Cognitive load and working memory in multimedia learning:

Conceptual and measurement issues

Øistein Anmarkrud¹, Anette Andresen¹, & Ivar Bråten²

¹Department of Special Needs Education, University of Oslo, Norway.

²Department of Education, University of Oslo, Norway.
Abstract

This article reviews contemporary research on multimedia learning that uses cognitive load theory as the major theoretical framework. In particular, we address the extent to which working memory has been conceptualized and measured in this research, what kind of subjective measures of cognitive load have been used and whether such measures are combined with other measures of cognitive load, and how results from subjective measures have been related to learning and achievement. The findings show that most of the reviewed studies did not include any clear conceptualization or measurement of working memory, used only general subjective measures containing one or very few items, and did not report findings consistent with the hypothesized relationship between cognitive load and multimedia learning. The findings are discussed in relation to the broader goal of improving research on cognitive load in the context of multimedia learning.

Keywords: Cognitive load, working memory, multimedia learning, subjective measures.
Cognitive Load and Working Memory in Multimedia Learning:
Conceptual and Measurement Issues

The 21st-century classroom is unprecedented in its use of multimedia, which involves combinations of various representations such as text (written or spoken), static graphics (e.g., pictures, illustrations, graphs, and diagrams), and dynamic graphics (e.g., animations and video; Butcher, 2014; Mayer, 2014). That is, new technologies, for example in the form of multimedia encyclopedias, simulation software, intelligent tutoring systems, or e-Courses, support the use of multiple formats that can be used to present information to students in flexible ways, with potential benefits for learning. Presenting instructional materials in the form of multimedia is often justified by referring to “the multimedia principle,” basically saying that learning with words and pictures is more effective than learning with words or pictures alone (e.g., Butcher, 2014).¹

Of note is that “multimedia” may be conceptualized more narrowly, such as simple combinations of text and graphic content within the same document, as exemplified by the work of Mayer and colleagues (e.g., Mayer, 1989; Mayer, Bove, Bryman, Mars, & Tapangco, 1996; Mayer & Gallini, 1990). However, multimedia also has come to be conceptualized more broadly, involving complex combinations of verbal (written or spoken) and pictorial (static or dynamic graphics) representations (Butcher, 2014; Mayer, 2014). Thus, in studies of multimedia learning, the complexity of multimedia may range from combinations of text and static graphics (e.g., Acarturk & Ozcelik, 2017; Chiu & Mok, 2017) to combinations of text,

¹ Importantly, the conditions of applicability of this basic multimedia principle are quite complex. Some of these conditions are discussed in a subsequent section termed Design Principles of Multimedia Learning.
audio files, and video files (e.g., Andrade, Huang, & Bohn, 2015; Chung, Cheon, & Lee, 2015) to combinations of texts, static graphics, video files, and several web pages (e.g., Cheng, Lu, & Yang, 2015). In particular, multiple representations distributed across multiple source documents may represent a complex combination in multimedia learning (List & Alexander, 2018). Multimedia learning may thus involve acquiring or constructing knowledge by processing combinations of verbal and pictorial representations at varying levels of complexity.

A substantial amount of educational research on multimedia learning uses cognitive load theory (Sweller, Ayres, & Kalyuga, 2011) to explain the potential benefits of using multimedia to enhance learning. At the core of cognitive load theory is the suggestion that multimedia learning can be best explained by examining how information is processed and stored, with particular attention being given to working memory and its role in learning (Schüler, Scheiter, & van Genuchten, 2011). Although research repeatedly has shown that multimedia may have several advantages for learning, quite a few studies fail to demonstrate positive outcomes (e.g., Dillon & Jobst, 2005; Lunts, 2002; Scheiter & Gerjets, 2007). In their extensive review, Scheiter and Gerjets (2007) identified several reasons for the ambiguity of results, arguing that the same features of digital environments that can promote learning may also be detrimental. For example, information represented in different ways has to be integrated by the learner. Previous work on the integration of verbal and visual information when reading traditional texts has shown that this integration can be challenging (Chan & Unsworth, 2011), and there are indications that meaning-making in multimedia reading environments poses even greater processing demands on the reader (Chan & Unsworth, 2011; Hartman, Morsink, & Zheng, 2010; Kamil & Chou, 2009). In brief, multimedia learning can represent unique opportunities for learning but can also challenge the learner’s working-memory capacity (Kalyuga, 2007; Scheiter & Gerjets, 2007).
The purpose of this article is to review contemporary research on multimedia learning that uses cognitive load theory as the major theoretical framework, with an emphasis on the conceptualization and measurement of working memory and the use of subjective self-report measures of cognitive load in this research. Specifically, we wanted to examine the extent to which working memory is clearly conceptualized and measured, what kind of subjective measures are used and whether such measures are combined with other measures of cognitive load, and how results from subjective measures of cognitive load are related to learning and achievement.

The reason we focus on the conceptualization and measurement of working memory is that working memory is considered a theoretical cornerstone in this area of research (Schüler et al., 2011; Sweller et al., 2011), and that careful attention to this variable presumably will facilitate the interpretation of results (e.g., by controlling for potential differences in working memory capacity between intervention and control groups; de Jong, 2010). Also, we specifically focus on the use of subjective measures of cognitive load because the psychometric properties of such measures may be open to question, at least when consisting of very few items and not combined with other types of cognitive load measures (de Jong, 2010; Naismith, Cheung, Ringsted, & Cavalcanti, 2015). In this regard, reliable and valid measures of cognitive load may be needed to better understand the cognitive mechanisms underlying effects (or lack of effects) of trying to reduce cognitive load.

In summary, our objective is to contribute to an improvement of research on cognitive load in the context of multimedia learning. That is, identifying potential issues with the conceptualization and measurement of working memory and the use of subjective measures of cognitive load in the latest multimedia research framed by cognitive load theory may provide an impetus for researchers in this area to refine their research designs and measurements by taking such lingering issues into consideration. In the following, we first discuss the concept
of working memory and, in particular, how working memory is represented within cognitive load theory. Next, we describe and discuss approaches to measuring cognitive load in multimedia learning. Based on this background analysis, we further specify the research questions guiding our review, present our results, and discuss their implications for research on cognitive load in the context of multimedia learning.

**Working Memory and Cognitive Load**

**Working Memory**

Working memory, often conceived as a “mental workspace”, can be defined as a processing resource with limited capacity involved in the storage of information while simultaneously manipulating information for brief periods of time (Alloway, 2009; Baddeley & Logie, 1999; Swanson & Alloway, 2012). There is an extensive research literature documenting a significant relationship between working memory capacity and academic achievement in domains such as reading (e.g., Cantin, Gnaedinger, Gallaway, Hesson-McInnis, & Hund, 2016; Sesma, Mahone, Levine, Eason, & Cutting, 2009; Swanson & Jerman, 2007), mathematics (e.g., Gathercole, Pickering, Knight, & Stegmann, 2004; Purpura & Ganley, 2014; Swanson, Jerman, & Zheng, 2008), and multimedia learning (e.g., Austin, 2009; Pazzaglia, Toso, & Cacciamani, 2007; Skuballa, Schwonke, & Renkl, 2012). Although several models of working memory have been proposed in the literature (for a review, see Miyake & Shah, 1999), Baddeley’s (e.g., Baddeley, 1992, 1999; Baddeley & Logie, 1999) model of working memory is widely used in research linking working memory to academic performance (Swanson & Alloway, 2012). According to this model, working memory consists of four components: the central executive, the phonological loop, the visuo-spatial sketchpad, and the episodic buffer (Baddeley, 2000).

The central executive functions as an attentional control system that regulates the three other subsystems. It has several tasks in Baddeley’s model: (a) it monitors and directs
processing in the other subsystems and links the results of this processing to long-term memory, (b) it distributes attention between tasks and assigns attention to incoming stimuli, (c) it transfers information to subsystems, (d) it updates and regulates the content of working memory, and (e) it codes representations for time and place. The phonological loop is the subsystem processing auditory information, including processing of both spoken and written words. The visuo-spatial sketchpad contains two components: the visual component addresses the characteristics of objects such as size, color, and shape, while the spatial component handles relational or spatial information and the control of movements. Finally, the episodic buffer is considered a limited-capacity storage system that temporarily stores and integrates information from the phonological loop and the visuo-spatial sketchpad with each other and with prior knowledge. Of note is that the idea of separate processing systems for verbal and visual information is not unique to Baddeley’s model. For example, Paivio’s (1971, 1986) classic dual coding theory provides a detailed explanation of how verbal and visual information is processed, integrated, stored, and recalled in the human mind.

