IMPROVE YOUR SCORE; THINK ALOUD!

THE EFFECTS OF THINKING ALOUD REVEALED IN AN EYE TRACKER STUDY

Author: Rutger Tjallema

Tutor: Dr. I.T.C. Hooge

Master Thesis Applied Cognitive Psychology (TCP)

Faculty Social Science

Utrecht University

August 2012

INTRODUCTION

Technological improvements have led the computer change from a heavy desk machine in the attic, to a lightweight tablet in the briefcase or a telephone in the front side pocket of the pants. People use these computers anywhere at any moment throughout the day and so the interaction between human and computer has increased dramatically. For developers of computer interfaces it is a challenge to make this interaction as smooth as possible by designing a user friendly interface. That this is not always an easy task can be illustrated by the ribbon-menu that Microsoft introduced in their Office 2007 package. *Do you remember the frustration when you were unable to relocate your toolbars?* According an online survey (Kyd, 2009), experienced users' productivity dropped with an average of 20 percent because of the new menu and 55 percent of the responders 'hate' the ribbon-menu. This is a clear example of the effects a graphical user interface (GUI) can have on the usability and user experience of the product. No wonder the interaction between human and computer has become a scientific area of high interest. One broadly used technique to study this human-computer interaction (HCI) is the Think Aloud (TA) method (Nielsen, Clemmensen and Yssing, 2002). In this research method tested subjects speak out their thoughts while executing a task, with as goal to get insight in the cognitive process of the subject. In the HCI field the TA method is used to reveal usability problems in the designed interface. Currently, there is a lack of conclusive research about the effects of the TA method on the users' performance. Therefore the present study will focus on the effects of thinking aloud on the performance of the subject. This introduction will first cover the work of Ericsson and Simon (1980, 1997) about the TA method, then the effects of thinking aloud, followed by the goal of the present study.

The Think Aloud Method

The TA method is a qualitative research method which uses verbalized thoughts as data to study the cognitive process of a subject. As the subject speaks out loud all the information that comes into its mind, the verbalized thoughts reflect the information that is attended to – the 'heeded' information (Ericsson and Simon, 1980) – at that moment. When the subject thinks out loud while performing a task, the verbal report reveals the cognitive process needed for the task completion. The main goal of the think aloud method is to get insight into this process. Two variants of the TA method can be distinguished: concurrent think aloud – where the subject verbalizes its thoughts while performing a task – and retrospective think aloud – where the subject verbalizes its thoughts retrospective, often while it is viewing a video of its task performance. The present study restricts itself to the concurrent think aloud method.

Before the TA method raised in popularity, the most used qualitative verbal research method to map cognitive processes was introspection (Van Someren, Barnard and Sandberg, 1994; Ericsson and Simon, 1997). Introspection though, was questioned heavily on its validity, as people tend to have difficulties with reporting their cognitive process and rather report judgments and internal theories (Nisbett and Wilson, 1977). As an alternative to introspection Duncker (1945) introduced speaking out loud the thoughts while performing on a task. This was the first investigator who mentioned TA as a verbal research method (Nielsen *et al,* 2002), but it was until the historical work of Ericsson and Simon (1980) before thinking aloud gained in popularity. In their widely cited review Ericsson and Simon (1980) covered over a hundred studies to build a founded theory where thinking aloud got divided into three levels of verbalization. The general principle was that thinking aloud is the verbalization of the short term memory. Their conclusion was that 'verbal reports, elicited with care and interpreted with full understanding of the circumstances under which they were obtained, are a valuable and thoroughly reliable source of information about cognitive processes' (Ericsson and Simon, 1980, pp 247). As the work of Ericsson and Simon (1980, 1997) is of significant importance to the present study, the next section will explain the fundamental principles of their theory.

Thinking aloud can be performed at three distinguishable levels. It depends on the type of task which level is required to verbalize the thoughts.

Level 1 verbalization is articulation of verbal coded information (read aloud). Because the information is presented in verbal form, no additional cognitive processing is required for the verbalization.

Level 2 verbalization is articulation of information that is not presented verbally. Recoding the information into a verbal form requires some cognitive processing, but no additional information has to get attended for the verbalization.

Level 3 verbalization requires scanning, filtering, retrieval of memory and processing before it can be verbalized.

In contrast with level 1 and level 2 verbalization, level 3 verbalization does require heavy additional processing to recode the information into a verbal form. Because of this processing the cognitive capacity available to the execution of the primary task gets diminished. Ericsson and Simon explain this process by means of the information processing theory.

This theory is based on a model first introduced by Atkinson and Shiffrin (1968). The model divides human memory in three components: the sensory memory, the short term memory (STM) and the long term memory (LTM). The sensory memory has a very limited capacity and holds the sensory input for a short time. The STM

or working memory contains the attended information of a specific moment and is responsible for processing this information. The LTM has a large capacity and is responsible for permanent memory storage. As the STM contains the attended information and is responsible for information processing, the content of this memory reflects the cognitive process that is needed to execute a specific task. The goal of the TA method is to extract this information. So, how does that work? According to Ericsson and Simon, thinking aloud is the articulation of the STM contents and results in a report of the ongoing cognitive process. As level 1 verbalization is articulation of verbal presented information, the information does not utilize the STM and henceforth the verbalization does not alter the cognitive process nor slows it down. Level 2 verbalization, however, does require STM processing to recode the presented information into a verbal form. This leads to a prolonged time to accomplish a task and the verbal report may be incomplete, but the cognitive processing will not be altered by the verbalization. At level 3 verbalization, the additional processing needed to verbalize takes place in the STM. The cognitive load of level 3 verbalization is comparable with introspection (Nielsen *et al*, 2002) which changes the course and structure of the process (Ericsson and Simon, 1980). As the goal of the TA method is to get insight in one's cognitive process, it must leave the process unchanged. Level 1 and level 2 verbalization meet this requirement. Level 3 verbalization does alter the cognitive process and henceforth it would not be valid to study the cognitive process with the TA method if the task required level 3 verbalization.

Effects of thinking aloud

According the theory of Ericsson and Simon (1997), level 1 and level 2 leave the cognitive process unchanged, but level 2 verbalization does slow down the process which results in longer response times. Longer response times are measured at making word puzzles and geometric puzzles, Raven's matrices (Rhenius and Deffner, 1990; Fox and Charness, 2009), gambles and anagrams (Russo, Johnson and Stephens, 1989) and seeking information on a website (Hertzum, Hansen and Andersen, 2009), whereas the criteria of level 2 verbalization are met. The prolonged response time can be caused by the ability to process visual information faster than verbalize this information (Rayner, 1998, 2009). Conversely, Van der Haak, De Jong and Schellens (2003, 2004) did not find prolonged response times in information seeking tasks in an online library catalogue. At this study, however, the subjects had the option to not complete the tasks, which is a possible explanation for the nonexistence of prolonged response time. The literature concerning prolonged response times compiles with three-level model of verbalization of Ericsson and Simon (1980, 1997).

