Cognitive Load Measurement:

Different instruments for different types of load?

Geert Wernaart

3340422

Master Onderwijskundig Ontwerp en Advisering

Universiteit Utrecht

Begeleider	Femke Kirschner
Tweede beoordelaar	Jeroen Janssen
Datum	18 juni 2012

Abstract

This study focused on the measurement of two types of cognitive load: intrinsic and extraneous load. In the past, a variety of measurement instruments have been used to indicate the overall cognitive load, but distinguishing between intrinsic and extraneous load remains difficult. Three different kinds of instruments to measure cognitive load were compared on what type of cognitive load they were able to measure: a subjective rating scale, a thinking aloud dual task and an eye tracker. Participants were given a number of puzzles to solve, which varied in intrinsic (high vs. low) and extraneous (high vs. low) cognitive load. Results show that all instruments were able to capture intrinsic load, but only the eye tracking parameter fixation duration was able to capture extraneous load. These findings are discussed in light of the used experimental material and its implications for future research on cognitive load and the distinction between intrinsic and extraneous load.

Introduction

Cognitive Load Theory is a theory on information processing, that relates working memory restraints to the effectiveness of instruction. In order for learning activities to occur, a too high cognitive load should be avoided. However, the ability to measure this load is the basis for making statements about the load imposed by a task. This study focuses on what aspects of cognitive load are measured by a variety of commonly used mental effort measurement instruments and how this can influence future measurement of cognitive load and further research on the theory.

Cognitive Load Theory (CLT) is a theoretical framework based upon human cognitive architecture, which consists of a long-term memory and a working memory (Sweller, 1988). CLT emphasizes working memory constraints as determinants of instructional design effectiveness. Since working memory is very limited in both capacity (4±1 elements) (Cowan, 2001) and duration (Peterson & Peterson, 1959), it should be used as effectively as possible so that schema construction in the long-term memory, whose capacity is unlimited, is optimized and relevant learning activities occur. Once information has been stored in long term memory, it can be retrieved when needed. Therefore, if learning has occurred over a long period of time, working memory limitations will be reduced.

CLT argues that there are two critical learning mechanisms: schema acquisition and automation (Sweller, 1994). Schema acquisition refers to the assumption that knowledge in long-term memory is stored in cognitive schemas that incorporate multiple elements of information hierarchically into one single element. An element is anything that needs to be learned, most often a single information entity (Sweller, et al., 1998). Once elements are stored in schemas in long-term memory, working memory has more capacity to process new elements. Automation refers to the process of processing elements with less effort (Sweller, 1994). Schema acquisition and automation can only occur when working memory is effectively used to process new information. If this does not occur, false or incomplete information or no information at all is stored in long term memory. Thus, for relevant learning activities to occur, working memory needs to be used effectively.

Cognitive load occurs when someone needs to process information in their working memory, for example during a task, a test or instruction. Since working memory is limited, a high cognitive load can occur and schema acquisition and automation are hindered when someone has to put too much mental effort in a learning activity. When the load is too high, this may result in cognitive overload. Cognitive load can be divided over a group of people, so the individual load is lower (Kirschner, et al. (2009). However, since most learning activity occurs individually, most research focuses on individual cognitive load and how it can be influenced. Cognitive load is seen as a very important factor to consider when designing effective instruction because instructional control of cognitive load is critically important to meaningful learning. When there is high cognitive load, working memory is used less effectively and less relevant learning activities occur. To obtain control over the imposed cognitive load, the load needs to be measured. The ability to measure and influence the cognitive load a task imposes on a learner is relevant for the design of effective instruction and it can gain more insight into the processes that underlie cognitive load. Although CLT distinguishes three different types of cognitive load (intrinsic, extraneous and germane load), it is usually measured as one concept. Therefore, the measurement of cognitive load is the main criticism on CLT. To verify that there are different types of cognitive load, they need to be individually measurable. Up to now, overall cognitive load was measured and different types of load were distinguished through assumptions on the research design. Being able to distinguish them, regardless of the design, would make research on cognitive load more relevant.

In a recent study, Jarodzka et al. (2012) suggest that different kinds of measurement techniques might measure different aspects of cognitive load. This study investigates whether this assumption holds ground and focuses on what type of cognitive load is actually measured by different, common-used, cognitive load measurement instruments. This gains more insight into the measurement of cognitive load, so future research on CLT can be more effective.

Types of cognitive load

In most current descriptions of CLT, three types of cognitive load are distinguished (van Merriënboer & Sweller, 2005). First, intrinsic cognitive load, which is caused by the number of

elements in a task and the degree to which these elements have to be related to each other. Intrinsic load may also be referred to as the task complexity. Second, extraneous cognitive load, which is caused by mental activities and elements that do not directly support learning. These activities and information use working memory capacity that does not result in schema acquisition or automation. Extraneous load is caused by 'bad' instructional design and should, therefore, be decreased as much as possible. Third, germane cognitive load, which is caused by mental activities and information that do support learning. Unlike extraneous cognitive load, germane cognitive load should be increased by instructional design. Intrinsic cognitive load is mainly imposed by the task complexity, while extraneous and germane cognitive load are mainly imposed by the way something has to be learned: the instructional design. Since intrinsic load cannot be influenced by the instructional design, the combination of extraneous and germane load should be as effective as possible. The greater the proportion of germane cognitive load, the greater the potential for learning is. Therefore, learning something requires a shift from extraneous load to germane load. The three types of cognitive load are additive and the total load imposed by a task cannot exceed working memory capacity. The total germane and extraneous load combined, is assumed to equal the total cognitive load minus intrinsic load.

In CLT, the main mechanism of intrinsic load is considered to be element interactivity (Sweller, et al., 1998). Element interactivity refers to what degree an element needs to be learned by integrating other elements. An element is easy to learn when it has low element interactivity, because it imposes a low working memory load. An often-used example is learning the vocabulary of a foreign language. A new word is relatively easy to learn since it can be learned with minimal reference to other elements. However, learning a foreign language's grammar is far more difficult, because it has high element interactivity. When learning to apply a grammatical rule, the learner needs knowledge about other related elements, e.g. tense of a sentence. Elements must then be learned simultaneously because they interact and therefore induce a higher intrinsic cognitive load. Instructional design has little influence on element interactivity because it is caused by what has to be learned and not by how it is presented. Element interactivity is dependent on the learner's expertise, which is the knowledge stored in the long term memory, and determined by learner characteristics. For a person with high

expertise, low-element interactivity might occur during a task, but for a person with less expertise doing the same task, high-element interactivity might occur. Therefore, instructional design should take in account the level of expertise of the learner, in order to make assumptions about the amount of imposed intrinsic load on the learner (Sweller, et al., 1998).