The multimodal storage system described by Baddeley can be assumed to play a pivotal role in processing multimedia content. More specifically, three implications of Baddeley’s model of working memory can be derived with respect to learning from multimedia (Schüler et al., 2011): (a) a reduced or limited working-memory capacity should have negative effects on multimedia learning, (b) these negative effects should hinge on which of the subsystems is affected by the resource limitations, and (c) limitations in the central executive and episodic buffer should have a negative effect on the integration of verbal and pictorial information. Given that working memory is an essential concept in cognitive load theory and research on multimedia learning, it is relevant to question the extent to which influential models of working memory, such as Baddeley’s (e.g., 1999, 2000) model, are represented within cognitive load theory.
**Cognitive Load Theory**

Cognitive load theory was developed to examine the limitations of cognitive processing in relation to learning (Chandler & Sweller, 1991; Paas, Renkl, & Sweller, 2004; Sweller et al., 2011). At the core of the theory is an information-processing model of human cognitive architecture in which novel information must be processed by working memory before it can be stored in long-term memory. According to the central principle of cognitive load theory (Sweller, 2010; Sweller et al., 2011; van Merriënboer & Sweller, 2005), working memory capacity is limited, and instructional design should therefore aim to reduce unnecessary working memory load to free capacity for learning-related processing. Otherwise, exceeding working memory capacity could result in inhibited learning due to cognitive overload (Kalyuga, 2011; Leppink, Paas, van der Vleuten, van Gog, & van Merriënboer, 2013).

More specifically, cognitive load theory postulates that any learning task imposes three different categories of cognitive load on working memory: intrinsic cognitive load, extraneous cognitive load, and germane cognitive load (Paas, Renkl, & Sweller, 2003; Paas et al., 2004; Sweller et al., 2011; van Merriënboer & Sweller, 2005). Intrinsic cognitive load refers to the inherent difficulty and complexity of any learning task. Some tasks (e.g., multiplication with three-digit numbers) have a higher inherit difficulty than other tasks (e.g., the calculation of “3 - 1”), regardless of the instructional approach. Hence, complex learning tasks impose a higher cognitive load on working memory than less demanding tasks. However, intrinsic cognitive load also depends on the prior knowledge of the learner, hence it is not solely a feature of the instructional content. Extraneous cognitive load, on the other hand, does not concern the complexity of a task but how the learning material is presented or the type of instructional activities undertaken. Extraneous cognitive load is thus strongly influenced by the instructional approach, for example the completeness of explanations or the
integration of instructional materials. Finally, germane cognitive load refers to the load placed on working memory by learning, as when relating information from long-term memory to new information.

Of note is that the term “germane cognitive load” has been questioned lately. Thus, in their recent update of cognitive load theory, Sweller and colleagues (2011) preferred the term “germane resources” because unlike intrinsic and extraneous cognitive load, germane load is not imposed by the nature or structure of the learning material. Germane cognitive load instead arises from the process in which the learner relates relevant information from long-term memory to information from the present learning task (Sweller, 2010; Sweller et al., 2011). Hence, it pertains to the working memory resources allocated to dealing with intrinsic cognitive load (Kalyuga, 2011; Sweller, 2010).

**Design Principles in Multimedia Learning**

Empirical studies have suggested that working memory load in learning with multimedia could be reduced by instructions that follow the principles of cognitive load theory (Sweller et al., 2011). Although people can be assumed to learn more from words and pictures than from words alone, not all multimedia presentations are considered equally effective (Mayer, 2005, 2009). According to Mayer (2005, 2009), instructional material should build on the knowledge of how the mind works and how people process information. Thus, evidence-based theory should guide the construction of multimedia instructional messages.

Important principles are that meaningful learning from multimedia is more likely to occur when corresponding words and pictures are presented near each other instead of far from each other on the page or screen (the spatial contiguity principle) and that the learning material should include cues to highlight how it is organized (the signaling principle; Mayer, 2005, 2009; Mayer & Fiorella, 2014). For example, splitting learners’ attention between two
or more sources that have to be integrated is likely to have a negative effect on learning due to higher cognitive load compared to an environment where the information sources are spatially integrated (Schnotz & Kurschner, 2007). In addition, the redundancy principle posits that including additional information sources superfluous or unnecessary for understanding is likely to be detrimental to performance (Mayer, Heiser, & Lonn, 2001). Of note is that this effect may be moderated by learners’ prior knowledge and expertise. For example, Kalyuga, Chandler, and Sweller (1998) found that high-expertise learners using a diagram did not need text-based information to be physically integrated with the diagram; in fact, they also experienced reduced cognitive load when the text was eliminated. Thus, within multimedia learning, the redundancy principle states that learning material should not combine text and pictures for learners for whom one of the representations is sufficient to build a mental representation because the redundant information requires mental effort without enhancing learning (Mayer, 2005). That experts may suffer from what novices find beneficial in this regard is called the expertise reversal effect. Of note is that this effect is less likely to appear at the extreme ends of the complexity spectrum, that is, when learning tasks represent very low or very high intrinsic cognitive load (Sweller et al., 2011).

**Working Memory within Cognitive Load Theory**

Although cognitive load theory often refers to Baddeley’s (1992, 1999; Baddeley & Logie, 1999) model of working memory as an important foundation, it has also been argued that the link between cognitive load theory and models of working memory is “… loose and appears at least partly inconsistent” (Schüler et al., 2011, p. 392). In particular, this inconsistency concerns the central executive that is the core of Baddeley’s model. Cognitive load theory has considered the existence of two modes of processing (i.e., auditory and visual) in working memory since the mid-1990s (e.g., Mousavi. Low, & Sweller, 1995; Tindall-Ford, Chandler, & Sweller, 1997). Reflecting this distinction, in the latest extensive presentation of
cognitive load theory, Sweller et al. (2011) described working memory as consisting of partly independent processors “… that correspond to the modality of information” (p. 44). However, Sweller et al. (2011) argued that knowledge stored in long-term memory acts as a de facto central executive. If the notion of a knowledge-independent central executive were right, what would then determine its actions, asked Sweller et al. (2011). In their view, the answer to this question would be a hierarchy of central executives where the second controls the first, the third controls the second, and so forth. Hence, the idea of a knowledge-independent central-executive is not considered logically viable. As an alternative, Sweller et al. (2011) suggested that only the combination of knowledge stored in long term memory and a “… random generate and test procedure is required for a functioning natural information system” (p. 35). It is therefore reasonable to conclude that although Sweller et al.’s (2011) perspective on working memory is in line with Baddeley’s model as far as modality specific processing components are concerned, there are substantial differences regarding how these subcomponents are regulated and governed.

**Measurement of Cognitive Load**

In the early days of cognitive load theory, cognitive load was predominantly measured using indirect measures such as time on task or error rates (Sweller et al., 2011). Although these approaches have been replaced by more direct measures to a large degree, they are still occasionally used to assess cognitive load in multimedia contexts (e.g., Mahdjoubi & A-Rahman, 2012). Three main categories of direct measures of cognitive load have been established in the literature: dual-task methodology, physiological measures, and subjective ratings (Leppink et al., 2013; Naismith & Cavalcanti, 2015; Paas, Tuovinen, Tabbers, & van Gerven, 2003; Sweller et al., 2011).

**Dual-Task Methodology**
Dual-task methodology includes a secondary task in combination with a primary task. For example, to examine the load imposed by a primary task (e.g., understanding UV radiation by studying a website containing multiple representations), participants are asked to press the space key on the computer keyboard every time a specific sound is played (secondary task). If the primary task imposes a high cognitive load, performance on the secondary task (e.g., response time) is expected to deteriorate if the primary and secondary tasks tap the same working memory resources (Brünken, Plass, & Leutner, 2003, 2004; DeLeeuw & Mayer, 2008). In the majority of research on cognitive load using dual-task methodology, the secondary task is dissimilar and requires less working memory than does the primary task (Sweller et al., 2011). Examples include pressing a pedal when a particular tone sounds (Brünken et al., 2004; Marcus, Cooper, & Sweller, 1996), pressing a key on a computer keyboard when a letter changes color (Brünken, Steinbacher, Plass, & Leutner, 2002), or noticing when a button is illuminated (van Gerven, Paas, van Merriënboer, & Schmidt, 2006). Sweller (1988) described and examined another, more complex, secondary task. He proposed that using problem solving as a strategy for learning involves two separable tasks: solving the problem and learning from problem solving. Sweller (1988) hypothesized that if learners treat problem solving as the primary task, learning becomes a secondary task, and poorer learning could result from high-demand problem solving.

