



A Systematic Meta-analysis of the Reliability and Validity of Subjective Cognitive Load Questionnaires in Experimental Multimedia Learning Research

Felix Krieglstein¹ · Maik Beege² · Günter Daniel Rey¹ · Paul Ginns³ · Moritz Krell⁴ · Sascha Schneider⁵

Accepted: 5 May 2022 / Published online: 20 May 2022
© The Author(s) 2022

Abstract

For more than three decades, cognitive load theory has been addressing learning from a cognitive perspective. Based on this instructional theory, design recommendations and principles have been derived to manage the load on working memory while learning. The increasing attention paid to cognitive load theory in educational science quickly culminated in the need to measure its types of cognitive load — intrinsic, extraneous, and germane cognitive load which additively contribute to the overall load. In this meta-analysis, four frequently used cognitive load questionnaires were examined concerning their reliability (internal consistency) and validity (construct validity and criterion validity). Results revealed that the internal consistency of the subjective cognitive load questionnaires can be considered satisfactory across all four questionnaires. Moreover, moderator analyses showed that reliability estimates of the cognitive load questionnaires did not differ between educational settings, domains of the instructional materials, presentation modes, or number of scale points. Correlations among the cognitive load types partially contradict theory-based assumptions, whereas correlations with learning-related variables support assumptions derived from cognitive load theory. In particular, results seem to support the three-factor model consisting of intrinsic cognitive load, extraneous cognitive load, and germane cognitive load. Results are discussed in relation to current trends in cognitive load theory and recommendations for the future use of cognitive load questionnaires in experimental research are suggested.

Keywords Cognitive load theory · Multimedia learning · Measurement issues · Scales and questionnaires · Meta-analysis

Felix Krieglstein and Maik Beege shared first authorship.

✉ Felix Krieglstein
felix.krieglstein@phil.tu-chemnitz.de

Extended author information available on the last page of the article

Introduction

In psychological research, subjective measurements are often used to assess constructs that are not directly observable. Such scales include multiple items each of which aims to provide information about the construct under study (McNeish, 2018). From a psychometric viewpoint, measuring can be defined “as assigning of numbers to observations in order to quantify phenomena” (Kimberlin & Winterstein, 2008, p. 2276). Within educational psychology, cognitive load theory, postulating that learning is associated with a cognitive burden imposed on the learner’s working memory (e.g., Mayer & Moreno, 2010; Sweller, 2020), is seen as one of the most influential frameworks. In recent years, there has been ongoing debate surrounding how to measure types of cognitive load reliably and validly in experimental settings (e.g., Brünken et al., 2003, 2010; Naismith et al., 2015; Schmeck et al., 2015). This meta-analysis aimed to examine the quality of the four most frequently used cognitive load questionnaires measuring types of cognitive load. This is done by examining the two quality criteria of reliability and validity — methodological requirements that need to be met before they can be classed as high-quality measuring instruments (Kimberlin & Winterstein, 2008). Furthermore, the goal of this work is to quantitatively verify theoretical assumptions (i.e., the types of cognitive load and their interrelationships; Kalyuga, 2011; Sweller, 2010) of the cognitive load theory (CLT).

The Construct of Cognitive Load

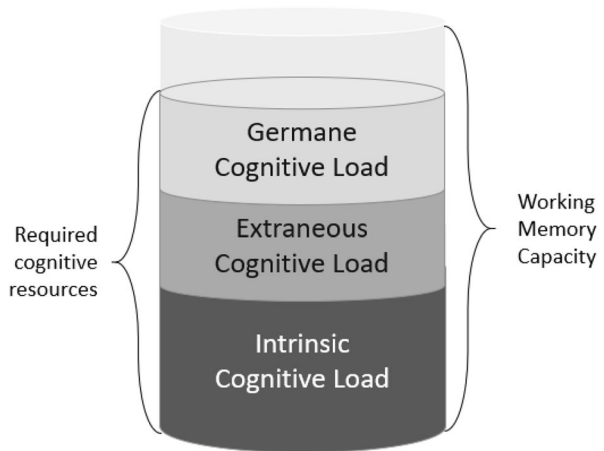
Working Memory and Cognitive Load

CLT, introduced in the 1980s by John Sweller (1988), is an established theoretical framework in educational psychology research, which applies our knowledge of human cognitive architecture and evolutionary educational psychology to instructional design (Sweller et al., 1998, 2019; Sweller, 2020, 2021). A central assumption of this cognitive theory is that learning arises from the interplay of working memory and long-term memory processes (Cowan, 2008; Sweller, 2016). Based on Cowan’s (1999) *embedded-processes model of working memory*, both cognitive systems are not to be considered separately. Working memory is argued to be the activated part of the long-term memory indicating that a focus of attention is paid to learning processes (for an overview, see Schweppe & Rummer, 2014). Relatively uncontroversial is the assumption that the working memory system (or short-term memory system) is limited in its capacity implicating that only a limited number of elements can be processed simultaneously (Baddeley, 1986; Cowan, 2001; Miller, 1956). Furthermore, empirical findings suggest that novel information that is unknown to the learner is lost after a certain time if not repeated (Jonides et al., 2005; Peterson & Peterson, 1959). These constraints on capacity and duration hamper information processing because only a certain

amount of information can be processed in the working memory simultaneously. In contrast, long-term memory stores retrievable information organized into schemata (Plass & Kalyuga, 2019; Schweppe & Rummer, 2014). Such schemata can help to overcome working memory limitations by chunking a certain amount of information into one element (Paas et al., 2003). To build such a schema, new information has to be selected, organized, and integrated into a coherent model (i.e., the SOI model; Mayer, 1996). Finally, schemata are stored in long-term memory and can be retrieved, if necessary, into the working memory in order to facilitate learning with a complex task. Assuming that each element would have to be processed individually, this would exceed the capacity of the working memory. The automation of interacting elements leads to the fact that these can be processed unconsciously in the future and thus reduces the load on the working memory (Paas et al., 2003).

Cognitive load can be viewed as a multidimensional construct involving both mental load and mental effort (Paas et al., 2003). Both constructs play an important role in our understanding of cognitive load, though it is still unclear how or even whether these are related. For instance, Krell (2017) developed a questionnaire that explicitly distinguishes between mental effort and mental load. Generally, it is assumed that mental effort is assignable to the active investment of germane resources when learning (Klepsch & Seufert, 2021). Therefore, to transfer knowledge to long-term memory and to integrate it with previously gained prior knowledge, learners must themselves become active by directing their cognitive resources toward learning-relevant activities (Klepsch & Seufert, 2020, 2021; Krell, 2017). This process can be encouraged by the design of the learning material. In contrast, learning materials' inherent characteristics like the complexity and the presentation format are experienced passively by the learner through what Krell (2017) described as task-related load or mental load. Following this line of reasoning, Sweller et al. (2011) argue that mental load and mental effort should be seen as two distinct constructs, which usually correlate positively. As previously mentioned, the cognitive load imposed on working memory is caused by the learning task. To complete a task, that is, to learn, requires an amount of invested mental effort. According to Paas (1992), mental effort is characterized by the usage and allocation of cognitive resources, indicating that the amount of mental effort is a reliable estimate of someone's motivation to acquire new information. Instructional designs and procedures should support the learner to efficiently use the available working memory resources for schemata acquisition while optimizing the information processing ability (Chen & Kalyuga, 2020; Korbach et al., 2018). For this purpose, CLT proposes design principles for instructional materials and procedures that aim at reducing unnecessary load on working memory and freeing up capacity for learning-related processing (Anmarkrud et al., 2019). As pointed out by Sweller et al. (2019), cognitive load is increased when unnecessary demands that tend to impede effective learning need to be processed. For instance, distracting elements within the learning environment can be a source of unnecessary cognitive load. However, working memory's efficacy can be enhanced when the learner has a certain level of domain-specific prior knowledge. Consequently, the learner is cognitively less burdened by the task (Feldon, 2007). The overriding goal formulated within the CLT is therefore to avoid

Fig. 1 Schematic illustration of the cognitive load theory



cognitive overload which manifests itself in “that the processing demands evoked by the learning task may exceed the processing capacity of the cognitive system” (Mayer & Moreno, 2003, p. 45).

Types of Cognitive Load

Perhaps the best-known version of CLT stipulates three additive types of cognitive load (cf. Sweller et al., 1998). In contrast, the original version of CLT did not distinguish between the different types of cognitive load (e.g., Sweller, 1988). The CLT was consequently further developed by undertaking a subdivision into the intrinsic cognitive load (ICL) and extraneous cognitive load (ECL) in order to better explain the phenomenon that some learning materials are more difficult to learn than others (cf. Sweller & Chandler, 1994). The three-factor model including intrinsic load, extraneous load, and germane load was then developed by Sweller et al. (1998) in the mid- to late 1990s (see Fig. 1). The germane cognitive load (GCL) type was added to the model in order to meet some findings in which germane load was increased for schema construction processes. Recent approaches have suggested that intrinsic and germane loads share the same theoretical foundation and can be classified as one type of load (e.g., Kalyuga, 2011; Sweller, 2010). Both versions, however, distinguish between intrinsic and extraneous cognitive load (Sweller, 2010; Sweller et al., 2019). In this vein, a factor analysis by Jiang and Kalyuga (2020) revealed that the intrinsic–extraneous model is suitable for assessing cognitive load. In contrast, a recent confirmatory analysis (Zavgorodniaia et al., 2020) has found strong support for the three-component model of intrinsic (ICL), extraneous (ECL), and GCL.¹ Furthermore, the most commonly used cognitive load questionnaires refer to the three-factor model and accordingly measure the three types of load separately. Therefore, for the aim of this work, the three-factor model was used.

¹ In the following, the abbreviations ICL, ECL, and GCL are used.

Intrinsic Cognitive Load

Intrinsic cognitive load (ICL) describes the learning tasks inherent complexity (Klepsch et al., 2017). Accordingly, this load is determined by the learning material's level of element interactivity and the learner's domain-specific prior knowledge (Leppink et al., 2013). Hereby, the element interactivity can be described as the number of elements that have to be processed in the working memory at the same time (Chen & Kalyuga, 2020). It is assumed that a low prior knowledge linked with high element interactivity results in a high ICL. In contrast, learners with high prior knowledge have already formed schemata, which can be used as prior knowledge to help them solve problems or learning tasks without this leading to an excessive load on their working memory. In conclusion, ICL can only be changed by modifying the complexity that has to be learned or by enhancing the learner's domain-specific prior knowledge (Sweller et al., 2019).

Extraneous Cognitive Load

Extraneous cognitive load (ECL) is determined by how the learning materials are presented and organized (Sweller et al., 2019). When information is more difficult to process (e.g., through task-irrelevant details; Sundararajan & Adesope, 2020), not enough cognitive resources might be available for processing information relevant for learning as working memory capacity is exceeded. In this case, the learner is forced to compensate for the unfavorable presentation by additional cognitive effort (e.g., search processes to overcome split-attention effects; Schroeder & Cenkci, 2018). Consequently, a CLT-based recommendation is to keep the ECL as low as possible in order to enable successful learning. However, the additive character of the cognitive load types suggests that ECL in particular becomes important when the learning material generates a large amount of intrinsic load (Paas et al., 2003). In contrast, if ICL is low, the learner will have enough cognitive resources to also handle higher levels of ECL.

Germane Cognitive Load

However, over the years, the concept of germane cognitive load (GCL) has been undergoing revision (Kalyuga, 2011; Sweller, 2010). Because learning aims to build up schemata in long-term memory, the GCL refers to the working memory's resources needed to handle the intrinsic cognitive load imposed (Sweller et al., 2019). Accordingly, the learner should carry out activities like self-explanation or note-taking, which in turn contribute to learning. In contrast to the other two loads, the GCL represents a productive load (Moreno & Park, 2010). Following these assumptions, a high GCL is an indicator for engaged learners directing their cognitive resources to the learning process (Klepsch et al., 2017). In this vein, Kalyuga (2011) argues that the germane load is indistinguishable from the intrinsic load as both categories share the same theoretical background. Thus, the GCL does not represent a load in itself but rather has a distributive function, so that available working memory resources are free to handle the complexity of the learning material. The

proposed intrinsic-extraneous cognitive load model removes the GCL as an autonomous type of cognitive load indicating that learning-relevant activities can be attributed to the ICL (Kalyuga, 2011; Sweller, 2010). With this assumption in mind, this work takes up the academic discourse regarding the number of cognitive load types in multimedia learning research.

Cognitive Load in Multimedia Learning

Empirical findings have shown that learning with multimedia can be enhanced when instructions follow the principles of CLT (Mayer & Moreno, 2003, 2010; Sweller et al., 2011). Hereby, multimedia learning is generally defined as learning from both pictures and words (Mayer, 2014). Learning is hence more effective when people actively construct coherent models from verbal and pictorial representations (Fletcher & Tobias, 2005). However, learning is not automatically encouraged just because instructions may include words and pictures. Thus, not all multimedia learning settings are considered equally conducive to learning. As pointed out by Mayer (2014), instructional designers should build on assumptions of human cognitive architecture and thus also consider CLT when providing multimedia learning environments and materials for learners.

On the one hand, several design recommendations have been made regarding how best to support learners to transfer new information into long-term memory and integrate it with pre-existing knowledge (Sweller et al., 2019). These recommendations primarily focus on reducing extraneous processing while encouraging learners to manage any ICL induced by the difficulty of the learning material (Mayer & Moreno, 2010). To assist learners in managing the inherent complexity of certain tasks, the principles of segmenting (Rey et al., 2019) and pre-training (Mayer et al., 2002) have been formulated within the CLT framework. Presenting the learning content in learner-paced segments or providing learners with relevant information for the upcoming learning material are assumed to make it easier for the learner to learn, even when the learning materials induce a high degree of complexity. In addition, the isolated elements effect describes the learning-beneficial effect when information with a high element interactivity are learned in an isolated form in the first step (Pollock et al., 2002). Once the elements are stored in long-term memory, interactions between them can be learned in order to create a coherent model (Sweller et al., 2019).

On the other hand, several design principles have also addressed how to reduce ECL as it does not support and can even impede learning (Sweller, 2010). Because extraneous processing may occur due to unnecessary search processes, the split-attention effect describes the learning-hindering effect when corresponding information sources need to be cognitively integrated by the learner (Ayres & Sweller, 2014). To counter such additional cognitive processes, CLT recommends following the principles of spatial and temporal contiguity (Ginns, 2006). Accordingly, related pieces of information should be presented as close to each other as possible with respect to spatial and temporal proximity. A more integrated format thus facilitates integration of learning-relevant information. Another way to save working memory

resources is to avoid redundancies by, for instance, presenting the same information both aurally and visually. Similarly, the redundancy effect (Kalyuga & Sweller, 2014) refers to the negative impact on learning of multiple ways presenting the same information.

Connections Among Cognitive Load Types

The CLT is an instructional framework finding wide application in multimedia learning research (Brünken et al., 2003; Paas & Sweller, 2014). In general, the different cognitive load types (ICL, ECL, and GCL) are presumed to add to the overall load (additivity hypothesis; Moreno & Park, 2010). In this vein, the different types of cognitive load form in sum the total cognitive load, whereby this assumption only applies if the capacity of working memory is not exceeded (Paas et al., 2003). When the cognitive load is approaching the limit of the working memory's capacity, the cognitive load types and their relationships can dynamically change. Following the theoretically based expectation that one load decreases when the other increases makes it easier to understand why learning materials are easy or difficult to learn. For example, when cognitive resources are depleted and ECL increases, fewer resources are available for germane processes and, thus, GCL decreases. Consequently, inconsistent connections between cognitive load facets can be assumed depending on the learning task including its complexity and presentation. Furthermore, the assumptions formulated very clearly in the CLT may differ from the subjective perceptions of the learners. It tends to be questionable whether learners are able to differentiate between the different types of cognitive load and whether questionnaires can differentiate between the types of cognitive load because of item construction and formulation. Consequently, methodological issues may cause the types of cognitive load to be related differently than formulated in the additivity hypothesis.

For instance, it can be assumed that the ICL and ECL should not be correlated since both loads are associated with different aspects of the learning materials (Sweller et al., 2019). Learners should therefore be able to differentiate between the tasks' inherent complexity (ICL) and the presentation of the learning material (ECL). Nevertheless, it can be argued that both sources of cognitive load cannot be assessed in a differentiated manner by learners. In this vein, it seems plausible that a complex learning content (e.g., biochemical processes) cannot be represented in a simple way and, thus, increases the ECL because of the complex presentation.

Based on the assumption that the ICL and GCL share a common theoretical background (Kalyuga, 2011), both variables should show an interdependency (measurable as correlation). The GCL is therefore not a load in itself but rather allocates available working memory resources to activities relevant to learning what is dealing with the intrinsic load (Sweller, 2010). However, this assumption is also questionable in light of the active load vs. passive load perspective (Klepsch & Seufert, 2021). This argues that the ICL results from the complexity of learning materials and is experienced passively by the learner, while the GCL relates to the allocation

of cognitive resources and is, therefore, of an active nature. The distinction between passive and active load could result in both variables not correlating with each other.

Predicting relationships between the ECL and GCL is difficult at first glance. As the GCL refers to the allocation of cognitive resources to learning-relevant activities (Bannert, 2002), its active character is evident. Thus, learners are responsible for investing cognitive resources in germane processes actively (i.e., active load). In contrast, learners experience ECL as a result of how learning materials are presented in a passive way (i.e., passive load; Klepsch & Seufert, 2021). This distinction should result in the two loads not correlating with each other. In contrast, a learning material not optimally designed (causing higher ratings of the ECL) could be related to a lower GCL because learners are less motivated to learn and hence make less of an effort. In this vein, the cognitive load caused by the learning material can be categorized as a motivational cost (e.g., Feldon et al., 2019). To sum up, the additivity hypothesis of the CLT can probably hardly be found in reality since methodical, as well as theoretical restrictions, have to be considered.

