert a merely modest effect on participants' behaviour, consisting mostly of invalidating measures of task completion times. Classic thinking aloud is, however, slightly more difficult to learn and practice than relaxed thinking aloud and not readily compatible with a lightweight approach to usability evaluation. Some practitioners and researchers may welcome the increased formality involved in consistently adopting classic thinking aloud (Cockton and Woolrych 2002); others may consider it misconstrued (Wixon 2003).

Abandoning Ericsson and Simon (1993) as the framework for understanding thinking aloud. In practical usability evaluation thinking aloud is at times more like an interview based on concurrent hands-on experience with a system than like classic thinking aloud (Nørgaard and Hornbæk 2006). While the former is essentially a communication process, the latter is meticulously set up as a solitary activity in which the person thinking out loud suspends any awareness of listeners. Thus, another framework is required to understand relaxed thinking aloud. Boren and Ramey (2000) suggest speech genres, but though their approach is only slightly more relaxed than classic thinking aloud it affects participants' task performance (Krahmer and Ummelen 2004).

Abandoning thinking aloud. Much of the mental activity that is of interest in usability evaluations is manifest not only through participants' verbalizations but also through their interactions with the system. Accordingly, some studies find that verbalization may add little value to usability evaluation (Lesaigle and Biers 2000, Van den Haak *et al.* 2003, 2004). Abandoning thinking aloud in favour of having participants perform in silence will remove an ecological gap in usability evaluations. Further, a retrospective interview, possibly supported by a video recording of the session, may provide a useful separation between performance and commentary (Frøkjær and Hornbæk 2005).

5.3 Limitations

This study has five limitations that should be remembered in interpreting the results. First, the experimental design allows no direct comparison between classic and relaxed thinking aloud. The two variants of thinking aloud were compared to performing in silence, not to each other. Second, participants solved pairs of very similar tasks, not the same tasks, in the thinking-aloud condition and the silent condition. The use of carefully paired tasks was a result of the within-subjects design in which the same participants performed in both the thinking-aloud condition and the silent condition. Paired tasks could have been avoided by a between-subjects design, but only at the expense of introducing an equivalent need for carefully pairing participants. Third, the number of participants was fairly small. As a consequence some effects of thinking aloud may have been missed because they did not reach significance. Fourth, the tasks were similar to tasks commonly used in usability evaluations and involved navigation on large web sites, formulation of queries, and assessment of text and illustrations. This made it prohibitively complex to define the possible strategies for solving the tasks and analyse whether thinking aloud affects participants' choice of strategy. Considerably simpler tasks are required to enable such analysis. Fifth, in solving the tasks participants used operational systems, not prototypes at various stages of development. Prototypes are likely to be rougher and thereby require more from participants in terms of maintaining their focus on progressing through a task in spite of disruptions caused by various usability problems. Thus, our study may underestimate the effects of thinking aloud compared to its use in usability evaluations of prototypes. The difference between operational systems and prototypes is, however, more important in studies aiming to investigate whether thinking aloud leads to the detection of different amounts or types of usability problems, something we have not investigated.

6 Conclusion

Thinking aloud is widely used as a method for usability evaluation. The method is, however, generally applied in a relaxed manner that conflicts with the prescriptions of Ericsson and Simon's classic model for obtaining valid verbalizations of thought processes. Descriptions of thinking aloud in the methodological literature often display a similar failure to consistently distinguish between classic thinking aloud (corresponding to verbalization at levels 1 and 2) and relaxed thinking aloud (corresponding to verbalization at level 3). In this study, we have investigated the effects of thinking aloud over performing in silence for both classic and relaxed thinking aloud.

Our results confirm that classic thinking aloud has little effect on participants' behaviour and mental workload, except for prolonging tasks. Hence, valid data about the use of a system can be obtained at the price of precise instructions and minimal interaction between the user that thinks aloud and the evaluator that listens in on the user's thoughts. Conversely, relaxed thinking aloud led to longer task completion times, a larger part of tasks spent on general distributed visual behaviour, more commands issued to navigate both within and between the pages of the web sites used for solving the tasks, and higher perceived mental workload. Hence, the relaxed approach to thinking aloud threatens the validity of the method and indicates that this approach, common in practical usability evaluation, may not be the authoritative yardstick it is often assumed to be.