**Physiological Measures**

The development of more sophisticated technologies has led to increased interest in physiological measures of cognitive load, and evidence has started to emerge that techniques such as cognitive papillary response, functional magnetic resonance imaging (fMRI), and electroencephalography (EEG) have considerable merit (Paas, Ayres, & Pachman, 2008; Sweller et al., 2011). Within the context of multimedia learning, there is also an increased use of eye-tracking methodology to examine changes in cognitive processing, and it has been
argued that an increased length of eye fixation and greater pupil size are related to increased cognitive load (Chuang & Liu, 2012; Underwood, Jebbet, & Roberts, 2004; van Gog & Jarodzka, 2013). Yet, eye-tracking data are not self-explanatory; there can be several reasons why a participant was looking somewhere for a certain amount of time or in a certain order (e.g., task instructions). Eye-tracking data should consequently be used in combination with other measures of cognitive load (e.g., Kok & Jarodzka, 2017; van Gog & Jarodzka, 2013).

**Subjective Measures**

The use of subjective measures is by far the most common approach to measuring cognitive load in the literature. In their recent systematic review of the validity of cognitive-load measures in simulation-based training, Naismith et al. (2015) found that 71% (34 of 48) of the studies reviewed used subjective measures of cognitive load, corroborating findings from previous reviews of the literature (e.g., Paas, Tuovinen, et al., 2003). The assumption underlying the use of subjective measures is that people are able to evaluate their own cognitive processes and provide a valid estimate of the cognitive load imposed by a task. This evaluation typically involves asking participants to rate their cognitive load on one or more items with Likert-type scales, with the Paas scale (1992) of mental effort being the most used and cited subjective measure (Naismith & Cavalcanti, 2015). The original Paas (1992) scale consists of one single item where participants rate their mental effort on a nine-point scale ranging from very, very low mental effort to very, very high mental effort. Although Paas and colleagues reported good reliability and sensitivity for this measure when it was introduced (Paas, 1992; Paas & van Merriënboer, 1994; Paas, van Merriënboer, & Adam, 1994), more recent studies of cognitive load using subjective measures have been criticized for not reporting any reliability or sensitivity data (Naismith & Cavalcanti, 2015; Naismith et al., 2015; Paas, Tuovinen, et al., 2003). Moving beyond such reliance on a single item measure, recent attempts to measure different types of cognitive load (i.e., intrinsic, extraneous, and
germane load) with multiple-item scales have showed promising results (e.g., Leppink et al., 2013; Sewell, Boscardin, Young, Cate, & O’Sullivan, 2016).

Another issue related to the use of subjective measures is how cognitive load has been conceptualized and operationalized in empirical research. Both perceived difficulty and expended effort have been widely used as proxies for cognitive load in various subjective measures without any clarification of the difference between the two constructs. For example, de Jong (2010) found in a review that perceived difficulty and mental effort were used interchangeably in several studies, with perceived difficulty being the term used in measures of cognitive load, and mental effort being the term used when presenting and discussing the results. In addition, de Jong (2010) found instances in which both terms were used separately in questionnaires but then combined into one construct in the analyses without further clarification. The relationship between the two constructs has been empirically examined to a minor degree. However, there are some studies indicating that perceived difficulty and mental effort are different constructs with distinct consequences for the measurement of cognitive load (Ayres & Youssef, 2008; Schmeck, Opfermann, van Gog, Paas, & Leutner, 2015). For example, although measures of perceived difficulty and effort are often correlated, the learner may be unable or unwilling to put in substantial effort if a task is too demanding (Sweller et al., 2011). Cognitive load has also been operationalized through terms such as pressure during task (Cheng et al., 2015), mental load (Chen & Wu, 2015), frustration (Huang et al., 2015) and mental demand (Kizilcec et al., 2015). In other words, there is a lack of consistency in how cognitive load has been operationalized in subjective measures, which makes it difficult to interpret empirical findings.

Issues of timing also are evident in that subjective measures of cognitive load have been administered at different points of time during problem solving (Schmeck et al., 2015; van Gog, Kirschner, Kester, & Paas, 2012). Some studies (e.g., Dindar, Yurdakul, & Dönmez,
have applied one item examining the perceived difficulty or expended effort repeatedly after every task in a learning situation, with this approach often referred to as immediate ratings of cognitive load. In other studies (e.g., Inan et al., 2015; Kalyuga, Chandler, Touvinen, & Sweller, 2001; Leutner, Leopold, & Sumfleth, 2009), one or more items have been administered after the end of the full learning session in an approach referred to as delayed ratings of cognitive load. Studies that have compared immediate and delayed ratings of cognitive load have consistently found that delayed ratings yield ratings higher than the average of repeated immediate ratings made during problem solving (Schmeck et al., 2015; van Gog et al., 2012).

A single delayed item at the end of a learning session is easy to administer, but it can be challenging to determine what such a rating really measures. Some participants might try to estimate an average of their cognitive load across tasks, while others judge their cognitive load based on the last task, and still others base their judgment of cognitive load on the most demanding or complex task they worked on. Although all of these possibilities (and different combinations of them) are plausible, there is little empirical evidence available. However, Schmeck et al. (2015) found that delayed ratings seemed to be more influenced by the most challenging tasks in a learning situation, rather than by when these challenging tasks are administered (e.g., at the beginning or the end of a learning session).

**The Potential Intrusiveness of Subjective Measures**

To be able to give accurate delayed judgments of cognitive load after a series of tasks, learners would have to allocate working memory capacity to monitoring cognitive load for the various tasks while trying to solve them, store information from this monitoring in long-term memory, and retrieve this information from long-term memory back into immediate memory to answer the question(s) about cognitive load. On the other hand, the use of repeated
measures of cognitive load after every task requires participants to consider the cognitive load of the task they just finished, drawing on information that is probably still in working memory (Schmeck et al., 2015; van Gog et al., 2012). A question that remains open, however, is the potential intrusiveness of repeated measures (Schmeck et al., 2015). If participants are told they will be asked questions about perceived difficulty or expended effort after every task, the measurement of cognitive load could be considered a secondary task. This means that explicitly encouraging participants to allocate working memory capacity to monitor their cognitive load during problem solving could lead to at least two outcomes when working on complex tasks: (a) little available working memory for the accurate judgment of cognitive load, resulting in inaccurate judgments of task difficulty or expended effort; or (b) poorer learning because the combined load from the task and the monitoring of cognitive load exceeds available working memory capacity.

Moreover, it can be assumed that the risk of intrusiveness of repeated measures of cognitive load would be particularly pronounced for participants with low working memory capacity because the working memory resources of these participants would be more easily depleted by the combined requirements of task completion and cognitive load monitoring (cf., Chen, Castro-Alonso, & Paas, 2018). As a consequence, it seems essential that studies using subjective measures of cognitive load measure and control for participants’ working-memory capacity, as well as describe the procedures and instructions used for the subjective measures of cognitive load very carefully to make it clear whether there is any extraneous load imposed by the measure of cognitive load itself. Of note is that the potential intrusiveness of the cognitive load measures discussed above may be less dependent on the number of items than on the secondary task introduced by such measures. Thus, although only one single question about mental effort may be asked, cognitive load measurement could intrude upon judgment
of cognitive load and/or learning because it may require that learners allocate working memory resources to monitor their cognitive load during learning task completion.

The Present Review

On this backdrop, we set out to review recent research using subjective measures of cognitive load when examining learning in a multimedia context. For several reasons, we restricted our review to studies published between 2014 and 2017. First, although subjective measures of cognitive load are widely used, there has been a discussion in the literature about whether such measures are sufficiently sensitive to discriminate between the different types of cognitive load and whether the subjective perceptions of task difficulty and mental effort are related to the actual cognitive load (e.g., Ayres, 2006; Brünken et al., 2003; van Gog & Paas, 2008). We therefore sought to determine to what extent the most recent published research on multimedia learning using subjective measures of cognitive load had responded to this critique by using scales with greater specificity and/or more items and by evaluating how cognitive load measured by means of subjective rating was related to learning.