Connections of Cognitive Load Types with Theory-Related Concepts

It is common to conduct cognitive load research in connection with learning tests to understand to what extent the intervention has contributed to successful learning. In line with Mayer (2001), knowledge gained in multimedia learning can be divided into two categories — retention and transfer. Retention is defined as remembering when the information explicitly mentioned in the learning material is asked for. In contrast, transfer is related to the application of acquired knowledge, for example, in new contexts (Mayer, 2001). Learning in both categories is typically assessed in experimental studies through multiple-choice or open question formats, among others. Accordingly, the retention-transfer differentiation is adopted in a wide range of experiments in multimedia learning research (e.g., Albus et al., 2021; Beege et al., 2019a; Beege et al., 2019b; Bender et al., 2021; Schneider et al., 2019b; Stárková et al., 2019).

Theoretical foundations of CLT postulate direct relationships between the cognitive load types and learning outcomes. Thus, it is assumed that learning materials that are difficult to encode lead to more extraneous processing, which in turn reduces learning outcomes because additional cognitive resources, irrelevant for learning, are wasted. Concerning ICL, learning materials should be designed so that the task's inherent element interactivity is easier to handle (Mayer & Moreno, 2010). Furthermore, it can be hypothesized that ECL negatively affects learning performance and that this can be justified on the basis of theoretical assumptions. Thus, inappropriately designed or organized learning materials require additional cognitive resources, which are consequently no longer available for actual learning (Sweller, 2010). In terms of GCL, it can be derived that a higher GCL leads to higher learning outcomes because this instead represents an active load (Klepsch & Seufert, 2021; Sweller, 2010). In contrast, instructional designs and procedures should challenge and motivate the learner to invest cognitive resources for understanding (Mayer & Moreno, 2010). In line with this reasoning, attempts have been made within CLT to

increase GCL (Paas & Van Merriënboer, 1994). Because increasing learning performance is the goal, greater GCL could lead to higher learning scores.

It is further assumed that the learner's domain-specific prior knowledge affects cognitive load and learning outcomes (Chen et al., 2017; Zu et al., 2021). Hereby, it is common to classify learners as novices or experts depending on the amount of their prior domain knowledge (Kalyuga & Renkl, 2010). In this vein, the expertise reversal effect states that the learner's domain-specific knowledge has a moderating effect on the effectiveness of CLT-based design recommendations (Chen et al., 2017). Consequently, the expertise reversal effect can be an additional source of ECL. Design decisions can enhance or reduce ECL perceptions in dependence of the prior knowledge of the learner. Accordingly, the interaction between the learner's expertise and the instructional procedures can lead to a reversal effect indicating that novices benefit more from an instructional intervention (reduced ECL), whereas experts may not benefit due to redundancies and associated inferences (no change or even enhanced ECL; Kalyuga, 2007).

Following the generally accepted definition of the ICL (e.g., Leppink et al., 2013), the domain-specific prior knowledge should correlate negatively with this cognitive load type. The more prior knowledge someone has, the less complex the learning material is perceived and vice versa. In this vein, one can assume that experts (with high prior knowledge) would assess a task involving a high ICL as less complex than novices (with low prior knowledge; Artino, 2008). Furthermore, learners with a high domain-specific prior knowledge can use already formed schemata while learning, making them less susceptible to poorly formatted learning materials that would tend to induce a high ECL (Paas et al., 2003). Accordingly, the domain-specific prior knowledge and the ECL should show a negative correlation indicating that learners with relatively high expertise report fewer ECL perceptions. Lastly, relationships between prior knowledge and GCL can also be postulated. With the assumption in mind that the GCL is indirectly related to the element interactivity of the learning material (Zu et al., 2020), it can be assumed that the prior knowledge and the germane load should correlate positively with each other. The more domain-specific prior knowledge (in the form of schemata) the learner has, the easier it is to allocate germane resources to learning-relevant activities.

Measuring Cognitive Load with Subjective Scales

Because working memory load is a key component of the CLT framework, measuring this load has been a high priority for researchers (Paas et al., 2003; Sweller, 2018; Sweller et al., 2011). However, cognitive load measurement is still an ongoing challenge in educational research (e.g., Ayres, 2018; de Jong, 2010; Kirschner et al., 2011; Moreno, 2010). In recent years, cognitive load research has adopted several measurement methods, with approaches divided into subjective scales and objective measures (Brünken et al., 2003). While self-reports are highly subjective, dual-task paradigms, learning outcomes, and physiological data are relatively objective methods. For example, cognitive load can be measured by asking learners to estimate their perceived cognitive load based on a Likert scale (direct) or by

measuring indicators that are assumed to be related to cognitive load (indirect). In this vein, dual-task approaches and physiological measures are promising alternatives to rating scales for measuring cognitive load but are beyond the scope of the current study.

Unidimensional Measurement of Cognitive Load

Subjective measures are still the most frequently used approach in educational research (e.g., Schmeck et al., 2015). Hereby, the learner is asked to assess and self-report the perceived amount of cognitive load while learning or working on a task (Sweller et al., 2011). This assessment is usually made after learning has taken place (Jiang & Kalyuga, 2020; Paas & Van Merriënboer, 1994). Accordingly, such instruments are applied on the assumption that individuals can give an accurate assessment of their experienced cognitive load — even if the questionnaire is conducted with a time delay (Ayres, 2006). In practical research, cognitive load is typically measured with numerical Likert-type rating scales in order to carry out statistical analyses (Ouweland et al., 2021). The most popular rating scale for subjective measurement of cognitive load for educational purposes was proposed in the early 1990s by Paas (1992). Hereby, learners are asked to rate their invested mental effort while learning on a nine-point single-item scale ranging from “very, very low mental effort” to “very, very high mental effort.” It is assumed that mental effort is an indicator of cognitive load. It has been shown that the Paas scale is sensitive for measuring differences in intrinsic cognitive load while learning (Naismith et al., 2015; Sweller et al., 2011). It should be noted that several studies adapted the scale by asking participants to rate the difficulty of the learning task (van Gog & Paas, 2008). This difference between invested mental effort and perceived task difficulty can quickly lead to problems of interpretation because learners are less motivated to invest mental effort when the learning is perceived as extremely difficult (e.g., Cennamo, 1993). Nevertheless, this scale enjoys frequent use as it is easy to implement and fast to handle for learners (Sweller, 2018). However, while it seems to be methodologically economical to measure this variable with one item, this is questionable from a psychometric point of view (e.g., Jiang & Kalyuga, 2020; Klepsch et al., 2017). Moreover, a measuring instrument consisting of only one item makes it impossible to calculate internal consistency, an important indicator for an instrument’s reliability. With the proviso that the questionnaire is applied several times in one experiment (e.g., after each chapter of the learning material), it is possible to calculate test–retest reliability. However, since cognitive load can vary during learning and is therefore dynamic in nature, calculating test–retest reliability could lead to misleading reliability values. In this vein, test–retest reliability is only valid when constructs stable over time are examined (e.g., Baumeister, 1991).

To avoid such methodological problems, Leppink et al., (2013, 2014), for example, have recommended using multiple items that allows for a more precise cognitive load measurement. Accordingly, measuring cognitive load without differentiating between the individual cognitive load types seems to be insufficient when evaluating

the effectiveness of multimedia learning environments or interventions (e.g., van Gog & Paas, 2008).

Multidimensional Measurement of Cognitive Load

Taking up this criticism, several studies have introduced measurements that target the cognitive load types separately (e.g., Eysink et al., 2009; Klepsch et al., 2017; Leppink et al., 2013, 2014). What these scales have in common is that they focus on certain cognitive load types and can therefore differentiate load more precisely. The instrument by Eysink et al. (2009) consists of six items (see Appendix A). What makes this questionnaire unique is that besides targeting ICL, ECL, and GCL, one additional item is included that measures the perceived overall cognitive load. However, this item asks learners to indicate the amount of effort they invested in following the learning material. The constructs ICL and GCL are only measured with one item each, so that no conclusions can be drawn about their internal consistency. While ICL asked participants to estimate the perceived difficulty of the learning material, GCL is related to the question of how easy or difficult it was to understand the learning content. In addition, three items concerning extraneous load refer both to the navigation and to design of the learning task, as well as to the accessibility of information.

Four years later, Leppink et al. (2013) developed a multidimensional scale (see Appendix B) including ten items referring to ECL (three items), ICL (three items), and GCL (four items). The authors conducted four studies with participants learning statistics to validate the questionnaire. Concretely, items representing ICL asked participants to estimate the complexity of presented topics, the learning activity, and covered formulas and definitions. In addition, the items representing ECL refer to the instruction and explanations in terms of their unclearness and ineffectiveness. The items concerning the GCL asked the participants to assess the extent to which the learning activity has enhanced their understanding and knowledge of the learning topic. Generally, the three-factorial structure with ten items was supported in the study by Leppink et al. (2013), though with some limitations. In particular, the GCL and learning outcomes did not correlate — a finding which contradicts theoretical assumptions of CLT. Moreover, the proposed model could only be partially supported in two studies. These issues encouraged the authors to review their proposed measurement (cf. Leppink et al., 2014; Appendix C). Since the learning topic of statistics was chosen in the first validation approach, two follow-up studies were conducted in order to examine whether the instrument is also reliable for other learning contexts. In addition, these studies should provide further evidence that the instrument can distinguish between the three types of cognitive load. First of all, the two studies supported the differentiation between items measuring intrinsic and extrinsic cognitive load. However, in line with the recent reconceptualization of GCL (e.g., Kalyuga, 2011; Sweller, 2010), “the assumption that the third factor in the psychometric instrument represents or closely relates to germane cognitive load is limited” (Leppink et al., 2014, p. 40) indicating that the three-factor model may not be fully adopted. In addition, Leppink et al. (2014) criticize the measurement by arguing

that item responses on ICL, ECL, and GCL give no indication of how much mental effort the learners invest. Thus, no conclusions can be drawn about the load imposed on the working memory. Addressing this problem, one item was added to each type of cognitive load, which targets the mental effort invested in each factor to examine more directly the relationship between cognitive load and learning outcomes. The added mental effort items increase the reliability of the intrinsic and extraneous load factors, but not for germane load. Both versions of the measurement (Leppink et al., 2013, 2014) enjoy great popularity and are frequently used in experimental studies dealing with multimedia learning settings (as shown in a review by Mutlu-Bayraktar et al., 2019). Klepsch et al. (2017) introduced another cognitive load self-report measurement trying to eliminate potential inconsistencies (see Appendix D). Hereby, ICL (two items) and ECL (three items) items refer to the task's complexity and the design of the learning material, while the GCL (three items) "should focus on the additional investment of cognitive processes into learning" (Klepsch et al., 2017, p. 5). In contrast to the Leppink questionnaires (2013, 2014), the instrument from Klepsch et al. (2017) is not specific to the learning topic and can therefore be easily adapted to the material used in a specific study (e.g., animation, video, or text). Like the scales from Leppink and colleagues (2013, 2014), the self-report measures differentiate all three cognitive load types. The authors validated the instrument with two different strategies. First, they used an informed rating: Students were trained to understand and differentiate between the types of cognitive load. The training consisted of a PowerPoint lecture that introduced students to the notion of cognitive load. After the training, the students were expected to be able to detect cognitive load types and to distinguish them from one another. Second, they used a naïve rating: students had to rate the same learning situations without being informed about the cognitive load types beforehand. Overall, informed ratings seem to be a promising instrument to measure the different types of cognitive load. However, giving participants an introduction to the CLT framework is not always possible in experimental studies. The naïve rating scale is much easier to handle and less time-consuming. The authors emphasize the broad possibilities for applying this method in several learning domains and studies. However, the results suggested that the GCL items should be used with caution because of varying levels of understandings on the part of respondents. The authors hence recommend conducting a reliability test. In the past few years, the naïve rating has been frequently used in experimental studies and is therefore part of this analysis.

Reliability and Validity of Subjective Cognitive Load Measurements

In general, the quality of a psychological test or measurement can be evaluated by means of three primary indicators — objectivity, reliability, and validity (Adams, 1936; Moosbrugger & Kelava, 2020). These quality criteria must be met in order to adequately measure a psychological construct using a questionnaire. For the aim of this work, the reliability and validity (Drost, 2011) of the cognitive load questionnaires are of central importance. To approach these constructs, several methods have been suggested in recent research.

The Internal Consistency of Subjective Cognitive Load Measurements

Reliability describes the consistency of a measuring instrument (Schuurman & Hamaker, 2019). Test theory assumes that a reliable instrument contains as little measurement error as possible, which maximizes the proportion of variance that arises due to actual differences in the construct to be measured (Kimberlin & Winterstein, 2008). Given a hypothetical situation in which the measurement is replicated, a reliable measurement should produce the same results under the premise that the measured construct remains unchanged (Heale & Twycross, 2015). Reliability scores can be calculated with the help of statistics that give an indication of the extent to which the instrument is free from measurement errors. Several well-known methods for estimating reliability have been established — internal consistency, parallel test, and test–retest (Schuurman & Hamaker, 2019), while various authors also rely on the split-half method to measure reliability (e.g., Cho, 2016; Thompson et al., 2010). To measure parallel test reliability, two different versions of an instrument measuring the same construct are presented to the participants several times with the two measuring instruments only differing in their wording. Concerning test–retest reliability, the procedure is similar to the parallel test method. However, the same instrument (with identical wordings) is given to participants more than once (Heale & Twycross, 2015). The third method, split-half, involves splitting the scale into two parallel halves which are then correlated (Cho, 2016). This procedure assumes that a test that is supposed to measure a construct should do this consistently across the scale with each item. The internal consistency, which is central to the aim of this work, is an estimate of the degree to which the items of the scale measure the same concept (Drost, 2011). Cronbach's alpha (α ; Cronbach, 1951) is the most frequently used indicator for internal consistency (Cho, 2016; Hogan et al., 2000; Osburn, 2000; Streiner, 2003). The alpha value represents the average of all split-half reliabilities (Cortina, 1993; Ferketich, 1990). The instrument is randomly split into two halves, whereby the correlation between the sum scores estimates the reliability of the half test (Warrens, 2015). To infer the reliability of the full test, an estimate correction is needed (Revelle & Zinbarg, 2009). The resulting Cronbach's alpha value should thus reach comparatively high values since it is equivariant with a high proportion of explained systematic variance. In general, it can have values between zero and one, but negative values are also possible when some items of the scale are negatively correlated (Vaske et al., 2017). Nevertheless, specifications are rare regarding how high the value must be in order to meet the requirement of representing a reliable test. Recommendations range from 0.65 to 0.80 (Green et al., 1977) to 0.90 (Streiner, 2003) needing to be reached before the reliability of the scale can be assessed as adequate. However, in the social sciences, a value above 0.70 is generally accepted (Nunnally, 1978; Taber, 2018). However, a clearly increased alpha value can quickly lead to redundancy between the items. Following Streiner (2003), an internal consistency of 0.90 and above indicating high correlations between the items may suggest that some items of the scale are redundant. These items are assumed to test the same question or statement in another guise (Tavakol & Dennick, 2011a). Concerning the cognitive

load types, reliable scales should be able to measure the different sub-types with high internal consistency. Accordingly, this work focuses on internal consistency.

The Validity of Subjective Cognitive Load Measurements

Given a reliable measurement, it should be not automatically assumed to be of high quality. It must also meet the standard of validity, which is generally defined “as the extent to which an instrument measures what it purports to measure” (Kimberlin & Winterstein, 2008, p. 2278, see also Kane, 2001). In this vein, Cook and Beckman (2006) as well as Kane (2013) pointed out that validity is not a property of the measuring instrument, but more the interpretation of what it measures. The results of a test or instrument are valid when the interpretations are justifiable in the context of the test’s intended use (Kimberlin & Winterstein, 2008). As outlined by Kane (2001), resulting evaluative judgments are based on the degree to which theoretical evidence supports the interpretation of the test. Hereby, competing interpretations should also be considered (e.g., Messick, 1989). Validating a measuring instrument is therefore not a routine task but rather a close linkage of theory-based assumptions and test data. Accordingly, the underlying construct (e.g., cognitive load while learning) can never be perfectly reflected by the test, where the aim is to achieve correlations that are as high as possible (Cook and Beckman, 2006).

It is generally accepted that validity is a multifaceted construct. Consequently, in the literature, there are three main approaches to investigate the validity of a psychological test — content validity, construct validity, and criterion validity (Heale & Twycross, 2015). Content validity describes the extent to which the items of a scale are representative of the targeted construct so that the scale measures all relevant aspects (Almanasreh et al., 2019). Assessing the content validity is mostly conducted with the help of expert opinions. In this vein, the content validity index (CVI) is calculated based on item relevance ratings made by experts (Polit & Beck, 2006). Consequently, content validity plays an important role in the instrument development phase (ideally on the basis of expert surveys and reviewers in the publication process) and is therefore not part of this analysis. Instead, construct validity and criterion validity (including explicitly measurable variables) were the main focus.

Construct Validity

Introduced by Cronbach and Meehl (1955), construct validity refers to the concordance between the results of the measurement and the underlying theory. In this vein, the instrument should measure all relevant facets of the concept adequately. With the introduction of the construct validity, our understanding of the concept validity has changed. The question was no longer whether a psychological test measures what it is supposed to measure, but how it fits into the nomological network with other theoretically related constructs (Borsboom et al., 2004; Colliver et al., 2012). Quantifying the construct validity of an instrument is mostly conducted by identifying correlations with several measures. The resulting correlation patterns provide information about the degree of conformity between the measure and theoretically

predictable variables (Westen & Rosenthal, 2003). Based on Campbell and Fiske (1959), construct validity can be divided into convergent and discriminant validity. Convergent validity is given when different instruments which aim to measure a common construct correlate highly with each other. In contrast, when instruments measuring different constructs do not correlate with each other, this effect is known as discriminant validity.

