*International Journal of Human-Computer Interaction*, vol. 29, issue 5 (2013), pp. 351-364.
Preprint version

# Thinking Aloud in the Presence of Interruptions and Time Constraints

Morten Hertzum and Kristin Due Holmegaard

Computer Science, Roskilde University, Roskilde, Denmark

mhz@ruc.dk, kdh@ruc.dk

**Abstract**. Thinking aloud is widely used for usability evaluation and its reactivity is therefore important to the quality of evaluation results. This study investigates whether thinking aloud (i.e., verbalization at levels 1 and 2) affects the behaviour of users who perform tasks that involve interruptions and time constraints, two frequent elements of real-world activities. We find that the presence of auditory, visual, audiovisual, or no interruptions interacts with thinking aloud for task solution rate, task completion time, and participants' fixation rate. Thinking-aloud participants also spend longer responding to interruptions than control participants. Conversely, the absence or presence of time constraints does not interact with thinking aloud, suggesting that time pressure is less likely to make thinking aloud reactive than previously assumed. Our results inform practitioners faced with the decision to either restrict verbalizations in usability evaluation to thinking aloud to avoid reactivity or relax the constraints on verbalization to obtain additional information.

**Keywords**: thinking aloud, verbalization, usability evaluation, user testing, interruption, time constraints, reactivity

## 1 Introduction

Evaluation is central to ensuring usable systems, and effective and non-reactive usability evaluation methods are, consequently, in high regard. Surveys repeatedly show that usability practitioners consider the thinking-aloud method one of their most important tools (Gulliksen, Boivie, Persson, Hektor, & Herulf, 2004; Venturi, Troost, & Jokela, 2006; Vredenburg, Mao, Smith, & Carey, 2002), and some researchers conclude that it may be the single most important usability evaluation method (Dumas & Fox, 2008; Nielsen, 1993). However, concerns remain about the reactivity of the thinking-aloud method because "We do not have basic information such as […] whether thinking aloud changes the way participants examine a product" (Dumas & Fox, 2008, p. 1140). Any effects of thinking aloud on a user's mental processes appear more likely to affect behaviour and performance in situations that impose high demands on the user, because such situations leave less capacity for additional activities that may mask or compensate for the changes in mental processes. Yet, it may be in these high-demand situations that usability and, hence, evaluation are most important. This study investigates thinking aloud on a mentally demanding task. We specifically investigate whether users' behaviour in response to interruptions and time constraints – two elements common in many real-world settings – is affected by thinking aloud.

In usability evaluations of consumer products, websites, and other information technologies thinking aloud is typically performed in a relaxed manner that allows for probing the users about their feelings, opinions, and the reasons for their actions. Such relaxed thinking aloud affects users' behaviour (Held & Biers, 1992; Hertzum, Hansen, & Andersen, 2009; Olmsted-Hawala, Murphy, Hawala, & Ashenfelter, 2010; Wright & Converse, 1992) but this reactivity tends to be considered secondary to the value of the extra information (Boren & Ramey, 2000; Rubin & Chisnell, 2008). The thinking-aloud method is, however, also applied in situations where it is considered imperative to keep any effects of thinking aloud on user behaviour at a minimum. According to Ericsson and Simon (1993), this can be achieved by applying their classic procedure for restricting thinking aloud to the verbalization of heeded information. This classic variant of thinking aloud is applied in some evaluations of information technologies and it is widespread in evaluations of systems for more mentally demanding tasks, more stressful work environments, and more safety-critical domains. In process control classic thinking aloud has, for example, long been applied in the analysis of operator performance in abnormal situations characterized by the presence of alarms and time pressure (e.g., Patrick, Gregov, Halliday, Handley, & O'Reilly, 1999).

This study is about the classic variant of thinking aloud. We have participants perform a code-breaking task multiple times and systematically vary the experimental situation by introducing interruptions and time constraints. *Interruptions* introduce a need for remaining receptive to their onset, which in our experiment is indicated by an auditory, visual, or audiovisual cue. Thinking aloud may reduce this receptiveness if it ties the user's attentional resources more fully to the code-breaking task. In addition, the user's mental process is disrupted when an interruption occurs. Because thinking aloud consists of giving verbal expression to this mental process, thinking aloud may delay the switch to the other task, ease resumption of the code-breaking task after the interruption, or affect behaviour in some other way. *The time constraint* entails that half of the code-breaking tasks are performed under time pressure. This likely interferes with thinking aloud, which has the acknowledged effect of slowing users down because verbalization is a slower process than thinking (Ericsson & Simon, 1993). In addition, time constraints increase mental workload and, thereby, the likelihood that any verbalization-induced changes in mental processes will affect behaviour and performance. We contend that evaluating the usability of systems for such demanding cognitive tasks, as opposed to mainly navigational or information-seeking tasks, is an important application area for the thinking-aloud method.

In the following, we account for related work (Section 2), describe our experimental method (Section 3), present our results (Section 4), and discuss their implications for the use of thinking aloud in usability evaluation (Section 5). The main finding of this study is that user behaviour in the presence of interruptions is affected by thinking aloud, in different ways for different types of interruptions.

## 2 Related work

Thinking aloud was introduced as a method for usability evaluation in the early 1980s (Lewis, 1982). In its essence, the thinking-aloud method consists of a user who thinks aloud while using a system, and an evaluator who observes the user and listens in on his or her thoughts. Thus, thinking aloud is a means of complementing and supplementing the observation of the user.

### *2.1 Thinking aloud*

The predominant theoretical model of thinking aloud is that of Ericsson and Simon (1980, 1993), who aim to establish verbal protocols as valid data. The essence of their model is to enable participants to remain focused on solving their task while merely giving verbal expression to the thoughts that emerge in attention. For this purpose Ericsson and Simon introduce a division of verbalizations into three levels:

Level 1 verbalization is the verbal expression of information that is already in attention in verbal form, for example the intermediate results produced during mental arithmetic. Ericsson and Simon propose that giving verbal expression to such information does not bring new information into attention and can, thus, be done without changing the stream of information to which a person attends.

Level 2 verbalization is the verbal expression of information that is already in attention but in nonverbal form, for example images and abstract concepts. To express such information verbally it must first be transformed into words by giving it a verbal label or creating some other verbal referent for it. Ericsson and Simon propose that this transformation does not bring new information into attention but may slow down task performance.

Level 3 verbalization is the verbal expression of information not currently in attention, such as descriptions of reasons for and feelings associated with current actions. This information must be retrieved from memory or created by mental processes initiated to establish, for example, the reasons for an action. That is, new information is brought into attention in place of the information otherwise involved in solving the task.

It is well-established that level 3 verbalization may distort thought processes and change behaviour (Nisbett & Wilson, 1977; Russo, Johnson, & Stephens, 1989), for example by shifting a person's focus from a search for the best option to a search for the option supported by the best reasons (Wilson & Schooler, 1991). This sets level 3 verbalization apart from verbalization at levels 1 and 2. While Ericsson and Simon (1993) emphasize this difference by excluding level 3 verbalization from their concept of thinking aloud, the thinking-aloud method in usability evaluation typically blurs the difference and invites verbalization at all three levels (Boren & Ramey, 2000; Dumas & Redish, 1999). Consistent with Ericsson and Simon (1993) we, henceforth, reserve the term thinking aloud for verbalization at levels 1 and 2.

### *2.2 Completeness and reactivity of thinking aloud*

On the basis of a review of 47 studies that compare performance while thinking aloud (i.e., verbalization at levels 1 and 2) with performance in the absence of verbalization, Fox, Ericsson, and Best (2011) find that thinking aloud does not affect performance except by prolonging task completion times. In contrast, the 27 reviewed studies of verbalization that includes explanation (i.e., verbalization at levels 1 to 3) show that the absence or presence of verbalization explains an average of 24% of the total variation in participants'

performance (Fox et al., 2011). These results are based on data from as many as 1995 (thinking aloud) and 875 (verbalization that includes explanation) participants and strongly support Ericsson and Simon's (1980, 1993) model of verbal reporting.