Second, although the construct of working memory is arguably a cornerstone of cognitive load theory, the correspondence between influential models of working memory, such as Baddeley’s (2000, 2001) model, and cognitive load theory is not straightforward (Schüler et al., 2011). In particular, Sweller et al.’s (2011) extensive and updated presentation of cognitive load theory presented a detailed and thorough explanation of the concept of working memory from a cognitive-load perspective. Therefore, we wanted to investigate whether the centrality of working memory within cognitive load theory was reflected in the most recent research on multimedia learning that used subjective measures of cognitive load.

Finally, in the last decade, cognitive load theory has been the subject of strong criticism regarding methodological and measurement issues (de Jong, 2010; Gerjets, Scheiter, & Cierniak, 2009) and conceptual clarity (de Jong, 2010; Schnitz & Kürschner, 2007). In
particular, de Jong (2010), based on a thorough review of the most cited cognitive load research, highlighted the need to include multi-item subjective measures of cognitive load, preferably in combination with other types of measures of cognitive load. Moreover, de Jong (2010) highlighted the need to include measures of working memory as a control variable because “it is not enough to know the demand (which could mean the effort applied); the system’s capacity must also be known in order to determine (over)load” (p. 123). Hence, we wanted to examine whether such critique and recommendations had been answered with more rigorous and comprehensive measurements as well as more conceptual clarity in the most recent research on cognitive load in a multimedia context.

Specifically, our review was guided by the following three questions.

1. How is working memory conceptualized in contemporary cognitive-load research on multimedia learning, and to what degree is working memory assessed in this research?

2. How are subjective measures of cognitive load used in this research, and are subjective measures combined with other measures of cognitive load?

3. How are the results from subjective measures of cognitive load related to learning and achievement?

Method

Search Strategy

We worked with a research librarian at our institution to develop a search protocol. The search was conducted in October 2017 and included four databases that represented both educational and psychological domains: ERIC, Social Sciences Citation Indexes, PsycINFO, and Web of Science. Search terms for cognitive load included cognitive load, mental load, workload, and mental effort. These search terms were combined with “multimedia” and/or “multimedia learning”. We also performed manual searches of highly relevant journals for articles about cognitive load and multimedia learning. These journals included Learning and

Study Selection and Data Extraction

We used a multistage approach to select relevant studies for the review. Whereas no particular study design or outcome measure was required, studies had to meet the following inclusion and exclusion criteria:

1. Citing of cognitive load theory: The studies had to explicitly cite and build upon cognitive load theory as a major theoretical framework.

2. Measure of cognitive load: A subjective measure of cognitive load would have to be used in the study, either as the only measure of cognitive load or in combination with other types of measures (e.g., secondary task or physiological measures of cognitive load).

3. The study would have to address learning in a formal educational context by means of multimedia (e.g., by using tablets, simulation software, or educational computer games).

4. Population: Participants should have no identified cognitive impairment or learning disability.

Next, we developed a coding scheme to extract the following information from the selected studies:

1. Study aim, including main research question(s).

2. Method, including description of participants (age, educational background, and sample size), study design, multimedia tasks, measure(s) of cognitive load, procedures for measuring cognitive load, and measure of working memory (if measured).
3. Conceptualization of working memory in literature review, including definition of working memory and references for this definition.

4. Findings regarding the relationship between cognitive load and learning, categorized as negative (i.e., high cognitive load was associated with low learning), positive, or neutral.

5. Findings regarding the relationship between cognitive load and working-memory capacity, categorized as negative (i.e., high working memory capacity was associated with low cognitive load), positive, or neutral.

Results

[Figure 1 about here]

Using the search terms described above, the initial search in the four databases yielded 381 results (see Figure 1). After the removal of 65 duplicates, the abstracts of the remaining 316 articles were examined by the first and second authors in light of the four inclusion and exclusion criteria, which resulted in 63 articles being removed from the review because they were theoretical papers or examined cognitive load in populations with learning disabilities. Based on a preliminary full-text examination of the remaining 253 articles, also conducted by the first and second authors in collaboration, 189 more articles were removed because they did not meet the inclusion and exclusion criteria. Specifically, 103 articles did not use subjective measures of cognitive load, 64 articles did not use cognitive load theory as an explicit theoretical framework, and 22 articles did not examine multimedia learning in a formal educational context. Manual searches of the journals mentioned above did not result in additional studies being identified. The final set of 64 articles, including 73 studies, were coded by the first author using the previously described coding system. Table 1 provides detailed descriptions of the 73 studies included in the review, whereas Table 2 summarizes the results across studies by year of publication. These studies represented all educational levels,
with 44 studies including college or university students, 19 studies including high-school students, and 10 studies including elementary or middle school students.

[Tables 1 and 2 about here]

**Working Memory**

As can be seen in Table 2, only 20 (31.25%) of the articles included any conceptualization or clear definition of working memory by referring to classic sources in the working memory literature, such as Baddeley and colleagues (e.g., Baddeley, 1986; Baddeley, 2001; Baddeley & Hitch, 1974), or to Paivio’s (e.g., 1971, 1986) work. Interestingly, there were no references to a cognitive-load perspective on working memory, such as that of Sweller et al. (2011). In the remaining 44 articles, working memory was often more or less explicitly described as limited in capacity through descriptions such as “… cognitive load exceeds working memory” (Cheng et al., 2015, p. 129), “… limited working memory resources” (Hsu, Gao, Liu, & Sweller, 2015, p. 114), or “… humans’ limited working memory” (Huang, Chen, Wu, & Chen, 2015, p. 159), without further clarification, explanation, or reference to theories of working memory. Only 10 (13.70%) of the 73 studies included a measure of working memory in the research design. In most instances, this working memory measure was used to control for systematic differences between experimental and control groups, and not to examine the relationship between working memory and cognitive load.

**Measurement of Cognitive Load**

Based on the inclusion criteria, all studies used some form of subjective self-report to measure cognitive load. The majority of studies (58, 79.45%) used delayed ratings; that is, participants were given one or more items targeting cognitive load after a learning or problem solving phase. The remaining 15 studies used immediate ratings of cognitive load; that is, participants were given one or more items during a learning or problem solving phase.
There was also some variation concerning the number of items used to measure cognitive load. As many as 31 studies (42.47%) used one single item. However, in 16 of these studies, this item was used repeatedly during the learning phase. Further, 21 (28.77%) of the studies used two items, 13 studies (17.81%) used three or four items, and seven studies (9.59%) used five or more items when measuring cognitive load. A total of 35 studies (47.95%) used the original Paas (1992) one-item scale, either alone or in combination with other measures.

Although the measures used in these studies were similar in that they consisted of one or more items that participants rated on a scale, there was large variation in how the concept of cognitive load was operationalized in the measures. The majority of the studies included one or more items asking participants to rate their “mental effort” (53, 72.60%) and/or “perceived difficulty of the task” (36, 49.32%). Of note is that some studies (e.g., Hsu et al., 2015) referred to the original Paas (1992) scale, in which “mental effort” is used in the item, but still used “cognitive load” in their own measures. As summarized in Table 2, participants were to a lesser degree asked to rate their “cognitive load”, “mental load”, “pressure during problem solving”, “frustration”, “mental demand”, “perceived complexity”, “clarity of learning material”, “perceived learning”, and “expended attention”.

Despite explicit critique in recent years towards the use of subjective measures of cognitive load without reporting any reliability or sensitivity data from own sample, only 15 of the 73 studies we reviewed (i.e., 20.55%) provided such information. In many instances, authors did refer to the results of reliability analyses conducted by the developers of particular measures, without discussing the validity of those analyses when measures were translated and used in other cultural contexts. Concerning the instructions given to participants before they rated their cognitive load, the information provided in the 73 studies was not clear. That is, in the majority of the studies, it was not possible to determine from the description of
procedures and instruments whether participants were told that they would receive questions about their cognitive load during or after the learning or problem solving phase. Only four of the studies (5.48%) provided information that made it possible to infer that participants actually knew that they should pay attention to their cognitive load during learning or problem solving. For example, although it was not explicitly stated, it was possible to infer that the participants in Dindar et al.’s (2015) study actually received such information, as the participants practiced “… how to respond to the Likert scale for reporting their CLs…” (p. 155).

