Human Factors, vol. 57, no. 1 (2015), pp. 101-109.
Preprint version

Thinking Aloud Influences Perceived Time

Morten Hertzum and Kristin Due Holmegaard

Roskilde University, Roskilde, Denmark

Abstract

Objective: We investigate whether thinking aloud influences perceived time.

Background: Thinking aloud is widely used in usability evaluation, yet it is debated whether thinking aloud influences thought and behavior. If thinking aloud is restricted to the verbalization of information to which a person is already attending, there is evidence that thinking aloud does not influence thought and behavior.

Method: In an experiment, 16 thinking-aloud participants and 16 control participants solved a code-breaking task 24 times each. Participants estimated task duration. The 24 trials involved two levels of time constraint (timed, untimed) and resulted in two levels of success (solved, unsolved).

Results: The ratio of perceived time to clock time was lower for thinking-aloud than control participants. Participants overestimated time by an average of 47% (thinking aloud) and 94% (control). The effect of thinking aloud on time perception also held separately for timed, untimed, solved, and unsolved trials.

Conclusion: Thinking aloud (verbalization at Levels 1 and 2) influences perceived time. Possible explanations of this effect include that thinking aloud may require attention, cause a processing shift that overshadows the perception of time, or increase mental workload.

Application: For usability evaluation, this study implies that time estimates made while thinking aloud cannot be compared with time estimates made while not thinking aloud, that ratings of systems experienced while thinking aloud may be inaccurate (because the experience of time influences other experiences), and that it may therefore be considered to replace concurrent thinking aloud with retrospective thinking aloud when evaluations involve time estimation.

Keywords: verbalization, thinking aloud, perceived time, usability evaluation, attention

INTRODUCTION

Thinking aloud enables usability evaluators to listen in on users' thoughts while observing their behavior and, thereby, enriches the basis for evaluating when and why users experience problems. This makes thinking aloud an important usability evaluation method (Dumas & Redish, 1999; Lewis, 2006; Rubin & Chisnell, 2008), provided that thinking aloud does not alter users' thoughts and behavior. Extensive work has found that under certain conditions thinking aloud does not influence thinking and behavior, except by possibly prolonging task performance (Ericsson & Simon, 1993; Fox, Ericsson, & Best, 2011). Central to these conditions is that thinking aloud be restricted to users' verbalization of information they are already heeding in order to perform their tasks, and that users do not add to this information by verbalizing reasons or other reflections on their behavior. Yet, concerns remain about whether these conditions are appropriate for usability evaluation (Boren & Ramey, 2000) and whether they succeed in preventing effects of thinking aloud on task performance

(Gilhooly, Fioratou, & Henretty, 2010; Haak, Jong, & Schellens, 2003). This study addresses the latter issue.

Multiple variants of thinking aloud are employed in usability evaluation, leading to the verbalization of different types of information. We focus on the variant specified by Ericsson and Simon (1993) as consisting of the verbalization of information that is already in attention in verbal or nonverbal form (i.e., verbalization at Levels 1 and 2, respectively). Ericsson and Simon (1993) argued that when thinking aloud is restricted to these types of verbalizations, it provides an accurate record of the thought process without altering it. Defined in this way, thinking aloud specifically excludes probing test users about their opinions, reasons, and other reflections because it is well-established that the verbalizations resulting from such probing (Level 3 verbalizations) distort thought processes and change behavior (Fox et al., 2011; Hertzum, Hansen, & Andersen, 2009; Wilson & Schooler, 1991). We hypothesize that thinking aloud while performing a task may influence mental processes and, thereby, task performance. Some evidence for this hypothesis has previously been provided by, for example, Gilhooly et al. (2010), who found that thinking aloud impaired participants' performance on spatial tasks. We specifically investigate whether thinking aloud influences perceived time.

The perception of time is a basic psychological experience. This study is about situations where people know beforehand that they will subsequently be asked to estimate time. In such situations most models of perceived time assume an internal clock or pulse that, when attended to, gives the experience of the passing of time (Brown, 1997, 2008; Grondin, 2010; Miller, Hicks, & Willette, 1978). In these models, perceived time is proportional to the number of attended pulses. This explains the common experience that the same interval of clock time may be perceived as long, when waiting for a webpage to download with nothing to do except attending to time pulses, and short, when absorbed in a computer game with little attention left for time pulses. Pulses may go unattended for multiple reasons and perceived time has, consequently, been proposed as an indicator of multiple constructs, including engagement (Larsen & von Eye, 2006), flow (Csikszentmihalyi, 1990), motivation (Conti, 2001), and mental workload (Block, Hancock, & Zakay, 2010). Thus, if thinking aloud results in more unattended time pulses then perceived time will be influenced in the same way as when a user experiences, for example, increased mental workload.