It should, however, be noted that this model is restricted in two ways important to the completeness of thinking aloud and contested in a third way important to whether it is reactive. First, thinking aloud aims to give verbal expression to the information to which a person attends. However, as tasks become highly practiced their execution becomes still more automated in the sense that increasing numbers of intermediate steps are performed without receiving conscious attention. This greatly speeds up performance but is also makes the intermediate steps unavailable for thinking aloud (Ericsson & Simon, 1993, p. 15). Second, Ericsson and Simon (1993, p. 223) explicitly "exclude feelings from the thoughts we will consider." This makes their concepts of thoughts narrower than, for example, James' (1890) stream of thought, which includes as an integral part a fringe of dimly perceived relations and objects. This fringe determines feelings and moods toward the present focus of attention, but is excluded from thinking aloud. Third, it is contested whether the transformation of attended information from nonverbal to verbal form entails that level 2 verbalization is reactive. Gilhooly, Fioratou, and Henretty (2010) had participants verbalize while solving verbal and spatial tasks and, thereby, attending to information in verbal and nonverbal form, respectively. Thus, the verbal tasks involved level 1 verbalization and the spatial tasks level 2 verbalization. The verbalizing participants had lower solution rates for spatial tasks than control participants but not for verbal tasks, indicating that level 2 verbalization impaired performance. A possible explanation of this finding is verbal overshadowing. Proponents of verbal overshadowing (Meissner & Brigham, 2001; Schooler & Engstler-Schooler, 1990) argue that verbally describing nonverbal stimuli, such as a face, can impair subsequent identification of the stimuli. Verbalization appears to produce a processing shift that gives preference to verbal thinking and, temporarily, dampens people's capacity for non-verbal thinking, such as identifying a face from a photo array (Schooler, 2002).

Of specific relevance to this study, Fox et al. (2011) analyze the effect of time constraints in the studies of thinking aloud. They find no difference in the relative performance of the thinking-aloud and control conditions between the 12 studies with time-constrained tasks and the 35 studies without time constraints. Fox et al. see this as evidence that the studies imposing time constraints provide participants with adequate time and, thereby, avoid mistaking insufficient time to solve tasks while thinking aloud for evidence that thinking aloud affects performance. This way of looking at time constraints presupposes that they are properties of the studies rather than inherent in tasks. Much real-world system use is, however, time-constrained in that actions must be taken at the correct time in relation to task demands rather than when the user feels ready. For example, Dickson, McLennan, and Omodei (2000) study verbalization in a computer simulation of the time-critical task of fighting a forest fire. They find that participants who verbalized reasons for their actions performed worse than participants who did not verbalize and that the performance of participants who thought aloud was intermediate between the two other conditions and no different from any of them. This result suggests that thinking aloud does not interact with time constraints, but the study gives no information to verify that the participants were under time pressure.

All the studies reviewed by Fox et al. (2011) concern participants who perform one task at a time. We are unaware of studies of verbalization in the presence of secondary, interrupting tasks. Karbach and Kray (2007) have, however, investigated whether thinking aloud affects the switching costs incurred by children when they have completed one task and turn to the next. They find that five-year-old children benefit from thinking aloud during task switching but that nine-year-old children do not. In explaining this finding they liken thinking aloud to egocentric speech and refer to Vygotsky's (1988) proposal that vocalized egocentric speech is a developmental stage preceding inner speech. Vygotsky proposes that between the ages of 3 and 7 egocentric speech gradually develops in structure and function and, thereby, becomes dissociated from external speech. As a result, the vocalization fades away and "in the end, it becomes inner speech" (Vygotsky, 1988, p. 183). Hence, five-year-old children are familiar with thinking aloud and likely to use it habitually in enhancing their task performance, whereas nine-year-old children tend to rely on inner speech and no longer benefit from thinking out loud.

### *2.3 Interruptions*

Interruptions are disruptive (Trafton & Monk, 2007), but the interval between the notification of a pending interruption and the user's response to the interruption provides an opportunity for preparing and, thereby, expediting the later resumption of the interrupted task (Trafton, Altmann, Brock, & Mintz, 2003). In such preparations inner speech is an effective means of self-instruction about how to resume a task; preventing inner speech drastically increases the time required to switch between tasks (Emerson & Miyake, 2003). The close links among thinking aloud, inner speech, and the mental processes involved in switching effectively between tasks suggest that performance may be sensitive to even small changes introduced by thinking aloud. Because interruptions are frequent in real work, thinking aloud will be severely limited if it can only be used for

uninterrupted tasks. Studies, for example, find that managers work uninterrupted for more than half an hour only once every two days (Mintzberg, 1975), that information workers spend about 11 minutes on events with a common goal before being interrupted or switching to another goal (González & Mark, 2004), and that physicians and nurse shift coordinators at emergency departments are interrupted an average of 16 and 25 times an hour, respectively (Spencer, Coiera, & Logan, 2004).

# 3 Method

To investigate whether and how thinking aloud affects people's behaviour in the presence of interruptions and time constraints we conducted an experiment. We have previously used the data from the control participants in an analysis of measures of mental workload (Hertzum & Holmegaard, 2012). Therefore, the description of the experimental method that follows resembles that of our previous article.

## *3.1 Participants*

A total of 32 participants (13 female, 19 male) took part in the experiment, see Table 1. The participants were experienced computer users with an average age of 25.3 years. In terms of background, 26 of the participants were students at a technical university, five were professionals, and one did not report his background. All participants had normal or corrected-to-normal vision, a requirement introduced by the eye-tracking equipment.

## *3.2 Thinking-aloud conditions*

The experiment involved two thinking-aloud conditions:

*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.

*Control*, in which participants were simply instructed to solve the tasks. Participants were neither instructed to verbalize, nor to be silent. This condition is similar to how people work when they are not enrolled in usability evaluations. In the control condition the experimenter remained silent.

## *3.3 Tasks*

The task, similar to the game of mastermind, consisted of breaking a four-digit code by making repeated guesses and receiving feedback for each guess. The code was restricted to the digits 1 through 6 (e.g., '2265'), and participants were provided up to eight guesses to break the code. These design choices were made on the basis of pilot tests aimed at finding a level of task difficulty where some codes were broken, others not, and the task remained challenging throughout the session.

The screen area for solving the code-breaking task occupied the right-hand side of the full-screen application used for running the experiment, see Figure 1. When participants made a guess they received feedback in terms of (a) the number of correct digits in their correct position in the code, (b) the number of correct digits not in their correct position, and (c) the number of incorrect digits. Importantly, the feedback gave only the number of digits in each of the three categories and was devoid of information about which digits belonged to which category. Once a guess had been made it could not be changed but the guess and the associated feedback remained visible on the screen. To solve the task participants had to merge the feedback from their guesses into an understanding that gradually narrowed down the possible digit combinations for the code.

We chose this task because it is a cognitive task and sufficiently demanding to impose considerable mental workload, because its brevity allows for multiple iterations within a single session, and because we hoped its game qualities would strengthen participants' motivation and help avoid fatigue. In addition, the task allowed for introducing a distinction between two levels of time constraint:

*Timed tasks*, during which participants had a maximum of 25 seconds for each guess. This time limit was set on the basis of pilot tests. The passing of the 25 seconds was impressed upon participants by a progress bar that visualized how the elapsed time filled still more of the 25-second interval. If a participant did not make a guess within the time limit that guess was lost, the participant was moved forward to the next guess, and the progress bar restarted.

*Untimed tasks*, during which no time limits were enforced and participants could spend as much time as they needed on each guess. There was no progress bar during untimed tasks.

## *3.4 Interruptions*

We assumed the code-breaking task was sensitive to interruptions because it involved keeping track of how feedback from new guesses fitted with or forced revision of the understanding built from the feedback from

earlier guesses. To investigate the effect of interruptions on thinking aloud, the code-breaking task was interrupted every 15-25 seconds. The interruptions occupied an area in the left-hand part of the screen, see Figure 1. In between interruptions this area contained an empty bar. When an interruption occurred, participants were notified in different ways depending on the interruption type:

*Auditory interruptions* were indicated by a one-second sound.

*Visual interruptions* were indicated by the appearance of a white square in the interruption bar.

*Audiovisual interruptions* were indicated by the one-second sound and the white square.

*No interruptions*; participants performed the code-breaking task without interruptions and the interruption bar was not present.