Criterion Validity

In terms of achieving criterion validity, psychological constructs like cognitive load should have a high degree of compliance with practically relevant external criteria (Dunn, 2020). Accordingly, the scale is not considered separately but in connection with other practically significant variables (Drost, 2011). Criterion validity can be classified into two types based on the timing of the measurements. If data is collected using the measuring instrument before data collection on the criterion, one speaks of predictive validity. Hereby, the scale's ability to predict the criterion variable is tested (Kimberlin & Winterstein, 2008). A measure can also be assessed in relation to relevant criterion variables at the same time (concurrent validity; Westen & Rosenthal, 2003).

The Present Work

Subjective judgments about perceived cognitive load during learning with multimedia can be associated with certain weaknesses. In this vein, there is an ongoing debate among researchers on how to assess cognitive load reliably and validly with self-rating scales (e.g., Brünken et al., 2003, 2010; Schmeck et al., 2015). Reliably and validly assessing the three types of cognitive load “has become the holy grail of CLT research” (Kirschner et al., 2011, p. 104). Nonetheless, subjective cognitive load measures are the most frequently used approach in educational research, as they can be easily implemented in experimental settings without taking up too much time. In order to verify the extent to which questionnaires provide reliable and valid insights about the construct of interest, the measuring instrument must meet the quality requirements already explained. The aim of this work is therefore to conduct a meta-analysis of cognitive load questionnaires with regard to their internal consistency and validity across studies. As questionnaires are always constructed in a theory-driven manner, it is also of great importance to examine the extent to which the cognitive load types correlate (1) among each other and (2) with important external criteria (e.g., learning outcomes) for validity purposes. In order to check the reliability and validity of cognitive load questionnaires, the four widely used instruments were chosen (see “[Measuring Cognitive Load with Subjective Scales](#)”): Eysink et al. (2009); Leppink et al., (2013, 2014); and Klepsch et al. (2017).

The first part of the analysis focuses on internal consistency as a sub-type of reliability. In addition, moderator analyses were conducted to examine whether and how the internal consistency value Cronbach's alpha is influenced by relevant CLT-related factors (Hall & Rosenthal, 1991). In this vein, the moderators of educational

setting, the domain of the instructional material, the presentation mode of the learning material, and the number of response options (i.e., the number of scale points) were considered. Based on the insights gained, recommendations will be formulated how on to use subjective cognitive load scales in experimental multimedia learning settings. The second part focuses on the validity of cognitive load questionnaires, specifically, construct validity (i.e., how can the different questionnaires represent the theoretical assumptions underpinning CLT) and criterion validity (i.e., how cognitive load relates to important external variables, which in our case are the learning measures known as retention and transfer, as well as domain-specific prior knowledge).

Methods

Search Strategy

This study focused on articles published in peer-reviewed journals, which used the cognitive load measurements from Eysink et al. (2009), Klepsch et al. (2017), and Leppink et al., (2013, 2014). These scales were selected as they measure all cognitive load types with multiple items. To find suitable studies, a literature search was carried out and finished on August 25, 2021. The “cited by” function of Google Scholar (for a review proving the adequacy for literature search, see Martin-Martin et al., 2017) was consulted to access the listing of the respective study. By this, an overview of all works could be gained which cited the respective scale validation’s studies (Eysink et al., 2009; Klepsch et al., 2017; Leppink et al., 2013, 2014). All of these studies ($N=1193$) were then evaluated in terms of the following inclusion criteria (for an overview of the search process, see Fig. 2).

To be included in the meta-analysis, experimental studies had to be carried out in the field of multimedia learning indicating that a learning material or setting was intentionally manipulated. In the fields of multimedia learning and cognitive load, controlled and randomized experiments are the common and ideal ways to research and sustainably improve instructional scenarios and materials (Borman, 2002; Cobb et al., 2003; Sweller, 2021). Consequently, only experimental studies were included. In this vein, it is important to rely on reliable and valid measurement methods and, therefore, experimental studies are part of the analysis. At least one multimedia learning medium had to be included in the experimental setting of the respective study. Thus, experimental studies were included in which a multimedia learning material was intentionally manipulated (e.g., varying the font of the learning text; Beege et al., 2021) or in which the handling of the learning material was intentionally varied (e.g., tracing the content of the learning material with the fingers; Tang et al., 2019). Only studies published in English were included to ensure transparency in the scientific community. Moreover, only peer-reviewed journal papers were included to ensure methodological quality (see Castro-Alonso et al., 2019). In addition, at least one cognitive load facet was measured with a pre-set list of subjective questionnaires (Eysink et al., 2009; Klepsch et al., 2017; Leppink et al., 2013, 2014). All articles included were scanned for relevant data (reliabilities and correlations;

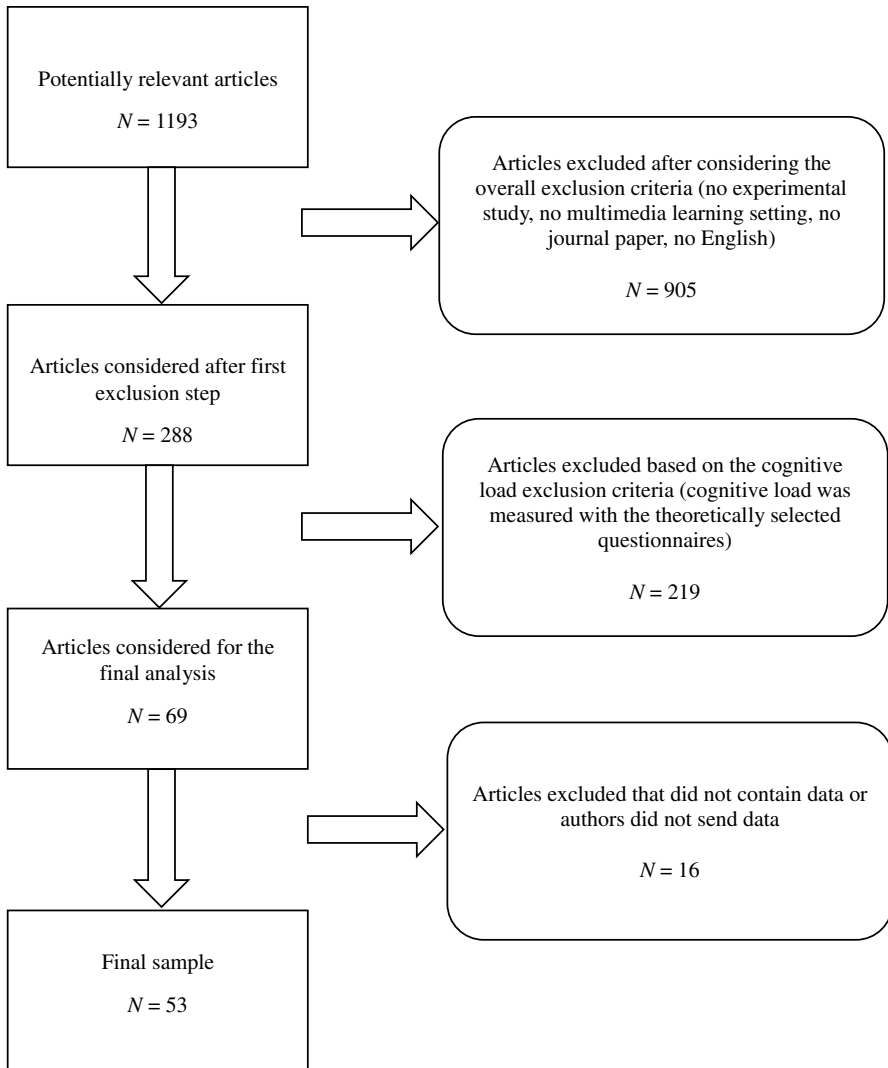


Fig. 2 Flow diagram of the selection of articles

see supplementary material). While most of the studies reported reliability values, correlation matrices were often missing. In the case of missing data, the corresponding author of the respective study was contacted by email and asked to fill in a prepared matrix and to send it back to the authors of this study. The matrix included a correlation matrix in which the authors should complete correlations on construct level between the concepts relevant for this work. Specifically, this matrix involved the constructs ICL, ECL, GCL, prior knowledge, retention, and transfer. When studies reported no relevant data within the manuscript and the supplementary material, or the authors did not reply to our email, the study was excluded from

the analysis. Even if data was incomplete (e.g., a study reported reliability values, but no correlations for validity analyses), it was nevertheless included in the analysis in order to collect as much data as possible. This resulted in the different numbers of effect sizes used in the calculation of the meta-analysis. An overview of all studies included in the meta-analysis is given in Table 1.

Sample Characteristics

Overall, 53 articles including 67 experiments ($N=7413$ participants) were considered for meta-analysis. Sample sizes, which were relevant for this work, ranged from $N=20$ to $N=485$. The mean age of the participants was 20.4 years, and the overall percentage of women was 63.5%. The mean sample size was $M=103.3$ ($SD=68.0$). Most experiments ($N=33$) included the questionnaire from Leppink et al. (2013) while the related questionnaire from Leppink et al. (2014) was used in seven experiments to measure cognitive load. Moreover, the Klepsch et al. (2017) questionnaire was used in 20 experiments while 10 experiments used the questionnaire from Eysink et al. (2009). In three studies (Skulmowski & Rey, 2018, 2020a; Thees et al., 2021), two different cognitive load questionnaires were used.

Measures of Reliability and Validity

For this meta-analysis, theory-based dependent variables (i.e., the reliability and validity) were defined in advance and related data was subsequently collected in the course of the literature search. Concerning reliability, this work focused on the internal consistency of the cognitive load questionnaires (Ferketich, 1990). Accordingly, the Cronbach's alpha values (Cronbach, 1951; Osburn, 2000; Streiner, 2003) of the respective cognitive load types were collected and meta-analytically calculated. Following the three-factor model (Klepsch et al., 2017; Sweller et al., 1998; Zavgorodniaia et al., 2020), the alpha values were collected separately for the ICL, ECL, and GCL. In terms of construct validity, correlations between the individual cognitive load types were collected and meta-analytically calculated. Following the already mentioned retention-transfer approach (Mayer, 2001), correlation calculations between the cognitive load types ICL, ECL, and GCL were conducted with retention and transfer performance to gain deeper insights into the criterion validity of the cognitive load questionnaires. Because retention and transfer have a different focus on knowledge gain (e.g., Mayer, 2001), correlations were calculated separately. Since the learner's domain-specific prior knowledge plays an important role within the CLT framework (e.g., Chen et al., 2017), correlations between this construct and the cognitive load types ICL, ECL, and GCL were also calculated meta-analytically.

Moderating Variables

Besides calculating main effects in a first step for the internal consistency meta-analysis, moderator analyses were conducted in a second step (Hall & Rosenthal, 1991). For this purpose, sub-groups were formed based on predefined criteria — the

Table 1 Descriptive overview of all studies considered for the meta-analysis

No	Authors	Educational setting	N	Proportion of females	Mean age (in years)	Learners prior knowledge (1 = low, 2 = high)	Instructional material domain	Presentation mode of the learning material	Used cognitive load questionnaire	Measured cognitive load types	Scale (number of response options)
1	Albus et al. (2021)	School education	107	43.0%	15.3	1	Natural sciences	Interactive	Klepsch et al. (2017)	ICL; ECL; GCL	7
2	Altmeyer et al. (2020)	Adult education	50	80.1%	26.0	1	Natural sciences	Mixed	Leppink et al. (2014)	ICL; ECL; GCL	6
3	Andrade et al. (2015)	Adult education	268	61.0%	-	1	Natural sciences	Mixed	Leppink et al. (2013)	ICL; ECL; GCL	-
4	Anggraini et al. (2020)	School education	30	-	-	-	Natural sciences	Dynamic	Leppink et al. (2014)	ECL; GCL	7
5	Becker et al. (2020)	School education	286	32.9%	15.6	1	Natural sciences	Mixed	Leppink et al. (2013)	ICL; ECL; GCL	-
6	Beege et al. (2019a) (exp. 1)	Adult education	73	75.3%	23.1	1	Natural sciences	Dynamic	Leppink et al. (2013)	ICL; ECL; GCL	11
7	Beege et al. (2019a) (exp. 2)	Adult education	126	79.4%	22.1	2	Natural sciences	Dynamic	Leppink et al. (2013)	ICL; ECL; GCL	11
8	Beege et al. (2021)	Adult education	138	85.5%	22.9	1	Natural sciences	Static	Klepsch et al. (2017)	ICL; ECL	7
9	Beege et al. (2017)	Adult education	84	73.8%	22.5	1	Natural sciences	Dynamic	Eysink et al. (2009)	ICL; ECL; GCL	9
10	Beege et al. (2020)	School education	118	46.1%	14.1	1	Natural sciences	Dynamic	Klepsch et al. (2017)	ICL; ECL; GCL	7
11	Beege et al. (2019b) (exp. 1)	Adult education	98	74.5%	22.7	1	Natural sciences	Static	Eysink et al. (2009)	ICL; ECL; GCL	9

Table 1 (continued)

No	Authors	Educational setting	N	Proportion of females	Mean age (in years)	Learners prior knowledge (1 = low, 2 = high)	Instructional material domain	Presentation mode of the learning material	Used cognitive load questionnaire	Measured cognitive load types	Scale (number of response options)
12	Beege et al. (2019b) (exp. 2)	Adult education	85	75.6%	23.5	1	Natural sciences	Static	Eysink et al. (2009)	ICL; ECL; GCL	9
13	Bender et al. (2021)	Adult education	194	77.32%	24.2	1	Natural sciences	Static	Klepsch et al. (2017)	ICL; GCL	5
14	Chung and Cheon (2020)	Adult education	100	58.0%	26.7	2	Natural sciences	Static	Leppink et al. (2013)	ECL; GCL	9
15	Colliot and Jamet (2018)	Adult education	43	62.8%	21.2	1	Natural sciences	Dynamic	Leppink et al. (2014)	ECL	-
16	Davis et al. (2019)	Adult education	172	49.4%	22.6	1	Social sciences	Dynamic	Leppink et al. (2013)	ICL; ECL; GCL	10
17	Debue and Van De Leemput (2014)	Adult education	92	87.0%	20.0	-	Others	Mixed	Leppink et al. (2013)	ICL; ECL; GCL	10
18	Dervic et al. (2019)	School education	49	53.7%	16.1	2	Natural sciences	Mixed	Leppink et al. (2013)	ICL; ECL; GCL	-
19	Eitel et al. (2019)	Adult education	84	69.0%	24.2	1	Natural sciences	Static	Klepsch et al. (2017)	ICL; ECL	5
20	Fanguy et al. (2019)	Adult education	110	26.4%	27.0	-	Social sciences	Dynamic	Leppink et al. (2013)	GCL	10
21	Glogger-Frey et al. (2017)	School education	99	81.8%	13.6	2	Natural sciences	Static	Leppink et al. (2013)	ECL	6
22	Greenberg et al. (2021)	Adult education	70	64.3%	22.7	1	Social sciences	Dynamic	Leppink et al. (2013)	ICL; ECL; GCL	11

Table 1 (continued)

No	Authors	Educational setting	N	Proportion of females	Mean age (in years)	Learners prior knowledge (1 = low, 2 = high)	Instructional material domain	Presentation mode of the learning material	Used cognitive load questionnaire	Measured cognitive load types	Scale (number of response options)
23	Gupta and Zheng (2020)	Adult education	160	73.1%	23.3	-	Logic and mathematics	Static	Leppink et al. (2013)	ICL; ECL; GCL	10
24	Klepsch and Seufert (2020) (exp. 1)	Adult education	62	34.4%	25.3	-	Logic and mathematics	Dynamic	Klepsch et al. (2017)	ICL; ECL; GCL	7
25	Klepsch and Seufert (2020) (exp. 2)	Adult education	62	33.9%	25.2	-	Others	Dynamic	Klepsch et al. (2017)	ICL; ECL; GCL	7
26	Klepsch and Seufert (2020) (exp. 3)	Adult education	40	82.5%	22.8	-	Logic and mathematics	Static	Klepsch et al. (2017)	ICL; ECL; GCL	7
27	Klepsch and Seufert (2020) (exp. 4)	Adult education	51	92.2%	23.2	-	Natural sciences	Static	Klepsch et al. (2017)	ICL; ECL; GCL	7
28	Klepsch and Seufert (2020) (exp. 5)	Adult education	31	54.8%	28.4	-	Natural sciences	Static	Klepsch et al. (2017)	ICL; ECL; GCL	7

Table 1 (continued)

No	Authors	Educational setting	N	Proportion of females	Mean age (in years)	Learners prior knowledge (1 = low, 2 = high)	Instructional material domain	Presentation mode of the learning material	Used cognitive load questionnaire	Measured cognitive load types	Scale (number of response options)
29	Klepsch and Seufert (2020) (exp. 6)	Adult education	36	91.7%	21.4	-	Logic and mathematics	Static	Klepsch et al. (2017)	ICL; ECL; GCL	7
30	Klepsch and Seufert (2021) (exp. 1)	Adult education	73	86.3%	22.3	-	Logic and mathematics	Static	Klepsch et al. (2017)	ICL; ECL; GCL	-
31	Klepsch and Seufert (2021) (exp. 2)	Adult education	72	83.3%	23.0	-	Natural sciences	Static	Klepsch et al. (2017)	ICL; ECL; GCL	-
32	Korbach et al. (2020)	Adult education	60	88.9%	23.7	2	Natural sciences	Static	Leppink et al. (2014)	ICL; ECL	10
33	Lehmann et al. (2019)	School education	167	85.6%	14.4	2	Social sciences	Static	Klepsch et al. (2017)	ECL; GCL	7
34	Liao et al. (2019)	School education	109	45.9%	-	-	Natural sciences	Interactive	Leppink et al. (2013)	ICL; ECL; GCL	11
35	Liao et al. (2020)	Adult education	20	70.0%	25.7	-	Natural sciences	Dynamic	Leppink et al. (2014)	ICL; ECL; GCL	10
36	Mikheeva et al. (2021)	Adult education	100	-	-	-	Logic and mathematics	Static	Leppink et al. (2013)	ICL; ECL	11
37	Miller et al. (2020)	Adult education	84	37.3%	22.2	-	Social sciences	Static	Klepsch et al. (2017)	ICL; ECL; GCL	10