On the other hand, instructional design can influence the amount of extraneous load to a large degree. When element interactivity is high, instructional design should attempt to reduce extraneous cognitive load because full working memory capacity is needed (Sweller, 1994). Much research has focused on how presentation formats increase extraneous load (and how this can be avoided). There are many findings regarding cognitive load effects of certain changes to the instructional design. The most influential effects are the worked example effect, the split-attention effect, the redundancy effect and the expertise-reversal effect (for an overview, see: Van Merrienboer & Sweller, 2005). The split-attention effect was used in this study to influence the extraneous load of tasks. This effect refers to the increase in extraneous load when information elements which are connected, are presented separately. An example of this effect is the case when a diagram is presented, while nothing about the solution of a problem can be revealed without certain statements, which are presented in a separate space. Learners must integrate these two sources of information, in order to get to a solution. This process can be cognitively demanding because it invokes working memory load (Sweller, et al., 1998). In order to connect the separately presented information elements, they need to be stored in working memory. Now working memory has less capacity to process all the information. All cognitive load effects can only be obtained under conditions of high intrinsic load (and thus high element interactivity). If intrinsic load is low, it would be unnecessary to reduce extraneous load, since enough working memory capacity is available.

Sweller (2010) argues that element interactivity is not only the main mechanism of intrinsic cognitive load, but it also underlies extraneous cognitive load. Sweller suggests that element interactivity is the major source of working memory load for both types of cognitive load. The load is extraneous when it can be reduced by altering the learning material. The load is intrinsic when it can be reduced by changing what needs to be learned. The various cognitive load effects (worked example, split-attention, redundancy and expertise-reversal effect) are usually discussed in isolation in the

context of extraneous load, but Sweller (2010) suggests they might have the same underlying cause: element interactivity. Sweller continues by stating that germane cognitive load can also indirectly be specified in terms of element interactivity, because it consists of the working memory resources that have to deal with the element interactivity associated with intrinsic load. When extraneous load increases, germane load decreases and less working memory resources are focused on the relevant element interactivity. Sweller suggests that "by defining both intrinsic and extraneous cognitive load in terms of element interactivity, it may be possible to analyze element interactivity prior to an experiment and so more easily predict experimental outcomes" (Sweller, 2010, p. 137).

Kalyuga (2011) continues this line of thought and argues that germane load and intrinsic load are indistinguishable. Kalyuga elaborates on the origins of CLT and underlines that the main focus was the reduction of extraneous load. This concept was soon followed by the concept of intrinsic load, because some empirical results could not be explained by extraneous load, but appeared to be caused by element interactivity. The concept of germane load was later introduced, but not because empirical results could not only be explained by extraneous and intrinsic load. At the time, these two loads were treated as needing to be reduced to foster learning. Therefore, a third concept was introduced to account for the cognitive effort that fosters learning. Current belief still stresses that extraneous load should be reduced, but intrinsic load simply cannot be reduced by instructional design and does not necessarily need to be reduced. The cognitive load that contributes to learning, by Kalyuga (2011) described as the 'good' load, may fit both definitions of intrinsic and germane load, while the 'bad load' (extraneous load) can still be clearly distinguished from it. Thus, the concept of germane load might be redundant and therefore, this study only focused on the differences between measuring intrinsic load and extraneous load. It is assumed that measuring germane load is not possible nor relevant, since it can be compared to measuring intrinsic load. The additive cognitive load could then only be divided into intrinsic load and extraneous load.

Measuring cognitive load

As mentioned before, the main criticism of CLT is that only the total cognitive load can be measured. Since CLT is a theoretical framework, it is difficult to observe and although various

measurement instruments have been designed, there appears to be no instrument that is able to measure the different types of load individually. Since cognitive load is additive, it is difficult to say what is caused by intrinsic load and what is caused by extraneous load. It is, however, important to be able to distinguish intrinsic load and extraneous load, in order to properly influence the amount of load that is imposed on a learner. Besides simply looking at performance outcomes, the most frequently used instruments can be categorized into three categories: subjective measures, task- and performance based measures and physiological measures (Sweller, et al., 1998; Van Mierlo et al., 2012). Indications of mental effort are commonly used to measure cognitive load, since an increased load requires increased mental effort and mental effort can be perceived by various measurement instruments.

Subjective measurement techniques refer to indications of experienced mental effort by participants. The subjective rating scale by Paas (1992) is the most well-known. On this subjective nine-point cognitive load rating scale, the learner can rate their mental effort from (1) very, very low to (9) very, very high. Subjective measures rely heavily on the monitoring of a participant's own mental efforts, which might be difficult to do for a participant. Furthermore, the subjective rating scale is not standardized and it is difficult to compare results from different studies, since it is commonly used to measure the differences between mental effort of participants within one study and not between studies.

Despite the fact that the subjective rating scale has proven to be valid and non-intrusive, Kirschner, Ayres and Chandler (2011) note that the subjective rating scale is not able to measure the different types of cognitive load individually but only cognitive load as one concept, even though the distinction between these types is commonly accepted and most research focuses on how to reduce extraneous load. Beckman (2010) explains that a decrease or increase in total cognitive load is difficult to interpret, since it could both be positive (when extraneous load is reduced or germane load is increased) and negative (when extraneous load is increased or germane load is reduced). This is usually accounted for by comparing mental effort to performance on a task. When the cognitive load is high and performance is low, it is assumed that the perceived load is extraneous and when the load is low and the performance is high, it is assumed that the load is germane. However, performance data

may not always be available and it is an assumption that cannot be falsified. Another option is to have controlled experimental settings where one of the loads is explicitly altered and where the difference in total cognitive load between conditions can be accounted for that one load.