The threats to the validity of relaxed thinking aloud make it less applicable to evaluations of details in users' interaction with a system. The increase in mental workload also means that if a task, or subtask, requires most of users' mental resources they will either be unable to verbalize or will modify their mental processes to free resources for verbalization. This may make relaxed thinking aloud most applicable to comparisons between situations in which participants have been thinking aloud, whereas caution in needed in making inferences from thinking-aloud sessions to silent, real-world use of systems. Similarly, task completion times for classic thinking aloud should only be used for comparing situations in which participants have been thinking aloud. For practitioners the implications of this study are to clarify the limitations of relaxed thinking aloud and provide evidence that classic thinking aloud is a more valid alternative. For researchers the implications are the importance of consistently distinguishing between classic and relaxed thinking aloud and of advancing – and further elaborating – valid methods while at the same time acknowledging the practical relevance of the extra comments that can be obtained through a more relaxed communication between user and evaluator.

Acknowledgements

We are grateful to Casper Højstrup who made the software for an initial eye-movement analysis at a critical stage of the project. Special thanks are due to the eight participants in the experiment.

Appendix A

The eight pairs of tasks used in the experiment. The task type (given in parentheses after each task, along with the web site) was not part of the task descriptions given to participants during the experiment.

Pairs of tasks in Session 1:

1a. What information can you find about today's weather on this site? Can you make www.tv2.dk your start page? (fact, www.tv2.dk)

1b. What information can you find about today's weather on this site? Can you make www.dr.dk your start page? (fact, www.dr.dk)

2a. Which city has the highest temperature today – Copenhagen or Aarhus? (fact, www.tv2.dk)

2b. Which city has the highest temperature today – Skagen or Nykøbing Falster? (fact, www.dr.dk)

3a. Based on the information you can find about it, would you be tempted to read/buy the book *The Devils*, volume 1, by Dostoyevsky? Make a voucher (e.g. to yourself) on an amount corresponding to the price of the book. (assessment, www.e-boghandel.dk)

3b. Based on the information you can find about it, would you be tempted to read/buy the book *Setting Free the Bears* by Irving? Make a voucher (e.g. to yourself) on an amount corresponding to the price of the book. (assessment, www.arnoldbusck.dk)

4a. Find a biography you consider interesting and order it. Type in the requested information about name, address, and so forth but do not submit it. (assessment, www.e-boghandel.dk)

4b. Find a cookbook you consider interesting and order it. Type in the requested information about name, address, and so forth but do not submit it. (assessment, www.arnoldbusck.dk)

Pairs of tasks in Session 2:

1a. What is the title of Jan Kjærstad's first book? (fact, www.arnoldbusck.dk)

1b. What is the title of Leif Panduro's longest book? (fact, www.e-boghandel.dk)

2a. Find an interesting book about psychological work conditions. The book must be from 2003 or 2004. Sort the search results by title; use advanced search. (assessment, www.arnoldbusck.dk)

2b. Find an interesting book about stress and how it can be managed. The book must be in Danish. Sort the search results by price; use advanced search. (assessment, www.e-boghandel.dk)

3a. How many of the Danish television channels – i.e. DR, DR2, TV2, TV2 Zulu, TV3, 3+, TV Danmark, and Kanal 5 – send a movie this Sunday evening? (fact, www.tv2.dk)

3b. How many of the Danish television channels – i.e. DR, DR2, TV2, TV2 Zulu, TV3, 3+, TV Danmark, and Kanal 5 – send cartoons this Sunday morning? (fact, www.dr.dk)

4a. What is the biggest domestic news story on the front page? (assessment, www.tv2.dk)

4b. What is the biggest international news story on the front page? (assessment, www.dr.dk)

References

BERRY, D.C., and BROADBENT, D.E., 1990, The role of instruction and verbalization in improving performance on complex search tasks. *Behaviour & Information Technology*, 9(3), 175-190.

BOREN, T., and RAMEY, J., 2000, Thinking aloud: Reconciling theory and practice. *IEEE Transactions on Professional Communication*, 43(3), 261-278.

CAMPBELL, D.J., 1988, Task complexity: A review and analysis. *Academy of Management Review*, 13(1), 40-52.

CHI, M.T.H., BASSOK, M., LEWIS, M., REIMANN, P., and GLASER, R., 1989, Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, 13(2), 145-182.

CHI, M.T.H., DE LEEUW, N., CHIU, M.-H., and LAVANCHER, C., 1994, Eliciting self-explanations improves understanding. *Cognitive Science*, 18(3), 439-477.

COCKTON, G., and WOOLRYCH, A., 2002, Sale must end: Should discount methods be cleared off HCI's shelves? *ACM Interactions*, 9(5), 13-18.

DESHON, R.P., CHAN, D., and WEISSBEIN, D.A., 1995, Verbal overshadowing effects on Raven's advanced progressive matrices: Evidence for multidimensional performance determinants. *Intelligence*, 21(2), 135-155.