However, concerning the task performance the findings are not so unanimous. Russo *et al* (1989) found that on a numeral task subjects have a lower proportion correct answers while thinking aloud, compared with a silent control condition. On a mental multiply task, though, the subjects had a higher proportion correct answers than the control condition. On the other two tasks (pictorial and verbally written) thinking aloud had no effect. Although the tasks meet the requirements of level 2 verbalization, thinking aloud affects the task performance of some tasks. Russo *et al* (1989) conclude, first, that the assumption about the nonexistence of reactivity at the think aloud method must be questioned and, second, it is difficult to determine a priori whether a task will be affected by thinking aloud or not. Van den Haak *et al* (2003) also found that subjects experienced more problems while thinking aloud. In their study the subjects had to search for specific information in an online library catalogue, either concurrent thinking aloud or retrospective thinking aloud. The test scores of the retrospective thinking aloud condition are comparable with a silent condition, as the subjects execute the tasks in both settings in silence. Even though the tasks met the criteria of level 1 and 2 verbalization, subjects in the TA condition were less successful in completing the tasks than subjects who executed the tasks in silence. Van den Haak *et al* (2003) concluded that the cognitive processing required for the task and the verbalization combined impaired both the task performance and the articulation. There is evidence that a more demanding task makes it harder for the subject to think aloud (Branch, 2000; Preece, Rogers and Sharp, 2007), which would indicate that both the primary task and the verbalization utilize the STM. However, in a succeeding study of Van den Haak *et al* (2004), they did not replicate the findings of their preceding study (2003) and did not find a difference in the amount of completed tasks between the TA condition and silent condition, while the study setup was the same. Hertzum *et al* (2009) asked to search for information on a website as well. Each participant had to think aloud at half of the tasks and execute the other half in silence. There was no difference in the proportion correct answers between the two conditions. These findings are congruent with the theory of Ericsson and Simon (1980, 1997) and other studies (Rhenius and Deffner, 1990; Bowers and Snyder, 1990). The results of Fox and Charness (2009) however reveal, that thinking aloud can improve the test performance as well. In their study, elderly people scored better on the Raven's matrices if they were thinking aloud, compared to the silent control condition. These fluctuating results indicate that the effect of thinking aloud is a delicate matter. As Russo *et al* (1989) concluded; it hard to tell a priori whether a task is reactive with thinking aloud or not. Clearly, though, the nonexistence of reactivity as stated by Ericsson and Simon (1997) is contradicted by several studies.

Another way to investigate whether thinking aloud alters the cognitive processing is using eye movements as variable. Earlier studies that used eye movements to investigate the think aloud method found 80% (Cooke,

2010) till 98% (Rhenius and Deffner, 1990) overlap between the gazes of the subject and its verbal report. As the eye movements correspond strongly with the attended information (Geiselman, 1977), the verbal reports produced with thinking aloud reflect the attended information. This makes eye movements interesting for studying the reactivity of thinking aloud. Hertzum *et al* (2009) recorded the eye movements of the subjects while they were seeking information on a website. The subjects executed half of the tasks in silence and the other half while thinking aloud. They looked at the fixation rate (fixation per second), fixation time, saccade duration and saccade length and found no difference in any of these variables between the silence condition and the think aloud condition. This would indicate that there is no difference in heeded information between executing a task while thinking aloud and executing a task in silence. However, this result is based on a relative small test group (N = 8) and the study used websites as stimuli. Using a website consists for 58% out of reading (Cooke, 2010), which is level 1 verbalization. As stated before, for most tasks in the usability research level 2 verbalization is required. At this moment there is a lack of studies that focus on the difference in eye movements of a subject performing a task that requires level 2 verbalization while thinking aloud and while in silence. The present study aims to fill in this gap by using pictorial tasks and measure the eye movements of the subject in both conditions. The main research question is:

RQ: What is the difference in the subject's eye movements between the think aloud condition and the silent condition?

This will be investigated by letting subjects execute tasks first in silence and second while thinking aloud. The first two hypotheses concerning the test scores are based on the theory of Ericsson and Simon (1980, 1997):

H1: The response time in the think aloud condition is longer than in the silence condition.

H2: There is no difference in proportion correct answers between the TA condition and the silent condition.

Hertzum *et al* (2009) found no difference in both fixation rate and fixation time, as they did not find a difference in saccade duration between the two conditions, while the RT in the TA condition was longer. This would indicate that the number of fixations is higher in the TA condition. In line with these findings, the following hypotheses concerning the eye movements are formulated:

H3: There is no difference in fixation rate between the TA condition and the silence condition.

H4: There is no difference in fixation time between the TA condition and the silence condition.

H5: There is no difference saccade length between the TA condition and the silence condition.

METHODS

Design and variables

The goal of the present study was to investigate whether a difference in the dependent variables is present between the think aloud (TA) condition and silent condition. The design was within subject 2 x 10: each subject was exposed to both conditions; first they were given 10 silent trials followed by 10 TA trials. The dependent variables are divided into response variables and eye movement variables. The response variables are the response time (RT) – which is the time between the stimulus onset and the subject's response – and the proportion correct answers. Concerning the eye movements, both the temporal and the spatial aspects have been taken into account. The temporal aspects are the number of fixations, fixation rate (fixations per second) and the average fixation time (FT). The spatial aspect are scanpath length (sum of the saccades) and average saccade length. All these variables are calculated for every trial, each trial falling within either the TA condition or the silent condition.

Subjects

In present study 21 women participated, who were all studying a master's degree or were graduated in a master's degree. None of the subjects had severe visual inabilities and they all had normal or corrected to normal sight. They were allowed to wear their glasses or contacts, if necessary.

Stimuli

The stimuli consisted of 24 multiple choice tasks, divided into four categories: cubicles which were unfolded (4 answer possibilities); logical series (4 answer possibilities); virtual jigsaw puzzles (4 answer possibilities) and visual series (5 answer possibilities). The tasks were mainly pictorial and meet the criteria of level 2 verbalization. See Appendix C for an overview of the tasks. The tasks were allocated to two blocks of 10 trials each (block A and B), both blocks had two practice trials. Each block was randomized for every participant and the blocks were presented in different sequences: the first subject had the sequence AB, the second subject had the sequence BA and so on.

Test setup

The trials were presented on a 17" [info] monitor with a resolution of 1280 x 1024 pixels, a refresh rate of 60 Hz and a colour depth of [info]. The eye movements were measured by an EasyGaze eye tracker, developed by Design Interactive. This eye tracker measured at a frequency of 50 Hz and was connected with a firewire cable to a PC. The PC was suited with an Intel [info processor] with 12 gigabyte DDR [info] RAM and had Windows

XP as operating system. The PC recorded both the eye movements and the subject's responses, which were entered on an American keyboard.

The monitor and keyboard where mounted in a wooden cubical of approximately 80 cm in all directions, with the inside painted black. The front side of the cubical was open, where a head rest was mounted. The whole test setup was shut off from daylight by a curtain to prevent interference from IR-light from outside the setup. There was a small light mounted within the cubical.

Think aloud

All the subjects were asked to think aloud, that is, to speak out all their thoughts that came into mind. The experimenter did not intervene while the test was running. If the subjects stopped talking during the TA condition, the experimenter asked the subject to 'Please think aloud'.

Procedure

For all subjects the experiment followed the same procedure. The first block they had to make the MC-tasks in silence, the second block they had to think aloud while executing the tasks. The blocks were organized as stated before; half of the subjects began with block A followed by block B; the other half of the subjects began with block B, followed by block A. Every block was preceded by two practice trials and introduced by an instruction screen. All trials were preceded by a fixation cross. The subjects had to fixate on the cross and then press the spacebar to proceed to the MC-task. They could enter their answer any moment, by pressing the corresponding key on the keyboard. Each trial had a time limit of 120 seconds. This was mentioned to the subjects, but the time was not presented during the trials. If the time limit got exceeded, the subject automatically proceeded to the next trial, without having entered an answer. The subjects were instructed to make the MC-tasks as good as possible. To collect qualitative good eye tracking data, it was important that the subjects did not move their head during the experiment. To help them keeping their head fixed, they were requested to position their head in the headrest..

RESULTS

To investigate whether there is a difference in the dependent variables between the TA and silent condition, the TA trials are compared with the silent trials in the next section. First the response variables will be covered, followed by the temporal aspects and the spatial aspects of the eye movements. At last the effects of the different task types will be examined.

The data of three subjects was excluded from the data analyses as a result of a measure-error (see Appendix B1 for an example of the measure error). As a result of a calibration-error the exact fixation points of the remaining subjects could not be determined. The data-analysis revealed that the fixation points were clustered at a particular distance of the point of interest, at a position where nothing was to be seen. See for an example of this deviation Appendix B2. Each trial this deviation had a different amplitude and direction and within each trial the deviation differed between the left and the right side of the screen. Henceforth a correction of the data could not be made. As a result, the analysis of the spatial aspects of the eye movements is limited to the scanpath length and saccade length.