Sweller (2010) suggests that the subjective rating scales by Paas (1992) can be used to measure extraneous load. As long as intrinsic load (and thus element interactivity) remains the same, the overall change in cognitive load can be contributed to a change in extraneous load because, as mentioned earlier, Sweller (2010) suggests the germane load is equal to the intrinsic load. Ayres (2006), on the other hand, argues that subjective rating scales can be used to measure intrinsic load, as long as germane and extraneous load are kept constant. In addition, Sweller et al. (2011) state that the types of load can be distinguished in research by using an appropriate experimental design, thus influencing one of the loads while keeping the others constant.

Hart & Staveland (1988) developed the NASA-TLX, which is a multi-dimensional subjective measure of workload. The measurement takes place after a sequence of learning tasks. One-dimensional measurements can take place during a sequence or task, more specifically after each sub-task (Paas, 1992). Another subjective mental effort rating scale was introduced by Kalyuga, Chandler and Sweller (1999), which consists of a seven-point scale, ranging from 'extremely easy' to 'extremely difficult'. Although this scale has proven to be valid in indicating cognitive load, it still does not make a distinction between the three loads. In a study by Huang (2010), participant's self-reports were used on two items, measuring the intrinsic load as mental effort investment and germane cognitive load as difficulty rating. Both scores turned out to be significantly different from each other. Consequentially, the subjective rating scale as designed by Paas (1992) remains to be the most reliable and most commonly used subjective measurement instrument.

Task- and performance- based techniques focus on either performance on the learning task or performance on a secondary task (a dual task). For example, Brünken, Plass & Leutner (2003) studied the feasibility of this approach by measuring the reaction time in a multimedia setting by asking the participant to click as soon as the color of the screen changed as the second task. They argue that a dual task can be used to indicate the cognitive load as it occurs. In comparison to the subjective

instruments, the dual task approach can be a more objective, less disruptive measurement. However, it needs to be a very low-cognitive task for the participant.

Thinking aloud, or concurrent verbal reporting, is sometimes used a dual task. The participant is asked to actively verbalize their thoughts as they perform a task. When the primary task is difficult, the performance on this secondary task usually decreases, resulting in silence, indicating high mental effort (Yin & Chen, 2007). One of the main advantages of the dual task approach is that in contrast to subjective techniques, they attempt to capture the amount of mental effort during task performance instead of afterwards. One of the main disadvantages is that the dual task might interfere with the primary task. However, besides using silent pauses as indicators of mental effort, explicit utterances can also be used as parameter. When the load is higher, participants often use more explicit utterances of mental effort (Jarodzka, et al., 2012).

Furthermore, physiological techniques are also a more objective way to measure mental effort, for example by measuring heart rate or brain activity. Although they provide insight in the pattern of the load and can measure the load as it occurs (instead of afterwards like subjective techniques), they are intrusive instruments. Participants might get uncomfortable, which might distract them from the task. However, a less intrusive physiological technique is eye tracking. This apparatus measures different movements of the eye balls. Even though eye tracking is often used in cognitive research, it has rarely been used to measure cognitive load. Parameters like the increase of pupil dilation (Klinger, Tversky & Hanrahan, 2011) and the decrease of fixation duration (Van Orden, Limbert, Makeig & Jung, 2001) have proven to be related to mental effort (Holmqvist et al., 2011). Van Gog et al. (2009) used eye tracking and concurrent verbal reporting in their study on CLT. They found that combining the two measurement instrument can gain greater insights into cognitive load than using only one of them, because they tend to measure different cognitive aspects.

In a recent study, Jarodzka et al. (2012) combined these three different measurement techniques (subjective, physiological and dual task) to analyze the split-attention effect on a computer based test. During the test, participants were asked to think aloud (as a dual task) and eye movements were recorded using eye tracking equipment. Afterwards, participants were asked to rate their perceived mental effort on a subjective rating scale. Eye movement was monitored to investigate what

students were looking at and for how long. Thinking aloud protocols where used to investigate how long participants would pause their speech, indicating the amount of mental effort and to count explicit utterances of increased mental effort. The findings led to the suggestion that a distinction can be made between explicit and implicit mental effort measurement, because it turned out that the different measurement instrument measured different concepts. Correlations were found between some rather implicit measurement parameters (eye tracking, silent pauses) and some rather explicit parameters (subjective ratings and explicit utterances). Although, no direct relation is laid between the suggestion of implicit and explicit aspects of cognitive load and the distinction between intrinsic load and extraneous load, it appears that different measurement instruments are able to measure different aspects of cognitive load.

Following Sweller (2010)'s and Kalyuga (2011)'s line of thought, it can be assumed that the concept of germane load is redundant and might be measured the same way as intrinsic load. Therefore, only the distinction between intrinsic and extraneous load seems relevant for this study. Jarodzka et al. (2012) suggest that further research on the different measurement instruments that they used needs to be conducted in order to make stronger statements about whether these instruments really do measure different aspects of cognitive load. Additionally, it is relevant to investigate whether they are able to distinguish intrinsic load and extraneous load, for further research on cognitive load purposes. If these instruments really do measure different aspects of cognitive load, this would shed new light on CLT and makes research on it more efficient. It would then perhaps be possible to measure more directly what the influence of either intrinsic or extraneous load is on the total load. The purpose of this study was finding an answer to the following question: *Do different cognitive load measurement instruments measure different types of cognitive load?*

Since the subjective rating scale by Paas (1992) has proven to be a valid measurement instrument and since it should be able to capture intrinsic and extraneous load separately as long as one of the two loads remains constant in the research design, it was expected that the subjective ratings would be able to measure both types of cognitive load individually. Furthermore, it was expected that the amount of pauses and explicit utterances in the concurrent verbal reporting would be higher when the cognitive load was higher. It was also expected that fixation duration would be shorter and pupil

dilation would be larger when the cognitive load was higher. Since intrinsic load and extraneous were manipulated in this study, it was expected that these parameters would be able to distinguish both types of load. Therefore, all these parameters were used in the study to investigate the differences between them in what they can actually measure. There were no assumptions as to what degree they would differ and thus, which parameter would be best to measure either intrinsic or extraneous load. However, it was expected that there would be differences between the parameters to what degree they would be able to capture intrinsic and extraneous load.

Method

Design

In this study, a subjective, a dual task and a physiological cognitive load measurement instrument were used to investigate what type of cognitive load they can each measure. This combination of instruments is equal to the ones used by Jarodzka et al. (2012), but this study focused specifically on whether they measure intrinsic load or extraneous load, by explicitly manipulating the intrinsic and extraneous load of the puzzles participants were given to solve. Two conditions were compared, each one with different combinations of intrinsic and extraneous load.