DUMAS, J.S., 2003, User-based evaluations. In *The Human-Computer Interaction Handbook*, J. Jacko and A. Sears (Eds.), pp. 1093-1117 (Mahwah, NJ: Erlbaum).

DUMAS, J.S., and REDISH, J.C., 1999, *A Practical Guide to Usability Testing. Revised Edition* (Exeter, UK: Intellect Books).

ERICSSON, K.A., and SIMON, H.A., 1980, Verbal reports as data. *Psychological Review*, 87(3), 215-251.

ERICSSON, K.A., and SIMON, H.A., 1993, *Protocol Analysis: Verbal Reports as Data. Revised Edition* (Cambridge, MA: MIT Press).

FRØKJÆR, E., and HORNBÆK, K., 2005, Cooperative usability testing: Complementing usability tests with user-supported interpretation sessions. In *Extended Abstracts of the CHI 2005 Conference on Human Factors in Computing Systems* (New York: ACM Press), pp. 1383-1386.

GOLDBERG, J.H., and KOTVAL, X.P., 1999, Computer interface evaluation using eye movements: Methods and constructs. *International Journal of Industrial Ergonomics*, 24(6), 631-645.

GUAN, Z., LEE, S., CUDDIHY, E., and RAMEY, J., 2006, The validity of the stimulated retrospective think-aloud method as measured by eye tracking. In *Proceedings of the CHI 2006 Conference on Human Factors in Computing Systems* (New York: ACM Press), pp. 1253-1262.

HART, S.G., and STAVELAND, L.E., 1988, Development of NASA-TLX (task load index): Results of empirical and theoretical research. In *Human Mental Workload*, P.A. Hancock and N. Meshkati (Eds.), pp. 139-183 (Amsterdam: Elsevier).

HELD, J.E., and BIERS, D.W., 1992, Software usability testing: Do evaluator intervention and task structure make any difference? In *Proceedings of the Human Factors Society 36th Annual Meeting* (Santa Monica, CA: HFS), pp. 1215-1219.

HENDY, K.C., HAMILTON, K.M., and LANDRY, L.N., 1993, Measuring subjective workload: When is one scale better than many? *Human Factors*, 35(4), 579-601.

JAMES, W., 1890, *The Principles of Psychology* (New York: Holt).

JUST, M.A., and CARPENTER, P.A., 1980, A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87(4), 329-354.

KRAHMER, E., and UMMELEN, N., 2004, Thinking about thinking aloud: A comparison of two verbal protocols for usability testing. *IEEE Transactions on Professional Communication*, 47(2), 105-117.

LANDIS, J.R., and KOCH, G.G., 1977, The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159-174.

LESAIGLE, E.M., and BIERS, D.W., 2000, Effect of type of information on real time usability evaluation: Implications for remote usability testing. In *Proceedings of the XIVth Triennial Congress of the International Ergonomics Association and 44th Annual Meeting of the Human Factors and Ergonomics Society* (Santa Monica, CA: HFES), pp. 6-585 - 6-588.

LEWIS, C., 1982, Using the 'thinking-aloud' method in cognitive interface design. Research Report RC9265 (Yorktown Heights, NY: IBM Watson Research Center).

MEISSNER, C.A., and BRIGHAM, J.C., 2001, A meta-analysis of the verbal overshadowing effect in face identification. *Applied Cognitive Psychology*, 15(6), 603-616.

MONK, A., WRIGHT, P., HABER, J., and DAVENPORT, L., 1993, *Improving Your Human-Computer Interface: A Practical Technique* (New York: Prentice Hall).

NIELSEN, J., 1993, *Usability Engineering* (Boston: Academic Press).

NISBETT, R.E., and WILSON, T.D., 1977, Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84(3), 231-259.

NYGREN, T.E., 1991, Psychometric properties of subjective workload measurement techniques: Implications for their use in the assessment of perceived mental workload. *Human Factors*, 33(1), 17-33.

NØRGAARD, M., and HORNBÆK, K., 2006, What do usability evaluators do in practice? An explorative study of think-aloud testing. In *Proceedings of the Sixth DIS Conference on Designing Interactive Systems* (New York: ACM Press), pp. 209-218.

RAYNER, K., 1998, Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3), 372-422.

RHENIUS, D., and DEFFNER, G., 1990, Evaluation of concurrent thinking aloud using eye-tracking data. In *Proceedings of the Human Factors Society 34th Annual Meeting* (Santa Monica, CA: HFS), pp. 1265-1269.

RUSSO, J.E., JOHNSON, E.J., and STEPHENS, D.L., 1989, The validity of verbal protocols. *Memory and Cognition*, 17(6), 759-769.