From the onset of a notification participants had five seconds to acknowledge the interruption by clicking the interruption bar, otherwise the interruption was cancelled. The acknowledgement caused the interruption bar to expand and reveal two target figures that differed in shape and colour and a reference figure that matched one target figure in shape and the other in colour. A text below the reference figure instructed participants to "Match by shape" or "Match by colour". Participants completed the interruption by clicking the target figure consistent with the instruction. The target figures, reference figure, and instruction differed across interruptions. The interruption task is loosely based on the Stroop effect (MacLeod, 1991) and was adopted from McFarlane (2002), who noted that it cannot be automated and thus requires attention.

### *3.5 Procedure*

Participants were initially introduced to the experiment and asked questions about their background. Then, participants were explained the task and the interruptions, followed by some training tasks during which participants performed both timed and untimed tasks and experienced the different types of interruptions. Participants were instructed to attempt to complete all tasks as well as to respond to all interruptions, and they were informed that they had five seconds to acknowledge interruptions. Participants in the thinking-aloud condition were instructed about how to think aloud and practiced thinking aloud on four training tasks: (1) What is the result of multiplying 11 × 12? (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, pp. 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 only. Next, participants were introduced to the task load index (TLX) (Hart & Staveland, 1988) and explained the definitions of its six subscales. 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.

Participants performed three blocks of eight tasks, each block consisting of one instance of every combination of time constraint and interruption type. The time constraint (i.e., timed or untimed) and interruption type (i.e., auditory, visual, audiovisual, or none) were indicated on the screen ahead of each task. Tasks appeared on the screen, and the experimenter kept silent except when participants stopped talking for more than 30 seconds in the thinking-aloud condition. When this happened the experimenter reminded participants to 'keep talking'. Upon completing a task participants rated their mental workload on the six TLX subscales. After each block participants were allowed a break before they commenced on the next block. After the third block participants were debriefed.

To minimize noise in the eye-tracking data, the experiment was run in a laboratory with controlled lighting conditions. External sunlight was blocked and it was ensured that the internal light sources did not produce glare in the computer screen. The experiment lasted an average of 2.1 hours per participant. As a token of our appreciation participants received a gift certificate of DKK 350.

### *3.6 Design*

The experiment employed a mixed design with thinking-aloud condition (thinking aloud, control) as a between-subject factor and two within-subject factors: time constraint (timed, untimed) and interruption type (auditory, visual, audiovisual, none). Each of the 32 participants performed three blocks of eight tasks. Thus, the experiment comprised a total of 32 participants × 3 blocks × 2 tasks × 4 interruption types = 768 tasks.

Participants alternated between timed and untimed tasks. Half of the participants in a condition started with a timed task, the other half with an untimed task. For two consecutive tasks (i.e., a timed and an untimed) participants received the same type of interruptions, then they proceeded to the next interruption type. The order of the interruption types within a block was determined by four balanced Latin squares, one for each group of

four participants. The assignment of participants to the rows of a Latin square was rotated for the second and third blocks. The code to be broken was randomly generated for each task.

### *3.7 Dependent variables*

We measured participants' task solution rate, task completion time, interruption performance, subtask behaviour, eye movements, and mental workload.

*Task solution rate* was the number of solved tasks in percent of the total number of tasks. A task was solved if the participant broke the code; that is, if one of the participant's eight guesses exactly matched the code.

*Task completion time*, *interruption performance*, and *subtask behaviour* were determined on the basis of the log files from the code-breaking application.

*Eye movements* were recorded by a remote eye tracker from SMI, mounted below the stimulus screen and sampling at 50 Hz. A calibration process, repeated for each block of tasks, ensured that the eye tracker accurately captured the participant's line of gaze. We used the eye-tracking data to determine participants' pupil diameter, which indicate mental workload (Beatty, 1982), and fixations, which indicate attended items and mental processing (Goldberg & Kotval, 1999). As in previous studies (e.g., Bailey & Iqbal, 2008; Bernhardt, Dabbs, & Riad, 1996), the pupil-diameter measurements were converted to percentages of the participant mean. Hence, a value below 100% represents a constriction and a value above 100% represents a dilation of the pupil, relative to its average diameter across the 24 tasks. 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 reported that the dispersion-based algorithm has very good accuracy and robustness. At a viewing distance of 60-70 cm, as recommended for the eye tracker, the deviation threshold was equivalent to a fixation area with a diameter of about 11 mm on the screen participants used for solving the tasks.

*Mental workload* was measured subjectively by TLX (Hart & Staveland, 1988), which consists of six subscales: mental demand, physical demand, temporal demand, effort, performance, and frustration. The subscales were rated from low (0) to high (100) in increments of five, except for performance where the anchors were good (0) and bad (100). Participants rated the six subscales with sliders on a pop-up screen that appeared immediately after completing each task. We left out the weighting procedure for combining the six subscales 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 (Hendy, Hamilton, & Landry, 1993; Nygren, 1991).

## 4 Results

Before analyzing the data, outliers were removed to avoid that patterns in the data were masked by a small number of tasks during which participants experienced fatigue or a drop in motivation. We removed 23 (3.0%) outlier tasks, which were more than three inter-quartile ranges above the upper quartile in task completion time.

### *4.1 Task solution rate*

Table 2 shows task solution rates for the remaining 745 tasks. Overall, the task solution rates were modest, indicating that the tasks were difficult, and the standard deviations were large, indicating considerable variation across participants. We found a significant effect of block on task solution rate, $F(2, 29) = 3.31$, $p < 0.05$. Helmert contrasts showed that the task solution rate for the first block was lower than the average task solution rate for the second and third blocks, suggesting a learning effect. There was also a significant effect of time constraint on task solution rate, $F(1, 30) = 40.08$, $p < 0.001$. Unsurprisingly, the task solution rate was lower for timed than untimed tasks. We found no effect of interruption on task solution rate, $F(3, 28) = 0.64$, $p = 0.6$.

With respect to thinking aloud, we found no effect of thinking aloud on task solution rate, $F(1, 30) = 1.26$, $p = 0.3$. In addition, we found no interaction between thinking aloud and time constraint, $F(1, 30) = 0.18$, $p = 0.7$, and no interaction between thinking aloud and block, $F(2, 29) = 0.23$, $p = 0.8$. There was, however, a significant interaction between thinking aloud and interruption on task solution rate, $F(3, 28) = 3.23$, $p < 0.05$, see Figure 2. Whereas tasks solved with auditory and audiovisual interruptions had similar task solution rates irrespective of whether participants were thinking aloud (both $p$s $> 0.8$), tasks solved with visual interruptions showed a significant 62% increase in task solution rates for thinking-aloud participants compared to control participants ($p < 0.05$). Tasks solved without interruptions had task solution rates similar to auditory and audiovisual interruptions for control participants and similar to visual interruptions for thinking-aloud participants; this 26% increase was however not significant ($p = 0.2$).

### 4.2 Task completion time

To avoid that success or failure at solving a task affected the analysis of task completion times, we analyzed task completion times for successfully solved tasks only. Table 3 shows task completion times for the 396 non-outlier, successfully solved tasks. We found no effect of block on task completion time, $F(2, 29) = 1.59$, $p = 0.2$, suggesting sufficient training and absence of fatigue. As for task solution rate, there was a significant effect of time constraint on task completion time, $F(1, 30) = 49.70$, $p < 0.001$. Unsurprisingly, task completion times were lower for timed than untimed tasks. The effect of interruption on task completion time merely approached significance, $F(3, 28) = 2.68$, $p = 0.05$.

With respect to thinking aloud, there was no effect of thinking aloud on task completion time, $F(1, 30) = 0.55$, $p = 0.5$. In addition, there was no interaction between thinking aloud and time constraint, $F(1, 30) = 0.07$, $p = 0.8$, and no interaction between thinking aloud and block, $F(2, 29) = 0.52$, $p = 0.6$. There was, however, a significant interaction between thinking aloud and interruption on task completion time, $F(3, 28) = 3.41$, $p < 0.05$, see Figure 3. While task completion times for thinking-aloud participants were near identical for tasks with auditory, visual, and no interruptions, they were spread out for control participants and 9%, 22%, and 32% lower, respectively. The 32% lower task completion times for tasks without interruptions approached a significant difference between control participants and thinking-aloud participants ($p = 0.06$). Task completion times for tasks with audiovisual interruptions were near identical for thinking-aloud participants and control participants, and they were 14% lower than thinking-aloud participants' task completion times for the other tasks.