We chose a mentally demanding task for our study because it may be for demanding tasks that usability and, hence, evaluation are most important. In addition, we included trials performed with and without time constraints in the study because both situations are common in practice and because perceived time may function differently in the presence and absence of time constraints, which alert participants to the passing of time (Hertzum & Holmegaard, 2013a). We conclude the study by discussing its implications for the use of thinking aloud in usability evaluation.

METHOD

To investigate whether thinking aloud influences perceived time we conducted an experiment with 32 participants, each performing a code-breaking task 24 times. We note that the data have previously been used in analyses of whether perceived time is indicative of mental workload (Hertzum & Holmegaard, 2013a) and whether thinking aloud affects behavior in situations with time constraints and interruptions (Hertzum & Holmegaard, 2013b). We have not previously analyzed whether thinking aloud influences perceived time.

Participants

The 32 participants (13 female, 19 male) were an average of 25.3 (*SD* = 3.7) years of age. In terms of background, 26 participants were students at a technical university, five were professionals, and one did not report his background. Participants were experienced computer users who used computers daily or near daily. Finally, 17 participants (8 thinking-aloud participants, 9 control participants) indicated that they played computer games.

Thinking-Aloud Conditions

The experiment involved two conditions: thinking aloud and control. Participants were assigned to either one or the other condition.

The 16 thinking-aloud 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.

The 16 control participants simply performed the tasks, and the experimenter remained silent. These participants received no information about thinking aloud and no instructions to think aloud. They were not instructed to keep silent because Kim (2002) found that remaining silent was unnatural to some people and degraded their task performance.

Task

The task, a computer version of the game of mastermind, consisted of breaking a four-digit code by making repeated guesses and receiving feedback for each guess. Participants had up to eight guesses to break the code, which was restricted to the digits 1 to 6 (e.g., 2612). These design choices were made on the basis of pilot testing aimed at finding a level of task difficulty where some codes were broken but the task remained challenging throughout the session.

When participants made a guess they received on-screen 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; it 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. 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. The difficulty of the task was increased further by periodic interruptions, the presence of which disrupted participants in working out the remaining code combinations. The interruptions consisted of a brief visual matching task.

We chose the code-breaking task because it was a nontrivial cognitive task, because its brevity allowed for multiple trials in a single session, and because we hoped its game qualities would help avoid fatigue. In addition, the task allowed for introducing distinctions between two levels of time constraint (timed, untimed) and two levels of task success (solved, unsolved):

Participants alternated between timed and untimed trials. During timed trials participants had a maximum of 25 seconds per guess, but they were not told that the time limit corresponded to 25 seconds. Rather, the passing of the 25 seconds was impressed upon participants by a progress bar that visualized how the elapsed time progressed toward the time limit. 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. During untimed trials participants could spend as much time as they needed on each guess. No time limit was enforced and the progress bar was not displayed during untimed trials. Participants experienced the timed trials as more demanding than the untimed trials, as indicated by higher mental-workload ratings and lower task-success rates (Hertzum & Holmegaard, 2013b). Mental workload was rated on the task load index (TLX; Hart & Staveland, 1988).

Participants' performance resulted in solved and unsolved trials. A trial was solved if the participant broke the code, that is, if the participant's guess exactly matched the code. A trial was unsolved if a participant had used all eight guesses and not broken the code. Task success was logged by the game application. We include task success in our analysis because previous work has shown that perceived time is affected by whether tasks are solved or unsolved (Hertzum & Holmegaard, 2013a).

Procedure

Participants were initially introduced to the experiment and asked questions about their background. Then, the code-breaking task was explained and participants tried performing it with and without the time constraint. Thinking-aloud participants were instructed about how to think aloud and practiced thinking aloud on four training tasks: (a) What is the result of 11 multiplied by 12? (b) Think of a friend. How many windows are there in your friend's house or flat? (c) Name 20 animals. (d) 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 give participants additional practice in verbalizing at Levels 1 and 2 only.