Table 1 (continued)

No	Authors	Educational setting	N	Proportion of females	Mean age (in years)	Learners prior knowledge (1 = low, 2 = high)	Instructional material domain	Presentation mode of the learning material	Used cognitive load questionnaire	Measured cognitive load types	Scale (number of response options)
38	Nebel et al. (2017a)	School education	56	48.2%	16.9	1	Social sciences	Interactive	Eysink et al. (2009)	ICL; ECL; GCL	-
39	Nebel et al. (2017b)	Adult education	87	75.9%	21.6	1	Logic and mathematics	Interactive	Eysink et al. (2009)	ICL; ECL; GCL	-
40	Nebel et al. (2016)	Adult education	115	70.3%	22.8	1	Logic and mathematics	Interactive	Eysink et al. (2009)	ICL; ECL; GCL	9
41	Peiko et al. (2020)	School education	58	55.2%	-	1	Natural sciences	Interactive	Leppink et al. (2013)	ICL; ECL; GCL	11
42	Schneider et al. (2018b) (exp. 2)	Adult education	86	74.4%	23.2	1	Social sciences	Static	Leppink et al. (2013)	ICL; ECL; GCL	11
43	Schneider et al. (2018b) (exp. 3)	School education	162	58.6%	17.6	1	Natural sciences	Static	Leppink et al. (2013)	ICL; ECL; GCL	11
44	Schneider et al. (2021) (Exp. 1)	Adult education	104	77.9%	22.1	1	Natural sciences	Static	Leppink et al. (2013)	ICL, ECL	7
45	Schneider et al. (2021) (exp. 2)	School education	153	63.9%	17.5	1	Natural sciences	Static	Leppink et al. (2013)	ICL, ECL	7
46	Schneider et al. (2018c) (exp. 2)	School education	102	48.5%	14.4	1	Natural sciences	Static	Leppink et al. (2013)	ICL; ECL; GCL	11

Table 1 (continued)

No	Authors	Educational setting	N	Proportion of females	Mean age (in years)	Learners prior knowledge (1 = low, 2 = high)	Instructional material domain	Presentation mode of the learning material	Used cognitive load questionnaire	Measured cognitive load types	Scale (number of response options)
47	Schneider et al. (2018d) (exp. 1)	School education	79	62.0%	17.3	1	Natural sciences	Static	Eysink et al. (2009)	ICL; ECL	9
48	Schneider et al. (2018d) (exp. 2)	School education	87	45.0%	14.2	1	Natural sciences	Static	Leppink et al. (2013)	ICL; ECL	-
49	Schneider et al. (2015)	School education	120	65.8%	18.2	1	Social sciences	Static	Eysink et al. (2009)	ICL; ECL; GCL	7
50	Schneider et al. (2019a) (exp. 1)	School education	87	49.4%	11.1	1	Social sciences	Static	Leppink et al. (2013)	ICL; ECL; GCL	11
51	Schneider et al. (2019a) (exp. 2)	School education	148	60.1%	14.1	1	Social sciences	Static	Leppink et al. (2013)	ICL; ECL; GCL	11
52	Schneider et al. (2019a) (exp. 3)	School education	162	58.4%	17.6	1	Social sciences	Static	Leppink et al. (2013)	ICL; ECL; GCL	11
53	Schneider et al. (2019b)	Adult education	100	72.0%	22.8	1	Social sciences	Static	Leppink et al. (2013)	ICL; ECL; GCL	11
54	Schrader et al. (2021)	Adult education	95	50.5%	23.5	1	Natural sciences	Dynamic	Klepsch et al. (2017)	ICL; ECL; GCL	7
55	Skulmowski et al. (2016)	Adult education	96	67.7%	23.9	1	Natural sciences	Mixed	Eysink et al. (2009)	ICL; ECL; GCL	-

Table 1 (continued)

No	Authors	Educational setting	N	Proportion of females	Mean age (in years)	Learners prior knowledge (1 = low, 2 = high)	Instructional material domain	Presentation mode of the learning material	Used cognitive load questionnaire	Measured cognitive load types	Scale (number of response options)
56	Skulmowski and Rey (2018)	Adult education	108	77.8%	22.3	-	Natural sciences	Static	Leppink et al. (2013) Eysink et al. (2009)	ICL; GCL ECL	11 10
57	Skulmowski and Rey (2020a)	Adult education	42	78.6%	-	-	Natural sciences	Static	Leppink et al. (2013) Klepsch et al. (2017)	ECL ECL	7 7
58	Skulmowski and Rey (2020b)	Adult education	50	74%	-	1	Natural sciences	Static	Klepsch et al. (2017)	ECL	-
59	Stark et al. (2018)	School education	148	57.0%	15.6	1	Natural sciences	Static	Leppink et al. (2013)	ICL; ECL; GCL	11
60	Stárková et al. (2019)	Adult education	181	-	22.2	-	Natural sciences	Static	Leppink et al. (2014)	ICL; ECL	7
61	Tang et al. (2019)	School education	44	-	10.7	1	Natural sciences	Mixed	Leppink et al. (2013)	ICL; ECL	11
62	Thees et al. (2021)	Adult education	95	15%	20.0	-	Natural sciences	Mixed	Klepsch et al. (2017) Leppink et al. (2013)	ICL, ECL, GCL ICL, ECL, GCL	7
63	Thees et al. (2020)	Adult education	74	26.9%	20.2	2	Natural sciences	Mixed	Leppink et al. (2013)	ICL; ECL; GCL	-
64	Wang et al. (2021a) (exp. 1)	School education	88	68.0%	17.5	2	Natural sciences	Static	Klepsch et al. (2017)	ICL; ECL; GCL	7

Table 1 (continued)

No	Authors	Educational setting	N	Proportion of females	Mean age (in years)	Learners prior knowledge (1 = low, 2 = high)	Instructional material domain	Presentation mode of the learning material	Used cognitive load questionnaire	Measured cognitive load types	Scale (number of response options)
65	Wang et al. (2021b) (exp. 1)	School education	93	47.3%	11.1	2	Logic and mathematics	Static	Leppink et al. (2013)	ICL; ECL	11
66	Wang et al. (2021b) (exp. 2)	School education	90	48.9%	20.8	2	Logic and mathematics	Static	Leppink et al. (2013)	ICL; ECL	11
67	Xiong (2017)	Adult education	485	84.1%	18.1	-	Natural sciences	Static	Leppink et al. (2014)	ICL; ECL; GCL	11

moderating variables. As cognitive load perceptions may depend on the learner's age and educational background, the educational setting in which the respective study was conducted was captured. Hereby, the classifications school education (involving primary and secondary education) and adult education (involving higher education and adult education) were chosen. Assuming that learners with increasing age are better able to make metacognitive assessments such as for the perceived cognitive load (van der Stel & Veenman, 2010), it could be hypothesized that the educational setting affects the consistency of the respective questionnaires. In this vein, Leahy (2018) pointed out that subjective questionnaires are problematic when these are used with children. As the second moderating variable, the domain of the instructional material was considered important as the CLT finds application in a wide range of learning domains (e.g., Alpizar et al., 2020; Brom et al., 2018; Rey et al., 2019; Schneider et al., 2019a). Hereby, four different sub-groups were defined: natural sciences, social sciences, logic and mathematics, and other. Within multimedia learning research, instructional material can be classified based on its presentation mode. Thus, four sub-groups were built describing the learner's intervention and control options of the learning material (e.g., Weidenmann, 2002). First, static learning material includes non-moving text and/or pictures (e.g., Schneider et al., 2015). Second, dynamic materials are characterized by moving images, as is the case with videos or animations. However, in some dynamic learning materials, the learner has no control over the progress indicating that the video or animation cannot be paused or rewind (i.e., system-paced materials). Third, multimedia learning materials presented in an interactive presentation mode allow learners to have full control over the progress (i.e., learner-paced materials). Hereby, the possibilities range from playing an educational video game (e.g., Nebel et al., 2016) to moving the learners head around in a virtual reality environment (e.g., Albus et al., 2021). Fourth, when a study experimentally manipulated the presentation mode (e.g., Andrade et al., 2015; Dervić et al., 2019), this study was classified as mixed. The same classification was used in a meta-analysis by Schroeder and Cenkcı (2018).

From a psychometric point of view, the number of response options (i.e., the number of scale points) plays an important role and has been examined in a variety of psychological studies (e.g., Lissitz & Green, 1975; Matell & Jacoby, 1972; Simms et al., 2019; Wakita et al., 2012). For instance, there is empirical evidence that scales involving an odd number of options are not suitable to measure a construct (Dalal et al., 2013). Nevertheless, there are no clear recommendations on how many answer options should be provided when learners are asked to rate their perceived cognitive load on a subjective scale. Therefore, the number of answer options was collected for each study included in this meta-analysis.

Analysis Methods

In general, a meta-analysis collects empirical studies addressing the same research question to calculate the mean and variance of a population effect (Field & Gillett, 2010). Meta-analyses from the research field of educational psychology usually report weighted average effect sizes (e.g., Cohen's *d*; Hedges' *g*) when examining

the effectiveness of learning formats or principles. The focus of this work, however, is to examine (a) the internal consistency via Cronbach's alpha and (b) the validity of the cognitive load questionnaires designed by Eysink et al. (2009), Klepsch et al. (2017), and Leppink et al., (2013, 2014). Both calculations were calculated following the same pattern based on correlations. The meta-analytical procedure was carried out using *JASP* version 0.15 (JASP Team, 2021). An effect size was confirmed as significant ($p < 0.05$) when the associated confidence interval (CI) does not include zero (Hedges et al., 1992; Nakagawa & Cuthill, 2007). In addition, moderator analyses were calculated with *SPSS* version 28.0 (IBM Corp, 2021). To compare the effect sizes (i.e., aggregated reliability and validity estimates) for significant differences, the 95% percent confidence intervals (CI) were consulted (Cumming & Finch, 2005). If an effect size was not included in the confidence interval of another effect size to be compared, it was assumed that a significant difference exists.

Internal Consistency and Methods of Generalization

Aggregating reliability estimates from different studies with meta-analytic methods is described as *reliability generalization* (Vacha-Haase, 1998). Hereby, a meta-analysis can also be used to identify moderators of the alpha value (Bonett, 2010). If alpha shows similar values across different samples and experimental conditions, strong evidence of reliability generalization can be provided. In practical applications, usually only sample estimates of Cronbach's alpha are provided by scientific articles. This is mainly because many measurements are conducted with too small sample sizes making it difficult to estimate alpha with adequate precision (Bonett, 2010). As this work includes a larger sample size, a more accurate estimate with confidence intervals can be calculated. Following previous studies that have cumulated estimates of reliability (e.g., Graham & Christiansen, 2009; Graham et al., 2011; Pentapati et al., 2020; Piqueras et al., 2017), internal consistency of cognitive load scales was calculated for ICL, ECL, and GCL separately. Hereby, the reliability generalization framework allows the comparison of reliability estimates across a variety of studies (Vacha-Haase, 1998). As Cronbach's alpha is variance-adjusted, it corresponds to the value of the squared correlation (Thompson & Vacha-Haase, 2000). In detail, the square root of the reliability coefficients was calculated to obtain a r -equivalent correlation (Graham et al., 2011). However, correlations are not normally distributed because the bounded value range $[-1, +1]$ can lead to a skewed sampling distribution (Silver & Dunlap, 1987). Accordingly, the Fisher's r -to- z transformation was then applied to this value to prepare it for further computations as is usual for meta-analyses (Hedges & Olkin, 1985). This way, the skewed distribution is transformed into a normal distribution:

$$z_r = 1/2 \times \ln \times \left[\frac{1+r}{1-r} \right]$$

The resulting values were then weighted using the inverse variance weight of the coefficients as suggested by Graham and Christiansen (2009). This procedure takes account of different sample sizes in the various studies:

$$V_z = \frac{1}{n - 3}$$

The standard error can therefore be calculated from this variance using the following formula (Borenstein et al., 2009):

$$SE_z = \sqrt{V_z}$$

On the assumption that reliability estimates represent different populations, a random-effects model was preferred to a fixed-effects model for the meta-analysis (Higgins et al., 2009). Besides, Field and Gillett (2010) recommend a random-effect model when conducting meta-analyses in social sciences. Because many studies failed to report reliability coefficients, it is even more important to determine the influence of those missing data on the overall mean of Cronbach's alpha. As pointed out by Graham and Christiansen (2009), researchers intentionally do not report coefficients that are of too low a reliability, and work that has reported non-significant studies because of poorly reliable measuring instruments has often not been published. Such circumstances result in a publication bias which has been quantitatively expressed using the rank correlation based on Kendall's tau (τ ; Begg & Mazumdar, 1994). This approach quantifies the relationship between the ranks of effect sizes and the ranks of their variances. The lower a correlation is, the more effect sizes are independent of the sample sizes of the studies. Since Fisher's z transformation was used for meta-analytic calculation of the summary effect and the confidence intervals' limits, these values were converted back into correlations (Hafdahl & Williams, 2009) using the formula by Rosenstein et al. (2009):

$$r = \frac{e^{2z} - 1}{e^{2z} + 1}$$

In the final step, the r values were transformed to the metric of coefficient alpha (α) in order to facilitate interpretation.

Validity

In order to gain deeper insights into how the individual cognitive load types are interrelated and how these relate to relevant criterion variables (domain-specific prior knowledge, retention, and transfer), corresponding correlations were analyzed (e.g., Field, 2005). By this, correlations between the variables of interest could be meta-analyzed. As suggested by Glass et al. (1981), all effect sizes were retrieved in the form of Pearson's product-moment correlations, a standardized and prominent effect size. Since not all studies in our sample reported r values, conversion procedures had to be carried out. Following Gilpin (1993), the raw effect sizes of Spearman's rho (ρ) were transformed to Pearson's r . When the study reported standardized beta coefficients (β), the correlation coefficient (r) was estimated with the formula proposed by Peterson and Brown (2005). Hereby, the mathematical relationship between the two coefficients is shown in a multiple regression model with two predictor variables:

$$r_{y1} = \beta_1 + r_{12}(r_{y2} - \beta_1 r_{12})$$

Since bivariate correlations were calculated including one predictor, it can be assumed that $r = \beta$. Afterward, the r values were transformed into Fisher's z for calculating average correlations because Pearson's correlation coefficient cannot be interpreted as an interval-scaled measure (Silver & Dunlap, 1987). This procedure was also applied by several meta-analyses that have aggregated correlation coefficients (e.g., Capaldi et al., 2014; Edwards & Holtzman, 2017; Richardson et al., 2012). Hereby, the same procedure as for reliability was used (cf. Hedges & Olkin, 1985). A random-effects model was also calculated in this meta-analysis (Field & Gillett, 2010; Higgins et al., 2009). In addition, the rank correlation for publication bias was calculated (Begg & Mazumdar, 1994). The means and confidence intervals were then back-transformed in the correlation coefficient r to simplify interpretation. Therefore, the same formula proposed by Borenstein et al. (2009) was used. To interpret the correlation coefficients, normative effect size guidelines for individual differences researchers were followed (cf. Gignac & Szodorai, 2016). Accordingly, $r = 0.10$ is relatively small, $r = 0.20$ is typical, and $r = 0.30$ is relatively large. Since prior knowledge is an important factor influencing cognitive load facets (i.e., element interactivity; Chen & Kalyuga, 2020 or expertise reversal effect; Chen et al., 2017) and, consequently, correlations among cognitive load facets as well as facets and learning scores, detailed analyses with regard to prior knowledge can be found in Appendix I.

Results

Internal Consistency of Subjective Cognitive Load Questionnaires

Overall Internal Consistency

With regard to the internal consistency of the cognitive load questionnaires, the analyses showed satisfactory results. First, the reliability values of all four questionnaires were accumulated to examine whether the questionnaires are capable of measuring cognitive load consistently (see Table 2). Hereby, the ICL showed an alpha value of $\alpha + = 0.823$, while the internal consistency of the ECL was lowest ($\alpha + = 0.773$). The GCL showed the highest value ($\alpha + = 0.860$). Considering confidence intervals of the effect sizes, significant differences between all three cognitive load types concerning the internal consistency estimates could be found. All rank correlations (Begg & Mazumdar, 1994) were not significant indicating that a publication bias does not seem to be present for the internal consistency of the four investigated cognitive load questionnaires.