### 4.3 Interruptions

We analyzed the 2720 interruptions that occurred during the 745 non-outlier tasks. Table 4 shows the percentage of interruptions to which participants responded, the response time, and the percentage of correct responses. Before conducting the statistical analysis the percentages of responses and correct responses were arcsine transformed because high percentage values are susceptible to ceiling effects and cannot be assumed normally distributed (Fleiss, 1981). There was a significant effect of thinking aloud on response time, $F(1, 30) = 4.67$, $p < 0.05$, with thinking-aloud participants needing more time to respond to interruptions than control participants. We found no effects of thinking aloud on response rate and solution rate, $F$s$(1, 30) = 0.47, 0.33$, respectively (both $p$s $> 0.4$). In addition, we found no interactions between thinking aloud and any of time constraint (all $p$s $> 0.6$), interruptions (all $p$s $> 0.1$), and block (all $p$s $> 0.2$) for response rate, response time, and solution rate.

### 4.4 Eye movements

To assign equal weight to each guess the eye movements were analyzed per guess. In addition, the number of fixations in a task was strongly correlated with task completion time ($r = 0.95$, $p < 0.001$), and 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. Eye movements were analyzed for the 745 non-outlier tasks, see Table 5. We found no effects of thinking aloud for fixation rate, fixation duration, pupil diameter, and saccade length, $F$s$(1, 30) = 0.20, 0.08, 1.43, 1.15$, respectively (all $p$s $> 0.2$). For fixation rate there was, however, a significant interaction between thinking aloud and interruption, $F(3, 28) = 2.76$, $p < 0.05$. While the fixation rates for auditory, audiovisual, and visual interruptions were similar for thinking-aloud (mean of 2.09 $s^{-1}$) and control (mean of 2.11 $s^{-1}$) participants, the fixation rate for no interruptions was 5% lower for thinking-aloud (1.99 $s^{-1}$) than control (2.09 $s^{-1}$) participants.

For the three other eye-movement measures – fixation duration, pupil diameter, and saccade length – we found no interactions between thinking aloud and interruption, $F$s$(3, 28) = 1.25, 0.54, 0.87$, respectively (all $p$s $> 0.2$). In addition, we found no interactions between thinking aloud and time constraint (all $p$s $> 0.3$) or block (all $p$s $> 0.08$) for any of the four eye-movement measures.

### 4.5 Subtask behaviour

To investigate differences in the participants' progress on the tasks we analyzed the number of code guesses submitted during a task and, for timed tasks only, the number of code guesses that timed out before the participant submitted a guess. Table 6 shows the mean values of these measures for non-outlier tasks. We found no effects of thinking aloud for number of code guesses and number of timed-out code guesses, $F$s$(1, 30) = 0.61, 0.009$, respectively (both $p$s $> 0.4$), and no interactions between thinking aloud and any of time constraint (both $p$s $> 0.3$), interruptions (both $p$s $> 0.1$), and block (both $p$s $> 0.5$).

To further illustrate the participants' subtask behaviour, Figure 4 shows their guess-by-guess performance in terms of guess duration, guess accuracy, pupil diameter, and fixation rate. The guess accuracy was determined by converting the feedback received for each guess to the percentage of code-guess combinations ruled out by this feedback. A correlation analysis of the 745 non-outlier tasks confirmed that thinking-aloud and control participants performed similarly on all four measures: The variation in the thinking-aloud participants'

performance explained 90%, 77%, 94%, and 68% of the variation in the control participants' performance for guess duration, guess accuracy, pupil diameter, and fixation rate, respectively (all $p$s < 0.05).

### *4.6 Mental workload*

Table 7 shows mental workload for the 745 non-outlier tasks. A multivariate analysis of variance of the six TLX subscales showed the intended increase in mental workload for timed compared to untimed tasks, Wilks' $\lambda$ = 0.22, $F(6, 25) = 14.81$, $p < 0.001$. It showed no effect of thinking aloud on mental workload, Wilks' $\lambda = 0.83$, $F(6, 25) = 0.86$, $p = 0.6$, and no interaction between thinking aloud and any of time constraint, Wilks' $\lambda = 0.89$, $F(6, 25) = 0.51$, $p = 0.8$, interruptions, Wilks' $\lambda = 0.48$, $F(18, 13) = 0.78$, $p = 0.7$, and block, Wilks' $\lambda = 0.72$, $F(12, 19) = 0.61$, $p = 0.8$. Similarly, we found no effect of thinking aloud for any individual TLX subscale, and for only one subscale (mental demand) did the interaction between thinking aloud and block approach significance, $F(2, 29) = 3.09$, $p = 0.06$. While mental demand for thinking-aloud participants was near constant during all three blocks, control participants appeared to experience a 13% decrease from the first to second block and a 6% increase from the second to third block.

## 5 Discussion

In the following we discuss how interruptions and time constraints affect the reactivity of thinking aloud, discuss implications of our study for the use of thinking aloud in usability evaluation, and note limitations of the study.

### *5.1 Thinking aloud is reactive in the presence of interruptions*

All significant effects of thinking aloud in this study are related to interruptions, none to time constraints. First, thinking-aloud participants spend more time from they are notified of an interruption to they complete it by clicking a target figure. This result appears consistent with the established finding that thinking aloud slows down performance (Ericsson & Simon, 1993; Fox et al., 2011) but it is inconsistent with Karbach and Kray's (2007) finding that children's response time to interruptions is either faster (5-year-old children) or unaffected (9-year-old children) when they are thinking aloud. Two possible explanations for this main effect of thinking aloud are that thinking-aloud participants take longer to notice interruptions or that they need longer to switch from the code-breaking task to the interruptions after having noticed them. While the former explanation suggests that thinking aloud make participants attend more exclusively to the code-breaking task, thereby leaving less attention for noticing interruptions, the latter explanation suggests that thinking aloud increases the task-switching costs. Increased task-switching costs (Monsell, 2003; Pashler, 2000) point toward additional mental processes, rather than merely slower performance, to unload the current task in a manner that facilitates later resumption and to load the new task. Our results combined with those of Karbach and Kray (2007) indicate that thinking aloud evolves from an aid in task switching for young children to the opposite for adults, suggesting that thinking aloud may be differentially applicable for children and adults in situations that require fast response to frequent interruptions.

Second, the presence or absence of thinking aloud affects performance on the code-breaking task differently depending on the interruption type. Thinking aloud and interruptions interact with regard to task solution rate, task completion time, and fixation rate. For task solution rate, thinking-aloud participants perform better than control participants during visual interruptions but display no difference for auditory, audiovisual, and no interruptions. For task completion time, auditory, audiovisual, and visual interruptions appear to dampen the slowdown incurred by thinking aloud compared to the slowdown experienced when tasks are solved without interruptions. This dampening is largest for audiovisual interruptions, smaller for auditory interruptions, and smallest for visual interruptions. For fixation rate, thinking-aloud participants make fewer fixations a second than control participants when they perform without interruptions but the same number of fixations when they perform in the presence of interruptions, irrespective of whether the interruptions are auditory, audiovisual, or visual. A candidate explanation for such interactions could be that auditory interruptions make thinking aloud reactive because both thinking aloud and the processing of these interruptions involve the auditory channel, which becomes overloaded. There is some, but not much, support for this explanation in the data in that the solution rate during auditory interruptions displays a downward trend and the completion time an upward trend for thinking-aloud compared to control participants, see Figures 2 and 3. Conversely, visual interruptions should not be expected to affect the reactivity of thinking aloud because the extra demand on participants' visual attention to check periodically whether an interruption has occurred can be performed in parallel with verbalizations. The task completion times provide some support for this explanation in that the completion times during visual interruptions are those most similar to the completion times during no interruptions, but the task solution rates do not support the explanation because thinking aloud improves task solution rates during visual interruptions. Interestingly, the effect of audiovisual interruptions on how thinking aloud affects performance is more like the effect of auditory than visual interruptions. Thus, participants are disrupted more by the auditory cues in audiovisual interruptions than they are able to benefit from the visual cues. Such a negative effect of

redundant modalities is frequent in studies of audiovisual alarms (e.g., Sanderson, Crawford, Savill, Watson, & Russell, 2004; Seagull, Wickens, & Loeb, 2001) and is normally explained by a human bias toward visual cues even when available auditory cues may result in better performance (Posner & Nissen, 1976). This explanation is, however, not consistent with our data.