Participants had to solve two Sudoku's and two logic puzzles. The reason for giving the participants both types of puzzles was to account for individual differences in preference for a numeric or a verbal puzzle. Since intrinsic load is partly influenced by learner characteristics, individual differences needed to be accounted for by using puzzles that were expected to differ in the amount of intrinsic load they would impose on the participants.

Participants

30 college students (14 males, 16 females; M = 22.10 years, SD = 2.25) participated in this study. Participants were selected on their level of expertise on Sudoku's and logic puzzles. They were expected to have some experience in solving the puzzles, but no regular experience. Participants were randomly assigned to one of two conditions: 1) low extraneous load (7 males, 8 females, M = 22.67 years, SD = 2.06 years) or 2) high extraneous load (7 males, 8 females, M = 21.53 years, SD = 2.36 years). All participants were given two low intrinsic and two high intrinsic puzzles to solve on a computer. These four puzzles were the same for all participants, but in the second condition, with increased extraneous load, a split format was used. This generated a between subject design for the extraneous load and a within subject design for the intrinsic load.

Material

Sudoku puzzle task. Participants were asked to solve two Sudoku's, one low intrinsic and one high intrinsic. Sudoku's are numeric puzzles in which missing numbers need to be filled in. The difference in intrinsic load was accounted for by using a low intrinsic puzzle (4x4 sized, 12 cells needed to be filled in) and a high intrinsic puzzle (9x9 sized, 56 cells needed to be filled in). The difference in extraneous load was accounted for by using an integrated and a split format. The split format increased the extraneous load by imposing a split-attention effect, by changing the numbers normally used in a Sudoku puzzle into symbols. A 'translation' for these images was given separately from the puzzle. However, the participants did have to solve the puzzle using numbers. Therefore, the participant had to split their attention between the puzzle and the translation of the symbols, in order to solve the puzzle.

Logic puzzle task. Participants were also asked to solve two logic puzzle tasks, one low intrinsic and one high intrinsic. These are verbal puzzles where participants have to solve a problem by combining given clues of information. The difference in intrinsic load was accounted for by the amount of cells in a table the participants had to fill in: nine in the low intrinsic puzzle and twenty in the high intrinsic puzzle; and by the amount of thinking steps they had to take to get to the final solution. Like the Sudoku puzzles, the difference in extraneous load was accounted for by using a split and an integrated format. In the split format, the clues were presented separately from the puzzle, so participants had to split their attention in order to solve the puzzle.

Filler and retention task. After each logic task, participants did a filler task of approximately 30 seconds to refresh their working memory and to avoid them from repeating the clues from the logic task. One filler task consisted of a couple of math problems and the other was a spot the differences task. After these tasks, the participants did a retention task, in which they were asked to recall as many clues from the logic task as they could. This retention task was used to indicate how much the participants learned from the task and weather the amount of intrinsic and extraneous load influenced this. Participants were instructed beforehand that they would be asked to recall the clues, so this did not come as a surprise for them.

Performance. Performance on each puzzle was measured by the amount of correct cells a participant had filled in, this was translated into a correct value percentage of the total amount of answers. The performance on the retention tasks was measured by the number of correct clues the participant could recall. The performance on the filler tasks was not measured.

Subjective instrument. Participants rated their perceived amount of mental effort on the subjective rating scale, as designed by Paas (1992), immediately after every Sudoku and logic puzzle. The question was how much effort it took them to solve the puzzle, which they could rate on a 9-point scale from 1 (very, very low effort) to 9 (very, very high effort).

Dual task. Participants were asked to think aloud during the tasks and an audio recorder was used to record all verbal utterances by the participants during the tasks. Verbal reporting parameters were silent pauses (of minimum 2 seconds) and explicit utterances of increased cognitive load.

Eye tracking. A Tobii T60 eye tracker was used to record all eye movements of the participants. Eye tracking parameters were fixation duration (in ms) and pupil dilation (in mm).

Mental effort parameters. To capture indicators of mental effort, mean pupil dilations (in mm) and fixation durations (in ms) were obtained from the eye tracking data, while silent pauses (minimum 2 seconds) and explicit utterances of increased mental effort were obtained from thinking-aloud. The utterances indicated doubt or difficulty when solving the puzzles, for example 'I'm stuck', 'I don't know what to do' or 'this is difficult'. Since the low intrinsic puzzles lasted up to only five minutes and the high intrinsic puzzles up to fifteen minutes, the mean amount of pauses and utterances per minute was calculated.

Procedure

Before the test started, participants received a brief instruction about the procedure of the test and were instructed to think aloud according to Ericcson & Simons (1993). A standard calibration for the eye tracker was used. After the audio recorder was switched on, the participants were able to ask questions regarding the procedure of the test. They were then instructed to start the test and to start thinking and reading aloud immediately. During the test, a researcher was seated behind a curtain and

looked on a second screen to see if the program ran smoothly and to give extra instructions when needed. Whenever participants forgot to think aloud they were instructed to 'keep talking'.

The test started with some general on-screen instruction about the test followed by instruction about the Sudoku puzzle. The instruction was longer for the participants in the high extraneous condition, since the split format had to be explained to them. This was followed by the low intrinsic Sudoku puzzle, instruction, the low intrinsic logic puzzle, a filler task and a retention task, instruction, the high intrinsic Sudoku puzzle, instruction, the high intrinsic logic puzzle, a filler task and a retention task. The participants had a maximum of five minutes to solve each low intrinsic puzzle and fifteen minutes for each high intrinsic puzzle. When they finished a task early, they were told to continue to the next part of the test. However, this only occurred with the low intrinsic tasks. No participant was able to finish either one of the high intrinsic tasks within fifteen minutes.

The test ended with a demographic questionnaire about the participant's sex, age, educational background and experience with Sudoku and logic puzzles. The test had a total duration of approximately 45 minutes. A gift voucher was rewarded to the person who performed best at solving the puzzles, in order to motivate the participants to participate and try their best.

Analysis

A 2 (intrinsic load: low vs. high) x 2 (extraneous load: low vs. high) analyses of variance (ANOVA) with repeated measures on the first factor was used to analyze the data. Mental effort (as measured by the six different parameters) was used as dependent variable. A significance level of .05 was used for all analyses. The two types of puzzles were analyzed separately for each parameter, because results show they differed in the amount of intrinsic load they imposed on participants.