Participants performed three blocks of eight trials, alternating between timed and untimed trials. Half of the participants in each condition started with a timed trial, the other half with an untimed trial. Ahead of each trial, the type of time constraint was indicated on the screen on which participants performed the code-breaking task. The experimenter kept silent, except when thinking-aloud participants stopped talking for more than 30 seconds. When this happened, the experimenter reminded participants to "keep talking". Upon completing a trial, participants estimated its duration on a pop-up screen. The estimate was made by positioning a slider on a scale from 0 to 10 minutes, with the possibility (used in only six instances) of shifting the scale to the interval from 10 to 20 minutes. After each block, participants were allowed a break before they performed the next block of trials.

The experiment lasted an average of 2.1 hr per participant. As a token of our appreciation participants received a gift certificate of DKK 350.

Measurement of Perceived Time

We expressed perceived time by the commonly used ratio of perceived time to clock time (Block et al., 2010). A ratio larger than unity meant that perceived time exceeded clock time, that is, overestimation. Conversely, a ratio smaller than unity meant underestimation. Perceived time was the participants' estimate of the duration of each trial. Clock time was the interval between the log events recording the start and end of a trial.

RESULTS

We analyzed the data using analysis of variance (ANOVA) with thinking-aloud condition (thinking aloud, control) as a between-subjects variable and block (first, second, third), time constraint (timed, untimed), and task success (solved, unsolved) as within-subjects variables. With the exception of the analysis of task success (see below), all 32 × 24 = 768 trials were included in the analyses.

Table 1 shows the average perceived time ratio, perceived time, and clock time for participants in the two conditions. Participants overestimated time by an average of 47% (thinking aloud) and 94% (control). The analysis is about the difference between thinking-aloud and control participants rather than about which group was more accurate, but we note that thinking-aloud participants made more accurate time estimates (i.e., overestimated less) than control participants. There was a significant effect of thinking-aloud condition on the perceived time ratio, $F(1, 30) = 5.77$, $p = .023$, with a lower perceived time ratio for thinking-aloud than control participants. We found no effect of thinking-aloud condition on perceived time, $F(1, 30) = 2.00$, $p = .17$. Likewise, there was no effect of thinking-aloud condition on clock time, $F(1, 30) = 0.04$, $p = .84$. Thus, thinking aloud did not prolong task completion.

Insert Table 1 about here

Table 2 shows a breakdown of the perceived time ratio onto block, time constraint, and task success.

There was no effect of block on the perceived time ratio, $F(2, 60) = 1.29$, $p = .28$, and no interaction between block and thinking-aloud condition, $F(2, 60) = 0.10$, $p = .88$. Rather, we found that the overall effect of thinking-aloud condition on the perceived time ratio also held separately for the second, $F(1, 30) = 6.12$, $p = .019$, and third, $F(1, 30) = 5.49$, $p = .026$, block but not for the first block, $F(1, 30) = 3.07$, $p = .090$. Learning possibly affected the first block, but fatigue did not appear to influence the study.

There was a significant effect of time constraint on the clock time spent completing the task, $F(1, 30) = 41.26$, $p < .001$. As intended, task completion time was shorter for timed ($M = 89$ sec) than untimed ($M = 159$ sec) trials, indicating that the alternation of timed and untimed trials produced an alternating series of trials at two different levels of temporal demand. For the perceived time ratio, we found no interaction between time constraint and thinking-aloud condition, $F(1, 30) = 2.57$, $p = .12$. However, with an observed power of .34 we cannot rule out that insufficient sample size masked that timed trials had a larger decrease in perceived time ratio from control to thinking-aloud participants than untimed trials. The overall effect of thinking-aloud condition on the perceived time ratio also held separately for timed, $F(1, 30) = 5.95$, $p = .021$, and untimed, $F(1, 30) = 4.97$, $p = .034$, trials. That is, thinking-aloud participants overestimated time less than control participants for untimed, low-workload trials as well as for timed, high-workload trials.

Insert Table 2 about here

Task success could not be controlled in the experimental design because it measured participants' performance. To analyze whether the perceived time ratio was affected by task success we, therefore, down-sampled the dataset by randomly selecting six solved and six unsolved trials for each participant. The down-sampling involved dropping four thinking-aloud and four control participants who either did not have six solved trials or six unsolved trials. This way, we arrived at a balanced dataset with 24 (participant) × 12 (trials per participant) = 288 trials. We note that an analysis of the full, but unbalanced, dataset showed the same significant effects as those reported below for the down-sampled dataset.