In order to gain deeper insights into whether the individual questionnaires can reliably measure cognitive load, the previous analysis was repeated separately for the four questionnaires (see Table 3). For the Leppink et al. (2013) questionnaire, the ICL showed an alpha value of $\alpha + = 0.845$, while an internal consistency of $\alpha + = 0.759$ was found for the ECL. Again, the GCL showed the highest alpha

Table 2 Aggregated effect sizes and confidence intervals for the internal consistency across the four cognitive load instruments

Measure	<i>k</i>	$\alpha+$	<i>p</i>	95% CI		Rank correlation	
				Lb	Ub	Kendall's τ	<i>p</i>
ICL	55	.823	<.001	.798	.845	0.103	.269
ECL	73	.773	<.001	.744	.800	−0.100	.214
GCL	42	.860	<.001	.828	.886	−0.188	.081

k=number of studies (or reliability coefficients); $\alpha+$ =accumulated Cronbach's alpha across studies; Lb and Ub=lower and upper bounds, respectively, of the 95% confidence interval around the overall reliability estimate; rank correlation=test for publication bias. For the Eysink et al. (2009) questionnaire, only the internal consistency for the ECL could be calculated, as it is measured with multiple items. ICL and GCL are measured with a single item

Table 3 Aggregated effect sizes and confidence intervals for the internal consistency of the cognitive load instruments

Measure	<i>k</i>	$\alpha+$	<i>p</i>	95% CI		Rank correlation	
				Lb	Ub	Kendall's τ	<i>p</i>
Leppink et al. (2013)							
ICL	28	.845	<.001	.818	.869	−0.074	.580
ECL	30	.759	<.001	.701	.807	−0.063	.630
GCL	24	.909	<.001	.887	.927	−0.201	.172
Klepsch et al. (2017)							
ICL	21	.776	<.001	.745	.804	0.291	.072
ECL	25	.798	<.001	.764	.829	−0.068	.638
GCL	14	.734	<.001	.686	.775	0.122	.546
Leppink et al. (2014)							
ICL	6	.851	<.001	.679	.935	>0.001	>.999
ECL	8	.788	<.001	.677	.866	−0.109	.708
GCL	4	.806	<.001	.775	.833	0.333	.750
Eysink et al. (2009)							
ECL	10	.740	<.001	.659	.804	−0.200	.484

k=number of studies (or reliability coefficients); $\alpha+$ =accumulated Cronbach's alpha across studies; Lb and Ub=lower and upper bounds, respectively, of the 95% confidence interval around the overall reliability estimate; rank correlation=test for publication bias. For the Eysink et al. (2009) questionnaire, only the internal consistency for the ECL could be calculated, as it is measured with multiple items. ICL and GCL are measured with a single item

value ($\alpha+ = 0.909$). Based on the effect sizes and the confidence intervals, it could be derived that all three cognitive load types differ significantly in their internal consistency. The questionnaire from Klepsch et al. (2017) also produced satisfactory Cronbach's alpha values for the individual cognitive load types. The internal consistency of the ICL amounts to a value of $\alpha+ = 0.776$ and $\alpha+ = 0.798$ for the ECL. The GCL showed a value of $\alpha+ = 0.734$. Consequently, the highest internal

consistency was found for the ECL. In addition, this value was significantly higher than the value of the GCL. The Cronbach's alpha of the ICL also differed significantly from the GCL. ICL and ECL did not differ significantly from each other. The cognitive load questionnaire by Leppink et al. (2014) joins the ranks of questionnaires with good internal consistency. Although comparatively few effect sizes were included in the analysis, the Cronbach's alpha values for ICL ($\alpha + = 0.851$), ECL ($\alpha + = 0.788$), and GCL ($\alpha + = 0.806$) can be considered satisfactory in terms of commonly used benchmarks (e.g., Nunnally, 1978). The three cognitive load types did not differ significantly from each other. Because the questionnaire from Eysink et al. (2009) only measures the ECL with multiple items, the internal consistency could only be calculated for this cognitive load type. Hereby, the ECL showed a satisfactory internal consistency across studies ($\alpha + = 0.740$). The non-significant rank correlations across all examined cognitive load scales seems to indicate that there is no publication bias (Begg & Mazumdar, 1994).

Internal Consistency — Moderating Variables

To investigate the influence of additional variables on the reliability of cognitive load questionnaires, moderator analyses were carried out. The results of the moderator analyses across all questionnaires are displayed in Table 4. According to the confidence intervals of the effect sizes, learners' age and educational background, the domain of the material, and the presentation mode were not significant moderators of the reliability of the ICL subscale. Only the number of scale points resulted in significant differences. According to the effect sizes, a 10-point Likert scale was most reliable. A 7-point Likert scale at least should be used to ensure stronger reliability. Concerning ECL, learners' age and educational background, as well as the presentation mode were not significant moderators. With respect to the domain of the learning material, the highest reliability was achieved in the field of mathematics and logic; reliability did not differ with regard to the other learning domains. Regarding the number of scale points, an odd number (i.e., 5 or 9 response options) resulted in higher reliabilities than an even number of scale points. Considering the GCL, learners' age and educational background, as well as the domain of the learning material were not significant moderators. Regarding the presentation mode, reliability was reduced when interactive learning media were used. Regarding GCL, the use of an 11-point Likert scale was associated with the highest reliability, whereas the use of a 7-point Likert scale led to the lowest reliability.

Validity of Subjective Cognitive Load Questionnaires

Construct Validity

To examine the construct validity of the subjective cognitive load questionnaires, the generally accepted definition was followed proposing an ideally high level of agreement between the measurement results and the underlying theoretical assumptions (Cronbach & Meehl, 1955; Westen & Rosenthal, 2003). In consequence, construct

Table 4 Aggregated effect sizes and confidence intervals for the internal consistency separated in terms of moderating variables

Measure	<i>k</i>	<i>α</i>	<i>p</i>	95% CI	
				Lb	Ub
ICL					
Educational setting					
School education	18	.826	<.001	.778	.865
Adult education	37	.821	<.001	.791	.848
Domain of the instructional material					
Natural sciences	43	.820	<.001	.788	.847
Social sciences	3	.827	<.001	.707	.901
Logic and mathematics	7	.840	<.001	.812	.865
Others	2	.832	<.001	.776	.875
Presentation mode					
Static	35	.813	<.001	.781	.842
Dynamic	7	.828	<.001	.776	.869
Interactive	6	.842	<.001	.703	.919
Mixed	7	.846	<.001	.787	.891
Number of response options					
5	6	.771	<.001	.733	.805
6	2	.712	<.001	.613	.790
7	15	.783	<.001	.721	.834
10	6	.881	<.001	.791	.933
11	19	.839	<.001	.802	.869
ECL					
Educational setting					
School education	24	.744	<.001	.672	.802
Adult education	49	.787	<.001	.758	.814
Domain of the instructional material					
Natural sciences	56	.771	<.001	.736	.801
Social sciences	6	.695	<.001	.598	.773
Logic and mathematics	9	.843	<.001	.782	.888
Others	2	.705	<.001	.616	.776
Presentation mode					
Static	46	.795	<.001	.761	.825
Dynamic	10	.770	<.001	.722	.812
Interactive	9	.749	<.001	.645	.826
Mixed	8	.644	<.001	.503	.754
Number of response options					
5	6	.844	<.001	.791	.884
6	3	.602	<.001	.403	.752
7	20	.766	<.001	.730	.798
9	6	.829	<.001	.743	.889
10	7	.720	<.001	.679	.757
11	18	.769	<.001	.689	.831

Table 4 (continued)

Measure	<i>k</i>	α	<i>p</i>	95% CI	
				Lb	Ub
GCL					
Educational setting					
School education	14	.880	<.001	.823	.919
Adult education	28	.848	<.001	.808	.881
Domain of the instructional material					
Natural sciences	30	.872	<.001	.834	.901
Social sciences	5	.869	<.001	.797	.916
Logic and mathematics	5	.795	<.001	.653	.883
Other	2	.762	<.001	.687	.822
Presentation mode					
Static	20	.866	<.001	.825	.898
Dynamic	9	.884	<.001	.810	.930
Interactive	6	.774	<.001	.613	.874
Mixed	7	.869	<.001	.790	.919
Number of response options					
6	2	.816	<.001	.682	.897
7	12	.732	<.001	.676	.781
9	1	.838	<.001	.768	.888
10	5	.854	<.001	.756	.914
11	16	.915	<.001	.889	.935

k=number of studies (or reliability coefficients); α =accumulated Cronbach's alpha across moderating variables; Lb and Ub=lower and upper bounds, respectively, of the 95% confidence interval around the overall reliability estimate. For the Eysink et al. (2009) questionnaire, only the internal consistency for the ECL could be calculated, as it is measured with multiple items. ICL and GCL are measured with a single item

validity was evaluated by considering the correlations between the cognitive load types of the investigated questionnaires. Regarding all questionnaires, meta-analytic correlations between the cognitive load types are displayed in Table 5. For the questionnaire by Leppink et al. (2013), a positive correlation between ICL and ECL and a negative correlation between ECL and GCL were found. ICL and GCL did not correlate with each other. When examining the questionnaire by Klepsch et al. (2017), a positive correlation between ICL and ECL as well as between ICL and GCL was found. ECL and GCL were negatively correlated with each other. Regarding the questionnaire by Leppink et al. (2014), a positive correlation between ICL and ECL was found. ECL and GCL as well as ICL and GCL were correlated negatively with each other. Concerning the questionnaire by Eysink et al. (2009), all types were positively correlated with each other. It should be mentioned that the number of included effect sizes from the questionnaires by Leppink et al. (2014) and Eysink et al. (2009) was relatively small. Across all four questionnaires, positive

Table 5 Construct validity of the cognitive load instruments across studies

Measure	<i>k</i>	<i>r</i> +	<i>p</i>	95% CI		Rank correlation	
				Lb	Ub	Kendall's τ	<i>p</i>
Leppink et al. (2013)							
ICL–ECL	25	.298	< .001	.221	.370	0.148	.304
ECL–GCL	20	–.186	.018	–.331	–.033	0.048	.770
ICL–GCL	19	–.086	.155	–.202	.033	0.271	.107
Klepsch et al. (2017)							
ICL–ECL	16	.403	< .001	.281	.512	–0.271	.155
ECL–GCL	10	–.135	.020	–.246	–.021	0.225	.369
ICL–GCL	10	.243	.002	.091	.383	–0.045	.857
Leppink et al. (2014)							
ICL–ECL	5	.350	< .001	.210	.477	0.738	.077
ECL–GCL	2	–.331	.006	–.529	–.099	1.000	> .999
ICL–GCL	2	–.246	.045	–.459	–.006	1.000	> .999
Eysink et al. (2009)							
ICL–ECL	9	.533	< .001	.375	.662	–0.222	.477
ECL–GCL	9	.597	< .001	.471	.700	–0.444	.119
ICL–GCL	9	.512	< .001	.377	.625	–0.278	.358

k=number of studies (or experiments); Lb and Ub=lower and upper bounds, respectively, of the 95% confidence interval around the overall correlation estimate; rank correlation=test for publication bias. All correlations are given in the Pearson's product moment correlation metric (*r*+

correlations between the constructs ICL and ECL were found. The largest correlation was found for the questionnaire by Eysink et al., (2009; *r* + = 0.53). Except for the questionnaire by Eysink et al. (2009), all questionnaires showed negative correlations between ECL and GCL. The largest negative correlation was found for the questionnaire by Leppink et al., (2014; *r* + = –0.33). The correlations between ICL and GCL were ambiguous across the investigated questionnaires.

Criterion Validity

Considering criterion validity, relevant variables in cognitive load research were included in order to investigate interrelationships of the cognitive load types with practically relevant external criteria (i.e., the domain-specific prior knowledge as well as learning outcomes of retention and transfer; Drost, 2011). Meta-analytic correlations between learning scales and cognitive load types of all questionnaires are displayed in Table 6. Meta-analytic correlations between domain-specific prior knowledge and cognitive load types of all questionnaires are displayed in Table 7. At first, the criterion validity of the questionnaire by Leppink et al. (2013) was investigated. ECL negatively correlated with both learning scales and GCL positively correlated with retention and transfer. ICL negatively correlated with transfer but not with retention performance. Furthermore, ICL and ECL negatively correlated with domain-specific prior knowledge. GCL did not correlate with prior knowledge.

Table 6 Criterion validity of the cognitive load instruments with learning performances retention and transfer across studies

Measure	<i>k</i>	<i>r</i> +	<i>p</i>	95% CI		Rank correlation	
				Lb	Ub	Kendall's τ	<i>p</i>
Leppink et al. (2013)							
ICL — retention	20	-.061	.137	-.140	.019	0.500	.002
ECL — retention	22	-.114	.010	-.198	-.027	0.004	.977
GCL — retention	15	.188	<.001	.112	.263	0.117	.550
ICL — transfer	16	-.188	<.001	-.254	-.112	0.127	.498
ECL — transfer	16	-.120	.013	-.213	-.025	-0.312	.095
GCL — transfer	12	.173	<.001	.111	.234	-0.339	130
Klepsch et al. (2017)							
ICL — retention	13	-.195	<.001	-.243	-.158	0.439	.044
ECL — retention	16	-.205	<.001	-.251	-.159	-0.149	.436
GCL — retention	7	.219	<.001	.130	.306	0.238	.562
ICL — transfer	12	-.138	.018	-.249	-.024	0.263	.253
ECL — transfer	12	-.261	<.001	-.308	-.213	0.164	.475
GCL — transfer	6	.217	<.001	.115	.316	-0.067	>.999
Leppink et al. (2014)							
ICL — retention	8	-.182	.088	-.376	.027	-0.206	.503
ECL — retention	9	-.172	.103	-.365	.035	-0.155	.582
GCL — retention	2	-.008	.966	-.359	.345	1.000	>.999
ICL — transfer	5	-.247	.010	-.416	-.061	-0.447	.296
ECL — transfer	6	-.179	.025	-.327	-.023	0.358	.330
GCL — transfer	1	-	-	-	-	-	-
Eysink et al. (2009)							
ICL — retention	9	-.232	<.001	-.312	-.149	0.389	.180
ECL — retention	10	-.223	<.001	-.295	-.150	0.689	.005
GCL — retention	9	-.223	<.001	-.294	-.150	0.611	.025
ICL — transfer	9	-.164	.007	-.278	-.045	0.611	.025
ECL — transfer	9	-.119	.067	-.245	.009	0.556	.045
GCL — transfer	9	-.182	<.001	-.265	-.095	0.333	.260

k=number of studies (or experiments); Lb and Ub=lower and upper bounds, respectively, of the 95% confidence interval around the overall correlation estimate; rank correlation=test for publication bias. All correlations are given in the Pearson's product moment correlation metric (*r*+

Second, the criterion validity of the questionnaire by Klepsch et al. (2017) was investigated. ICL and ECL negatively correlated with both learning scales and GCL positively correlated with retention and transfer. Furthermore, ICL and ECL negatively correlated with prior knowledge. GCL did not correlate with prior knowledge. Third, the criterion validity of the questionnaire by Leppink et al. (2014) was investigated. ICL negatively correlated with both learning scales. ECL negatively correlated with transfer but not retention performance. GCL did not correlate with retention. Meta-analytic correlation between GCL and transfer could not be conducted

Table 7 Criterion validity of the cognitive load instruments with prior knowledge

Measure	<i>k</i>	<i>r</i> +	<i>p</i>	95% CI		Rank correlation	
				Lb	Ub	Kendall's τ	<i>p</i>
Leppink et al. (2013)							
ICL — prior knowledge	20	-.109	< .001	-.165	-.052	-0.048	.770
ECL — prior knowledge	21	-.093	.005	-.158	-.028	-0.072	.650
GCL — prior knowledge	15	.073	.081	-.009	.153	0.059	.765
Klepsch et al. (2017)							
ICL — prior knowledge	15	-.179	< .001	-.270	-.086	0.170	.390
ECL — prior knowledge	15	-.164	< .001	-.212	-.117	-0.070	.724
GCL — prior knowledge	9	.014	.756	-.076	.105	0.222	.477
Leppink et al. (2014)							
ICL — prior knowledge	3	-.185	.165	-.423	.077	0.816	.221
ECL — prior knowledge	4	-.036	.676	-.204	.133	0.183	.718
GCL — prior knowledge	1	-	-	-	-	-	-
Eysink et al. (2009)							
ICL — prior knowledge	6	-.141	.215	-.350	.082	0.600	.136
ECL — prior knowledge	6	-.162	.116	-.350	.040	0.200	.719
GCL — prior knowledge	6	-.193	.024	-.349	-.025	0.467	.272

k=number of studies (or experiments); Lb and Ub=lower and upper bounds, respectively, of the 95% confidence interval around the overall correlation estimate; rank correlation=test for publication bias. All correlations are given in the Pearson's product moment correlation metric

because of the lack of data ($k = 1$). Furthermore, ICL and ECL did not correlate with prior knowledge. Meta-analytic correlation between GCL and prior knowledge could not be conducted because of the lack of data ($k = 1$). Overall, the number of studies that used the questionnaire from Leppink et al (2014) was very small. Thus, interpretation of the results is restricted. Finally, the criterion validity of the questionnaire by Eysink et al. (2009) was investigated. All cognitive load types negatively correlated with retention and transfer. Furthermore, GCL negatively correlated with prior knowledge. In addition, prior knowledge did not correlate with ICL and ECL. Summarizing the results, significant negative correlations between the sub-facet ICL and both learning scores occurred across all four questionnaires. Regarding retention, the largest correlation was found for the questionnaire by Eysink et al., (2009; $r+ = -0.23$). Regarding transfer, the largest correlation was found for the questionnaire by Leppink et al., (2014; $r+ = -0.25$). Furthermore, across all questionnaires, significant negative correlations between the sub-facet ECL and both learning scores occurred. Regarding retention, the largest correlation was found for the questionnaire by Eysink et al., (2009; $r+ = -0.22$). Regarding transfer, the largest correlation was found for the questionnaire by Leppink et al., (2014; $r+ = -0.18$). These results are in line with the theoretical implications derived from the CLT but effect sizes were rather moderate (cf. Gignac & Szodorai, 2016). Correlations between GCL and learning scores as well as cognitive load types and prior knowledge were rather ambiguous across the questionnaires. Consistent with predictions based on

CLT, positive correlations between GCL and learning scores and negative correlations between prior knowledge and ICL as well as ECL could be observed regarding the questionnaires by Leppink et al. (2013) and Klepsch et al. (2017). Missing correlations between prior knowledge and ICL as well as ECL and negative correlations between GCL and prior knowledge as well as learning scales could be observed regarding the questionnaire by Eysink et al. (2009).

Discussion

The aim of this meta-analysis was the investigation of subjective cognitive load questionnaires in terms of their validity and reliability in experimental multimedia learning settings. In the following, the results of these analyses will be discussed with regard to whether they comply with the assumptions of CLT.