Results

All means and standard deviations of the scores on the parameters are presented in Table 1 for the low intrinsic puzzles and in Table 2 for the high intrinsic puzzles. All six parameters were used as dependent variables: subjective rating scale, fixation duration, pupil dilation, silent pauses, explicit utterances and performance. The independent variables were the amount of intrinsic load and the amount of extraneous load, which varied per condition. For each parameter, a separate 2 (intrinsic load: low and high) x 2 (extraneous load: low and high) ANOVA with repeated measures on extraneous load was performed.

First, a 2 (intrinsic load: low and high) x 2 (extraneous load: low and high) x 2 (puzzle: Sudoku and logic) ANOVA with repeated measures on extraneous load was used to analyze main effects and interaction effects for all six cognitive load parameters on the Sudoku's and the logic puzzles. As expected, significant main effects were found for the type of puzzle on the rating scale, F (1, 28) = 9.02, MSE = 1.96, $p < .05$ on the fixation duration, F (1, 28) = 15.38, MSE = 493.25, $p < .05$ and on the pupil dilation, F (1, 28) = 4.60, MSE = .01, $p < .05$. There were significant interaction effects between intrinsic load and the type of puzzle as well for the rating scale, F (1, 28) = 15.93, MSE = 1.93, $p < .05$, for fixation duration, F (1,28) = 19.79, MSE = 300.00, $p < .05$ and for pupil dilation, F (1, 28) = 7.97, MSE = .01, $p < .05$, indicating that the intrinsic load was different as measured by these three parameters according to the type of puzzle. There was only a significant interaction effect between extraneous load and the type of puzzle for fixation duration, F (1, 28) = 5.80, $p < .05$. There were no significant main and interaction effects for the type of puzzle on silent pauses, explicit utterances and performance, indicating that these parameters did not differ according to the type of puzzle.

Due to these findings, the scores on the Sudoku's and the logic puzzles were not combined in further analyses, but analyzed separately, in order to account for the differences in intrinsic load between the puzzles as perceived by the subjective rating scale and the eye tracking parameters. Next, each parameter was analyzed in a 2 (intrinsic load) x 2 (extraneous load) ANOVA for the Sudoku's and the logic puzzles separately. The type of puzzle was not the main focus of this study, but since

they differed in intrinsic load, they had to be analyzed separately for each of the six cognitive load measurement parameters.

Subjective Rating Scale. Regarding the subjective rating scale, there was a significant main effect for intrinsic load on the Sudoku's, F (1, 28) = 136.81, MSE = 2.53, $p < .05$, and for the logic puzzles, F (1, 28) = 50.26, MSE = 2.45, $p < .05$, indicating that subjective ratings were different on the low and high intrinsic versions of both puzzles. There was no significant main effect for extraneous load on both types of puzzles, indicating that subjective ratings on the low and high extraneous puzzles were in general the same, with F (1, 28) = 1.17, *ns* for the Sudoku's and F (1, 28) < 1, *ns* for the logic puzzles. There was also no significant interaction effect between intrinsic load and extraneous load on both the Sudoku, F (1, 28) < 1, *ns* and the logic puzzles, F (1, 28) = 1.58, *ns*, indicating that the intrinsic load did not differ according to the extraneous load.

Fixation Duration. Regarding the fixation duration, there was a significant main effect for intrinsic load on the Sudoku's, F (1, 28) = 17.02, MSE = 504.97, $p < .05$, but not on the logic puzzles, F (1, 28) < 1, *ns*, indicating that subjective ratings were different on the low and high intrinsic versions of the Sudoku's, but not on the logic puzzles. There was a significant main effect for extraneous load on the Sudoku's, F (1, 28) = 5.93, MSE = 504.97, $p < .05$, indicating that fixation duration on the low and high extraneous Sudoku's were different. There was no significant main effect for extraneous load on the logic puzzles, F (1, 28) < 1, *ns*, indicating that fixation duration on the low and high extraneous logic puzzles was in general the same. There was a significant interaction effect between intrinsic load and extraneous load on the Sudoku's, F (1, 28) = 6.10, $p < .05$, but not on the logic puzzles, F (1, 28) < 1, *ns*, indicating that the intrinsic load did differ according to the extraneous load on the Sudoku's, but not on the logic puzzles.

Pupil Dilation. Regarding the pupil dilation, there was a significant main effect for intrinsic load on the Sudoku's, F (1, 28) = 42.12, MSE = .01, $p < .05$, and the logic puzzles, F (1, 28) = 13.86, MSE = .01, $p < .05$, indicating that pupil dilation was different on the low and high intrinsic versions of both puzzles. There was no significant main effect for extraneous load on both types of puzzles, indicating that pupil dilation on the low and high extraneous puzzles was in general the same, with F (1, 28) < 1, *ns*, for the Sudoku and F (1, 28) < 1, *ns*, for the logic puzzles. There was also no

significant interaction effect between intrinsic load and extraneous load on both the Sudoku, F (1, 28) < 1, *ns* and the logic puzzles, F (1, 28) = 2.39, *ns*, indicating that the intrinsic load did not differ according to the extraneous load.

Silent Pauses. Regarding the silent pauses, there was a significant main effect for intrinsic load on the Sudoku's, F (1, 28) = 5.43, MSE = .21, p < .05 and the logic puzzles, F (1, 28) = 8.30, MSE = .17, p < .05, indicating that the amount of silent pauses was different on the low and high intrinsic versions of both puzzles. There was no significant main effect for extraneous load on both types of puzzles, indicating that the amount of silent pauses on the low and high extraneous puzzles was in general the same, with F (1, 28) < 1, *ns*, for the Sudoku and F (1, 28) < 1, *ns*, for the logic puzzles. There was also no significant interaction effect between intrinsic load and extraneous load on both the Sudoku, F (1, 28) < 1, *ns* and the logic puzzles. F (1, 28) < 1, *ns*, indicating that the intrinsic load did not differ according to the extraneous load.