We found a significant effect of task success on the perceived time ratio, $F(1, 22) = 49.41$, $p < .001$, with a higher perceived time ratio for solved ($M = 1.96$) than unsolved ($M = 1.42$) trials. However, the interaction between task success and thinking-aloud condition merely approached significance, $F(1, 22) = 4.27$, $p = .051$. The observed power of the test for an interaction was .51 and a larger decrease in perceived time for solved than unsolved trials may, thus, be masked by the modest size of the down-sampled dataset. There were significant effects of thinking-aloud condition on the perceived time ratio for both solved, $F(1, 22) = 11.11$, $p = .003$, and unsolved, $F(1, 22) = 6.57$, $p = .018$, trials.

DISCUSSION

Thinking-aloud as well as control participants overestimated time. A general overestimation of time is not unusual in studies of time perception. For example, Thomas, Handley, and Newstead (2004, Experiment 2) and Loftus, Schooler, Boone, and Kline (1987, Experiment 2) found overestimations of an average of 13-32% and 407%, respectively. Miller et al. (1978) proposed that the pulses based on which time is perceived arise from an overlearned association with change in mental content. It

appears that participants' time perception is often not well calibrated to experimental settings, which may be more intense than day-to-day activities.

Interpretation of Results

We found a lower ratio of perceived to clock time for thinking-aloud than control participants. A possible explanation of this finding is that thinking aloud occupies attention. To obtain thoughts for verbalization, people may need to monitor their thought process continuously in a way that is not necessary for the thinking itself to take place. The attention occupied by such monitoring would be specific to thinking aloud and imply less attention left for time pulses, resulting in more unattended pulses, which, in turn, yields the lower ratio of perceived to clock time. According to this explanation the way in which thinking aloud requires attention to produce accurate verbalizations resembles the way in which time pulses must be attended to produce accurate time estimates. This explanation is consistent with predominant models of perceived time (Brown, 1997, 2008), which explain variations in perceived time by the allocation of attention to either time pulses or another task, such as thinking aloud.

Another possible explanation is verbal overshadowing (Meissner & Brigham, 2001; Schooler, 2002), which posits that verbalization induces a processing shift that interferes with the processing of nonverbal information. However, most work on verbal overshadowing has involved the verbalization of reasons (Level 3 verbalization) and thus not been about thinking aloud. An exception is the study by Gilhooly et al. (2010), who found lower task completion rates for thinking-aloud than control participants on spatial tasks but not on verbal tasks. Thinking aloud during spatial tasks involves a transformation of the nonverbal information into verbal form (Level 2 verbalization). In contrast, thinking aloud can proceed without such a transformation during verbal tasks (Level 1 verbalization). Thus, the transformation appears to impair performance. To the extent that perceived time involves the processing of nonverbal information it may be similarly impaired by a processing shift introduced by thinking aloud.

In spite of the work linking perceived time to mental workload (e.g., Block et al., 2010; Hart, 1975), we hesitate to interpret the lower ratio of perceived to clock time as an indication of increased mental workload during thinking aloud. We hesitate for two reasons. First, an interpretation in terms of mental workload would suggest a lower perceived time ratio for timed, high-workload trials than for untimed, low-workload trials but we found a higher perceived time ratio for timed than untimed trials (see Table 2). Hertzum and Holmegaard (2013a) propose that this finding may be explained by the explicit way in which the timed trials draw attention to time. Conversely, studies of whether perceived time is an indicator of mental workload have manipulated workload by varying the number of digits in mental arithmetic tasks or by other non-temporal manipulations. Our finding of a higher perceived time ratio for timed than untimed trials suggests caution in linking perceived time to mental workload. Second, our previous analysis of the participants' pupil diameters (a physiological measure of mental workload) and TLX ratings (a subjective measure of mental workload) found no difference between thinking-aloud and control participants (Hertzum & Holmegaard, 2013b). Similarly, Kammerer and Gerjets (2013) and Hertzum et al. (2009) also report no effect of thinking aloud on mental workload, as measured by pupil diameters and TLX, respectively. These other sources of data do not support an interpretation of our results as indicating increased mental workload during thinking aloud. We recognize, however, that perceived time may be sensitive to smaller changes in mental workload than pupil diameters and TLX ratings.