Only 10 of the studies (13.70%) combined subjective rating of cognitive load with another method of measurement. The results concerning the relationship between the subjective measure of cognitive load and the additional measure were ambiguous. Thus, five of the studies (e.g., Dindar et al., 2015; Korbach et al., 2017) did not find any statistically significant relationship between scores on their subjective measure and an additional measure. For example, Korbach et al. (2017) combined their subjective measure of cognitive load with a secondary task where participants tapped a predefined rhythm with their feet. The results showed that the accuracy of the tapping was related to learning, but this was not the case for the scores on the subjective measure of cognitive load. Further, there was no statistically significant relationship between scores on the subjective measure and the secondary task. In one of the studies (Lee, 2014), the results were less clear. In that study, a subjective measure was combined with electroencephalography (EEG), and a statistically significant correlation was found between perceived difficulty of task (subjective measure) and beta waves in the T3 area, while no such relationship was found between mental effort (subjective measure) and beta waves. In the four remaining studies, statistically significant relations between subjective and additional measures were found. For example, Marefat et al. (2016) combined their subjective measure of cognitive load with a secondary task where participants pressed the
space key on a computer keyboard whenever a large letter at the bottom of the computer screen changed color, finding that both measures of cognitive load predicted learning.

**Relationship between Cognitive Load and Other Measured Variables**

As described previously, a central idea in research of multimedia learning from a cognitive-load perspective is that a good instructional design should reduce cognitive load, and learning should consequently improve because more working memory capacity is available. Among the 73 studies, 32 (43.84%) presented results consistent with this expectation. As an example, Yung and Paas (2015) examined the effect of adding visual representations to mathematical problems in a computer-based multimedia instructional context developed to teach students arithmetic operations. A control group used the same multimedia materials, but in the control condition, there were no visual representations. The authors hypothesized that supporting students’ emerging arithmetic understanding with visual representations would lead to reduced cognitive load and better processing of the learning materials with consequent higher learning performance. The results were consistent with their hypothesis: participants in the condition with visual representations rated their cognitive load as lower during the learning phase and obtained better results on the posttest compared to participants in the control condition.

In other studies, however, the intervention in question resulted in better learning but there was no statistically significant effect on cognitive load. As an example, Inan et al. (2015) investigated the effects of modality on learning from multimedia instruction. In that study, participants used an interactive computer-presented diagram to learn about places of articulation in human speech. Two versions of the computer diagram were developed: one version in which the diagram (cross-section of the human speech system) included text descriptions explaining each point in the interactive diagram and one version using audio descriptions. The results showed that condition had a statistically significant effect on learning
(participants in the text condition outperformed those who studied with audio descriptions), but there was no statistically significant effect on cognitive load.

In yet other studies, the experimental condition did not result in better learning but did have an effect on cognitive load. Of note, in Dindar et al. (2015), there was no effect of condition on the subjective ratings of cognitive load, but there was a statistically significant effect of condition on cognitive load measured using a secondary task. Additionally, there was no statistically significant correlation between the two measures of cognitive load.

Discussion

In this review, we examined a sample of recent studies of multimedia learning framed by cognitive load theory that measured cognitive load by means of subjective ratings. In the following, we lay out and discuss the key findings and relate them to the broader goal of improving research on cognitive load in the context of multimedia learning.

Working Memory in Recent Cognitive Load Research

The vast majority of the studies in this review did not refer to any conceptualization of working memory or provide any explicit definition of this construct. In one sense, this is somewhat surprising because working memory and cognitive load theory are by nature highly related (de Jong, 2010). Indeed, the idea of using research about human cognitive architecture to design instruction that facilitates learning despite the limited capacity of our working memory was described as “… the ultimate aim of the theory” (p. vii) in Sweller et al.’s (2011) extensive updated version of cognitive load theory. Moreover, Sweller et al. (2011) provided a well-grounded and thorough conceptualization of working memory, and both the common ground and the differences between a cognitive-load perspective on working memory and Baddeley’s (1992, 1999, 2000) model of working memory were discussed. The differences between a cognitive-load perspective and Baddeley’s model of working memory were also reflected in Schüler et al.’s (2011) review of research on working memory in multimedia
instruction, in which those authors concluded that the two theories at least appear to be inconsistent. It is noteworthy that the majority of studies in this review that did provide an explicit and clear definition of working memory based this definition on Baddeley’s model and not on Sweller et al.’s (2011) conceptualization of working memory.

It is also noteworthy that the procedures for measuring cognitive load were described superficially in the reviewed studies. Without an explicit and detailed description of the instructions participants received prior to problem solving or the instructional phase, it is impossible to determine whether the measurement imposed a cognitive load. That is, if participants prior to problem solving or learning are told to pay attention to task difficulty, mental effort, or cognitive load, the continuous attention directed towards this task could arguably impose an extraneous load during problem solving.

In studies using immediate ratings of cognitive load, the extraneous load imposed by the measure could potentially be even more profound, as the learner would receive a prompt to allocate attention, and thus working-memory capacity, towards monitoring cognitive load every time a cognitive-load item is administered during learning. The measurement of cognitive load could thus be conceived as a secondary task, and the accuracy of the ratings could become lower when tasks become more challenging, particularly among learners with low working memory capacity. On the other hand, if participants do not receive any instruction about the measure of cognitive load that will follow, the cognitive load of the task will be less well monitored, and participants would thus have to base their rating of cognitive load on a retrospective estimate of invested mental effort or task difficulty, a task they did not prepare for. In both cases, participants’ working memory capacity must be considered.

When de Jong (2010) reviewed the 35 most frequently cited studies with “cognitive load” as a descriptor, he found that working memory was rarely measured. Our review of the most recent research in the area corroborates this finding. Based on his review, de Jong (2010)
strongly recommend that working memory should be included as a control variable to ensure that differences between intervention and control groups cannot be attributed to differences in working memory, which is particularly important in quasi experimental work.

In addition, measures of working memory capacity are needed to assess the extent to which the potential intrusiveness of subjective cognitive load measures may vary with working memory capacity. In this regard, a stronger relationship between ratings of cognitive load and performance among learners with higher working-memory capacity compared to participants with poorer working-memory capacity might suggest that subjective ratings of cognitive load impose an extraneous load. Yet another reason to measure working memory capacity in future research on multimedia learning framed by cognitive load theory is that such measures may be used to validate subject measures of cognitive load by examining to what extent lower ratings of cognitive load are associated with higher working memory capacity and vice versa.

Finally, operationalization and measurement of working memory should be based on a clear conceptualization of this construct. For example, a conceptualization of working memory that does not consider a central executive component but rather brings prior domain knowledge to the forefront (Sweller et al., 2011) may make it less pertinent to address executive functioning skills per se (cf., Follmer, 2018) and more pertinent to include the variable of prior domain knowledge. Attention to prior domain knowledge in the research designs is, of course, essential to investigate the expertise reversal effect and the implications of this effect for adapting instructional methods to learners at different levels of domain knowledge (Sweller et al., 2011).

**Subjective Measures of Cognitive Load**

Despite the critique of very general measures of cognitive load including few items (e.g., de Jong, 2010; Naismith et al., 1015), our review showed that variations of the Paas (1992) scale were most frequently used to measure cognitive load. One issue concerning the
wide use and adaption of this scale into new contexts and cultures is the lack of psychometric validation. In their systematic review of the use of cognitive-load measures in simulation-based medical training, Naismith et al. (2015) found that the extensive use of adapted versions of the Paas (1992) scale (e.g., changing the number of points on the Likert-scale, or modifying the item to ask about perceived difficulty instead of invested mental effort) requires the reestablishment of the psychometric properties of the instrument, which was seldom performed in the research they reviewed.