Internal Consistency of the Examined Cognitive Load Scales

The cognitive load questionnaires from Klepsch et al. (2017) and from Leppink et al., (2013, 2014) showed satisfactory results indicating that the respective items for ICL, ECL, and GCL seem to have a high internal consistency. Hence, the questionnaires can all be recommended to measure the different cognitive load types because a high reliability can be observed although usually only three to five items per construct are used. All values are higher than 0.70 and consequently can be considered satisfactory. Moreover, it can be concluded that no items of the scales are redundant as the alpha value 0.90 did not occur for any cognitive load type (Tavakol & Dennick, 2011a). In addition, the satisfactory values of the cognitive load scales are associated with a small measurement error (Tavakol & Dennick, 2011b). For instance, the alpha value ($\alpha = 0.827$) for the ICL involves an error variance of 0.316 ($0.827 \times 0.827 = 0.684$; $1.00 - 0.684 = 0.316$). The alpha values of the ECL scales ($\alpha = 0.773$) as well of the GCL scales ($\alpha = 0.860$) also reaffirm the use of the scales in experimental multimedia learning research. However, it must be noted that for the Eysink et al. (2009) questionnaire, only the internal consistency for the ECL could be calculated, as the ICL and GCL are measured with single items. Besides the lack of possibility to calculate the internal consistency, it seems insufficient from the point of view of measurement theory to measure complex constructs like intrinsic and germane load with one single item.

Validity of the Examined Cognitive Load Scales

With respect to the validity of the cognitive load questionnaires, this work examined both construct validity and criterion validity (Cronbach & Meehl, 1955; Drost, 2011; Westen & Rosenthal, 2003). In this way, connections of the cognitive load types among each other can be calculated on the one hand (construct validity) and connections with theory-related concepts (i.e., domain-specific prior knowledge, retention, and transfer) on the other (criterion validity). The following conclusions

have to be interpreted under the assumption that the types of cognitive load can change dynamically, in particular when the working memory's capacity is exceeded. Furthermore, cognitive load is a latent, not directly observable construct, whereby theoretical assumptions cannot be transferred congruently in the real world.

Across all four cognitive load questionnaires, positive correlations between the ICL and ECL were found. Hence, the positive correlation supported by relatively large effect sizes between these two cognitive load types seems inconsistent with the additivity hypothesis of the CLT. While the ICL refers to the tasks' inherent complexity, the ECL is determined how the information is presented and formatted (Sweller et al., 2019). Therefore, the ICL and ECL should describe different aspects of the learning material. However, apparently, ICL and ECL cannot be completely separated from each other by the learner. When learners perceive a learning material to be high in ECL because of a poor presentation of information, they might also perceive the learning material to be more complex. In addition, it seems plausible that a complex learning content can also only be represented with a rather complex design. The positive correlation between the ICL and ECL thus tends to contradict the theoretically stated additivity hypothesis. On the one hand, this could be based on the already mentioned problem that CLT assumptions may differ from subjective evaluations. Otherwise, it could be a measurement problem, so that item formulations do not accurately reflect the types of cognitive load. This could be addressed by two possibilities. First, developing and validating a new cognitive load questionnaire could help to overcome the missing differentiation. Second, referring to Klepsch et al. (2017), learners could be informed before the experiment about which aspects of the learning material are described by the ICL and ECL and how to distinguish between them. Based on this "meta"-knowledge, more accurate judgments about perceived cognitive load could be made. In a similar vein, Zu et al. (2021) found empirical evidence that learners' ability to distinguish between the ICL and ECL depends on their domain-specific prior knowledge. In summary, the construct validity of the questionnaires examined may be relatively limited when considering underlying theoretical assumptions. Particularly salient are correlations between the ICL and ECL, which suggest that the two constructs cannot be grasped separately by learners. In this vein, the available cognitive load questionnaires seem to lack sufficient discriminant validity because ICL and ECL showed relatively high correlations. Tendency, the two, different labeled measures, ICL and ECL, assess a same construct (extrinsic convergent validity; Gonzalez et al., 2021). The construct validity of the questionnaires by Klepsch et al. (2017) as well as by Leppink et al., (2013, 2014) revealed a negative correlation between the cognitive load types ECL and GCL. In line with our understanding of cognitive load, a higher ECL could be accompanied by a lower GCL because cognitive resources are wasted to compensate for the sub-optimal design or presentation of the learning material (Sweller et al., 2019). This connection could be also based on motivational influences suggesting that unfavorably designed learning materials could lower an individual's motivation to learn (Feldon et al., 2019). The more that learners are motivated, the more germane (or learning-relevant) resources are invested by the learner to master the task. In this vein, there is empirical evidence that motivated learners reported higher levels of GCL (Cook et al., 2017; Costley & Lange, 2018). One could consequently

argue that higher ECL perceptions are related to lower investments of mental effort indicating that passive load affects active load (Klepsch & Seufert, 2021). However, these conclusions are limited by the fact that the questionnaire by Eysink et al. (2009) found a positive correlation between the ECL and GCL limiting the construct validity of this questionnaire. In recent years, the three-factor model of cognitive load has been widely discussed in view of its theoretical justifiability (Kalyuga, 2011; Sweller, 2010; Sweller et al., 2011, 2019). The results of this meta-analysis can add momentum to this discussion. Theoretical assumptions suggest that both the ICL and GCL have the same theoretical basis so that the notion of the GCL might be redundant (Kalyuga, 2011). Consequently, both variables should correlate with each other as the GCL distributes cognitive resources to handle the tasks element interactivity (described as ICL). However, the questionnaire from Leppink et al. (2013) found a non-significant correlation between these cognitive load types indicating that they are not statistically connected which seems to support the additivity hypothesis. This is another point that tends to support the three-factor model (e.g., as in the factor analysis by Zavgorodniaia et al., 2020). The Klepsch et al. (2017) as well as Eysink et al. (2009) questionnaires showed significant but rather small positive correlations between ICL and GCL. Even if correlations do not allow causal statements of course, it can, at least, be hypothesized that both variables measure two related but no uniform construct. In general, the questionnaire from Eysink et al. (2009) makes it difficult for the learner to differentiate between the cognitive load types as the item formulations dissolve the theoretical boundaries of the CLT. Thus, all items include the term “easy or difficult” which rather evokes associations with the ICL (task difficulty) and thus seems to be insufficient for the evaluation of the ECL and GCL.

Concerning criterion validity, correlations with theory-related concepts (i.e., domain-specific prior knowledge, retention, and transfer) revealed interesting insights. First, across all four cognitive load questionnaires, negative correlations were found between the ICL and the learning outcomes in both retention and transfer. However, the correlation between ICL and retention failed to reach significance in the Leppink et al. (2013) questionnaire. It can be assumed that higher ICL perceptions are related to worse learning outcomes. In the light of CLT, this unsurprising finding can be explained by the task complexity (Ayres, 2006; Chen & Kalyuga, 2020). The lower the learner estimates the task’s inherent complexity (refers to element interactivity), the better the result achieved in the learning test. This effect can probably be explained by domain-specific prior knowledge (Park et al., 2015). Learners with expertise in the domain relevant for the learning material can rely on previously generated (automated) schemata including interacting elements which help them to deal with the task’s complexity (Paas et al., 2003; Sweller et al., 2011). In contrast, learners with low prior knowledge have not acquired schemata, so that each element needs to be processed separately while learning. In terms of ECL, negative correlations between this cognitive load type and learning outcomes in retention and transfer were found. As correlations between ECL and retention (Leppink et al., 2014) as well as between the ECL and transfer (Eysink et al., 2009) failed to reach significance, the following conclusion must be made with some caution. In general, the results regarding the criterion validity support theoretical assumptions

of the CLT (Sweller et al., 2019). Therefore, the negative relationship between the ECL and learning outcomes indicates that learners who perceive the learning material as unfavorably designed for learning (causing higher ECL ratings) perform worse in the learning test. In this case, cognitive resources needed to cope with the complexity of the task are wasted for processing the poorly designed instructions (Klepsch & Seufert, 2020). This is in line with the common recommendation to reduce ECL while learning in order to enhance learning outcomes (Beckmann, 2010; Leppink & Heuvel, 2015). Reducing ECL can free cognitive resources to deal with the task's inherent element interactivity (Paas et al., 2003). In terms of GCL, positive correlations with the learning outcomes retention and transfer are also explainable based on CLT's tenets. However, correlations between GCL and the learning indicators retention and transfer were non-significant for the Leppink et al. (2014) questionnaire. In general, the positive correlations (questionnaires by Klepsch et al., 2017; Leppink et al., 2013) indicated that higher GCL perceptions go hand in hand with higher learning outcomes. As GCL is held to arise from learning-relevant activities such as taking notes or remembering previously acquired knowledge, this cognitive load can be seen as supporting successful learning. In line with Paas and van Gog (2006), a high GCL indicates that learners are engaged to learn and direct their mental resources to learning-relevant activities. Increasing GCL is thus a central challenge within CLT (Klepsch & Seufert, 2020; Moreno & Park, 2010; Paas & van Gog, 2006) — what is underlined by the results of this work. Although the positive correlation is a logical consequence of the theoretical assumptions of the CLT, the relatively low level of correlations is surprising. Accordingly, a higher correlation is to be expected, because learners who report a high GCL should also have a comparatively high learning gain. The small correlations indicate a gap between the learners' subjective evaluated GCL and their actual objective result in the learning test. Assuming that there is strong evidence for meta-cognitive beliefs and learning outcomes being related (e.g., Al Khatib, 2010; Nelson & Dunlosky, 1991; Sungur, 2007), cognitive load questionnaires should be able to better capture the relationship between the GCL and learning achievements. In contrast, the negative correlation between the GCL and learning outcomes, reported in the Eysink et al. (2009) questionnaires, gives a further indication that this questionnaire does not adequately measure the GCL due to unfavorably formulated items. It is important to add here that it is also possible for a learner to invest a high level of GCL, which ultimately does not pay off, so not much learning could be done. This could happen particularly if the ICL (i.e., the complexity of the learning material) is too high and/or the learner has too domain-specific little prior knowledge. Thus, a relatively high GCL is not necessarily associated with better learning performance. In this vein, motivational beliefs should not be neglected, but are, according to Feldon et al. (2019), a result of the instruction and could affect the GCL and related concepts such as mental effort.

Regarding domain-specific prior knowledge, negative correlations between the ICL and prior knowledge occurred in the questionnaires from Leppink et al. (2013) and Klepsch et al. (2017). In line with our current understanding of cognitive load, the learner's expertise affects ICL perceptions (Artino, 2008; Bannert, 2002). Learners with high expertise can draw on schemata during learning, which help to cope

with the complexity of the task (Kirschner et al., 2009; Leppink & Heuvel, 2015). In this vein, it could also be possible that learners reporting a high ICL have less domain-specific prior knowledge. Similar results were found between ECL and prior knowledge. When learners can rely on domain-specific prior knowledge, they perceive a lower ECL. Accordingly, learners are less susceptible to poorly formatted and designed learning materials, when enough expertise is available for learning. This is based on the additive relationship between ICL and ECL. Counterintuitively, non-significant correlations between the GCL and prior knowledge were found. In view of CLT, it could be assumed that learners with a certain level of prior knowledge are better able to allocate their cognitive resources to learning-relevant activities (Paas & van Gog, 2006; Paas et al., 2003).

Recommendations for Further Use of Cognitive Load Scales in Experimental Research

Subjective questionnaires play an important role in experimental cognitive load research, as they help us to better understand cognitive processes during learning. These findings can help to further improve learning materials or procedures. Therefore, cognitive load questionnaires should meet the highest psychometric requirements (Embretson, 2013). On one hand, the reliability analyses showed satisfactory Cronbach's alpha values justifying the use of these scales in experimental settings. On the other hand, based on the moderator analyses, recommendations can be made that should be considered in the future when using cognitive load questionnaires. In terms of the number of scale points, moderator analyses suggest that at least a 7-point scale could be used to ensure high reliability when ICL is to be measured. A scale with 10 response options was associated with the significantly highest reliability when measuring the ICL. The ECL could be measured with 5-point or 9-point scales to ensure high reliability. Scales with an even number of response numbers (i.e., 6 or 10 response numbers) were associated with lower reliability and should therefore not be used, which counteracts findings from Dalal et al. (2013). In terms of GCL, using a 11-point scale was associated with the highest reliability, but all numbers of scale points but a 7-point scale resulted in a high reliability. From a pragmatic perspective, however, researchers are likely to measure ICL, ECL, and GCL with the same number of scale points. This also makes it easier for the participant to understand the scale. Taken together, moderator analyses support using a 9-point scale. With respect to the domain of the instructional material and the presentation mode, the internal consistency of the ICL showed only slight differences indicating that this scale can be used in various learning settings. Furthermore, the ECL showed the highest reliability when the experimental studies took place in the instructional domain of logic and mathematics. Possibly, the perception of the presentation and format of the learning material is particularly sensitive when dealing with complex learning topics. In interactive learning environments, the GCL showed lower reliability indicating that learners are less able to monitor and report on their learning process. Across all cognitive load types, the absence of notable differences between school education and adult education seems to suggest that the

questionnaires can reliably measure cognitive load over a wide age range. However, researchers should ensure, prior to the experiment, that learners can understand the item formulations and can thus make suitable meta-cognitive judgments. In this vein, Leahy (2018) warns against using subjective cognitive load questionnaires with children. However, Wang, Ardasheva, et al. (2021a), Wang, Ginns, et al. (2021b)) were successful in using a cognitive load questionnaire with 10- to 12-year participants which outlines the need for future research. The literature review uncovered noticeable differences in the descriptions of the cognitive load scale. For example, there is often a lack of information on the number of levels of the scale, the labels of the scale points, or even reliability. However, because these points are quite essential to ensure the fit of the scale to the experimental purpose, researchers are encouraged to specify as precisely as possible which scales of cognitive load were used including the number of options (e.g., 10-point scale) and to report reliability values such as Cronbach's alpha (Cronbach, 1951), McDonald's omega (McDonald, 1999), or Revelle's omega (Revelle & Zinbarg, 2009). However, researchers should be aware of the ongoing debate concerning methodological weaknesses of Cronbach's alpha (Christmann & Aelst, 2006; Hayes & Coutts, 2020; McNeish, 2018; Panayides, 2013; Sijtsma, 2009; Taber, 2018) and should critically evaluate the reliability values also with regard to the construct to be measured. In this vein, Deng and Chan (2017) emphasize that Cronbach's alpha tends to misestimate true reliability unless τ -equivalent items are involved.

It is particularly noticeable that researchers tend to interchange the cognitive load questionnaires from Leppink and colleagues (2013; 2014). This circumstance is probably due to the similarity of the two questionnaires, which, however, differ significantly because of the additional mental effort items and the resulting higher number of items included in the questionnaire published in 2014, which, on the one hand, had an aggravating effect on the data synthesis for this work and, on the other hand, complicates the interpretation of the results in the respective primary study. Moreover, articles were found reporting the reliability for the three cognitive load types together, which is not coherent from a theoretical point of view. Although all three types measure the burden on the working memory (or at least the ECL and ICL), they concentrate on different aspects of the learning material. In addition, the individual cognitive load types can vary in terms of their reliability — if the alpha value of the entire scale is given and this is poor, the results of this measurement should not be used further for a generalization. If the reliability is calculated individually, this cognitive load type with a poor Cronbach's alpha could be removed, whereas categories with a more satisfactory value can be included for interpretation.

Limitations and Future Directions

Although the present work is the first of its kind in the history of CLT research, some remarks must be made that may partially limit the results but should also encourage researchers to follow up on this work. The Cronbach's alpha values of the various cognitive load questionnaires examined are satisfactory but should nonetheless be considered with caution. It is a more or less unwritten rule that very poor Cronbach's

alpha values might lead to studies not being published by (highly ranked) journals. In this vein, it is also possible that studies reporting non-satisfactory values for internal consistency are not even submitted by the authors and, with the subsequent “file-drawer problem” distorting the actual values seen in the published literature. Accordingly, there is evidence to assume a bias towards too high values. Moreover, the cognitive load questionnaires differed with respect to the number of items used. As Cronbach’s alpha is sensitive to the number of items in the scales (e.g., Tavakol & Dennick, 2011a, b; Vaske et al., 2017), comparing questionnaires with different numbers of items makes interpretation difficult. Furthermore, responses in Likert scales can lead to inflated inter-item correlations (i.e., estimates of internal consistency). These dependencies arise above all when people classify similar statements with similar values using a Likert scale. Positive correlations between the items thus lead to increased reliability which can be at the expense of the measurement’s validity (Eisinga et al., 2013). Unclear items or items that do not appear meaningful to the learner reinforce this tendency (Schuman et al., 1981).

The basis of these meta-analyses was primary studies from the field of multimedia learning. Consequently, the studies differ in terms of information presentation, multimedia design, as well as the learning topic covered what may influence cognitive load perception. To account for these influences, moderator analyses for reliability were calculated. However, problems concerning the heterogeneity of the primary studies cannot be fully ruled out (e.g., Thompson, 1994). Thus, slight changes in the multimedia design (and the learning content which has to be learned) could also have led to changes in the answers in the Likert scale and therefore cognitive load perceptions. However, not all possible influences can be calculated, which is a common problem in meta-analyses (Borenstein et al., 2021).

Probably the most important conclusion of this meta-analysis is that there is still a lot of research to be done in the field of CLT measurement. Particularly striking here was the age of the participants which ranged (when considering the standard deviation) between 16 and 24 years. Consequently, the majority of them were affiliated with a secondary school or university. Assuming that learners with increasing age are better able to estimate their perceived cognitive load (e.g., van der Stel & Veenman, 2010), more research with under-represented demographics is needed. For instance, developing reliable and valid cognitive load questionnaires for younger target groups (i.e., elementary school students) could be fruitful (Leahy, 2018). In general, the possibility to use questionnaires in different languages (i.e., test adaptation; Ercikan & Lyons-Thomas, 2013) is a desirable goal but associated with challenges. However, adapting questionnaires to different cultures is a difficult task (Hambleton & Patsula, 1998) requiring knowledge of the cultural as well as linguistic circumstances. In terms of the cognitive load questionnaires, the original versions of the scales are available in English (Eysink et al., 2009; Leppink et al., 2013, 2014) or bilingual in German and English (Klepsch et al., 2017). However, many studies in cognitive load research have been conducted in European or Asian countries with other mother tongues. It can be assumed that the original questionnaires were translated and re-interpreted. This might limit the results of this meta-analysis because it may also affect reliability and validity. Thus, researchers should stick to questionnaire translation rules since even small changes in the item formulation can affect

the respondents' understanding (Harkness et al., 2004). However, studies involved in this meta-analysis often fail to indicate if questionnaires were translated. In sum, it would be desirable to make translations accessible to the research community.