Explicit Utterances. Regarding the explicit utterances, there was a significant main effect for intrinsic load on the Sudoku's, F (1, 28) = 15.22, MSE = .02, p < .05 and the logic puzzles, F (1, 28) = 13.54, MSE = .01, p < .05, indicating that the amount of explicit utterances was different on the low and high intrinsic versions of both puzzles. There was no significant main effect for extraneous load on both types of puzzles, indicating that the amount of explicit utterances on the low and high extraneous puzzles were in general the same, with F (1, 28) < 1, *ns*, for the Sudoku and F (1, 28) < 1, *ns*, for the logic puzzles. There was also no significant interaction effect between intrinsic load and extraneous load on both the Sudoku, F (1, 28) < 1, *ns*, and the logic puzzles, F (1, 28) = 1.00, *ns*, indicating that the intrinsic load did not differ according to the extraneous load.

Performance. Regarding the participants' performance, there was a significant main effect for intrinsic load on the Sudoku's, F (1, 28) = 217.25, MSE =443, 85, p < .05 and the logic puzzles, F (1, 28) = 220.71, MSE = 375.90, p = .05, indicating that performance was different on the low and high intrinsic versions of both puzzles. There was no significant main effect for extraneous load on both types of puzzles, indicating that performance on the low and high extraneous puzzles were in general the same, with F (1, 28) < 1, *ns* for the Sudoku and F (1, 28) = 1.48, *ns* for the logic puzzles. There was also no significant interaction effect between intrinsic load and extraneous load on both the

Sudoku, F (1, 28) < 1, *ns* and the logic puzzles. F (1, 28) = 2.17 *ns*, indicating that the intrinsic load did not differ according to the extraneous load.

Retention. Regarding the retention on the logic puzzles, there was no significant main effect for intrinsic load, F (1, 28) < 1, *ns*, no significant main effect for extraneous load, F (1, 28) = 2.64, *ns* and no significant interaction effect between intrinsic and extraneous load, F (1, 28) < 1, *ns*, indicating that there were no differences in retention between the conditions.

Correlations. To see which instruments correlated with each other, correlations were calculated (all *p*s < .05). On the low intrinsic Sudoku, the subjective rating scale was significantly related to the silent pauses, r = .39, to explicit utterances, r = .59 and to performance, r = -.69. There was also a significant relation between explicit utterances and performance, r = -.36. On the low intrinsic logic puzzle, the subjective rating scale was significantly related to silent pauses, r = .39 and to performance, r = -.69. Pupil dilation was significantly related to fixation duration, r = -.56 and to performance, r = .40. On the high intrinsic Sudoku, only explicit utterances were significantly related to performance, r = -.38. On the high intrinsic logic puzzle, only the subjective rating scale was significantly related to the pupil dilation, r = -.37. Overall, there were no similar relationships between all four puzzles.

Conclusion and discussion

The present study tried to investigate whether different cognitive load measurement instruments measure different types of cognitive load. Recent findings (Jarodzka et al., 2012) led to the assumption that different instruments would be able to measure different aspects of cognitive load. This study specifically focused on distinguishing extraneous load and intrinsic load. Three different instruments were used, a subjective rating scale, a thinking aloud dual task and eye tracking. Also, performance was measured. Six parameters were compared to what degree they were able to distinguish intrinsic load and extraneous load in two different conditions were participants had to solve both low and high intrinsic puzzles, with low extraneous puzzles in the first condition and high extraneous puzzles in the second condition. The six parameters were the subjective rating scale, fixation duration, pupil dilation, silent pauses, explicit utterances and performance. Main and interaction effects were calculated to see to what degree the parameters could distinguish the intrinsic load end extraneous load from the total cognitive load and whether they interacted.

Since the Sudoku´s and the logic puzzle turned out to differ in intrinsic load, they were analyzed separately. This was expected, since a difference in the type of puzzle would result in different learner preferences for a numeric or a verbal puzzle, thus, influencing the intrinsic load. Almost all six parameters were able to capture intrinsic load, with the exception of fixation duration on the logic puzzles. Almost none of the parameters were able to capture extraneous load, with the exception of fixation duration on the Sudoku puzzles. This suggests that all used cognitive load measurement instruments are useful to measure intrinsic load, but that extraneous load remains difficult to distinguish.

Regarding the main effect for extraneous load on fixation duration, the method of the study, using a split-format, might suggest that participants would have shorter fixation duration in the high extraneous condition, since they had to look somewhere else in order to access additional information. Contrary to that assumption, the mean fixation duration is higher on the high extraneous condition. This is unexpected, since an increase in fixation duration would indicate lower mental effort. It also appeared that the intrinsic load was different according to the extraneous load. This is also somewhat

surprising, since these two loads are not expected to influence each other. They are assumed to be caused by different aspects of the task (i.e. the difficulty of the task and the way the task is presented). Fixation duration was also able to distinguish the amount of extraneous load between the two types of puzzles. This leads to the suggestion that fixation duration is able to measure extraneous load individually and that the other five parameters might not be.

Although there were some differences between the Sudoku's and the logic puzzles, it seems that all measurement instruments were able to distinguish intrinsic load. Ayres (2006) stated that this would be possible for the subjective rating scale, as long as there would be no changes in extraneous load. It seems that this goes for all six parameters. However, distinguishing extraneous load from the total cognitive load remains an issue. Only fixation duration was able to distinguish extraneous load, but only on the Sudoku's and not on the logic puzzles. Apart from the subjective rating scale on the Sudoku and performance on the logic puzzle, the other parameters were not heading in the expected direction, suggesting they are not likely to be able to distinguish extraneous load.

An answer to the question whether different measurement instrument measure different types of cognitive load has to be carefully formulated. It appears that all instruments were able to measure intrinsic load, but fixation duration was the only parameter able to capture the differences between the low extraneous and high extraneous condition on only one type of puzzle. This would imply that different measurement instruments are able to measure different types of cognitive load, but not to a very large degree. It was expected that at least the subjective rating scale would be able to measure extraneous load, as long as the intrinsic load remained even (Sweller, 2010). However, this was not the case. It was expected that with the current method of the study, at least the subjective rating scale would be able to measure both intrinsic and extraneous load and that other parameters would also be able to make a distinction, perhaps to a lesser degree. Therefore, it is surprising that of all six parameters, only fixation duration was able to measure extraneous load. This leads to the question why most of the used parameters were not able to measure extraneous load and why they only seem to be able to measure intrinsic load. There seem to be two possible explanations for these unexpected results. The first explanation for the fact that only one of the parameters was able to distinguish extraneous load on one type of puzzle is that perhaps the extraneous load was not manipulated

properly in this study. The second explanation is that extraneous load might be hard to measure and to distinguish from the total cognitive load and that the used measurement instruments are not sufficient enough. This would have some implications for future research on cognitive load.