SALVUCCI, D.D., and GOLDBERG, J.H., 2000, Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the Eye Tracking Research and Applications Symposium 2000* (New York: ACM Press), pp. 71-78.

SCHOOLER, J.W., 2002, Verbalization produces a transfer inappropriate processing shift. *Applied Cognitive Psychology*, 16(8), 989-997.

SCHOOLER, J.W., and ENGSTLER-SCHOOLER, T.Y., 1990, Verbal overshadowing of visual memories: Some things are better left unsaid. *Cognitive Psychology*, 22(1), 36-71.

SPOOL, J., SCANLON, T., SCHROEDER, W., SNYDER, C., and DEANGELO, T., 1999, *Web Site Usability: A Designer's Guide* (San Francisco, CA: Morgan Kaufmann).

VAN DEN HAAK, M.J., DE JONG, M.D.T., and SCHELLENS, P.J., 2003, Retrospective vs. concurrent think-aloud protocols: Testing the usability of an online library catalogue. *Behaviour & Information Technology*, 22(5), 339-351.

VAN DEN HAAK, M.J., DE JONG, M.D.T., and SCHELLENS, P.J., 2004, Employing think-aloud protocols and constructive interaction to test the usability of online library catalogues: A methodological comparison. *Interacting with Computers*, 16(6), 1153-1170.

WILSON, T.D., and SCHOOLER, J.W., 1991, Thinking too much: Introspection can reduce the quality of preferences and decisions. *Journal of Personality and Social Psychology*, 60(2), 181-192.

WIXON, D., 2003, Evaluating usability methods: Why the current literature fails the practitioner. *ACM Interactions*, 10(4), 29-34.

WOOD, R.E., 1986, Task complexity: Definition of the construct. *Organizational Behavior and Human Decision Processes*, 37(1), 60-82.

WRIGHT, R.B., and CONVERSE, S.A., 1992, Method bias and concurrent verbal protocol in software usability testing. In *Proceedings of the Human Factors Society 36th Annual Meeting* (Santa Monica, CA: HFS), pp. 1220-1224.

Table 1. Task completion times in session comparing classic thinking aloud with performing in silence, *N* = 63 tasks.

Task completion time	Classic thinking aloud		Silent	
	Mean	*SD*	*Mean*	*SD*
Fact tasks (seconds/task)	110	47	82	35
Assessment tasks * (seconds/task)	303	92	217	41

* $p < 0.05$

Table 2. Task completion times in session comparing relaxed thinking aloud with performing in silence, $N = 64$ tasks.

Task completion time	Relaxed thinking aloud		Silent	
	Mean	*SD*	*Mean*	*SD*
Fact tasks (seconds/task)	201	55	131	76
Assessment tasks ** (seconds/task)	319	148	114	49

** $p < 0.01$

Table 3. Eye-movement measures in session comparing classic thinking aloud with performing in silence, $N = 63$ tasks.

Eye-movement measure	Classic thinking aloud		Silent	
	Mean	*SD*	*Mean*	*SD*
Half of fixations (percent of task time) **	48	3	46	5
Fixation rate (fixations/second)	2.73	0.29	2.54	0.30
Fixation duration (milliseconds)	272	30	288	30
Saccade duration (milliseconds)	100	58	116	59
Saccade length (pixels)	76	7	79	10

** $p < 0.01$

Table 4. Eye-movement measures in session comparing relaxed thinking aloud with performing in silence, $N = 64$ tasks.

Eye-movement measure	Relaxed thinking aloud		Silent	
	Mean	*SD*	*Mean*	*SD*
Half of fixations (percent of task time) [+]	52	4	51	4
Fixation rate (fixations/second)	2.68	0.27	2.49	0.26
Fixation duration (milliseconds)	277	41	278	28
Saccade duration (milliseconds)	106	33	137	65
Saccade length (pixels)	73	8	77	11

[+] Interaction between thinking-aloud condition and task type, $p < 0.05$

Table 5. Focused and distributed visual behaviour in session comparing classic thinking aloud with performing in silence, $N = 63$ tasks.

Eye-movement measure	Classic thinking aloud		Silent	
	Mean	*SD*	*Mean*	*SD*
Focused: general [+] (percent of task time)	18	9	16	11
Focused: text (percent of task time)	10	7	10	6
Focused: illustration (percent of task time)	8	10	9	12
Distributed: general [+] (percent of task time)	12	7	9	7
Distributed: text (percent of task time)	29	14	33	17
Distributed: illustration (percent of task time)	1	1	1	2

[+] Interaction between thinking-aloud condition and task type, $p < 0.05$

Table 6. Focused and distributed visual behaviour in session comparing relaxed thinking aloud with performing in silence, $N = 64$ tasks.