The reason of this deviation became clear after the measurements of the present study were conducted. As a result of this deviation Hooge (internal paper) decided to test what effect the stimuli background luminance had on the accuracy of the eye tracker. In his study the eye tracker was first calibrated with a grey calibration screen. The subjects were then asked to follow a specific pattern with their gaze. This was repeated by different background luminance (along the grey-scale). In the trials with a low-luminance background the eye tracker measurement was quite accurate, but with a high-luminance background there was a deviation between the pattern and the measured fixation points. Hooge explained this effect by the smaller pupil dilation caused by the high luminance of the background. The direction of a smaller pupil seems harder to measure, which results in less accuracy. As the backgrounds of the stimuli of the present study were white, the same effect caused the deviation in the eye movement-data.

Response time and proportion correct

The response time (RT) in the TA condition (M = 43.22, SD = 21.86) is longer than in the silent condition (M = 26.83, SD = 14.51), $t(358) = -8.379$, $p < .001$ (see Table 1 and Figure 2). In the TA condition (M = .79, SD = .41) the proportion correct answers is larger than in the silent condition (M = .67, SD = .47), $t(358) = -2.623$, $p = .009$ (see Table 1, figure 2). Clearly the subject spends 61% more time to solve the task and respond 18% more correctly in the TA condition. Figure 1 visualizes the RT (lines) and proportion correct (bars) of both conditions for each task. Two tasks are answered correct in 100% of the trials and thus have no distinguishable ability. Of

Table 1: Response time (RT) and proportion correct answers in the silent condition and TA condition.

	Silent		TA		
	Mean	SD	Mean	SD	Significance
RT (s)	26.83	14.51	43.22	21.86	$p<.001$
Prop. correct	0.67	0.47	0.79	0.41	$p<.01$

the other 18 tasks, 14 tasks are answered more correctly in the TA condition, while the subjects took more time to come to an answer. As the RT and proportion correct both increase, it might be that these differences result from a speed-accuracy tradeoff. The correlation between RT and proportion correct ($r(18) = .60$, $p = .005$) points in this direction as well.

Eye Movements: Temporal aspects

The temporal aspects of the eye movements are number of fixations, fixation rate and fixation time. When the number of fixations of the two conditions is compared, the trials in the TA condition ($M = 77.97$, $SD = 41.42$) have a significant larger number of fixations than the trials in the silent condition ($M = 54.44$, $SD = 31.63$),$t(355) = -6.029$, $p < .001$ (Figure 4). The subjects have more fixations when thinking aloud. But as the number of fixations is affected by the RT, the fixation rate reflects the fixations corrected for the RT. In the TA condition ($M = 1.85$, $SD = .45$) the fixation rate is smaller than in the silent condition ($M = 2.09$, $SD = .52$), $t(355) = 4{,}817$, $p < .001$ (Figure 5). This reveals that, if corrected for the RT, the subject has fewer fixations in the TA condition than in the silent condition. The fixation time (FT) is significantly longer in the TA condition ($M = 538.03$, $SD = 166.15$) than in the silent condition ($M = 470.67$, $SD = 125.10$), $t(355) = -4.326$, $p < .001$ (Figure 6). The fixation rate and fixation time are related in such way that an increase in fixation time leads to a decrease of the fixation rate.

Eye movements: Spatial aspects

The spatial aspects of the eye movements are scanpath length and saccade length. When the scanpath length of the TA condition ($M = 18980.40$, $SD = 10294.83$) is compared to the scanpath length of the silent condition ($M = 13295.01$, $SD = 7526.85$), it is longer in the TA condition, $t(355) = -5.954$, $p < .001$(Table 3). This indicates that the subject's gaze covers a larger distance in the TA condition than in the silent condition. The saccade length

Table 2: temporal aspects - difference in number of fixations, fixation rate and fixation time (FT) between the silent and the TA condition

	Silent		TA		
	Mean	SD	Mean	SD	Significance
Number of fixations	54.44	31.63	77.97	41.42	$p < .001$
Fixation rate	2.09	0.52	1.85	0.45	$p < .001$
FT (ms)	470.67	125.1	538.03	166.15	$p < .001$

Table 3: difference in scanpath length (SPL) and mean saccade length SL between the silent and the TA condition

	Silent		TA		
	Mean	SD	Mean	SD	Significance
SPL (pixels)	13295.01	7526.85	18980.4	10294.83	$p < .001$
Saccade length (pixels)	255.09	57.09	248.57	52.41	*n.s.*

does not differ between the TA condition (M = 248.57, SD = 52.41) and the silent condition (M = 255.09, SD = 57.09), $t(355) = 1.124$, $p = .262$. As the scanpath length is larger in the TA condition while the saccade length does not differ between the two conditions, in the TA condition the subject's gaze appears to cover a larger distance, but with the same deviation as in the silent condition. This would indicate that subject – while thinking aloud – fixates at the same points of interest, but fixates more often at the same point.

Effects of the task type

The tasks used for this study can be divided into four categories; cubicles which are unfolded; logical series; virtual jigsaw puzzles and visual series. To test if there is an effect of the type of task on the dependent variables, several ANOVA's were conducted. The full results of these ANOVA's and the results of the Temhane post-hoc tests are displayed in Table A1 in Appendix A. In figures 7-9 is visualized how the task type affects the eye movements. Probably the eye movements are affected by the tasks' visual layout. As the logical serie-tasks and the visual serie-tasks both have a sequence that has to be discovered, the subject's gaze seems to follow the sequence, which results in relatively small saccades. The fixation time of the logical serie-tasks is the highest and of the visual serie-tasks the lowest, which would indicate that the fixations points of the logical serie-tasks contain the most information and the fixation points of the visual serie-tasks the least (Henderson, Weeks and Hollingworth 1999). The figures of the logical serie-tasks are indeed more complex. For the cubicle-tasks and puzzle-tasks the large saccade length indicates a larger distance between the points of interest, which indeed is the case at the layout of these tasks.

By means of several t-tests the effects of thinking aloud are examined per task type. See for the full results of the t-tests Table A2 in Appendix A. For most of the variables, there are no differences between the effects of TA on the individual tasks and the effects of TA on the complete test battery. Some effects are not significant though, which is most likely caused by the decrease in test data by selecting one test type.