Third, the fixation rate is related to the number of components a participant needs to process to solve a task (Goldberg & Kotval, 1999). The interaction between thinking aloud and fixation rate shows that participants who think aloud but are not interrupted process fewer components a second to solve the code-breaking tasks than participants who perform in silence, are interrupted, or both. As the code-breaking tasks are similar, processing fewer components probably means revisiting fewer components. It appears reasonable that interruptions increase the need to revisit components. The reduced need to revisit components during thinking aloud suggests a more systematic mental process. While a performance-enhancing systematization of the mental process is a frequently reported effect of verbalization at level 3 (Chi, De Leeuw, Chiu, & LaVancher, 1994; Fox et al., 2011), it discords with previous studies of thinking aloud (e.g., Hertzum et al., 2009). The interpretation of the fixation rate as an indication of a more systematic mental process while thinking aloud is, however, not corroborated by the measures of the subtask behaviour, see Figure 4. Thus, the fewer revisits are insufficient to increase guess accuracy, decrease guess duration, or reduce mental workload as measured by the pupil diameter.

### *5.2 Time constraints are unlikely to make thinking aloud reactive*

The absence or presence of time constraints had no effect on whether thinking aloud affected performance. Initially, it is worth noting that our time-constraint manipulation worked in that task solution rates were lower, task completion times lower, and mental workload higher for timed than untimed tasks. On this basis we contend that the absence of interactions between thinking aloud and time constraint cannot be explained away by claiming that participants were not under increased time pressure during timed tasks. Thus, our results strengthen those of Dickson et al. (2000), who also find that thinking aloud does not interact with time constraints but provide no manipulation check to verify that their time constraint increased time pressure. We agree with Fox et al. (2011) in their assertion that when participants are under time pressure, the prolonged task times associated with thinking aloud must be expected to degrade performance. It is, therefore, surprising that thinking aloud and time constraint do not interact. A candidate explanation is that the time pressure must be even higher before an interaction emerges. The ratio of timed to untimed task completion times is however 52%, indicating a substantial time pressure. An alternative explanation could be that at time pressures of this magnitude performance suffers irrespective of whether or not participants think aloud and that the additional time required for thinking aloud does not significantly change how much performance suffers. This explanation fits our data. It proposes that thinking aloud is only reactive in the presence of time constraints if participants have sufficient time for performing a task when they are not thinking aloud but insufficient time when it is performed while thinking aloud. As the interval satisfying this criterion is narrow, thinking aloud will in most cases not become reactive in the presence of time constraints.

### *5.3 Subtask behaviour and mental workload are unaffected by thinking aloud*

None of our measures of participants' subtask behaviour and mental workload are affected by thinking aloud. Rather, we find that for guess duration, guess accuracy, pupil diameter, and fixation rate the variation in thinking-aloud participants' performance explains as much as 68-94% of the variation in control participants' performance. In addition, mental workload neither shows an effect of thinking aloud for task-level mental workload measured by TLX, nor for moment-by-moment mental workload measured by pupil diameters. Our finding that subtask behaviour and mental workload are unaffected by thinking aloud for the demanding task of code breaking extends previous studies of thinking aloud in usability evaluation, in which Hertzum et al. (2009) report similar findings for the simpler task of web navigation and Olmsted-Hawala et al. (2010) find no evidence of reactivity, also for web-navigation tasks. Conversely, Haak, Jong, and Schellens (2003) report lower task solution rates for thinking-aloud participants in a study of an online library catalogue, but it appears that the participants received neither instructions nor training in how to think aloud. Cooke (2010) reports evidence of the accuracy of thinking-aloud data in that the majority of the verbalizations made during thinking aloud matched on-screen words at which the users were fixating during a usability evaluation. This suggests that in the absence of interruptions thinking aloud is not reactive, and it indicates that even in the presence of interruptions important aspects of the task process remain unaffected by thinking aloud.

### *5.4 Implications*

The results of our study have several implications for research and practice. First, interruptions and multitasking are rare in contemporary usability evaluation, which instead tends to have users perform tasks one at a time (Dumas & Redish, 1999; Rubin & Chisnell, 2008). This poses a threat to the validity of usability evaluation because interruptions and multitasking are common in many real-world settings. However, extending usability

evaluation with interruptions – by either simulating interruptions in the laboratory or moving evaluations to the field – makes thinking aloud reactive and thus poses a new threat to evaluation quality. To avoid this reactivity it may be considered to use retrospective thinking aloud in which users perform tasks without thinking aloud and thereafter think aloud while watching a video recording of their task performance. Retrospective thinking aloud appears to provide valid verbalizations for short tasks (Ericsson & Simon, 1993) and to be as effective as concurrent thinking aloud in supporting the identification of usability problems (Haak, Jong, & Schellens, 2007), but it comes at the cost of twice as long sessions.

Second, we propose that thinking aloud is not reactive in the presence of time constraints, except in the narrow interval in which participants have sufficient time when they are not thinking aloud but insufficient time when thinking aloud. This suggests that thinking aloud is less sensitive to time constraints than previously assumed and, thereby, that it can be applied when the usability of a system must be evaluated in situations characterized by time pressure. An extension of the applicability of thinking aloud to include timed tasks is important to practitioners because time pressure is frequent in many use situations, including computer games, emergency management, and process control. The importance of evaluating systems under realistic time pressures is emphasized by Kokini, Lee, Koubek, and Moon (forthcoming), who find that time constraints lead to lower perceived usability. In terms of research implications the proposed relation between thinking aloud and time constraints calls for investigating the possible reactivity of thinking aloud at different levels of time constraint to determine whether the presence and absence of reactivity follow the proposed pattern.

Third, the absence of main effects of thinking aloud, except for the response time to interruptions, accords with previous findings that thinking aloud is normally not reactive. This expounds the trade-off faced by practitioners when they decide between either restricting verbalizations in usability evaluation to thinking aloud or including verbalizations of reasons and feelings. The former can normally be assumed to be non-reactive, the latter is reactive but provides additional information. If verbalizations are to be restricted to thinking aloud, then proper instructions and user training (see, e.g., Ericsson & Simon, 1993; Fox et al., 2011) are essential. If verbalizations of reasons and feelings are imperative, practitioners may consider obtaining them retrospectively while users are watching a video recording of their task performance.

Fourth, interruptions can differ in multiple ways and future research should detail the interruption characteristics that make thinking aloud reactive. This study shows that it matters whether the interrupt notification is auditory, visual or audiovisual. Another important characteristic of interruptions is whether they are externally imposed by, for example, alarms and notifications or internally generated by users who interrupt themselves by switching between multiple tasks. If thinking aloud is also reactive in the presence of internally generated interruptions, the scope of the reactivity is considerably increased because both types of interruptions are frequent in practice (González & Mark, 2004; Spencer et al., 2004; Trafton & Monk, 2007). The possible effect of other characteristics of interruptions, such as their length and complexity, is also worth investigating. Finally, we are intrigued by the possibility that the applicability of thinking aloud in situations that require fast response to frequent interruptions may be age dependent because children and adults appear to differ in the mental costs they experience when switching between interruptions and interrupted tasks while thinking aloud.

### *5.5 Limitations*

Three limitations should be remembered in interpreting the results of this study. First, while our subjective experience from the experimental sessions is that participants complied with Ericsson and Simon's (1993) prescriptions for thinking aloud, we acknowledge the absence of a control variable to verify that participants verbalized at levels 1 and 2 only. We are unaware of a study that has defined and employed such a control variable. Second, participants experienced the code-breaking task 24 times. This is an artificial situation compared to most real-world settings, though similar to how games are often played. Our analysis of participants' performance across blocks shows that the task did not become trivial and that fatigue was not a problem. We contend that the game qualities of the code-breaking task were instrumental in maintaining participants' motivation. Third, the task in this study is a problem-solving task and thereby differs from common usability-evaluation tasks such as web navigation. Our results may not be directly transferable to evaluations of systems that aim to support users in navigation and information-seeking tasks. Instead, the demands of our task are in many respects similar to those faced by process-control operators, computer-game players, and people in many multi-tasking environments.