Conclusion

Over the years, CLT has become a major theory in educational psychology research. Results of this meta-analysis revealed that cognitive load during learning can be reliably measured with currently available subjective questionnaires. In contrast, significant correlations between cognitive load types might question the construct validity of the cognitive load questionnaires and/or the additivity hypotheses postulated by the CLT. Results of correlations among cognitive load types with relevant criterion variables tend to support the three-factor model of cognitive load comprising ICL, ECL, and GCL. Overall, multimedia learning researchers should be encouraged to use cognitive load questionnaires in their research while being aware of the concrete designation of the scale, the number of response options, and the correct indication of the scale's reliability.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s10648-022-09683-4>.

Acknowledgements We want to sincerely thank Jessica Gröninger for the assistance with data collection and data preparation.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Ethics Approval This article does not contain any studies with human participants performed by any of the authors. It has only meta-analyzed studies already conducted. Therefore, no ethical approval from an Ethics Committee was required.

Informed Consent In this work, a meta-analysis was performed and reported. For this purpose, no experimental study was conducted by the authors. Therefore, no informed consent was required for participation and publication.

Conflict of Interest The author Paul Ginns is member of the Editorial Board of the Educational Psychology Review. Besides that, the authors have no competing interests to declare that are relevant to the content of this article.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

References marked with an asterisk (*) indicate studies included in the meta-analysis.

- Adams, H. F. (1936). Validity, reliability, and objectivity. In W. R. Miles (Ed.), *Psychological studies of human variability* (pp. 329–350). American Psychological Association; Psychological Review Company. <https://doi.org/10.1037/13516-024>
- *Albus, P., Vogt, A., & Seufert, T. (2021). Signaling in virtual reality influences learning outcome and cognitive load. *Computers & Education*, *166*, 104154. <https://doi.org/10.1016/j.compedu.2021.104154>
- Al Khatib, S. A. (2010). Meta-cognitive self-regulated learning and motivational beliefs as predictors of college students' performance. *International Journal for Research in Education*, *27*, 57–71.
- Almanasreh, E., Moles, R., & Chen, T. F. (2019). Evaluation of methods used for estimating content validity. *Research in Social and Administrative Pharmacy*, *15*, 214–221. <https://doi.org/10.1016/j.sapharm.2018.03.066>
- Alpizar, D., Adesope, O. O., & Wong, R. M. (2020). A meta-analysis of signaling principle in multimedia learning environments. *Educational Technology Research and Development*, *68*, 2095–2119. <https://doi.org/10.1007/s11423-020-09748-7>
- *Altmeyer, K., Kapp, S., Thees, M., Malone, S., Kuhn, J., & Brünken, R. (2020). The use of augmented reality to foster conceptual knowledge acquisition in STEM laboratory courses—Theoretical background and empirical results. *British Journal of Educational Technology*, *51*, 611–628. <https://doi.org/10.1111/bjjet.12900>
- *Andrade, J., Huang, W. H. D., & Bohn, D. M. (2015). The impact of instructional design on college students' cognitive load and learning outcomes in a large food science and human nutrition course. *Journal of Food Science Education*, *14*, 127–135. <https://doi.org/10.1111/1541-4329.12067>
- *Anggraini, W., Sunawan, S., & Murtadho, A. (2020). The effects of the presence of tutor in the learning video on cognitive load and academic achievement. *Islamic Guidance and Counseling Journal*, *3*, 9–17. <https://doi.org/10.25217/igcj.v3i1.656>
- Anmarkrud, Ø., Andresen, A., & Bråten, I. (2019). Cognitive load and working memory in multimedia learning: Conceptual and measurement issues. *Educational Psychologist*, *54*, 61–83. <https://doi.org/10.1080/00461520.2018.1554484>
- Artino, A. R. (2008). Cognitive load theory and the role of learner experience. An abbreviated review for educational practitioners. *AACE Journal*, *16*, 425–439.
- Ayres, P. (2006). Using subjective measures to detect variations of intrinsic cognitive load within problems. *Learning and Instruction*, *16*, 389–400. <https://doi.org/10.1016/j.learninstruc.2006.09.001>
- Ayres, P. (2018). Subjective measures of cognitive load: What can they reliability measure? In R. Z. Zheng (Ed.), *Cognitive load measurement and application: A theoretical framework for meaningful research and practice* (pp. 9–28). Routledge.
- Ayres, P., & Sweller, J. (2014). The split-attention principle in multimedia learning. In R. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 206–226). Cambridge University Press. <https://doi.org/10.1017/CBO9781139547369.011>
- Baddeley, A. (1986). *Working memory*. Oxford University Press.
- Bannert, M. (2002). Managing cognitive load—Recent trends in cognitive load theory. *Learning and Instruction*, *12*, 139–146. [https://doi.org/10.1016/S0959-4752\(01\)00021-4](https://doi.org/10.1016/S0959-4752(01)00021-4)
- Baumeister, R. F. (1991). On the stability of variability: Retest reliability of metraits. *Personality and Social Psychology Bulletin*, *17*, 633–639. <https://doi.org/10.1177/0146167291176005>
- Beckmann, J. F. (2010). Taming a beast of burden—On some issues with the conceptualisation and operationalisation of cognitive load. *Learning and Instruction*, *20*, 250–264. <https://doi.org/10.1016/j.learninstruc.2009.02.024>
- *Beege, M., Nebel, S., Schneider, S., & Rey, G. D. (2019a). Social entities in educational videos: Combining the effects of addressing and professionalism. *Computers in Human Behavior*, *93*, 40–52. <https://doi.org/10.1016/j.chb.2018.11.051>

- *Beege, M., Nebel, S., Schneider, S., & Rey, G. D. (2021). The effect of signaling in dependence on the extraneous cognitive load in learning environments. *Cognitive Processing*, 22, 209–225. <https://doi.org/10.1007/s10339-020-01002-5>
- *Beege, M., Schneider, S., Nebel, S., Mittangk, J., & Rey, G. D. (2017). Ageism–age coherence within learning material fosters learning. *Computers in Human Behavior*, 75, 510–519. <https://doi.org/10.1016/j.chb.2017.05.042>
- *Beege, M., Schneider, S., Nebel, S., & Rey, G. D. (2020). Does the effect of enthusiasm in a pedagogical agent’s voice depend on mental load in the learner’s working memory? *Computers in Human Behavior*, 112, 106483. <https://doi.org/10.1016/j.chb.2020.106483>
- *Beege, M., Wirzberger, M., Nebel, S., Schneider, S., Schmidt, N., & Rey, G. D. (2019b). Spatial continuity effect vs. spatial contiguity failure. Revising the effects of spatial proximity between related and unrelated representations. *Frontiers in Education*, 4, 86. <https://doi.org/10.3389/feduc.2019b.00086>
- *Becker, S., Klein, P., Gößling, A., & Kuhn, J. (2020). Using mobile devices to enhance inquiry-based learning processes. *Learning and Instruction*, 69, 101350. <https://doi.org/10.1016/j.learninstruc.2020.101350>
- Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics*, 50, 1088–1101. <https://doi.org/10.2307/2533446>
- *Bender, L., Renkl, A., & Eitel, A. (2021). Seductive details do their damage also in longer learning sessions—When the details are perceived as relevant. *Journal of Computer Assisted Learning*, 37, 1248–1262. <https://doi.org/10.1111/jcal.12560>
- Bonett, D. G. (2010). Varying coefficient meta-analytic methods for alpha reliability. *Psychological Methods*, 15, 368–385. <https://doi.org/10.1037/a0020142>
- Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). Effect sizes based on correlations. In M. Borenstein, L. V. Hedges, J. P. T. Higgins, & H. R. Rothstein (Eds.), *Introduction to meta-analysis* (pp. 41–43). John Wiley & Sons Ltd. <https://doi.org/10.1002/9780470743386.ch6>
- Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2021). *Introduction to meta-analysis*. John Wiley & Sons.
- Borman, G. D. (2002). Experiments for educational evaluation and improvement. *Peabody Journal of Education*, 77, 7–27. https://doi.org/10.1207/S15327930PJE7704_2
- Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, 111, 1061–1071. <https://doi.org/10.1037/0033-295X.111.4.1061>
- Brom, C., Stárková, T., & D’Mello, S. K. (2018). How effective is emotional design? A meta-analysis on facial anthropomorphisms and pleasant colors during multimedia learning. *Educational Research Review*, 25, 100–119. <https://doi.org/10.1016/j.edurev.2018.09.004>
- Brünken, R., Moreno, R., & Plass, J. (2010). Current issues and open questions in cognitive load research. In J. Plass, R. Moreno, & R. Brünken (Eds.), *Cognitive load theory* (pp. 253–272). Cambridge University Press. <https://doi.org/10.1017/CBO9780511844744.014>
- Brünken, R., Plass, J. L., & Leutner, D. (2003). Direct measurement of cognitive load in multimedia learning. *Educational Psychologist*, 38, 53–61. https://doi.org/10.1207/S15326985EP3801_7
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81–105. <https://doi.org/10.1037/h0046016>
- Capaldi, C. A., Dopko, R. L., & Zelenski, J. M. (2014). The relationship between nature connectedness and happiness: A meta-analysis. *Frontiers in Psychology*, 5, 976. <https://doi.org/10.3389/fpsyg.2014.00976>
- Castro-Alonso, J. C., Wong, M., Adesope, O. O., Ayres, P., & Paas, F. (2019). Gender imbalance in instructional dynamic versus static visualizations: A meta-analysis. *Educational Psychology Review*, 31, 361–387. <https://doi.org/10.1007/s10648-019-09469-1>
- Cennamo, K. S. (1993). Learning from video: Factors influencing learners’ preconceptions and invested mental effort. *Educational Technology Research and Development*, 41, 33–45. <https://doi.org/10.1007/BF02297356>
- Chen, O., & Kalyuga, S. (2020). Cognitive load theory, spacing effect, and working memory resources depletion: Implications for instructional design. In S. Hai-Jew (Ed.), *Form, function, and style in instructional design: Emerging research and opportunities* (pp. 1–26). IGI Global. <https://doi.org/10.4018/978-1-5225-9833-6>

- Chen, O., Kalyuga, S., & Sweller, J. (2017). The expertise reversal effect is a variant of the more general element interactivity effect. *Educational Psychology Review*, 29, 393–405. <https://doi.org/10.1007/s10648-016-9359-1>
- Cho, E. (2016). Making reliability reliable: A systematic approach to reliability coefficients. *Organizational Research Methods*, 19, 651–682. <https://doi.org/10.1177/1094428116656239>
- Christmann, A., & Van Aelst, S. (2006). Robust estimation of Cronbach's alpha. *Journal of Multivariate Analysis*, 97, 1660–1674. <https://doi.org/10.1016/j.jmva.2005.05.012>
- *Chung, S., & Cheon, J. (2020). Emotional design of multimedia learning using background images with motivational cues. *Journal of Computer Assisted Learning*, 36, 922–932. <https://doi.org/10.1111/jcal.12450>
- Cobb, P., Confrey, J., diSessa, A., Lehrer, R., & Schauble, L. (2003). Design experiments in educational research. *Educational Researcher*, 32, 9–13. <https://doi.org/10.3102/0013189X032001009>
- *Colliot, T., & Jamet, E. (2018). Understanding the effects of a teacher video on learning from a multimedia document: An eye-tracking study. *Educational Technology Research and Development*, 66, 1415–1433. <https://doi.org/10.1007/s11423-018-9594-x>
- Colliver, J. A., Conlee, M. J., & Verhulst, S. J. (2012). From test validity to construct validity... and back?. *Medical Education*, 46, 366–371. <https://doi.org/10.1111/j.1365-2923.2011.04194.x>
- Cook, D. A., & Beckman, T. J. (2006). Current concepts in validity and reliability for psychometric instruments: Theory and application. *The American Journal of Medicine*, 119, 166e7–166e16. <https://doi.org/10.1016/j.amjmed.2005.10.036>
- Cook, D. A., Castillo, R. M., Gas, B., & Artino, A. R., Jr. (2017). Measuring achievement goal motivation, mindsets and cognitive load: Validation of three instruments' scores. *Medical Education*, 51, 1061–1074. <https://doi.org/10.1111/medu.13405>
- Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, 78, 98–104. <https://doi.org/10.1037/0021-9010.78.1.98>
- Costley, J., & Lange, C. (2018). The moderating effects of group work on the relationship between motivation and cognitive load. *Int Rev Res Open Distrib Learn*, 19, 68–90. <https://doi.org/10.19173/irrold.v19i1.3325>
- Cowan, N. (1999). An embedded-processes model of working memory. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 62–101). Cambridge University Press. <https://doi.org/10.1017/CBO9781139174909.006>
- Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, 24, 87–185. <https://doi.org/10.1017/S0140525X01003922>
- Cowan, N. (2008). What are the differences between long-term, short-term, and working memory? *Progress in Brain Research*, 169, 323–338. [https://doi.org/10.1016/S0079-6123\(07\)00020-9](https://doi.org/10.1016/S0079-6123(07)00020-9)
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334. <https://doi.org/10.1007/BF02310555>
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281–302. <https://doi.org/10.1037/h0040957>
- Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist*, 60, 170–180. <https://doi.org/10.1037/0003-066X.60.2.170>
- Dalal, D. K., Carter, N. T., & Lake, C. J. (2013). Middle Response Scale Options are Inappropriate for Ideal Point Scales. *Journal of Business and Psychology*, 29, 463–478. <https://doi.org/10.1007/s10869-013-9326-5>
- *Davis, R. O., Vincent, J., & Park, T. (2019). Reconsidering the voice principle with non-native language speakers. *Computers & Education*, 140, 103605. <https://doi.org/10.1016/j.compedu.2019.103605>
- *Debut, N., & Van De Leemput, C. (2014). What does germane load mean? An empirical contribution to the cognitive load theory. *Frontiers in Psychology*, 5, 1099. <https://doi.org/10.3389/fpsyg.2014.01099>
- de Jong, T. (2010). Cognitive load theory, educational research, and instructional design: Some food for thought. *Instructional Science*, 38, 105–134. <https://doi.org/10.1007/s11251-009-9110-0>
- Deng, L., & Chan, W. (2017). Testing the difference between reliability coefficients alpha and omega. *Educational and Psychological Measurement*, 77, 185–203. <https://doi.org/10.1177/0013164416658325>
- *Dervić, D., Nermin, Đ. A. P. O., Mešić, V., & Đokić, R. (2019). Cognitive load in multimedia learning: An example from teaching about lenses. *Journal of Education in Science Environment and Health*, 5, 102–118. <https://doi.org/10.21891/jeshe.481698>

- Drost, E. A. (2011). Validity and reliability in social science research. *Education Research and Perspectives*, 38, 105–123.
- Dunn, W. W. (2020). Validity. In L. J. Miller (Ed.), *Developing norm-referenced standardized tests* (pp. 149–168). Routledge. <https://doi.org/10.4324/9781315859811>
- Edwards, T., & Holtzman, N. S. (2017). A meta-analysis of correlations between depression and first person singular pronoun use. *Journal of Research in Personality*, 68, 63–68. <https://doi.org/10.1016/j.jrp.2017.02.005>
- Eisinga, R., Grotenhuis, M. T., & Pelzer, B. (2013). The reliability of a two-item scale: Pearson, Cronbach, or Spearman-Brown? *International Journal of Public Health*, 58, 637–642. <https://doi.org/10.1007/s00038-012-0416-3>
- *Eitel, A., Bender, L., & Renkl, A. (2019). Are seductive details seductive only when you think they are relevant? An experimental test of the moderating role of perceived relevance. *Applied Cognitive Psychology*, 33, 20–30. <https://doi.org/10.1002/acp.3479>
- Embretson, S. E. (2013). Test design: Developments in psychology and psychometrics. Academic Press.
- Ercikan, K., & Lyons-Thomas, J. (2013). Adapting tests for use in other languages and cultures. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I.C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), *APA handbook of testing and assessment in psychology, vol. 3. Testing and assessment in school psychology and education* (pp. 545–569). American Psychological Association. <https://doi.org/10.1037/14049-026>
- Eysink, T. H. S., De Jong, T., Berthold, K., Kollöffel, B., Opfermann, M., & Wouters, P. (2009). Learner performance in multimedia learning arrangements: An analysis across instructional approaches. *American Educational Research Journal*, 46, 1107–1149. <https://doi.org/10.3102/0002831209340235>
- *Fanguy, M., Costley, J., Baldwin, M., Lange, C., & Wang, H. (2019). Diversity in video lectures: Aid or hindrance? *International Review of Research in Open and Distributed Learning*, 20. <https://doi.org/10.19173/irrodl.v20i2.3838>
- Feldon, D. F. (2007). The Implications of research on expertise for curriculum and pedagogy. *Educational Psychology Review*, 19, 91–110. <https://doi.org/10.1007/s10648-006-9009-0>
- Feldon, D. F., Callan, G., Juth, S., & Jeong, S. (2019). Cognitive load as motivational cost. *Educational Psychology Review*, 31, 319–337. <https://doi.org/10.1007/s10648-019-09464-6>
- Ferketich, S. (1990). Internal consistency estimates of reliability. *Research in Nursing & Health*, 13, 437–440. <https://doi.org/10.1002/nur.4770130612>
- Field, A. P. (2005). Is the meta-analysis of correlation coefficients accurate when population correlations vary? *Psychological Methods*, 10, 444–467. <https://doi.org/10.1037/1082-989X.10.4.444>
- Field, A. P., & Gillett, R. (2010). How to do a meta-analysis. *British Journal of Mathematical and Statistical Psychology*, 63, 665–694. <https://doi.org/10.1348/000711010X502733>
- Fletcher, J. D., & Tobias, S. (2005). The multimedia principle. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 117–133). Cambridge University Press. <https://doi.org/10.1017/CBO9780511816819.008>
- Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences*, 102, 74–78. <https://doi.org/10.1016/j.paid.2016.06.069>
- Gilpin, A. R. (1993). Table for conversion of Kendall's Tau to Spearman's Rho within the context of measures of magnitude of effect for meta-analysis. *Educational and Psychological Measurement*, 53, 87–92. <https://doi.org/10.1177/0013164493053001007>
- Ginns, P. (2006). Integrating information: A meta-analysis of the spatial contiguity and temporal contiguity effects. *Learning and Instruction*, 16, 511–525. <https://doi.org/10.1016/j.learninstruc.2006.10.001>
- Glass, G. V., McGaw, B., & Smith, M. L. (1981). Meta-analysis in social research. Sage Publications.
- Gonzalez, O., MacKinnon, D. P., & Muniz, F. B. (2021). Extrinsic convergent validity evidence to prevent jingle and jangle fallacies. *Multivariate Behavioral Research*, 56, 3–19. <https://doi.org/10.1080/00273171.2019.1707061>
- *Gupta, U., & Zheng, R. Z. (2020). Cognitive load in solving mathematics problems: Validating the role of motivation and the interaction among prior knowledge, worked examples, and task difficulty. *European Journal of STEM Education*, 5, 5. <https://doi.org/10.20897/ejsteme/9252>
- *Glogger-Frey, I., Gaus, K., & Renkl, A. (2017). Learning from direct instruction: Best prepared by several self-regulated or guided invention activities? *Learning and Instruction*, 51, 26–35. <https://doi.org/10.1016/j.learninstruc.2016.11.002>