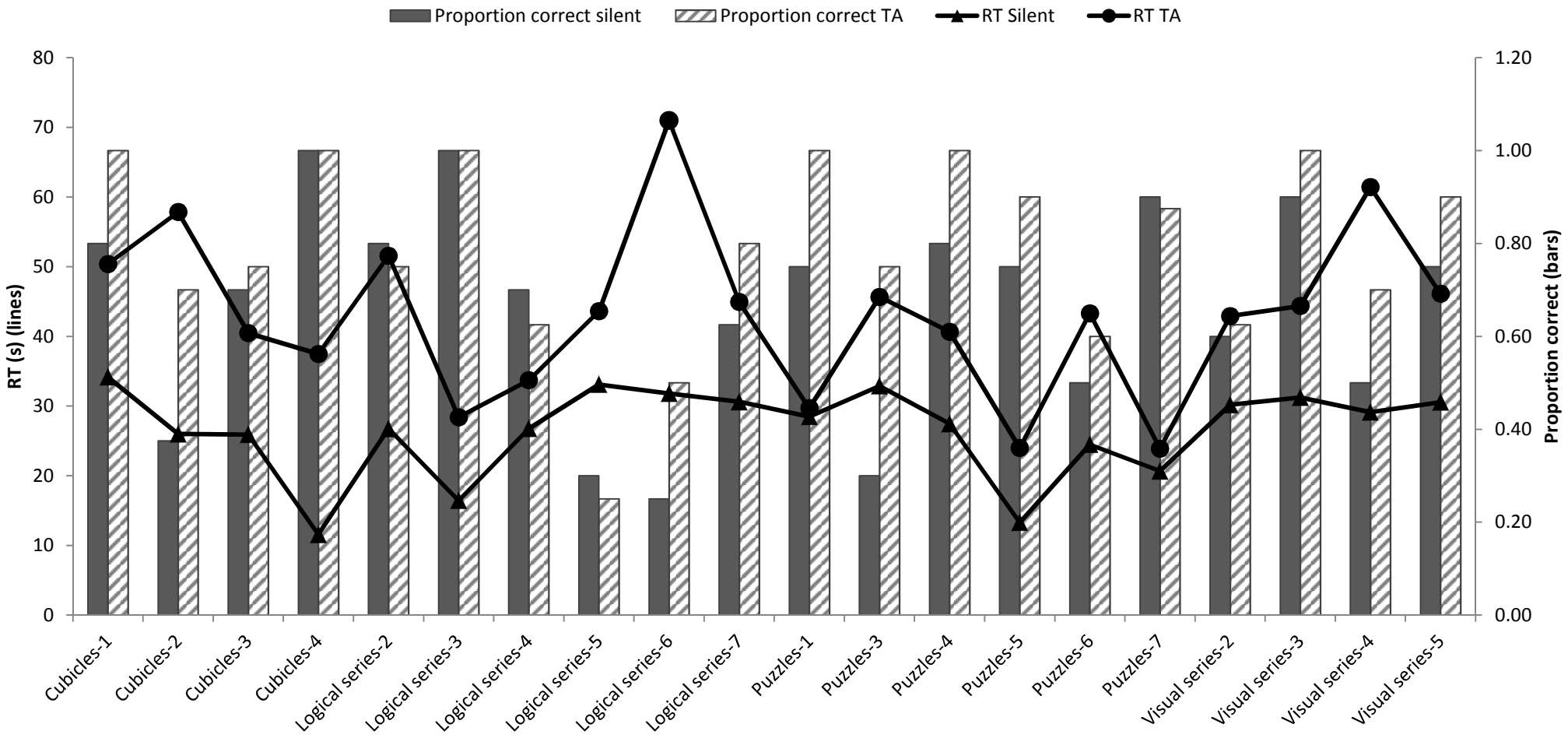


Figure 1: Response time (RT, line) plotted with proportion correct (bars) for each task. This visualizes that the tasks with a high proportion correct have a shorter RT.

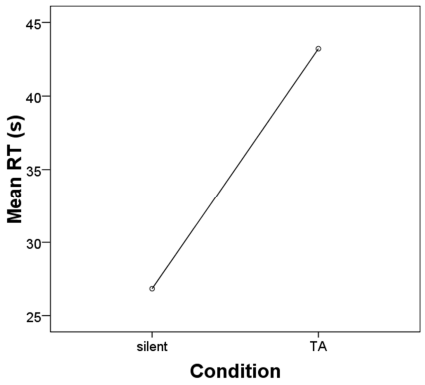


Figure 2: mean response time (RT) in both the silent and TA condition. Difference is significant $p < .001$.

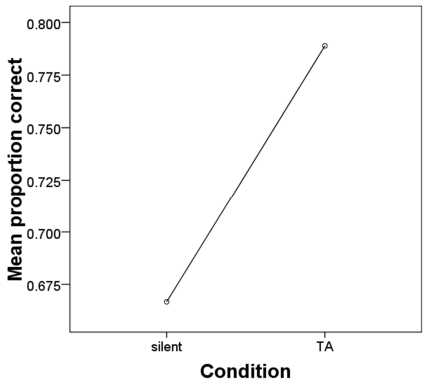


Figure 3: mean proportion correct in both the silent and TA condition. Difference is significant $p < .05$.

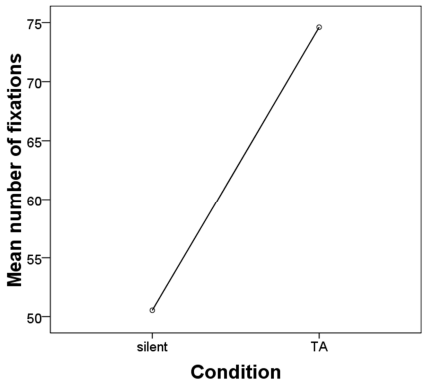


Figure 4: mean number of fixations in the silent and TA condition. Difference is significant $p < .001$.

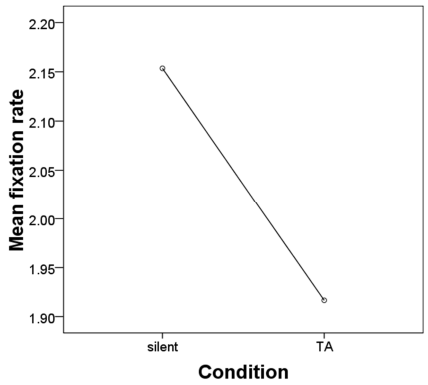


Figure 5: mean fixation rate in the silent and TA condition. Difference is significant $p < .001$.

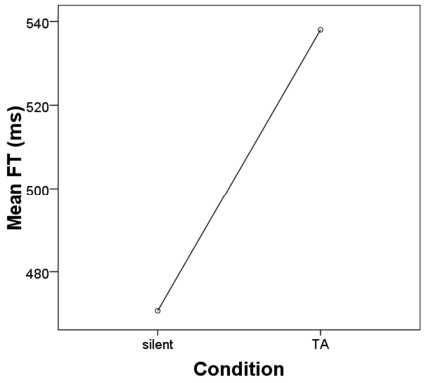


Figure 6: mean fixation time (FT) in both the silent and TA condition. Difference is significant $p < .001$.

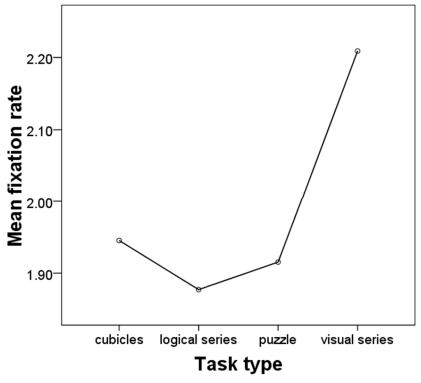


Figure 7: mean fixation rate per task type

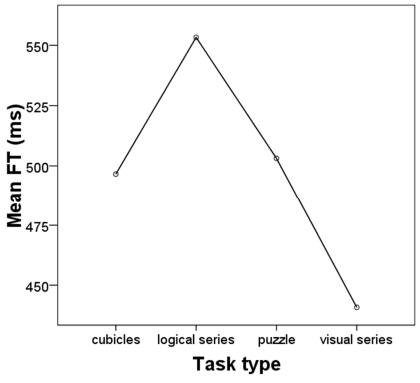


Figure 8:mean fixation time (FT) per task type

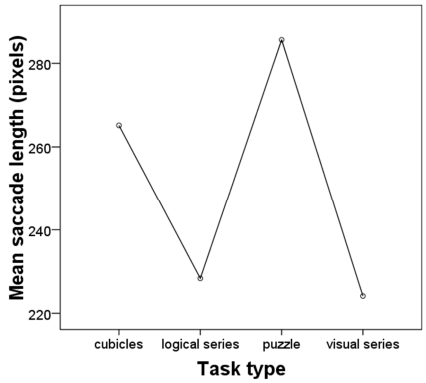


Figure 9: mean saccade length per task type

DISCUSSION

The present study shows that there are differences between the TA condition and the silent condition. First, the RT in the TA condition is longer than in the silent condition, which confirms the first hypothesis. These results are consistent with the theory of Ericsson and Simon (1980, 1997), as the tasks required mainly level 2 verbalization. In line with the findings of the present study, other studies with pictorial tasks (Rhenius and Deffner, 1990; Fox and harness, 2009) also found longer RT in the TA condition compared with the silent condition. The same effect is found in other tasks (Russo *et al*, 1989; Hertzum*et al*, 2009). Obviously thinking aloud causes a prolonged RT. The responsible mechanism for this effect will be explained later in this discussion.