## 6 Conclusion

We find that thinking aloud is reactive in the presence of interruptions. Thinking aloud interacts with interruptions on the two central performance measures task solution rate and task completion time. In addition, thinking aloud prolongs the time for responding to interruptions. Participants are disrupted by the simultaneous presence of thinking aloud and auditory interruptions, they benefit from the simultaneous presence of thinking

aloud and visual interruptions, and they experience audiovisual interruptions more like auditory than visual interruptions. With respect to time constraints, we find no reactivity of thinking aloud in the presence of time constraints. This is somewhat surprising because a slowdown in performance is an acknowledged effect of thinking aloud. We propose that time constraints may only make thinking aloud reactive in the usually narrow interval in which participants have sufficient time when not thinking aloud but insufficient time when thinking aloud. Further work is required to assess this proposition; in this study both thinking-aloud and control participants tended to have insufficient time. The main implication of this study is that thinking aloud appears to be reactive in situations with a realistic number of interruptions and, hence, deficient in evaluations of the usability of many systems in some of their frequent and critical use situations.

## Acknowledgements

The first author provided the initial idea for the study, analyzed the data, and wrote most of the article. The second author planned, set up, and ran the experimental sessions. The authors took equal part in refining the initial idea for the study, and both authors critically read and revised draft versions of the article. We are grateful to Alexandre Alapetite and Steen Weber, who helped solve several issues about the use of the eye-tracking equipment, and to Signe Arnklit, who recruited the participants for the experiment. Special thanks are due to the participants.

## References

Bailey, B. P., & Iqbal, S. T. (2008). Understanding changes in mental workload during execution of goal-directed tasks and its application for interruption management. *ACM Transactions on Computer-Human Interaction, 14*(4), 21:01-21:28.

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin, 91*(2), 276-292.

Bernhardt, P. C., Dabbs, J. M., & Riad, J. K. (1996). Pupillometry system for use in social psychology. *Behavior Research Methods, Instrumentation, & Computers, 28*(1), 61-66.

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

Chi, M. T. H., De Leeuw, N., Chiu, M.-H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18*(3), 439-477.

Cooke, L. (2010). Assessing concurrent think-aloud protocol as a usability test method: A technical communication approach. *IEEE Transactions on Professional Communication, 53*(3), 202-215.

Dickson, J., McLennan, J., & Omodei, M. M. (2000). Effects of concurrent verbalization on a time-critical, dynamic decision-making task. *Journal of General Psychology, 127*(2), 217-228.

Dumas, J. S., & Fox, J. E. (2008). Usability testing: Current practice and future directions. In A. Sears & J. A. Jacko (Eds.), *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies, and Emerging Applications* (Second ed., pp. 1129-1149). New York: Erlbaum.

Dumas, J. S., & Redish, J. C. (1999). *A practical guide to usability testing. Revised edition*. Exeter, UK: Intellect Books.

Emerson, M. J., & Miyake, A. (2003). The role of inner speech in task switching: A dual-task investigation. *Journal of Memory and Language, 48*, 148-168.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review, 87*(3), 215-251.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data. Revised edition*. Cambridge, MA: MIT Press.

Fleiss, J. (1981). *Statistical methods for rates and proportions*. New York: Wiley.

Fox, M. C., Ericsson, K. A., & Best, R. (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.

Gilhooly, K. J., Fioratou, E., & Henretty, N. (2010). Verbalization and problem solving: Insight and spatial factors. *British Journal of Psychology, 101*(1), 81-93.

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

González, V. M., & Mark, G. (2004). "Constant, constant, multi-tasking craziness": Managing multiple working spheres *Proceedings of the CHI 2004 Conference on Human Factors in Computing Systems* (pp. 113-120). New York: ACM Press.

Gulliksen, J., Boivie, I., Persson, J., Hektor, A., & Herulf, L. (2004). Making a difference - A survey of the usability profession in Sweden *Proceedings of the NordiCHI 2004 Conference on Human-Computer Interaction* (pp. 207-215). New York: ACM Press.

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

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

Hendy, K., Hamilton, K. M., & Landry, L. N. (1993). Measuring subjective workload: When is one scale better than many? *Human Factors, 35*(4), 579-601.

Hertzum, M., Hansen, K. D., & Andersen, H. H. K. (2009). Scrutinising usability evaluation: Does thinking aloud affect behaviour and mental workload? *Behaviour & Information Technology, 28*(2), 165-181.

Hertzum, M., & Holmegaard, K. D. (2012). Perceived time as a measure of mental workload: Effects of time constraints and task success. *To appear in International Journal of Human-Computer Interaction*.

Haak, M. J. v. d., Jong, M. D. T. d., & 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.

Haak, M. J. v. d., Jong, M. D. T. d., & Schellens, P. J. (2007). Evaluation of an informational web site: Three variants of the think-aloud method compared. *Technical Communication, 54*(1), 58-71.

James, W. (1890). *The principles of psychology*. New York: Dover.

Karbach, J., & Kray, J. (2007). Developmental changes in switching between mental task sets: The influence of verbal labeling in childhood. *Journal of Cognition and Development, 8*(2), 205-236.

Kokini, C. M., Lee, S., Koubek, R. J., & Moon, S. K. (forthcoming). Considering context: The role of mental workload and operator control in users' perceptions of usability. *International Journal of Human-Computer Interaction*.

Lewis, C. (1982). Using the "thinking-aloud" method in cognitive interface design, RC 9265 (#40713). Yorktown Heights, NY: IBM Thomas Watson Research Center.

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin, 109*(2), 163-203.

McFarlane, D. C. (2002). Comparison of four primary methods for coordinating the interruption of people in human-computer interaction. *Human-Computer Interaction, 17*(1), 63-139.

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

Mintzberg, H. (1975). The manager's job: Folklore and fact. *Harvard Business Review, 53*(4), 49-61.

Monsell, S. (2003). Task switching. *Trends in Cognitive Sciences, 7*(3), 134-140.

Nielsen, J. (1993). *Usability engineering*. Boston, MA: Academic Press.

Nisbett, R. E., & 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.

Olmsted-Hawala, E. L., Murphy, E. D., Hawala, S., & Ashenfelter, K. T. (2010). Think-aloud protocols: A comparison of three think-aloud protocols for use in testing data-dissemination web sites for usability *Proceedings of the CHI 2010 Conference on Human Factors in Computing Systems* (pp. 2381-2390). New York: ACM Press.

Pashler, H. (2000). Task switching and multitask performance. In S. Monsell & J. Driver (Eds.), *Control of Cognitive Processes: Attention and Performance XVIII* (pp. 277-307). Cambridge, MA: MIT Press.

Patrick, J., Gregov, A., Halliday, P., Handley, J., & O'Reilly, S. (1999). Analysing operators' diagnostic reasoning during multiple events. *Ergonomics, 42*(3), 493-515.

Posner, M. I., & Nissen, M. J. (1976). Visual dominance: An information-processing account of its origins and significance. *Psychological Review, 83*(2), 157-171.

Rubin, J., & Chisnell, D. (2008). *Handbook of usability testing: How to plan, design, and conduct effective tests* (Second ed.). Indianapolis, IN: Wiley.

Russo, J. E., Johnson, E. J., & Stephens, D. L. (1989). The validity of verbal protocols. *Memory & Cognition, 17*(6), 759-769.

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

Sanderson, P., Crawford, J., Savill, A., Watson, M., & Russell, W. J. (2004). Visual and auditory attention in patient monitoring: A formative analysis. *Cognition, Technology & Work, 6*(3), 172-185.

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

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

Seagull, F. J., Wickens, C. D., & Loeb, R. G. (2001). When is less more? Attention and workload in auditory, visual, and redundant patient-monitoring conditions *Proceedings of the Human Factors and Ergonomics Society 45th Annual Meeting* (pp. 1395-1399). Santa Monica, CA: HFES.