The first explanation would be that the extraneous load was not manipulated properly. The main argument for this assumption is that there were no differences in performance between the low and high extraneous conditions. The absence of performance effects might indicate that there was no perceivable difference in extraneous load. It appears that this is caused by ceiling effects. When a value is already very high, adding something has little influence on the total effect. As indicated by the subjective rating scale scores, the cognitive load was very high on the high intrinsic load puzzles, regardless of the amount of extraneous load. Looking closer to the fixation duration scores on the high extraneous Sudoku's, a large increase (from the low extraneous Sudoku's) in mental effort can be seen for the low intrinsic puzzle and a small increase can be seen for the high intrinsic puzzle. There might have been too little room for extraneous load to increase the total mental effort when the intrinsic load was high. This would explain the interaction effect between intrinsic and extraneous load, but also the fact that the other parameters were not able to measure extraneous load at all.

There often was a larger difference between the low and high extraneous condition on the low intrinsic puzzles than on the high intrinsic puzzles. The high intrinsic puzzles might have had a too high intrinsic load, which could have caused such high mental effort that the extraneous load could not noticeably increase the total cognitive load. Therefore, it would be hard for the instruments to distinguish between intrinsic load and extraneous load when the extraneous load was not able to add something substantial to the total cognitive load. It seems that this was the case for both types of puzzles, but slightly less for the Sudoku's. There was after all a significant main effect on fixation duration and although the main effect for the subjective rating scale was not significant, it is in the expected direction.

The finding that a ceiling effect makes it harder to distinguish different types of cognitive load has implications for future research on cognitive load. Not only does it seem difficult to distinguish between different types of cognitive load, but they might also be overshadowed by each other, which makes it even more difficult to take them apart. Further research should take in account the amount of

intrinsic load that is imposed by the used task. One of the weaknesses of this study was that the used puzzles had not been tested on the amount of intrinsic load they imposed. It would have been useful to measure the amount of intrinsic load before adding extraneous load, in order to check whether the intrinsic load is not too high for the extraneous load to be able to add something substantial to the total load. Therefore, further research should take in account the amount of intrinsic load that is imposed by a task. This study revealed that all six parameters can be used for this, but the strongest main effects were on the subjective rating scale and performance. A combination of these two should lead to a good indication of intrinsic load. If the intrinsic load is not too high, the extraneous load might be easier to be measured for the instruments. Although the Sudoku's appear to be good tasks to measure cognitive load, they should impose a lower intrinsic load, in order to add substantial extraneous load. However, the intrinsic load should not be too low, since extraneous load can only appear when there is high intrinsic load. This explains why the parameters were also not able to capture the extraneous load in the low intrinsic tasks, since it can only impose cognitive load when there is a substantially high enough intrinsic load. Perhaps a difficult 6x6 Sudoku or a moderate 9x9 Sudoku could be used for further research, instead of the very difficult 9x9 puzzle used in this study.

The second explanation for the findings of this study would be that extraneous load is indeed very hard to measure and that the used instruments and parameters may not be sufficient enough. The most unexpected finding is that the subjective rating scale was not able to distinguish extraneous load, as opposed to what Sweller (2010) states. This would imply that the concept of extraneous load might need to be revised, since it would not be measurable. However, further research on the first explanation needs to be conducted in order to make further statements about the concept of intrinsic and extraneous load, since the subjective rating scale has been verified to be able to measure extraneous load in various studies.

In addition to the focus on intrinsic and extraneous load, there were also correlations calculated to investigate which parameters correlated with each other. However, there are no clear relationships between the parameters, since the significant correlations varied heavily between all four given puzzles. This does imply that there really are differences in what is measured by the different instruments. Some of the instruments correlated with each other on some puzzles, indicating that

Jarodzka et al. (2012)'s suggestion about explicit and implicit aspects of mental effort might be on to something. Some correlations were found between the subjective rating scale and the thinking aloud parameters. However, in this study no clear pattern can be found in the correlations. This could be due to the before mentioned ceiling effects of the puzzles, which might have caused the parameters to be more alike in what they have measured.

Moreover, it is notable that there were some individual differences between participants in the data, especially in thinking aloud. Although all participants received the same instruction, there were differences between the ways they thought aloud. Some participants were very quiet and had to be reminded to keep thinking several times, while others kept on talking with very little pauses. Also, some participants were very explicit in their utterances, while others kept focusing on the task, even though they indicated high mental effort through the subjective rating scale. These differences might influence the data, so perhaps participants should be selected or trained more carefully on their thinking aloud skills for further research on these parameters. There were also some differences in the eye tracking data, especially with pupil dilation. It seems that some participants have larger pupils than others. This is not a problem for the within subject factor extraneous load, but for the between subject factor intrinsic load, this could cause some unaccounted for differences in the data. Perhaps relative pupil dilation data could be used, in which the increase or decrease from a neutral state is captured.

Based on the current study we argue that the used cognitive load measurement instruments are able to distinguish intrinsic load from the total cognitive load and that only a few might be able to measure extraneous load, especially fixation duration. However, further research, in which the intrinsic load of the task is not too high, should give more insight in these results and make out which instrument would be best to use. Due to the poor manipulation of extraneous load in this study, the findings were not as expected. Since performance did not differ between the low and high extraneous conditions, further research should make sure that the intrinsic load of a task is neither too high, nor too low for the extraneous load to be able to add something to the total cognitive load. However, the findings of this study do support the suggestion that different cognitive load measurement instruments measure different aspects of cognitive load, but further research needs to make out to what degree they differ and how they can be used effectively.

References

Ayres, P. (2006). Using subjective measures to detect variations of intrinsic cognitive load within problems. *Learning and Instruction, 16,* 389-400. doi: 10.1016/j.learninstruc.2006.09.001

Beckman, J.F. (2010). Taming a beast of burden – On some issues with the conceptualization and operationalisation of cognitive load. *Learning and Instruction, 20,* 250-264. doi:10.1016/j.learninstruc.2009.02.024

Brünken, R., Plass, J.L., & Leutner, D. (2003). Direct measurement of cognitive load in multimedia learning. *Educational Psychologist, 38(1),* 53-61.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of metal storage capacity. *Behavioral and Brain Sciences, 24,* 87-114.