Second, the proportion correct is higher in the TA condition than in the silent condition. This is a remarkable result and it rejects the second hypothesis (H2) that there would be no difference in proportion correct between the two conditions. These findings contradict the theory of Ericsson and Simon (1997), since they indicate the presence of reactivity. Ericsson and Simon pointed out that if the criteria of level 2 verbalization are met, the cognitive process will not be altered. However, the higher proportion correct indicates that the cognitive process is affected by TA in such way that the subject performed better on the tasks. Although there are several studies indicating the presence of reactivity, only Russo *et al*(1989) and Fox and Charness (2010) found an increase in test score in some conditions. Russo *et al* (1989) conducted 4 types of tests, each test had TA trials and silent trials. The task where the subject had to choose between two simple gambles, the proportion correct was 20% higher in the TA condition than in the silent condition. The other tasks had either no difference in proportion correct or a lower proportion correct (numeric task). Although at one of the 4 tasks the difference in proportion correct was comparable with the findings of the present study, this gamble task is very dissimilar to the tasks of present study. Russo *et al* (1989) concluded that the effects of TA vary between tasks and that it is difficult to determine a priori whether a task is reactive. Fox and Charness (2010) used Raven's matrices, which are more alike to the tasks of the present study. But they found only a higher proportion correct at the elderly adults who participated in the study, while the young adults did not show this effect. Therefore, results of this study cannot be generalized to the present study. In the experiment of Van den Haak *et al* (2003) the subject experienced more problems with seeking information at a website in the TA condition. This effect is opposite to the effect of the present study. In both studies there was reactivity, but whereas in the present study TA led to a higher proportion correct, in Van den Haak *et al* (2003) it led to more problems experienced by the subject. Van den Haak *et al*

(2003) concluded that the added cognitive load of thinking aloud had a negative effect on the performance. This indicates that both thinking aloud and executing the primary task are utilizing the STM (Van Someren *et al*, 1994). However, in the present study TA appears to improve the performance, what contradicts this conclusion. Clearly there are several studies that found reactivity, which is in contradiction with the theory of Ericsson and Simon (1997), but none of those studies can explain the higher proportion correct answers found by the present study.

Third, concerning the temporal aspects of the eye movement behavior, the fixation rate is lower, the fixation time is longer and the number of fixations is higher in the TA condition. This rejects the third hypothesis (H3) that there is no difference in fixation rate and the fourth hypothesis (H4) that there is no difference in fixation time. Hertzum *et al* (2009) did not find difference in fixation rate and fixation time between the two conditions.They did not mention the number of fixations, but as they found a higher RT with the same fixation rate, it is plausible that the number of fixation was higher as well. This effect would be secondary though: the RT is a covariate for the number of fixations. That is in line with the findings of the present study, where the higher number of fixation is a result of a longer RT. The different findings in fixation rate and fixation time between Hertzum *et al* (2009) and the present study can be a result of the tasks. In the study of Hertzum *et al* the subject had to search information on a website, whereas the tasks of the present study consisted of solving a pictorial task.There are large differences in eye movements between the two studies; Hertzum *et al* having a higher fixation rate and a shorter fixation time than the present study. Using a website – that consists for the largest part out of reading (Cooke, 2010) – affects the eye movements in another way than a pictorial task. Again, it appears that the effects of thinking aloud differ strongly along the task types (Russo *et al*,1989). The outcomes of the present study point out, however, that the subject has a lower fixation rate and a longer fixation time in the TA condition. These outcomes are related: as the fixation time increases, there will automatically be fewer fixations per second.

Fourth, concerning the spatial aspects of the eye movement behavior, there is no difference in saccade length between the TA condition and the silent condition. This finding confirms the fifth (H5) hypothesis and is in line with Hertzum *et al* (2009). This indicates that thinking aloud does not alter the distribution of the fixation points. Unfortunately, as a result of the calibration error, nothing can be said about the location of the fixation points.

Effect of task type

The four task types had distinguishable effects on the eye movements of the subject. The differences in in eye movements within the task types are effected by the visual layout of the tasks. Shifting attention over the points

of interests is reflected in the fixation points and saccades (Geiselman, 1977). The visual layout of the stimulus clearly affects the eye movements. Possibly, each task type requires a specific problem solving strategy and this is reflected in the eye movements. This is beyond the scope of the present study though.

Thinking aloud does not have different effects on each of the four task types. Although there are small differences and some effects lack significance, in general the effects of TA on each task matches with the effects of the whole test battery. In the study of Russo *et al* (1989) the reactivity varied over the different tasks. The contrast between their study and the present study is most likely caused by the similarity of the tasks used by the present study.

Putting it together

In the TA condition the RT is longer, the proportion correct is higher, the fixation rate is lower and the fixation time is longer than in the control condition. The saccade length does not vary between the two conditions. What is the relation between these variables? The very basic difference between the TA condition and the silent control condition is the articulation of the heeded information. Whereas the fixation time in the silent condition reflects the amount of time needed to process the information in a particular visual area (Henderson *et al*, 1999; Rayner, 1998; 2009), the fixation time in the TA condition reflects the time needed to articulate the information in a particular visual area (Geiselman, 1977). Clearly, the articulation causes longer fixation time in the TA condition (see for the same effect in reading aloud: Rayner, 1998; 2009). Longer fixation time results in a lower fixation rate.

Further there is a higher number of fixations in the TA condition, while the saccade length does not vary between the two conditions. This indicates that the subject fixates at the same points more often while thinking aloud. This could imply that the subject was looking for more information to come to a solution for the task. This hypothesis is in line with the correlation between RT and proportion correct. It seems justified to conclude that the subjects of the present study took more time to gather information – which effects a higher number of fixations with the same deviation – trying to answer correctly in the TA condition. Indeed this resulted in a higher proportion correct answers. It is plausible that the subjects were more motivated to give a correct answer because their reasoning was overheard by the experimenter. Subjects that know on forehand that they will get examined about their test results, tend to score higher on these tests (Tetlock and Kim, 1987). In other words, while thinking aloud, people move along the speed-accuracy tradeoff to more accurate, which diminishes the speed and results in longer RT's and higher proportion correct answers.

Conclusion and implication

Getting back to the research question – what is the difference in eye movements between the TA condition and the silent condition – only one variable is different between the two conditions as result of thinking aloud: the fixation time. The fixation time is longer in the TA condition because of the time needed to articulate the heeded information. Other differences found are secondary effects that are not a result of thinking aloud, but a result of another variable: lower fixation rate as result of a longer fixation time; higher number of fixation, longer scanpath length and higher proportion correct as result of a longer RT. Possibly the longer RT is a result of a motivation shift of the subject, which could be affected by the fact that the subject's reasoning got overheard by the experimenter.

The longer fixation time in the TA condition does fit in the theory of Ericsson and Simon (1980, 1997). In their theory they have not included eye movements, but a prolonged fixation time as a result of the articulation does not contradict in any way with the three-level model of verbalization. In the very basics, the longer RT is also in line with the theory. Ericsson and Simon state that the longer RT is a result of processing the information into verbal form and that this processing slows down the subject, but does not alter the cognitive process. The present study, however, concludes that the prolonged RT is a result of a motivational shift. This leads to alteration of the cognitive process in such way that it improves the test score. Thinking aloud does affect the subject's performance, but it is a secondary effect.

With the findings of the present study in mind, it seems that the validity of the think aloud method is dependent on the goal of the study. If the investigator is interested in the cognitive process in which no test score is involved – that is, the subject cannot do it right or wrong – the think aloud method might be a good choice. However, if there are test scores involved, thinking aloud tents to motivate subjects to perform better, which leads to unrepresentative test outcomes.

Limitations and future studies

The present study has two major limitations. The first issue is the inexperience of the subjects. The subjects never participated in a think aloud setting before, they had their first encounter with this research method during the present study. This resulted in an awkward experience, a wrong way of speaking out their thoughts (more like introspection) and lack of verbalization. This lack of verbalization is the second limitation. It was not monitored at what moments and how many times the subjects stopped talking, how long the silence moments were and how often the experimenter had to ask the subject to think out load. As a more demanding task leads to verbalization difficulties (Preece, 2007), the silent moments reflect an interesting cognitive process. The silent moments in the verbal reports can be an interesting subject for a future study. Last the uncertainty about the

effects of motivational shift and experimenter presence: The design of the present study makes it impossible to – first – test whether motivational shift indeed affected the RT and – second – whether the motivational shift was affected by the experimenter overhearing the subject. These would be fascinating questions for future studies.