Spencer, R., Coiera, E., & Logan, P. (2004). Variation in communication loads on clinical staff in the emergency department. *Annals of Emergency Medicine, 44*(3), 268-273.

Trafton, J. G., Altmann, E. M., Brock, D. P., & Mintz, F. E. (2003). Preparing to resume an interrupted task: Effects of prospective goal encoding and retrospective rehearsal. *International Journal of Human-Computer Studies, 58*(5), 583-603.

Trafton, J. G., & Monk, C. A. (2007). Task interruptions. *Reviews of Human Factors and Ergonomics, 3*(1), 111-126.

Venturi, G., Troost, J., & Jokela, T. (2006). People, organizations, and processes: An inquiry into the adoption of user-centred design in industry. *International Journal of Human-Computer Interaction, 21*(2), 219-238.

Vredenburg, K., Mao, J.-Y., Smith, P. W., & Carey, T. (2002). A survey of user-centered design practice *Proceedings of the CHI 2002 Conference on Human Factors in Computing Systems* (pp. 471-478). New York: ACM Press.

Vygotsky, L. S. (1988). On inner speech. In M. B. Franklin & S. S. Barten (Eds.), *Child Language: A Reader* (pp. 181-187). Oxford, UK: Oxford University Press.

Wilson, T. D., & 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.

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

**Table 1**. Participants

| | Control | Thinking aloud |
|---|---|---|
| Gender | | |
| Female | 6 | 7 |
| Male | 10 | 9 |
| Years of age | | |
| Mean | 25.4 | 25.1 |
| Range | 21 - 33 | 20 - 37 |
| Background | | |
| Student | 13 | 13 |
| Professional | 2 | 3 |
| Not reported | 1 | 0 |
| Computer use | | |
| Every day | 14 | 15 |
| Weekly | 1 | 1 |
| Not reported | 1 | 0 |
| Play computer games | | |
| Yes | 9 | 8 |
| No | 6 | 8 |
| Not reported | 1 | 0 |

**Table 2**. Task solution rates (in percent), $N = 745$ non-outlier tasks

| | Control | | Thinking aloud | |
|---|---|---|---|---|
| | *Mean* | *SD* | *Mean* | *SD* |
| Block | | | | |
| Block 1 | 43 | 24 | 53 | 24 |
| Block 2 | 50 | 26 | 55 | 27 |
| Block 3 | 53 | 23 | 64 | 23 |
| Time constraint | | | | |
| Timed | 38 | 20 | 44 | 24 |
| Untimed | 62 | 27 | 72 | 25 |
| Interruption * | | | | |
| Auditory | 55 | 27 | 52 | 27 |
| Visual | 39 | 26 | 63 | 27 |
| Audiovisual | 52 | 28 | 51 | 27 |
| None | 50 | 26 | 63 | 27 |

* $p < 0.05$ (interaction effect of thinking aloud and interruption)

**Table 3**. Task completion times (in seconds), $N = 396$ non-outlier, correctly solved tasks

| | Control | | Thinking aloud | |
|---|---|---|---|---|
| | *Mean* | *SD* | *Mean* | *SD* |
| Block | | | | |
| Block 1 | 97.04 | 29.83 | 104.78 | 47.26 |
| Block 2 | 81.03 | 46.07 | 100.00 | 48.92 |
| Block 3 | 85.75 | 36.23 | 97.90 | 57.15 |
| Time constraint | | | | |
| Timed | 55.23 | 17.11 | 61.39 | 19.98 |
| Untimed | 105.02 | 42.10 | 119.01 | 55.27 |
| Interruption * | | | | |
| Auditory | 96.88 | 42.47 | 106.36 | 61.40 |
| Visual | 82.76 | 38.45 | 106.37 | 72.82 |
| Audiovisual | 90.50 | 41.21 | 92.32 | 56.46 |
| None | 73.77 | 43.37 | 108.23 | 53.85 |

* $p < 0.05$ (interaction effect of thinking aloud and interruption)

**Table 4**. Interruption measures, $N = 2720$ interruptions (row 1) and 2475 interruption responses (rows 2, 3)

| | Control | | Thinking aloud | |
|---|---|---|---|---|
| | *Mean* | *SD* | *Mean* | *SD* |
| Response rate (percent) | 96 | 7 | 91 | 10 |
| Response time (seconds) * | 3.38 | 0.51 | 3.71 | 0.37 |
| Solution rate (percent) | 98 | 5 | 99 | 2 |

* $p < 0.05$ (main effect of thinking aloud)

**Table 5**. Eye-movement measures, $N = 745$ non-outlier tasks

| | Control | | Thinking aloud | |
|---|---|---|---|---|
| | *Mean* | *SD* | *Mean* | *SD* |
| Fixation rate (fixations/s) * | 2.10 | 0.21 | 2.06 | 0.20 |
| Fixation duration (ms) | 441 | 49 | 436 | 37 |
| Pupil diameter (percent) | 100.5 | 1.5 | 99.9 | 1.3 |
| Saccade length (pixels) | 100 | 9 | 105 | 13 |

* $p < 0.05$ (interaction effect of thinking aloud and interruption)

**Table 6**. Subtask-behaviour measures, $N = 745$ non-outlier tasks (row 1) and 384 non-outlier, timed tasks (row 2)

| | Control | | Thinking aloud | |
|---|---|---|---|---|
| | *Mean* | *SD* | *Mean* | *SD* |
| Guess count | 6.66 | 0.76 | 6.49 | 0.73 |
| Guess timeouts | 0.28 | 0.30 | 0.29 | 0.43 |

**Table 7**. Mental workload (measured by TLX), $N = 745$ non-outlier tasks

| | Control | | Thinking aloud | |
|---|---|---|---|---|
| | *Mean* | *SD* | *Mean* | *SD* |
| Mental demand | 62 | 10 | 56 | 18 |
| Physical demand | 9 | 9 | 4 | 8 |
| Temporal Demand | 43 | 11 | 41 | 15 |
| Effort | 55 | 13 | 51 | 16 |
| Performance | 40 | 13 | 40 | 14 |
| Frustration | 47 | 15 | 41 | 18 |