Finally, the effect of thinking aloud on perceived time also holds separately for timed, untimed, solved, and unsolved trials. This result is noteworthy because Hertzum and Holmegaard (2013a) showed that perceived time is influenced by whether tasks are timed or untimed and whether they are solved or unsolved. Thus, the effect of thinking aloud on perceived time is robust across other factors that also influence perceived time.

Implications

The primary research implication of this study is that further studies are needed to explain the mechanism through which thinking aloud influences perceived time. A better understanding of this mechanism is necessary to determine the scope of the effect. We have discussed three possible mechanisms: attention, overshadowing, and mental workload. An explanation in terms of attention would suggest that our results may generalize from perceived time to other tasks that must be continuously attended for accurate performance. One group of such tasks may be those for which situation awareness is a main concern. Conversely, an explanation in terms of verbal overshadowing would suggest that the effect of thinking aloud is about the temporary blocking of some types of processing, rather than about the depletion of attentional resources. In this case the effect may be present in a wider range of situations and largely unrelated to the high level of mental demand imposed by the code-breaking task used in this study. Finally, an explanation in terms of mental workload would call for reconciling perceived time with workload measures such as pupil diameters and TLX ratings.

We want to point to three implications of this study for the practice of usability evaluation. First, the effect of thinking aloud on perceived time invalidates comparison with duration estimates made while not thinking aloud. For example, users who have been thinking aloud in a usability evaluation should not be asked whether the evaluated system was fast or slow because this implies a comparison with systems they have experienced outside the evaluation, that is while not thinking out loud.

Second, the experience of time influences other experiences, which then become tainted by the effect of thinking aloud on perceived time. For example, Ramsay, Barbesi, and Preece (1998) found that webpages with short download times were rated as more interesting than pages with long download times. Post-test ratings of systems experienced while thinking aloud may, therefore, be inflated because task completion times were perceived shorter than they would have been in the absence of thinking aloud. Such inflation may lead to misinterpretation of the ratings, especially if they are compared with ratings made while not thinking aloud.

Third, in evaluations that involve time estimation a way of counteracting the above implications may be to replace thinking aloud while performing tasks with thinking aloud retrospectively after the tasks have been performed. This way, tasks and their duration are experienced without the influence of thinking aloud. Retrospective thinking aloud may be stimulated by viewing a video recording of the performance of the tasks and has been found to produce usability results comparable to those of concurrent thinking aloud (Haak et al., 2003). The cost of retrospective over concurrent thinking aloud is, however, that test sessions become twice as long.

Limitations

Three limitations should be remembered in interpreting the results of this study. First, we used Ericsson and Simon's (1993) procedure for instructing participants about how to think aloud, and our subjective experience from the experimental sessions confirmed that participants verbalized at Levels 1 and 2 only. We acknowledge, however, the absence of a control variable to verify this. Second, participants displayed a general tendency to overestimate time. We acknowledge that the 0-to-10-minute scale used for indicating perceived time may in part be responsible for this overestimation because the midpoint of the scale (300 seconds) was well above the average clock time for a task (124 seconds). The 0-to-10-minute scale was chosen to accommodate all tasks. Third, participants performed the code-breaking task 24 times and might be subject to an anchoring effect (Tversky & Kahneman, 1974), thereby failing to adjust their time estimates sufficiently from one trial to the next. The alternation of timed and untimed trials aimed to minimize such anchoring.

CONCLUSION

In an experiment with 32 participants, we found that thinking aloud influences perceived time. There was no difference in the clock time spent solving tasks but the perceived time ratio was lower for thinking-aloud than control participants. According to the predominant models of perceived time this effect implies that thinking aloud occupies attention. Accounts of verbal overshadowing propose the alternative explanation that the effect is caused by a processing shift that interferes with the perception of time. Further work is required to explain the mechanism underlying the effect of thinking aloud on perceived time and, thereby, its wider implications for usability evaluation. An immediate implication of this study is that time estimates made during thinking aloud cannot be validly compared with time estimates made while not thinking aloud.

KEY POINTS

- Thinking aloud influences perceived time, suggesting that it occupies attention.
- Alternative explanations of the finding include that thinking aloud may overshadow nonverbal processing or increase mental workload.
- Usability evaluators must treat time estimates, and ratings affected by the experience of time, cautiously if they are made while thinking aloud.

ACKNOWLEDGEMENTS

We are grateful to Signe Arnklit, who recruited the participants for the experiment. Special thanks are due to the participants.