REFERENCES

Kyd, C., (2009). Excel 2007's Ribbon Hurts Productivity, Survey Shows. *Online article at www.exceluser.com*

Atkinson, R.C., & Shiffrin, R.M. (1968). Human memory: A proposed system and its control processes. *The psychology of learning and motivation, 2,* 89-195

Bowers, V.H., & Snyder, H.L. (1990). Concurrent versus Retrospective Verbal Protocol for Comparing Window Usability. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting October 1990, 34,* 1270-1274

Branch, J.L., (2000). Investigating the Information-Seeking Processes of Adolescents: The Value of Using Think Alouds and Think Afters. *Library & Information Science Research, 22 (4),* 371-392

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

Duncker, K., (1945). On problem solving. *Psychological Monographs, 58 (5),* 270

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 (Rev. ed.).* Cambridge : MIT Press.

Fox, M.C., & Charness, N., (2009). How to Gain Eleven IQ Points in Ten Minutes: Thinking Aloud Improves Raven's Matrices Performance in Older Adults. *Aging, Neuropsychology, and Cognition, 17 (2),* 191-204

Geiselman, R.E., & Bellezza F.S., (1977). Eye-movements and overt rehearsal in word recall. *Journal of Experimental Psychology: Human Learning and Memory. 3 (3),* 305-315

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

Hooge, I.T.C., (Unpublished Internal paper). Faculty Social Science, Utrecht University, The Netherlands

Nielsen, J., Clemmensen, T., & Yssing, C., (2002). Getting access to what goes on in people's heads?: reflections on the think-aloud technique. *NordiCHI '02 Proceedings of the second Nordic conference on Human-computer interaction* 101-110

Nisbett, R.E., & Wilson, T.D., (1977). Telling More Than We Can Know: Verbal Reports on Mental Processes. *Psychological Review, 84 (3),* 232-259

Preece, J., Rogers, Y., & Sharp, H., (2007). *Interaction design: beyond human-computer interaction (2nd ed.)*. Chichester, West Sussex : Wiley

Rayner, K., (1998). Eye Movements in Reading and Information Processing: 20 Years of Research. *Psychological Bulletin, 124 (3),* 372-422

Rayner, K., (2009). Eye movements and attention in reading, scene perception, and visual search. *The Quarterly Journal of Experimental Psychology, 62 (8),* 1457-1506

Rhenius, D., & Deffner, G., (1990). Evaluation of concurrent thinking aloud using eye-tracking data. *Proceedings of the human factors society annual meeting, 34 (17),* 1265-1269

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

Tetlock, P.E., & Kim, J.I., (1987). Accountability and judgment processes in a personality prediction task. *Journal of personality and social psychology, 52 (4),* 700-709

Van der Haak, M.J., De Jong, M.D.T. & Schellens, P.J., (2003). Retrospective vs. concurrent think-aloud protocols: testing the usability of an online library catalogue. *Behaviour & Information Technology, 22 (5),* 339-351

Van der Haak, M.J., De Jong, M.D.T. & Schellens, P.J., (2004). Employing think-aloud protocols and constructive interaction to test the usability of online library catalogues: a methodological comparison. *Interacting with Computers 16*, 1153–1170

Van Someren, M.W., Barnard, Y.F. & Sandberg, J.A.C., (1994). *The think aloud method: A practical guide to modelling cognitive processes*. London : Acadamic Press

APPENDIX A: Tables

Table A1: Effects of tasktype on response time (RT), proportion correct answers, number of fixations, fixations rate, fixation time (FT) and saccade length. Significant differences are flagged with a star (*).

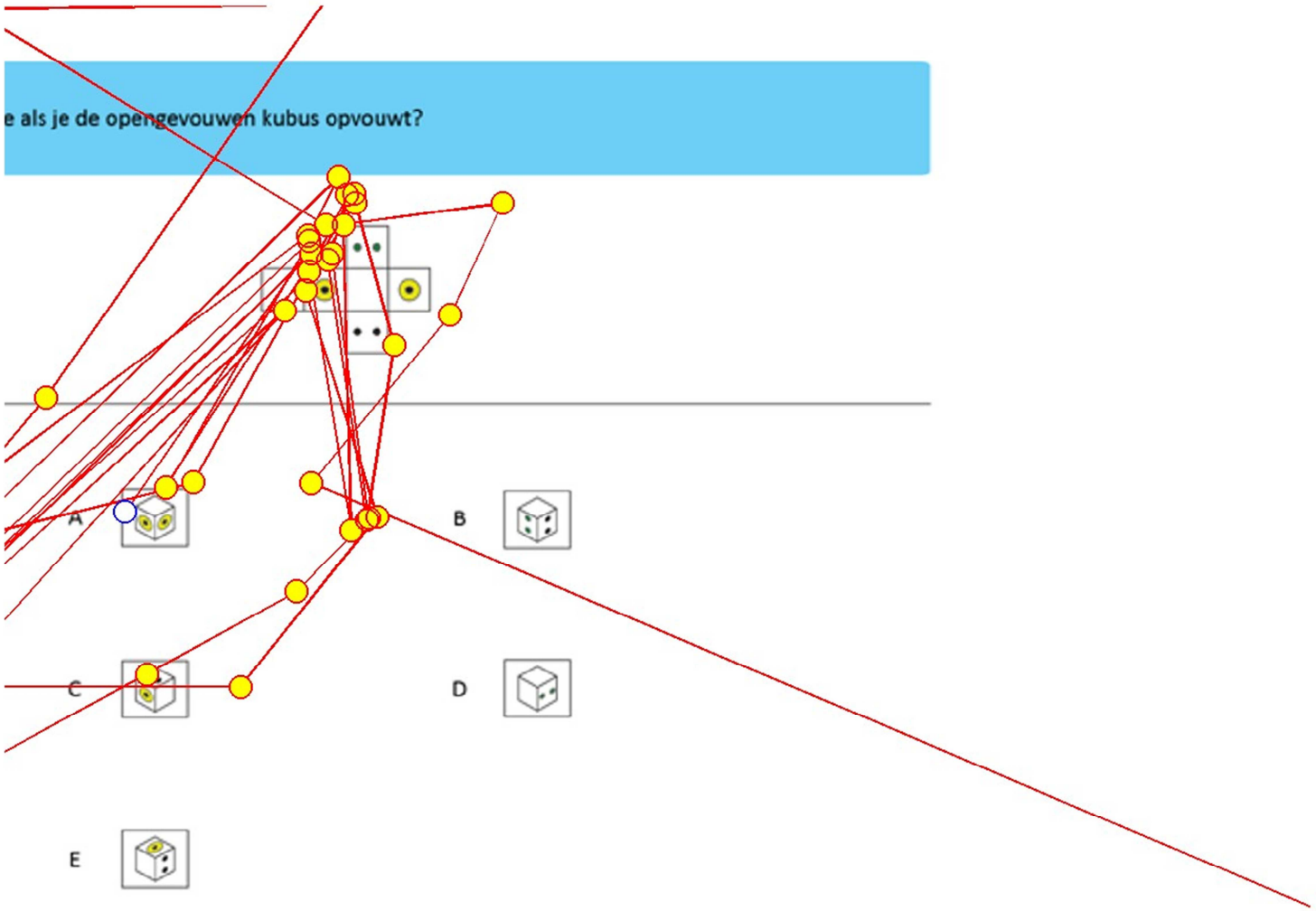
	cubicles		logical series		puzzles		visual series			
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	ANOVA	Significance
RT (s)	**35.85**	**24.09**	**36.77**	**20.55**	**29.55**	**14.76**	**39.80**	**21.37**	***F*(3,356) = 4.386, *p* = .005**	***p* < .01**
cubicles	-		*n.s.*		*n.s.*		*n.s.*			
logical series	*n.s.*		-		7.22*		-10.25*			
puzzles	*n.s.*		-7.22*		-		*n.s.*			
visual series	*n.s.*		10.25*		*n.s.*		-			
Proportion correct	**0.79**	**0.41**	**0.64**	**0.48**	**0.76**	**0.43**	**0.75**	**0.44**	***F*(3,356) = 2.186, p = .089**	***n.s.***
cubicles	-		*n.s.*		*n.s.*		*n.s.*			
logical series	*n.s.*		-		*n.s.*		*n.s.*			
puzzles	*n.s.*		*n.s.*		-		*n.s.*			
visual series	*n.s.*		*n.s.*		*n.s.*		-			
Number of fixations	**68.21**	**45.97**	**65.05**	**36.07**	**53.80**	**25.92**	**84.38**	**43.40**	***F*(3,353) = 9.700 *p* < .001**	***p* < .001**
cubicles	-		*n.s.*		*n.s.*		*n.s.*			
logical series	*n.s.*		-		*n.s.*		-19.57*			
puzzles	*n.s.*		*n.s.*		-		-30.57*			
visual series	*n.s.*		19.57*		30.57*		-			
Fixation rate	**1.95**	**0.41**	**1.88**	**0.59**	**1.92**	**0.39**	**2.21**	**0.50**	***F*(3,353) = 8.658 *p* < .001**	***p* < .001**
cubicles	-		*n.s.*		*n.s.*		-.26*			
logical series	*n.s.*		-		*n.s.*		-.33*			
puzzles	*n.s.*		*n.s.*		-		-.29*			
visual series	.26*		.33*		.29*		-			
FT (ms)	**496.61**	**112.77**	**553.30**	**203.93**	**503.16**	**115.91**	**440.75**	**105.05**	***F*(3,353) = 7.661 *p* < .001**	***p* < .001**
cubicles	-		*n.s.*		*n.s.*		55.86*			
logical series	*n.s.*		-		*n.s.*		112.55*			
puzzles	*n.s.*		*n.s.*		-		62.40*			
visual series	-55.86*		-112.55*		-62.40*		-			
Saccade length (pixels)	**265.15**	**45.81**	**228.34**	**53.26**	**285.62**	**47.95**	**224.12**	**42.68**	***F*(3,353) = 35.611 *p* < .001**	***p* < .001**
cubicles	-		-36.81*		-20.47*		41.03*			
logical series	36.81*		-		-57.28*		*n.s.*			
puzzles	20.47*		57.28*		-		61.50*			
visual series	-41.03*		*n.s.*		-61.50*		-			