Eye-movement measure	Classic thinking aloud		Silent	
	Mean	*SD*	*Mean*	*SD*
Focused: general (percent of task time)	23	8	21	7
Focused: text (percent of task time)	3	4	5	7
Focused: illustration (percent of task time)	3	3	2	2
Distributed: general (percent of task time) **	10	4	5	5
Distributed: text (percent of task time)	38	7	39	17
Distributed: illustration (percent of task time)	1	1	1	2

** $p < 0.05$

Table 7. Hand-movement measures in session comparing classic thinking aloud with performing in silence, $N = 63$ tasks.

Hand-movement measure	Classic thinking aloud		Silent	
	Mean	*SD*	*Mean*	*SD*
Mouse clicks (pr task)	6.38	4.06	5.22	4.09
Scrolling instances (pr task)	15.53	10.69	13.19	6.79
Writing instances (pr task)	8.31	8.60	6.81	7.16

Table 8. Hand-movement measures in session comparing relaxed thinking aloud with performing in silence, $N = 64$ tasks.

Hand-movement measure	Relaxed thinking aloud		Silent	
	Mean	*SD*	*Mean*	*SD*
Mouse clicks (pr task) *	5.91	4.90	4.00	2.61
Scrolling instances (pr task) **	26.56	21.49	10.91	8.86
Writing instances (pr task)	2.44	1.81	1.91	0.97

* $p < 0.05$, ** $p < 0.01$

Table 9. Mental workload, measured by TLX, in session comparing classic thinking aloud with performing in silence, $N = 64$ tasks.

TLX subscale	Classic thinking aloud		Silent	
	Mean	*SD*	*Mean*	*SD*
Mental demand * (0: low – 100: high)	41	26	31	17
Physical demand (0: low – 100: high)	20	22	14	13
Temporal demand (0: low – 100: high)	23	22	20	13
Effort (0: low – 100: high)	28	24	21	14
Performance (0: good – 100: poor)	30	26	29	23
Frustration (0: low – 100: high)	26	19	18	13

* $p < 0.05$

Table 10. Mental workload, measured by TLX, in session comparing relaxed thinking aloud with performing in silence, $N = 64$ tasks.

TLX subscale		Relaxed thinking aloud		Silent	
		Mean	*SD*	*Mean*	*SD*
Mental demand (0: low – 100: high)	**	30	15	19	10
Physical demand (0: low – 100: high)		16	13	10	7
Temporal demand (0: low – 100: high)	*	18	12	10	7
Effort (0: low – 100: high)	*	25	12	14	8
Performance (0: good – 100: poor)	* +	21	18	17	16
Frustration (0: low – 100: high)	*	21	15	10	7

* $p < 0.05$, ** $p < 0.01$ $^{+}$ Interaction between thinking-aloud condition and task type, $p < 0.05$

Table 11. Correlations between measures, $N = 31$ classic thinking-aloud tasks.

Correlations (Pearson's *r*)	Task completion time	Half of fixations	General focused visual behaviour	General distributed visual behaviour	Mouse clicks	Scrolling instances	Mental workload[1]
Task completion time	1.00	-0.05	0.85**	0.75**	0.80**	0.71**	0.23
Half of fixations		1.00	-0.06	0.27	-0.19	-0.11	0.07
General focused visual behaviour			1.00	0.62**	0.78**	0.59**	0.36*
General distributed visual behaviour				1.00	0.56**	0.48**	0.11
Mouse clicks					1.00	0.64**	0.34
Scrolling instances						1.00	0.46**
Mental workload[1]							1.00

* $p < 0.05$, ** $p < 0.01$, [1] Combined mental-workload value calculated as mean of the six TLX subscales

Table 12. Correlations between measures, $N = 32$ relaxed thinking-aloud tasks.

Correlations (Pearson's *r*)	Task completion time	Half of fixations	General focused visual behaviour	General distributed visual behaviour	Mouse clicks	Scrolling instances	Mental workload[1]
Task completion time	1.00	-0.07	0.92**	0.71**	0.91**	0.76**	0.44*
Half of fixations		1.00	-0.02	-0.26	0.07	-0.07	-0.01
General focused visual behaviour			1.00	0.63**	0.96**	0.65**	0.27
General distributed visual behaviour				1.00	0.66**	0.25	0.14
Mouse clicks					1.00	0.59**	0.17
Scrolling instances						1.00	0.49**
Mental workload[1]							1.00

* $p < 0.05$, ** $p < 0.01$, [1] Combined mental-workload value calculated as mean of the six TLX subscales