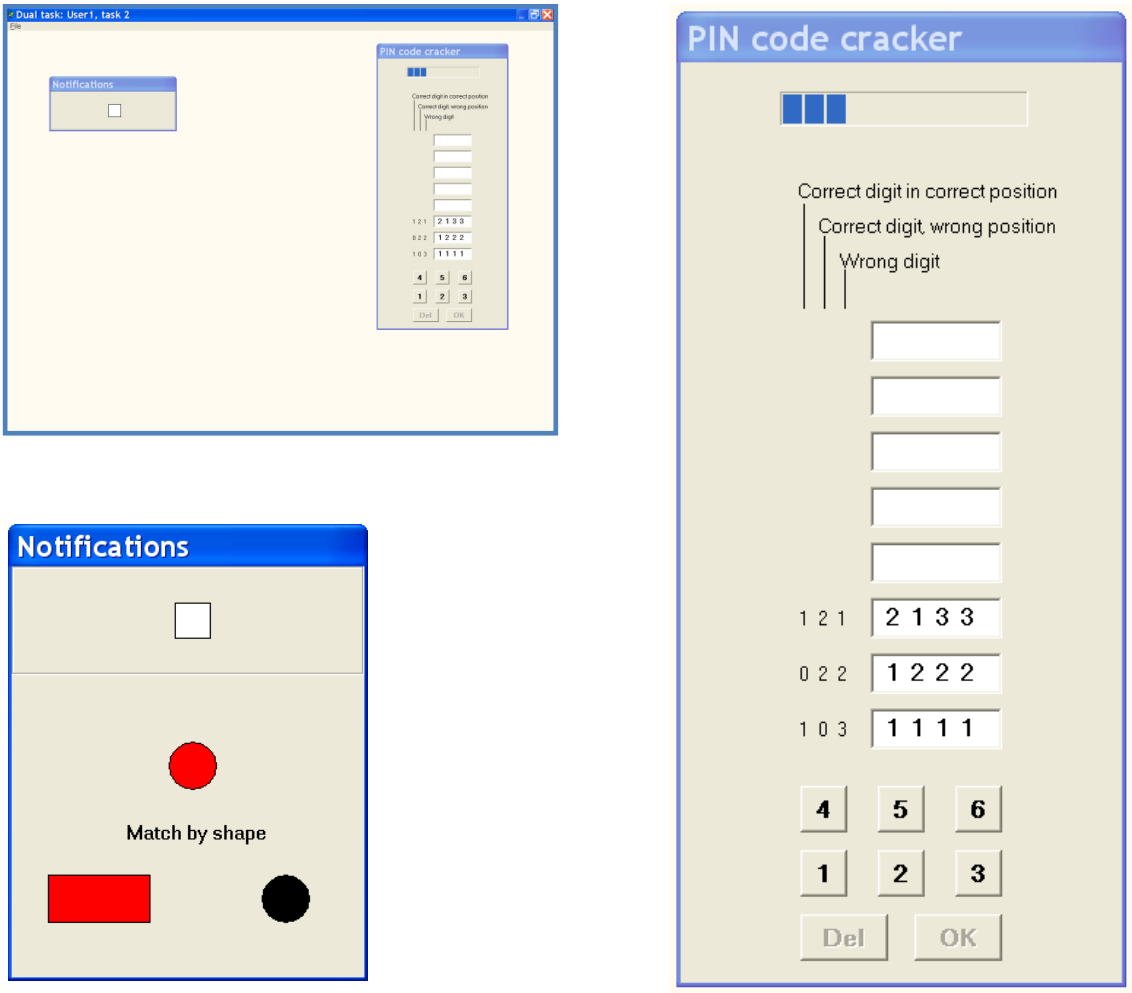


**Figure 1**. The full-screen application (top left) with the bar on the left visually notifying the participant of an interruption and the window on the right for the code-breaking task. The interrupting task (bottom left) after the participant has acknowledged the interruption. The code-breaking task (right) after the participant has just completed the third guess of a timed task with the code 2361.

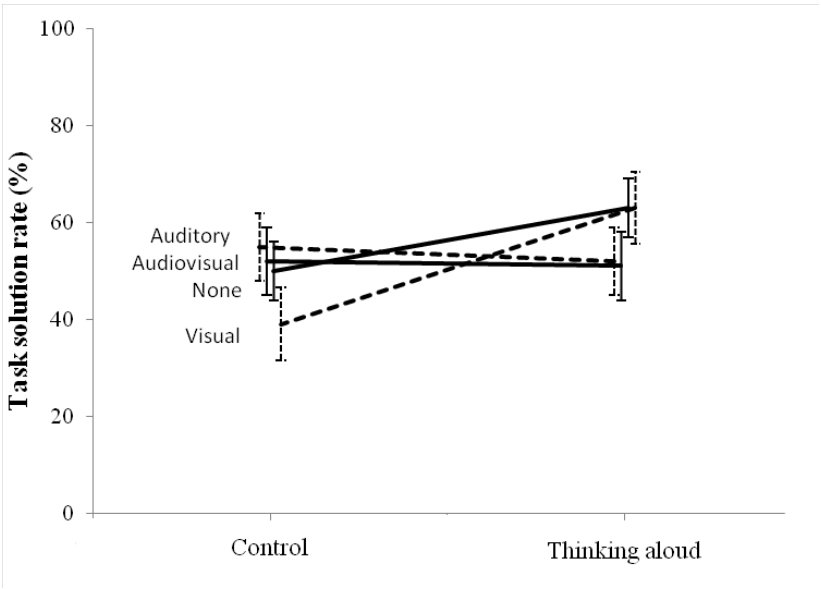


**Figure 2**. Task solution rates; error bars show standard error of the mean, $N = 745$ non-outlier tasks

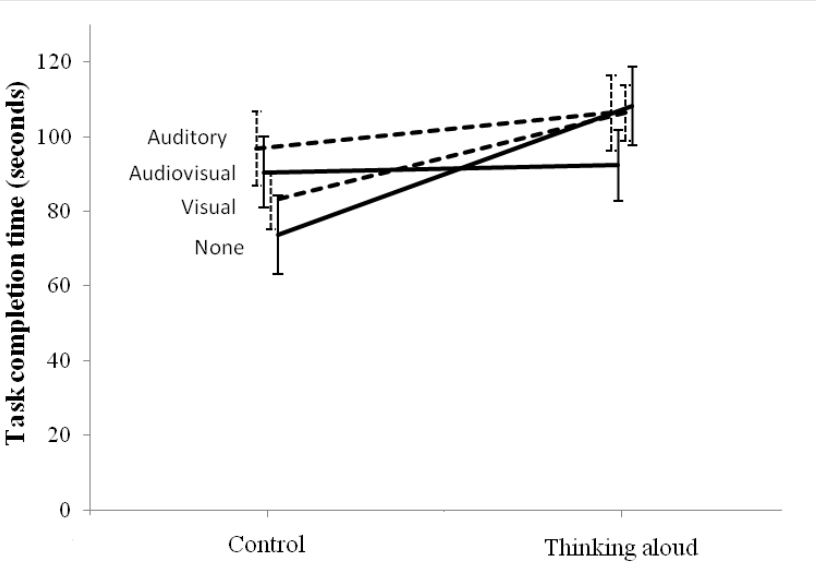


**Figure 3**. Task completion times; error bars show standard error of the mean, $N = 396$ tasks

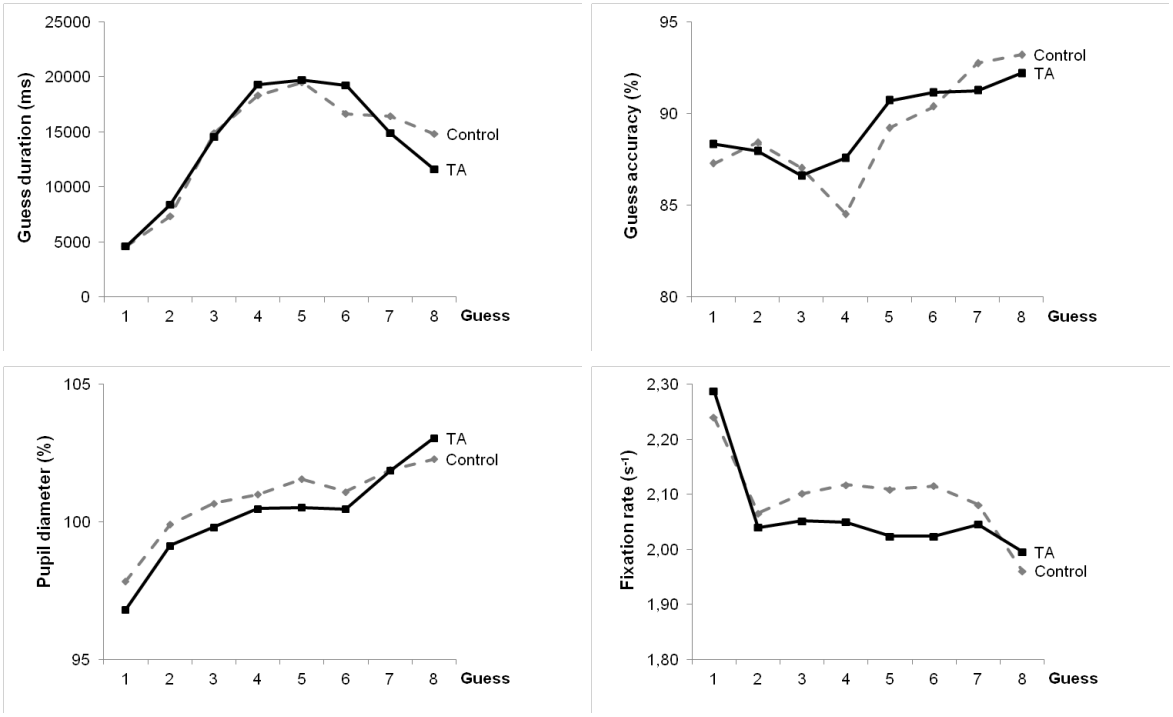


**Figure 4**. Subtask behaviour as it evolved for thinking-aloud (TA) and control participants over the eight guesses in terms of guess duration (top left), guess accuracy (top right), pupil diameter (bottom left), and fixation rate (bottom right), $N = 745$ non-outlier tasks.