APPENDIX A: Tables

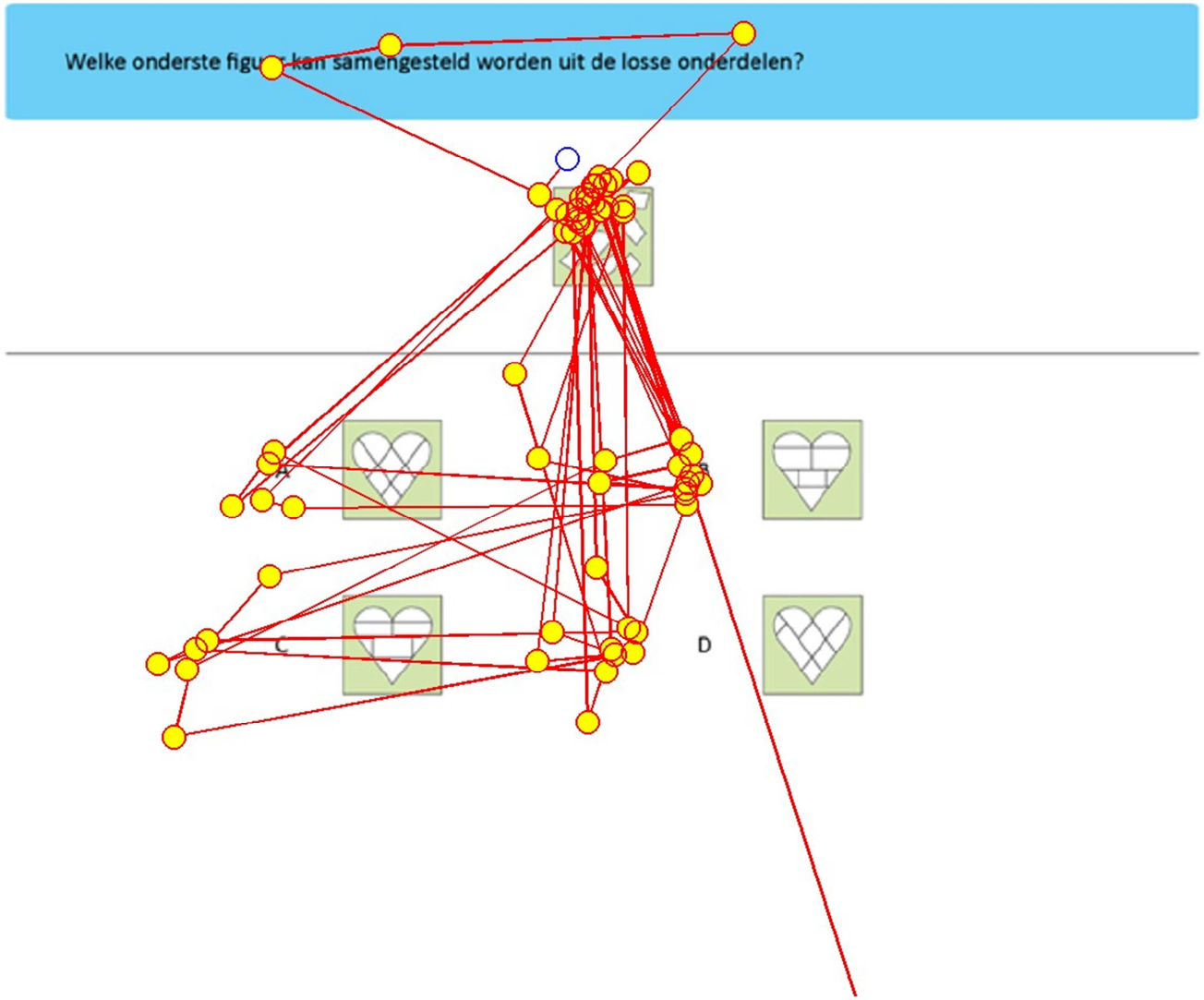
Table A2: difference in response time (RT), proportion correct answers, fixation rate , fixation time (FT) and saccade lngth for each tasktype. Tested for significance with t-tests.