Hart, S.G., & Staveland, L.E. (1988). Development of NASA-TLX (Task Load Index): Results of experimental and theoretical research. In P. A. Hancock & N. Meshkati (eds.), *Human mental workload* (pp. 139-183). Amsterdam: North Holland.

Holmqvist, L., Nyström, M., Andersson, R., Dewhurst, R., Jardoza, H., & Van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures.* Oxford, UK: Oxford University Press.

Huang, W. (2010). Evaluating learners' motivational and cognitive processing in an online game-based learning environment. *Computers in Human Behavior, 27,* 694-704. doi:10.1016/j.chb.2010.07.021

Jarodzka, H., Janssen, N., Kirschner, P.A., & Erkens, G. (2012). Avoiding split attention in computer-based testing: Is neglecting additional information facilitative? *In print.*

Kalyuga, S., Chandler, P., & Sweller, J. (1999). Managing split-attention and redundancy in multimedia instruction. *Applied Cognitive Psychology, 13,* 351-371.

Kalyuga, S. (2011). Cognitive load theory: How many types of cognitive load does it really need? *Educational Psychology Review, 23,* 1-19. doi: 10.1007/s10648-010-9150-7

Kirschner, F., Paas, F., & Kirschner, P.A. (2009). A cognitive load approach to collaborative learning: United brains for complex tasks. *Educational Psychology Review, 21*, 31-42. doi: 10.1016/j.chb.2008.12.008

Kirschner, P.A., Ayres, P., & Chandler, P. (2011). Contemporary cognitive load theory: the good, the bad and the ugly. *Computers in Human Behavior, 27,* 99-105. doi: 10.1016/j.chb.2010.06.025

Klinger, J., Tversky, B., & Hanrahan, P. (2011). Effects of visual and verbal presentation on cognitive load in vigilance, memory, and arithmetic tasks. *Psychophysiology, 48,* 323-332. doi: 10.1111/j.1469-8986.2010.01069.x

Paas, F. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive load approach. *Journal of Educational Psychology, 84,* 429-434.

Peterson, L. R., & Peterson, M. J. (1959). Short-term retention of individual verbal items. *Journal of Experimental Psychology, 58,* 193-198.

Sweller, J. (1988). Cognitive load during problem solving. *Cognitive Science, 12,* 257-285.

Sweller, J. (1994). Cognitive load theory, learning difficulty and instructional design. *Learning and Instruction, 4,* 295-312.

Sweller, J. (2010). Element interactivity and intrinsic, extraneous and germane cognitive load. *Educational Psychology Review, 22,* 123–138. doi: 10.1007/s10648-010-9128-5

Sweller, J., van Merrienboer, J., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10(3),* 251-296.

Sweller, J., Ayres, P., & Kalyuga, S. (2011). Measuring Cognitive Load. In: J.M. Spector & S. LaJoie (eds.), *Cognitive Load Theory* (pp. 71-85). New York: Springer.

Van Gog, T., Kester, L., Nievelstein, F., Giesbers, B., & Paas, F. (2009). Uncovering cognitive processes: Different techniques that can contribute to cognitive load research and instruction. *Computers in Human Behavior, 25,* 325-331. doi:10.1016/j.chb.2008.12.021

Van Merrienboer, J. J. G., & Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. *Educational Psychology*

Review, 17, 147–177. doi: 10.1007/s10648-005-3951-0

Van Mierlo, C.M., Jarodzka, H., Kirschner, F., & Kirschner, P.A. (2012). Cognitive load theory and E-Learning: How to optimize E-learning using cognitive load theory. *In print.*

Van Orden, K. F., Limbert, W., Makeig, S., & Jung, T. P. (2001) Eye activity correlates of workload during a visuospatial memory task. *Human Factors, 43(1),* 111-121. doi: 10.1518/001872001775992570

Yin, B., & Chen, F. (2007). Towards automatic cognitive load measurement from speech analysis. *Human-Computer Interaction: Interaction, Design, and Usability, 4550,* 1011-1020.

Table 1

Means and Standard Deviation of the Low Extraneous Puzzles

	Low extraneous	High extraneous
Sudoku low intrinsic		
Subjective Rating Scale	3.47 (2.17)	4.27 (2.28)
Fixation Duration	299.87 (38.08)	341.67 (37.96)
Pupil Dilation	3.22 (0.48)	3.21 (0.33)
Silent Pauses	0.61 (0.64)	0.83 (084)
Explicit Utterances	0.03 (0.08)	0.09 (0.01)
Performance	87.78 (27.07)	88.89 (27.03)
Logic low intrinsic		
Subjective Rating Scale	5.93 (1.79)	5.27 (1.62)
Fixation Duration	285.67 (25.24)	295.93 (31.06)
Pupil Dilation	3.09 (0.45)	3.18 (0.29)
Silent Pauses	0.59 (0.51)	0.69 (0.60)
Explicit Utterances	0.04 (0.10)	0.04 (0.12)
Performance	80.00 (32.85)	94.07 (16.19)
Retention	0.87 (0.74)	1.27 (0.59)

Table 2

Means and Standard Deviations of the High Intrinsic Puzzles

	Low extraneous	High extraneous
Sudoku high intrinsic		
Subjective Rating Scale	8.60 (0.74)	8.73 (0.46)
Fixation Duration	290.27 (30.94)	303.40 (31.25)
Pupil Dilation	3.00 (0.42)	3.04 (0.26)
Silent Pauses	0.95 (0.35)	1.04 (0.53)
Explicit Utterances	0.19 (0.17)	0.18 (0.12)
Performance	10.83 (8.39)	5.48 (3.05)
Logic high intrinsic		
Subjective Rating Scale	8.53 (0.83)	8.40 (1.18)
Fixation Duration	292.20 (27.04)	297.80 (25.81)
Pupil Dilation	3.04 (0.42)	3.04 (0.30)
Silent Pauses	0.98 (0.47)	0.92 (0.45)
Explicit Utterances	0.18 (0.14	0.13 (0.13)
Performance	13.00 (13.34)	12.33 (11.93)
Retention	1.07 (0.59)	1.20 (0.56)