While Naismith et al. (2015) examined measurements of cognitive load within simulations in medical training, de Jong (2010) found similar results in his review of studies from a broader set of educational contexts: There were several adaptions of the Paas (1992) scale, including changes in the number of points on the Likert scale, differences in the anchoring terms of the Likert scale, or differences in the wording of the items (e.g., amount of effort, how difficult, or how easy) or the timing of the measure (when it was administered during a task), without a full discussion of how such changes may affect the reliability and validity of the measure. Our results clearly show that the lack of data on the psychometric properties of subjective measures of cognitive load is still a reason for concern in contemporary research on multimedia learning.

It is also interesting that some of the reviewed studies discussed their use of general subjective measures as a limitation. For example, Hsu et al. (2015), who used the Paas (1992) scale, discussed whether the combination of subjective measures of cognitive load, challenging learning tasks, and the complex simulation technology often used in instructional design research makes it difficult for learners to estimate their cognitive load. This is because the complex simulation technology imposes some cognitive load, and it may be difficult for learners to discern how much cognitive load comes from the learning task versus the technology. Hsu et al. (2015) therefore recommended that the specificity of questionnaires is
increased and that more use is made of immediate ratings of cognitive load. Similarly, Park (2015) argued that the use of subjective single-item measures, such as the Paas (1992) scale, makes it impossible to differentiate between different types of cognitive load.

Inan et al. (2015) discussed additional issues with subjective measures of cognitive load, for example whether actual cognitive load is reflected in an individual’s subjective perception of mental effort. Accordingly, Inan et al. (2015) concluded that there is a need to combine subjective measures with other ways to assess cognitive load, such as a secondary task. With this recommendation as a backdrop, it is problematic that among the studies we reviewed, only 10 studies used multiple measurements of cognitive load, typically combining subjective ratings with a secondary task. While we caution against drawing conclusions from a small number of studies, it is also noteworthy that half of those 10 studies did not obtain any statistically significant correlation between their two measures of cognitive load. However, although not studying multimedia learning, Szulewski, Gegenfurtner, Howes, Sivilotti, and van Merriënboer (2017) found that a subjective measure of cognitive load (i.e., the Paas scale) correlated substantially with a physiological measure of cognitive load (i.e., pupil size).

Our review also showed that the majority of the studies operationalized cognitive load through items asking participants to rate their mental effort or perceived difficulty in relation to a task or learning phase. In approximately half of the studies, this was the only way in which cognitive load was operationalized. Mental effort, perceived difficulty, and cognitive load are clearly related, but they are not identical (Ayres & Youssef, 2008; de Jong, 2010; Schmeck et al., 2015). That is, whereas cognitive load is experienced by the learner, mental effort is exerted by the learner (de Jong, 2010). Therefore, it seems problematic that these terms are used as synonyms in the majority of the studies we reviewed. In Paas’ (1992) highly cited article, for example, cognitive load was described as a construct containing two dimensions: mental effort and mental load. In relation to this conceptualization, studies
operationalizing cognitive load with mental effort as their sole indicator could be said to be based on a too narrow conceptualization, whereas studies operationalizing cognitive load through a combination of indicators such as mental effort, perceived task difficulty, frustration during task, and pressure during task could be said to be based on a too broad conceptualization, calling the construct validity of such measurements into question.

**Relationships between Cognitive Load and Learning**

Our review showed that fewer than half of the studies (32 out of 73) reported findings consistent with the hypothesized relationship between cognitive load and multimedia learning, which implies that good instructional designs lead to reduced levels of cognitive load, as well as improved multimedia learning due to reduced processing demands on working memory. In other studies, however, the instructional designs that were tested either showed improvement in learning with no reduction in cognitive load or reduced cognitive load with no effect on learning. There are several possible reasons for the lack of consistency between the theoretical assumptions and the empirical findings in the majority of the reviewed studies. For example, it is possible that in some instances the observed effects on learning were caused by something other than a reduction in cognitive load, and that in some instances a reduction in cognitive load was not sufficient to bring about improvement in learning; both possibilities are problematic from a theoretical point of view.

Another possibility is that the subjective measures that were used lacked the sensitivity required to detect relevant changes in experienced cognitive load. For example, there seem to be reliability and validity issues with the widely used Paas (1992) scale. As highlighted by de Jong (2010), the evidence for claims about the psychometric properties of this measure is scarce, and it has not been thoroughly compared with other measures of cognitive load in order to establish construct validity. Of note is also that only four of the studies we reviewed presented precise and detailed information about the administration of the subjective self-
report measure used, making it difficult to preclude the possibility that the administration and
the procedures of the cognitive load measurements might have influenced the results. Finally,
it is conceivable that the lack of consistency between hypotheses and findings in the majority
of the reviewed studies in some instances might be due to the complexity and difficulty of the
tasks used to measure learning and achievement. In any case, without reliable, valid, and
sensitive measurements of cognitive load, the interpretation of findings becomes difficult in
terms of the cognitive mechanisms underlying effects (or lack of effects) on multimedia
learning. On the other hand, a more precise understanding of these cognitive mechanisms may
not only allow for further theoretical validation of the core assumptions of cognitive load
theory but also lead to improved instructional designs targeting these mechanisms.

**Conclusion**

Cognitive load theory has been important in research on the potential benefits of using
multimedia for learning. Research conducted within this perspective has driven instructional
design in technology-rich educational contexts forward. Cognitive load theory has also been
important in building connections between cognitive science and instructional practice. At the
same time, cognitive load theory has been the subject of critique, in particular regarding
measurement. The current review of research on multimedia learning framed by cognitive
load theory demonstrates that there are lingering issues with the attention to working memory
and the use of subjective measures of cognitive load in the most recent research in this area.

First, the vast majority of the studies we reviewed did not include a measure of
working memory, nor did they provide a clear conceptualization of this core construct within
cognitive load theory. This is problematic because working memory can be considered an
important control variable whose inclusion also allows for the validation of subjective
cognitive load measures. Second, the use of subjective measures of cognitive load was still
categorized by general measures including very few items, lack of psychometric
information, lack of combination with other types of cognitive load measures, and inconsistent operationalization, as well as lack of clear descriptions of the procedures used for measuring cognitive load. This is problematic because reliable, valid, and sensitive measures of cognitive load are needed to understand the cognitive mechanisms underlying the effects (or lack of effects) of instructional designs.

That said, measuring cognitive load in a reliable, valid, and sensitive way is a challenging and complex task, and some researchers within cognitive load theory have invested great effort in developing more objective measures of cognitive load and comparing them to subjective measures (Korbach, Brünken, & Park, 2018; Korbach et al., 2017; Park & Brünken, 2015). Regarding subjective measures, in particular, multi-item measures targeting different types of cognitive load have recently been developed and validated (Leppink et al., 2013; Sewell et al., 2016). These include a measure attempting to capture the sources of different types of cognitive load in complex learning tasks involving multiple documents (Cerdan, Candel, & Leppink, 2018). However, these promising avenues for cognitive load measurement were represented in the body of research that we reviewed only to a limited extent.

The implications of our review seem straightforward. Future empirical work in this area needs to include measures of working memory, not only as a control variable but also to validate subjective measures of cognitive load. For example, without attention to working memory, it is not possible to determine to what extent the potential intrusiveness of subjective measures may vary with learners’ working memory capacity, which seems like an interesting and important issue. Moreover, the development and testing of reliable, valid, and sensitive measures of cognitive load need to be prioritized in this area of research, for example by further developing and testing multi-item subjective measures targeting different types of cognitive load and comparing such measures with other types of measures (e.g., physiological
measures). By including reliable, valid, and sensitive measures of cognitive load in future research, more precise understanding of the cognitive mechanisms underlying the effects (or lack of effects) of instructional designs may be achieved. In turn, such understanding may have theoretical implications because it allows for further validation of assumptions within cognitive load theory, and instructional implications because it allows for more precision in designing instructional interventions.
References

References marked with an asterisk (*) are included in the review


**Chang, C.-C., Liang, C., Chou, P.-N., & Lin, G.-Y. (2017).** Is game-based learning better in flow experience and various types of cognitive load than non-game based learning? Perspective from multimedia and media richness. *Computers in Human Behavior, 71*, 218-227


*Educational Psychology Review, 23*, 1-19.

*Human Factors, 40*, 1-17.


based on the interaction of multiple levels of knowledge representation. *Journal of Chemical Education, 91*, 1409-1416.