		Silent		TA			
		Mean	SD	Mean	SD	t-test	Significance
Cubicles	RT (s)	25.03	17.09	46.67	25.39	$t(70) = -4.241$, $p < .001$	$p < .001$
	Proportion correct	0.72	0.45	0.86	0.35	$t(70) = -1.452$, $p = .151$	*n.s.*
	Fixation rate	2.10	0.40	1.79	0.36	$t(69) = 3.450$, $p < .001$	$p < .001$
	FT (ms)	455.13	97.10	536.94	113.51	$t(69) = -3.259$, $p = .002$	$p < .01$
	Saccade length (pixels)	278.31	48.66	252.35	39.44	$t(69) = -2.473$, $p = .016$	*n.s.*
Logical Series	RT (s)	27.72	14.95	45.81	21.50	$t(106) = -5.075$, $p < .001$	$p < .001$
	Proportion correct	0.61	0.49	0.67	0.48	$t(106) = -0.596$, $p = .552$	*n.s.*
	Fixation rate	2.10	0.60	1.66	0.49	$t(106) = 4.205$, $p < .001$	$p < .001$
	FT (ms)	481.12	148.39	625.48	226.54	$t(106) = -3.917$, $p < .001$	$p < .001$
	Saccade length (pixels)	230.15	52.14	226.53	54.78	$t(106) = -.946$, $p = .726$	*n.s.*
Puzzles	RT (s)	24.81	13.69	34.29	14.38	$t(106) = -3.507$, $p < .001$	$p < .001$
	Proportion correct	0.67	0.48	0.85	0.36	$t(106) = -2.284$, $p = .024$	$p < .05$
	Fixation rate	1.93	0.44	1.90	0.35	$t(104) = .385$, $p = .701$	*n.s.*
	FT (ms)	504.10	120.97	502.21	111.77	$t(104) = .083$, $p = .934$	*n.s.*
	Saccade length (pixels)	288.29	53.81	282.95	41.62	$t(104) = .571$, $p = .569$	*n.s.*
Visual Series	RT (s)	30.32	11.77	49.27	24.56	$t(70) = -4.176$, $p < .001$	$p < .001$
	Proportion correct	0.69	0.47	0.81	0.40	$t(70) = -1.082$, $p = .283$	*n.s.*
	Fixation rate	2.31	0.54	2.10	0.45	$t(70) = 1.779$, $p = .008$	$p < .01$
	FT (ms)	420.86	101.21	460.65	106.42	$t(70) = -1.626$, $p = .109$	*n.s.*
	Saccade length (pixels)	221.03	37.48	227.21	47.65	$t(70) = -.611$, $p = .543$	*n.s.*

APPENDIX B1: Example of the eye tracker measure error

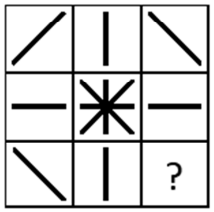


APPENDIX B2: Example of the eye tracker calibration error



Appendix C: Tasks

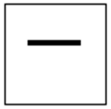
Welke afbeelding hoort op de plek van het vraagteken in het bovenste figuur?



A 

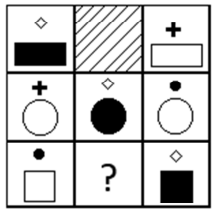
B 

C 

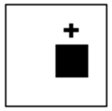
D 

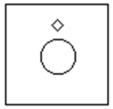
Appendix C: Tasks

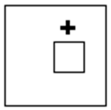
Welke afbeelding hoort op de plek van het vraagteken in het bovenste figuur?



A 

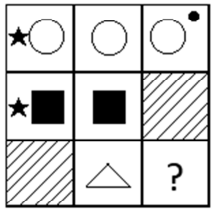
B 

C 

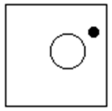
D 

Appendix C: Tasks

Welke afbeelding hoort op de plek van het vraagteken in het bovenste figuur?



A 

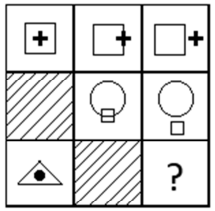
B 

C 

D 

Appendix C: Tasks

Welke afbeelding hoort op de plek van het vraagteken in het bovenste figuur?



A 

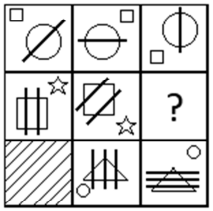
B 

C 

D 

Appendix C: Tasks

Welke afbeelding hoort op de plek van het vraagteken in het bovenste figuur?



A 

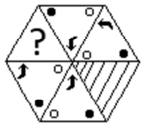
B 

C 

D 

Appendix C: Tasks

Welke afbeelding hoort op de plek van het vraagteken in het bovenste figuur?



A 

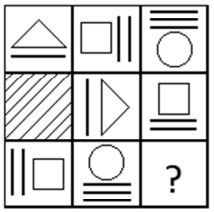
B 

C 

D 

Appendix C: Tasks

Welke afbeelding hoort op de plek van het vraagteken in het bovenste figuur?



A 

B 

C 

D 

Appendix C: Tasks

Welke onderste figuur kan samengesteld worden uit de losse onderdelen?

A 

B 

C 

D 

Appendix C: Tasks

Welke onderste figuur kan samengesteld worden uit de losse onderdelen?

A 

B 

C 

D 

Appendix C: Tasks

Welke onderste figuur kan samengesteld worden uit de losse onderdelen?

A 

B 

C 

D 

Appendix C: Tasks

Welke onderste figuur kan samengesteld worden uit de losse onderdelen?

A 

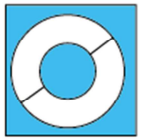
B 

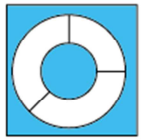
C 

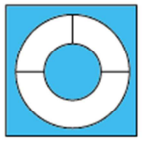
D 

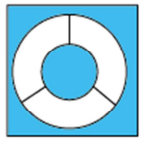
Appendix C: Tasks

Welke onderste figuur kan samengesteld worden uit de losse onderdelen?

A 

B 

C 

D 

Appendix C: Tasks

Welke onderste figuur kan samengesteld worden uit de losse onderdelen?

A 

B 

C 

D 

Appendix C: Tasks

Welke onderste figuur kan samengesteld worden uit de losse onderdelen?

A 

B 

C 

D 

Appendix C: Tasks

Welke onderste figuur moet logischerwijs op de plaats van het vraagteken staan in de bovenste reeks?

A 

B 

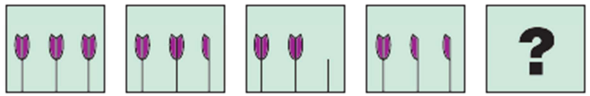
C 

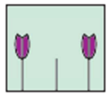
D 

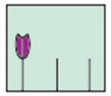
E 

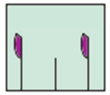
Appendix C: Tasks

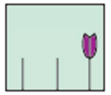
Welke onderste figuur moet logischerwijs op de plaats van het vraagteken staan in de bovenste reeks?

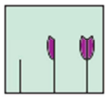


A 

B 

C 

D 

E 

Appendix C: Tasks

Welke onderste figuur moet logischerwijs op de plaats van het vraagteken staan in de bovenste reeks?

A 

B 

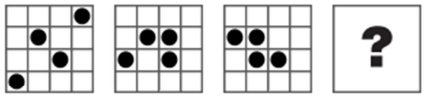
C 

D 

E 

Appendix C: Tasks

Welke onderste figuur moet logischerwijs op de plaats van het vraagteken staan in de bovenste reeks?



A 

B 

C 

D 

E 

Appendix C: Tasks

Welke onderste figuur moet logischerwijs op de plaats van het vraagteken staan in de bovenste reeks?

A 

B 

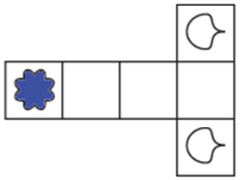
C 

D 

E 

Appendix C: Tasks

Welke kubus krijg je als je de opengevouwen kubus opvouwt?



A 

B 

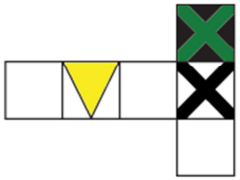
C 

D 

E 

Appendix C: Tasks

Welke kubus krijg je als je de opengevouwen kubus opvouwt?



A 

B 

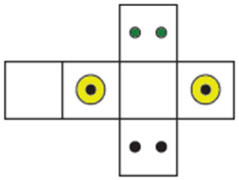
C 

D 

E 

Appendix C: Tasks

Welke kubus krijg je als je de opengevouwen kubus opvouwt?



A 

B 

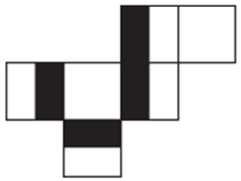
C 

D 

E 

Appendix C: Tasks

Welke kubus krijg je als je de opengevouwen kubus opvouwt?



A 

B 

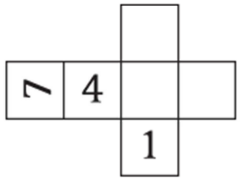
C 

D 

E 

Appendix C: Tasks

Welke kubus krijg je als je de opengevouwen kubus opvouwt?



A 

B 

C 

D 

E