REFERENCES

Block, R. A., Hancock, P. A., & Zakay, D. (2010). How cognitive load affects duration judgments: A meta-analytic review. *Acta Psychologica, 134*(3), 330-343.

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

Brown, S. W. (1997). Attentional resources in timing: Interference effects in concurrent temporal and nontemporal working memory tasks. *Perception & Psychophysics, 59*(7), 1118-1140.

Brown, S. W. (2008). Time and attention: Review of the literature. In S. Grondin (Ed.), *Psychology of Time* (pp. 111-138). Bingley, UK: Emerald.

Conti, R. (2001). Time flies: Investigating the connection between intrinsic motivation and the experience of time. *Journal of Personality, 69*(1), 1-26.

Csikszentmihalyi, M. (1990). *Flow - The psychology of optimal experience*. New York: Harper.

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

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

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.

Grondin, S. (2010). Timing and time perception: A review of recent behavioral and neuroscience findings and theoretical directions. *Attention, Perception & Psychophysics, 72*(3), 561-582.

Hart, S. G. (1975). Time estimation as a secondary task to measure workload. In *Proceedings of the 11th Annual Conference on Manual Control* (pp. 64-77). Washington, DC: US Government Printing Office.

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.

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. (2013a). Perceived time as a measure of mental workload: Effects of time constraints and task success. *International Journal of Human-Computer Interaction, 29*(1), 26-39.

Hertzum, M., & Holmegaard, K. D. (2013b). Thinking aloud in the presence of interruptions and time constraints. *International Journal of Human-Computer Interaction, 29*(5), 351-364.

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.

Kammerer, Y., & Gerjets, P. (2013). The role of thinking-aloud instructions and prior domain knowledge in information processing and source evaluation during web search. In M. Knauff, M. Pauen, N. Sebanz & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 716-721). Austin, TX: Cognitive Science Society.

Kim, H. S. (2002). We talk, therefore we think? A cultural analysis of the effect of talking on thinking. *Journal of Personality and Social Psychology, 83*(4), 828-842.

Larsen, E., & von Eye, A. (2006). Predicting the perceived flow of time from qualities of activity and depth of engagement. *Ecological Psychology, 18*(2), 113-130.

Lewis, J. R. (2006). Usability testing. In G. Salvendy (Ed.), *Handbook of Human Factors and Ergonomics. Third Edition* (pp. 1275-1316). New York: Wiley.

Loftus, E. F., Schooler, J. W., Boone, S. M., & Kline, D. (1987). Time went by so slowly: Overestimation of event duration by males and females. *Applied Cognitive Psychology, 1*(1), 3-13.

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.

Miller, G. W., Hicks, R. E., & Willette, M. (1978). Effects of concurrent verbal rehearsal and temporal set upon judgments of temporal duration. *Acta Psychologica, 42*(3), 173-179.

Ramsay, J., Barbesi, A., & Preece, J. (1998). A psychological investigation of long retrieval times on the world wide web. *Interacting with Computers, 10*(1), 77-86.

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

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

Thomas, K., Handley, S., & Newstead, S. (2004). The effects of prior experience on estimating the duration of simple tasks. *Current Psychology of Cognition, 22*(2), 83-100.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124-1131.

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.

TABLE 1: Perceived time ratio, *N* = 768 trials

	Thinking aloud		Control	
	M	*SD*	*M*	*SD*
Perceived time ratio *	1.47	0.45	1.94	0.64
Perceived time (seconds)	160	80	198	75
Clock time (seconds)	126	54	122	44

* $p < .05$

TABLE 2: Breakdown of perceived time ratio

	Thinking aloud		Control	
	M	*SD*	*M*	*SD*
Block				
Block 1 (*N* = 256 trials)	1.43	0.47	1.85	0.86
Block 2 (*N* = 256 trials) *	1.46	0.48	1.96	0.64
Block 3 (*N* = 256 trials) *	1.53	0.52	2.02	0.65
Time constraint				
Timed (*N* = 384 trials) *	1.56	0.48	2.11	0.76
Untimed (*N* = 384 trials) *	1.39	0.43	1.78	0.55
Task success				
Solved (*N* = 144 trials) **	1.58	0.42	2.34	0.67
Unsolved (*N* = 144 trials) *	1.20	0.36	1.63	0.47

* $p < .05$, ** $p < .01$