Table 1

Description of all studies included in the review

<table>
<thead>
<tr>
<th>#</th>
<th>Article</th>
<th>Subjective measure of cognitive load</th>
<th>Rating</th>
<th>Wording of item(s)</th>
<th>No. of items</th>
<th>Other measure of CL</th>
<th>Measure of WM</th>
<th>Def. of WM</th>
<th>Representations</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Andrade, Huang, &amp; Bohn (2015)</td>
<td>Leppink et al. (2013)</td>
<td>I</td>
<td>PC, CLM, PL</td>
<td>10 (R)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>VW, VS, SG, DG</td>
</tr>
<tr>
<td>7</td>
<td>Chen &amp; Wu (2015)</td>
<td>Sweller et al. (1998)</td>
<td>D</td>
<td>ME, ML</td>
<td>4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>VW, VS, SG, DG</td>
</tr>
<tr>
<td>8</td>
<td>Cheng, Huang, Shadiev, Hsu &amp; Chu (2014)</td>
<td>Sweller et al. (1998)/NASA TLX</td>
<td>D</td>
<td>Not reported</td>
<td>Not reported</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>VS, DG</td>
</tr>
<tr>
<td>10</td>
<td>Chiu &amp; Mok (2017)</td>
<td>Paas (1992)</td>
<td>D</td>
<td>ME</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>VW, SG</td>
</tr>
<tr>
<td>12</td>
<td>Costley &amp; Lange (2017)</td>
<td>Adapted version of Leppink et al. (2013)</td>
<td>D</td>
<td>PL</td>
<td>4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>VW, VS, SG, DG</td>
</tr>
<tr>
<td></td>
<td>Authors</td>
<td>Paas (1992)</td>
<td>Exper.</td>
<td>Method</td>
<td>Condition</td>
<td>Comparison</td>
<td>Design</td>
<td>Annotation</td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>----------------------------------</td>
<td>-------------</td>
<td>--------</td>
<td>--------</td>
<td>-----------</td>
<td>------------</td>
<td>--------</td>
<td>------------</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>Experiment 1</em></td>
<td>Paas (1992)/Cierniak et al. (2009)</td>
<td>D</td>
<td>ME/DT</td>
<td>1/1=2</td>
<td>-</td>
<td>-</td>
<td>VW, SG</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>Experiment 2</em></td>
<td>Paas (1992)/Cierniak et al. (2009)</td>
<td>D</td>
<td>ME/DT</td>
<td>1/1=2</td>
<td>-</td>
<td>-</td>
<td>VW, SG</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>Experiment 3</em></td>
<td>Paas (1992)/Cierniak et al. (2009)</td>
<td>D</td>
<td>ME/DT</td>
<td>1/1=2</td>
<td>-</td>
<td>X</td>
<td>VW, SG</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>Experiment 4</em></td>
<td>Paas (1992)/Cierniak et al. (2009)</td>
<td>D</td>
<td>ME/DT</td>
<td>1/1=2</td>
<td>-</td>
<td>-</td>
<td>VW, SG</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Fenesi &amp; Kim (2014)*</td>
<td>No reference</td>
<td>D</td>
<td>DT</td>
<td>1 (R)</td>
<td>-</td>
<td>-</td>
<td>X</td>
<td>VW, VS, SG</td>
</tr>
<tr>
<td></td>
<td><em>Experiment 1</em></td>
<td>Paas (1992)</td>
<td>I</td>
<td>ME</td>
<td>1 (R)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>VW, VS, SG</td>
</tr>
<tr>
<td></td>
<td><em>Experiment 2</em></td>
<td>Paas (1992)</td>
<td>I</td>
<td>ME</td>
<td>1 (R)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>VW, VS, SG</td>
</tr>
<tr>
<td>21</td>
<td>Huang, Chen, &amp; Ho (2014)</td>
<td>Adapted from Cerpa et al. (1996) and Kalyuga (2000)</td>
<td>D</td>
<td>Not reported</td>
<td>5</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>VW, VS, SG, DG</td>
</tr>
<tr>
<td>23</td>
<td>Huang &amp; Mayer (2016)</td>
<td>Paas (1994)</td>
<td>I</td>
<td>ME</td>
<td>1(R)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>VW, VS, DG</td>
</tr>
<tr>
<td></td>
<td>Study</td>
<td>Methodology</td>
<td>Measure</td>
<td>Target</td>
<td>Mean</td>
<td>95% CI</td>
<td>Result</td>
<td>VW</td>
<td>VS</td>
</tr>
<tr>
<td>---</td>
<td>-----------------------------------------</td>
<td>-------------</td>
<td>---------</td>
<td>--------</td>
<td>------</td>
<td>--------</td>
<td>--------</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>26</td>
<td>Izmirili &amp; Kurt (2016)</td>
<td>Paas (1992)</td>
<td>I</td>
<td>ME</td>
<td>1 (R)</td>
<td>-</td>
<td>X</td>
<td>VW</td>
<td>VS</td>
</tr>
<tr>
<td>27</td>
<td>Jan, Chen, &amp; Huang (2016)</td>
<td>Marcus et al. (1996)</td>
<td>D</td>
<td>ME, ML</td>
<td>4</td>
<td>-</td>
<td>-</td>
<td>VW</td>
<td>SG</td>
</tr>
<tr>
<td>29</td>
<td>Johnson, Reisslein, &amp; Reisslein (2014)</td>
<td>No reference</td>
<td>D</td>
<td>ME, DT</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>VW</td>
<td>VS</td>
</tr>
<tr>
<td>33</td>
<td>Kizilcec, Bailenson, &amp; Gomez (2015)</td>
<td>NASA TLX</td>
<td>D</td>
<td>ME, MD</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>VW</td>
<td>VS</td>
</tr>
<tr>
<td>34</td>
<td>Korbach, Brünken, &amp; Park (2016)*</td>
<td>Paas (1992)</td>
<td>D</td>
<td>ME, DT</td>
<td>2 (R)</td>
<td>X</td>
<td>X</td>
<td>VW</td>
<td>SG</td>
</tr>
<tr>
<td>35</td>
<td>Korbach, Brünken, &amp; Park (2017)*</td>
<td>Paas (1992)</td>
<td>D</td>
<td>ME, DT</td>
<td>2 (R)</td>
<td>X</td>
<td>X</td>
<td>VW</td>
<td>SG</td>
</tr>
<tr>
<td>37</td>
<td>Leahy &amp; Sweller (2016)</td>
<td>No reference</td>
<td>D</td>
<td>DT</td>
<td>1</td>
<td>-</td>
<td>X</td>
<td>VW</td>
<td>SG</td>
</tr>
<tr>
<td>38</td>
<td>Lee (2014)</td>
<td>Bralfish et al. (1972)</td>
<td>D</td>
<td>ME, DT</td>
<td>2</td>
<td>X</td>
<td>-</td>
<td>VS</td>
<td>DG</td>
</tr>
<tr>
<td>Paper Reference</td>
<td>Experiment</td>
<td>Design</td>
<td>ME, DT</td>
<td>X</td>
<td>VS, DG</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------------</td>
<td>------------</td>
<td>--------</td>
<td>--------</td>
<td>---</td>
<td>-------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lee &amp; Mayer (2015)</td>
<td>Experiment 1</td>
<td>D</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>VS, DG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brünken et al. (2010)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brünken et al. (2010)</td>
<td>Experiment 2</td>
<td>D</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>VS, DG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brünken et al. (2010)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lehmann, Goussios, &amp; Seufert (2016)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Klepsch &amp; Seufert (2012)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Liew, Zin, &amp; Sahri (2017)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kalyuga et al. (2000)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lin &amp; Yu (2016)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chang, Lei &amp; Tseng (2011)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lin, Lee, Wang, &amp; Lin (2016)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adapted version of Paas (1992)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marefat, Rezaee, &amp; Naserieh (2016)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paas (1992)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mattis (2015)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paas (1992)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mihalca, Mengelkamp, Schnottz, &amp; Paas (2015)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ayres (2006)/Paas (1992)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Milenko, Segedin, &amp; Hrin (2014)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No reference</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nebel et al. (2017)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adaption from Eysink et al. (2009)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Park (2015)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paas (1992)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Park &amp; Brünken (2015)*</td>
<td>Experiment 1</td>
<td>D</td>
<td>1 (R)</td>
<td>-</td>
<td>X</td>
<td>VS, DG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paas (1992)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Park &amp; Brünken (2015)*</td>
<td>Experiment 2</td>
<td>D</td>
<td>1 (R)</td>
<td>X</td>
<td>X</td>
<td>VS, DG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paas (1992)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Park, Korbach, &amp; Brünken (2015)*</td>
<td></td>
<td>D</td>
<td>CL</td>
<td>1 (R)</td>
<td>-</td>
<td>X</td>
<td>VS, DG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paas (1992)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Park, Knörzer, Plass, &amp; Brünken (2015)</td>
<td></td>
<td>D</td>
<td>DT/ME</td>
<td>1/1=2</td>
<td>-</td>
<td>X</td>
<td>VS, DG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kalyuga et al. (2000)/Paas (1992)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Park, Münzer, Seufert, &amp; Brünken (2016)*</td>
<td></td>
<td>D</td>
<td>ME</td>
<td>1 (R)</td>
<td>-</td>
<td>-</td>
<td>VS, DG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No.</td>
<td>Authors &amp; Year</td>
<td>Reference</td>
<td>Method &amp; Description</td>
<td>DT</td>
<td>ME</td>
<td>CL</td>
<td>PL</td>
<td>EA</td>
<td>PC</td>
</tr>
<tr>
<td>-----</td>
<td>----------------</td>
<td>-----------</td>
<td>----------------------</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>54</td>
<td>Radulović, Stojanović, &amp; Županec (2016)</td>
<td>No reference</td>
<td>I DT 1 (R) - - -</td>
<td>DG, NS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>55</td>
<td>Renkel, Skuballa, Schwonke, Harr, &amp; Leber (2015)</td>
<td>No reference</td>
<td>I ME, DT 2 (R) - - -</td>
<td>VW, SG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>56</td>
<td>Roelle, Lehmkühl, Beyer, &amp; Berthold (2015)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Experiment 1</td>
<td>Swaak &amp; de Jong (2001)</td>
<td>D DT 2 - - X</td>
<td>VW, SG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Experiment 2</td>
<td>Eysink et al. (2009)</td>
<td>D ME, CL 7 - - -</td>
<td>VW, VS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>57</td>
<td>Schneider, Nebel, Pradel, &amp; Rey (2015)</td>
<td>Paas et al. (1994)</td>
<td>D ME, CL 4 X - -</td>
<td>VW, VS, SD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>59</td>
<td>Sithole, Chandler, &amp; Abeysekera (2017)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Experiment 1*</td>
<td>Paas (1992)</td>
<td>D ME 1 (R) - - -</td>
<td>VW, VS, SG, DG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Experiment 2*</td>
<td>Paas (1992)</td>
<td>D ME 1 (R) - - -</td>
<td>VW, VS, SG, DG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>62</td>
<td>Wang, Sundararajan, Adesope, &amp; Ardasheva (2017)</td>
<td>No reference</td>
<td>D DT, MD 2 - - -</td>
<td>VW, VS, SG, DG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Note.** D = Delayed rating, I = Immediate rating, ME = Mental effort, CL = Cognitive load, ML = Mental load, DT = Difficulty of task, P = Pressure during task, F = Frustration, MD = Mental demand, PC = Perceived complexity, CLM = Clarity of learning material, PL = Perceived learning, EA = Expended attention, R = Repeated measurement, VW = Verbal written, VS = Verbal spoken, SG = Static graphic, DG = Dynamic graphic, NS = Not specified

*The reason these studies are categorized as involving delayed as well as repeated rating is that participants were asked to rate their cognitive load once in the middle of the learning phase and once after the last task.*
### Table 2

**Summarization of the results**

<table>
<thead>
<tr>
<th></th>
<th>2014</th>
<th>2015</th>
<th>2016</th>
<th>2017</th>
<th>N (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of studies(^1)</td>
<td>17</td>
<td>29</td>
<td>13</td>
<td>14</td>
<td>73 (100)</td>
</tr>
<tr>
<td>Number of items in subjective measure(^2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 item</td>
<td>5</td>
<td>13</td>
<td>6</td>
<td>7</td>
<td>31 (42.47)</td>
</tr>
<tr>
<td>2 items</td>
<td>9</td>
<td>8</td>
<td>3</td>
<td>1</td>
<td>21 (28.77)</td>
</tr>
<tr>
<td>3-4 items</td>
<td>1</td>
<td>5</td>
<td>2</td>
<td>5</td>
<td>13 (17.81)</td>
</tr>
<tr>
<td>5 (\rightarrow) items</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>7 (9.59)</td>
</tr>
<tr>
<td>Wording of items</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mental effort</td>
<td>12</td>
<td>24</td>
<td>8</td>
<td>9</td>
<td>53 (72.60)</td>
</tr>
<tr>
<td>Cognitive load</td>
<td>-</td>
<td>4</td>
<td>-</td>
<td>-</td>
<td>4 (5.48)</td>
</tr>
<tr>
<td>Mental load</td>
<td>-</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>4 (5.48)</td>
</tr>
<tr>
<td>Difficulty of task</td>
<td>11</td>
<td>11</td>
<td>7</td>
<td>7</td>
<td>36 (49.32)</td>
</tr>
<tr>
<td>Pressure during task</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>1 (1.37)</td>
</tr>
<tr>
<td>Frustration</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>1 (1.37)</td>
</tr>
<tr>
<td>Mental demand</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>3 (4.13)</td>
</tr>
<tr>
<td>Perceived learning</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>2</td>
<td>3 (4.13)</td>
</tr>
<tr>
<td>Clarity of learning material</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>1</td>
<td>2 (2.74)</td>
</tr>
<tr>
<td>Expended attention</td>
<td>-</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>2 (2.94)</td>
</tr>
<tr>
<td>Perceived complexity</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>1</td>
<td>2 (2.74)</td>
</tr>
<tr>
<td>Delayed measurement of CL</td>
<td>13</td>
<td>23</td>
<td>9</td>
<td>13</td>
<td>58 (79.45)</td>
</tr>
<tr>
<td>Immediate measurement of CL</td>
<td>4</td>
<td>6</td>
<td>4</td>
<td>1</td>
<td>15 (20.55)</td>
</tr>
<tr>
<td>Information about the instruction participants received about the measurement of CL prior to learning phase</td>
<td>-</td>
<td>2</td>
<td>2</td>
<td>-</td>
<td>4 (5.48)</td>
</tr>
<tr>
<td>Additional measure of CL</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>10 (13.70)</td>
</tr>
<tr>
<td>Provides data from reliability analysis of subjective measure of CL in own sample</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>15 (20.55)</td>
</tr>
<tr>
<td>Explicit definition of working memory</td>
<td>6</td>
<td>9</td>
<td>2</td>
<td>3</td>
<td>20 (31.25)</td>
</tr>
<tr>
<td>Measure of working memory</td>
<td>2</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>10 (13.70)</td>
</tr>
<tr>
<td>Expected relationship between CL and learning</td>
<td>9</td>
<td>12</td>
<td>3</td>
<td>8</td>
<td>32 (43.84)</td>
</tr>
<tr>
<td>Participants</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>University</td>
<td>11</td>
<td>18</td>
<td>7</td>
<td>8</td>
<td>44 (60.27)</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>6</td>
<td>6</td>
<td>4</td>
<td>19 (26.03)</td>
</tr>
<tr>
<td>------------------</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>------------</td>
</tr>
<tr>
<td>High school</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Elementary or middle school</td>
<td>3</td>
<td>5</td>
<td>-</td>
<td>2</td>
<td>10 (13.70)</td>
</tr>
</tbody>
</table>

Note: ¹The number of studies are higher than the number of articles included in this review since some articles include more than one study. ²One study does not provide information about number of items. ³The percentage is based on number of articles, not number of studies.
Figure caption

Figure 1. *Overview of the coding process*
Total papers identified: 381

Removal of duplicates: 65 duplicates removed

Review of 316 abstracts: 63 papers removed
- 56 theoretical or review papers
- 5 papers examining populations with learning disabilities
- 2 papers written in other languages than English or German

Preliminary coding of 253 full texts: 189 papers removed
- 103 did not use a subjective measure of cognitive load
- 64 did not use cognitive load theory as an explicit theoretical framework
- 22 did not examine multimedia learning in a formal educational context

Final coding: 64 papers