- Graham, J. M., & Christiansen, K. (2009). The reliability of romantic love: A reliability generalization meta-analysis. *Personal Relationships*, *16*, 49–66. <https://doi.org/10.1111/j.1475-6811.2009.01209.x>
- Graham, J. M., Diebels, K. J., & Barnow, Z. B. (2011). The reliability of relationship satisfaction: A reliability generalization meta-analysis. *Journal of Family Psychology*, *25*, 39–48. <https://doi.org/10.1037/a0022441>
- Green, S. B., Lissitz, R. W., & Mulaik, S. A. (1977). Limitations of coefficient alpha as an index of test unidimensionality. *Educational and Psychological Measurement*, *37*, 827–838. <https://doi.org/10.1177/001316447703700403>
- *Greenberg, K., Zheng, R., Gardner, M., & Orr, M. (2021). Individual differences in visuospatial working memory capacity influence the modality effect. *Journal of Computer Assisted Learning*, *37*, 735–744. <https://doi.org/10.1111/jcal.12519>
- Hafdahl, A. R., & Williams, M. A. (2009). Meta-analysis of correlations revisited: Attempted replication and extension of Field's (2001) simulation studies. *Psychological Methods*, *14*, 24–42. <https://doi.org/10.1037/a0014697>
- Hall, J. A., & Rosenthal, R. (1991). Testing for moderator variables in meta-analysis: Issues and methods. *Communications Monographs*, *58*, 437–448. <https://doi.org/10.1080/03637759109376240>
- Hambleton, R. K., & Patsula, L. (1998). Adapting tests for use in multiple languages and cultures. *Social Indicators Research*, *45*, 153–171. <https://doi.org/10.1023/A:1006941729637>
- Harkness, J., Pennell, B. E., & Schoua-Glusberg, A. (2004). Survey questionnaire translation and assessment. In R. M. Groves, G. Kalton, J. Rao, N. Schwarz, C. Skinner, S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, & E. Singer (Eds.), *Methods for testing and evaluating survey questionnaires* (pp. 453–473). John Wiley & Sons Inc. <https://doi.org/10.1002/0471654728.ch22>
- Hayes, A. F., & Coultts, J. J. (2020). Use omega rather than Cronbach's alpha for estimating reliability. But.... *Communication Methods and Measures*, *14*, 1–24. <https://doi.org/10.1080/19312458.2020.1718629>
- Heale, R., & Twycross, A. (2015). Validity and reliability in quantitative studies. *Evidence-Based Nursing*, *18*, 66–67. <https://doi.org/10.1136/eb-2015-102129>
- Hedges, L. V., Cooper, H., & Bushman, B. J. (1992). Testing the null hypothesis in meta-analysis: A comparison of combined probability and confidence interval procedures. *Psychological Bulletin*, *111*, 188–194. <https://doi.org/10.1037/0033-2909.111.1.188>
- Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Academic Press.
- Higgins, J. P. T., Thompson, S. G., & Spiegelhalter, D. J. (2009). A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society: Series A (statistics in Society)*, *172*, 137–159. <https://doi.org/10.1111/j.1467-985X.2008.00552.x>
- Hogan, T. P., Benjamin, A., & Brezinski, K. L. (2000). Reliability methods: A note on the frequency of use of various types. *Educational and Psychological Measurement*, *60*, 523–531. <https://doi.org/10.1177/001316440021970691>
- IBM Corp. (2021). *IBM SPSS Statistics for Windows, Version 28.0* [Computer software]. IBM Corp.(2021). Retrieved October 22, 2021, from <https://www.ibm.com/dede/analytics/spss-statistics-software>
- JASP Team (2021). JASP Version 0.15. [Computer software]. Retrieved October 22, 2021, from <https://jasp-stats.org/>
- Jiang, D., & Kalyuga, S. (2020). Confirmatory factor analysis of cognitive load ratings supports a two-factor model. *Tutorials in Quantitative Methods for Psychology*, *16*, 216–225. <https://doi.org/10.20982/tqmp.16.3.p216>
- Jonides, J., Lacey, S. C., & Nee, D. E. (2005). Processes of working memory in mind and brain. *Current Directions in Psychological Science*, *14*, 2–5. <https://doi.org/10.1111/j.0963-7214.2005.00323.x>
- Kalyuga, S. (2007). Expertise reversal effect and its implications for learner-tailored instruction. *Educational Psychology Review*, *19*, 509–539. <https://doi.org/10.1007/s10648-007-9054-3>
- Kalyuga, S. (2011). Cognitive load theory: How many types of load does it really need? *Educational Psychology Review*, *23*, 1–19. <https://doi.org/10.1007/s10648-010-9150-7>
- Kalyuga, S., & Renkl, A. (2010). Expertise reversal effect and its instructional implications: Introduction to the special issue. *Instructional Science*, *38*, 209–215. <https://doi.org/10.1007/s11251-009-9102-0>

- Kalyuga, S., & Sweller, J. (2014). The redundancy principle in multimedia learning. In R. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 247–262). Cambridge University Press. <https://doi.org/10.1017/CBO9781139547369.013>
- Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, 38, 319–342. <https://doi.org/10.1111/j.1745-3984.2001.tb01130.x>
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50, 1–73. <https://doi.org/10.1111/jedem.12000>
- Kimberlin, C. L., & Winterstein, A. G. (2008). Validity and reliability of measurement instruments used in research. *American Journal of Health-System Pharmacy*, 65, 2276–2284. <https://doi.org/10.2146/ajhp070364>
- Kirschner, P. A., Ayres, P., & Chandler, P. (2011). Contemporary cognitive load theory research: The good, the bad and the ugly. *Computers in Human Behavior*, 27, 99–105. <https://doi.org/10.1016/j.chb.2010.06.025>
- 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. <https://doi.org/10.1007/s10648-008-9095-2>
- Klepsch, M., Schmitz, F., & Seufert, T. (2017). Development and validation of two instruments measuring intrinsic, extraneous, and germane cognitive load. *Frontiers in Psychology*, 8, 1997. <https://doi.org/10.3389/fpsyg.2017.01997>
- *Klepsch, M., & Seufert, T. (2020). Understanding instructional design effects by differentiated measurement of intrinsic, extraneous, and germane cognitive load. *Instructional Science*, 48, 45–77. <https://doi.org/10.1007/s11251-020-09502-9>
- *Klepsch, M., & Seufert, T. (2021, April). Making an effort versus experiencing load. *Frontiers in Education*, 6, 645284. <https://doi.org/10.3389/educ.2021.645284>
- Korbach, A., Brünken, R., & Park, B. (2018). Differentiating different types of cognitive load: A comparison of different measures. *Educational Psychology Review*, 30, 503–529. <https://doi.org/10.1007/s10648-017-9404-8>
- *Korbach, A., Ginns, P., Brünken, R., & Park, B. (2020). Should learners use their hands for learning? Results from an eye-tracking study. *Journal of Computer Assisted Learning*, 36, 102–113. <https://doi.org/10.1111/jcal.12396>
- Krell, M. (2017). Evaluating an instrument to measure mental load and mental effort considering different sources of validity evidence. *Cogent Education*, 4, 1280256. <https://doi.org/10.1080/2331186x.2017.1280256>
- Leahy, W. (2018). Case studies in cognitive load measurement. In R. Z. Zheng (Ed.), *Cognitive load measurement and application: A theoretical framework for meaningful research and practice* (pp. 199–223). Routledge/Taylor & Francis Group.
- *Lehmann, J. A. M., Hamm, V., & Seufert, T. (2019). The influence of background music on learners with varying extraversion: Seductive detail or beneficial effect? *Applied Cognitive Psychology*, 33, 85–94. <https://doi.org/10.1002/acp.3509>
- Leppink, J., Paas, F., Van der Vleuten, C. P., Van Gog, T., & Van Merriënboer, J. J. (2013). Development of an instrument for measuring different types of cognitive load. *Behavior Research Methods*, 45, 1058–1072. <https://doi.org/10.3758/s13428-013-0334-1>
- Leppink, J., Paas, F., Van Gog, T., van Der Vleuten, C. P., & Van Merriënboer, J. J. (2014). Effects of pairs of problems and examples on task performance and different types of cognitive load. *Learning and Instruction*, 30, 32–42. <https://doi.org/10.1016/j.learninstruc.2013.12.001>
- Leppink, J., & van den Heuvel, A. (2015). The evolution of cognitive load theory and its application to medical education. *Perspectives on Medical Education*, 4, 119–127. <https://doi.org/10.1007/s40037-015-0192-x>
- *Liao, C. W., Chen, C. H., & Shih, S. J. (2019). The interactivity of video and collaboration for learning achievement, intrinsic motivation, cognitive load, and behavior patterns in a digital game-based learning environment. *Computers & Education*, 133, 43–55. <https://doi.org/10.1016/j.compedu.2019.01.013>
- *Liao, S., Kruger, J. L., & Doherty, S. (2020). The impact of monolingual and bilingual subtitles on visual attention, cognitive load, and comprehension. *The Journal of Specialised Translation Issue*, 33, 70–98.
- Lissitz, R. W., & Green, S. B. (1975). Effect of the number of scale points on reliability: A Monte Carlo approach. *Journal of Applied Psychology*, 60, 10–13. <https://doi.org/10.1037/h0076268>

- Martin-Martin, A., Orduña-Malea, E., Harzing, A. W., & López-Cózar, E. D. (2017). Can we use Google Scholar to identify highly-cited documents? *Journal of Informetrics*, *11*, 152–163. <https://doi.org/10.1016/j.joi.2016.11.008>
- Matell, M. S., & Jacoby, J. (1972). Is there an optimal number of alternatives for Likert-scale items? Effects of testing time and scale properties. *Journal of Applied Psychology*, *56*, 506–509. <https://doi.org/10.1037/h0033601>
- Mayer, R. E. (1996). Learning strategies for making sense out of expository text: The SOI model for guiding three cognitive processes in knowledge construction. *Educational Psychology Review*, *8*, 357–371. <https://doi.org/10.1007/BF01463939>
- Mayer, R. E. (2001). *Multimedia learning*. University Press.
- Mayer, R. E. (2014). Cognitive theory of multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 43–71). Cambridge University Press. <https://doi.org/10.1017/CBO9781139547369.005>
- Mayer, R. E., Mathias, A., & Wetzell, K. (2002). Fostering understanding of multimedia messages through pre-training: Evidence for a two-stage theory of mental model construction. *Journal of Experimental Psychology: Applied*, *8*, 147–154. <https://doi.org/10.1037/1076-898X.8.3.147>
- Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist*, *38*, 43–52. https://doi.org/10.1207/S15326985EP3801_6
- Mayer, R. E., & Moreno, R. (2010). Techniques that reduce extraneous cognitive load and manage intrinsic cognitive load during multimedia learning. In J. L. Plass, R. Moreno, & R. Brünken (Eds.), *Cognitive load theory* (pp. 131–152). Cambridge University Press. <https://doi.org/10.1017/CBO9780511844744.009>
- McDonald, R. P. (1999). Test theory: A unified treatment. Lawrence Erlbaum.
- McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, *23*, 412–433. <https://doi.org/10.1037/met0000144>
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13–103). American Council on Education and Macmillan.
- *Mikheeva, M., Schneider, S., Beege, M., & Rey, G. D. (2021). The influence of affective decorative pictures on learning statistics online. *Human Behavior and Emerging Technologies*, *3*, 401–412. <https://doi.org/10.1002/hbe2.250>
- Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*, 81–97. <https://doi.org/10.1037/h0043158>
- Miller, R. A., Stenmark, C. K., & Itersum, K. V. (2020). Dual computer displays reduce extraneous cognitive load. *Journal of Computer Assisted Learning*, *36*, 890–897. <https://doi.org/10.1111/jcal.12442>
- Moosbrugger, H., & Kelava, A. (2020). Qualitätsanforderungen an Tests und Fragebogen („Gütekriterien“) [Quality requirements for tests and questionnaires (“quality criteria“)]. In H. Moosbrugger & A. Kelava. (Eds.), *Testtheorie und Fragebogenkonstruktion* (pp. 13–38). Springer. https://doi.org/10.1007/978-3-662-61532-4_2
- Moreno, R. (2010). Cognitive load theory: More food for thought. *Instructional Science*, *38*, 135–141. <https://doi.org/10.1007/s11251-009-9122-9>
- Moreno, R., & Park, B. (2010). Cognitive load theory: Historical development and relation to other theories. In J. L. Plass, R. Moreno, & R. Brünken (Eds.), *Cognitive load theory* (pp. 9–28). Cambridge University Press. <https://doi.org/10.1017/CBO9780511844744.003>
- Mutlu-Bayraktar, D., Cosgun, V., & Altan, T. (2019). Cognitive load in multimedia learning environments: A systematic review. *Computers & Education*, *141*, 103618. <https://doi.org/10.1016/j.compedu.2019.103618>
- Naismith, L. M., Cheung, J. J., Ringsted, C., & Cavalcanti, R. B. (2015). Limitations of subjective cognitive load measures in simulation-based procedural training. *Medical Education*, *49*, 805–814. <https://doi.org/10.1111/medu.12732>
- Nakagawa, S., & Cuthill, I. C. (2007). Effect size, confidence interval and statistical significance: A practical guide for biologists. *Biological Reviews*, *82*, 591–605. <https://doi.org/10.1111/j.1469-185x.2007.00027.x>
- *Nebel, S., Schneider, S., Beege, M., Kolda, F., Mackiewicz, V., & Rey, G. D. (2017a). You cannot do this alone! Increasing task interdependence in cooperative educational videogames to encourage collaboration. *Educational Technology Research and Development*, *65*, 993–1014. <https://doi.org/10.1007/s11423-017-9511-8>


- *Nebel, S., Schneider, S., Schledjewski, J., & Rey, G. D. (2017b). Goal-setting in educational video games: Comparing goal-setting theory and the goal-free effect. *Simulation & Gaming, 48*, 98–130. <https://doi.org/10.1177/1046878116680869>
- *Nebel, S., Schneider, S., & Rey, G. D. (2016). From duels to classroom competition: Social competition and learning in educational videogames within different group sizes. *Computers in Human Behavior, 55*, 384–398. <https://doi.org/10.1016/j.chb.2015.09.035>
- Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "delayed-JOL effect." *Psychological Science, 2*, 267–271. <https://doi.org/10.1111/j.1467-9280.1991.tb00147.x>
- Nunnally, J. C. (1978). *Psychometric theory*. McGraw-Hill.
- Osburn, H. G. (2000). Coefficient alpha and related internal consistency reliability coefficients. *Psychological Methods, 5*, 343–355. <https://doi.org/10.1037/1082-989X.5.3.343>
- Ouweland, K., van der Kroef, A., Wong, J., & Paas, F. (2021). Measuring cognitive load: Are there more valid alternatives to Likert rating scales? *Frontiers in Education, 6*, 702616. <https://doi.org/10.3389/educ.2021.702616>
- 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. <https://doi.org/10.1037/0022-0663.84.4.429>
- Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist, 38*, 1–4. https://doi.org/10.1207/S15326985EP3801_1
- Paas, F., & Sweller, J. (2014). Implications of cognitive load theory for multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 27–42). Cambridge University Press. <https://doi.org/10.1017/CBO9781139547369.004>
- Paas, F., & van Gog, T. (2006). Optimising worked example instruction: Different ways to increase germane cognitive load. *Learning and Instruction, 16*, 87–91. <https://doi.org/10.1016/j.learninstruc.2006.02.004>
- Paas, F., & Van Merriënboer, J. J. G. (1994). Variability of worked examples and transfer of geometrical problem-solving skills: A cognitive-load approach. *Journal of Educational Psychology, 86*, 122–133. <https://doi.org/10.1037/0022-0663.86.1.122>
- Panayides, P. (2013). Coefficient alpha: Interpret with caution. *Europe's Journal of Psychology, 9*, 687–696. <https://doi.org/10.5964/ejop.v9i4.653>
- Park, B., & Brünken, R. (2015). The rhythm method: A new method for measuring cognitive load—An experimental dual-task study. *Applied Cognitive Psychology, 29*, 232–243. <https://doi.org/10.1002/acp.3100>
- Park, B., Korbach, A., & Brünken, R. (2015). Do learner characteristics moderate the seductive-details-effect? A cognitive-load-study using eye-tracking. *Journal of Educational Technology & Society, 18*, 24–36.
- Pentapati, K. C., Yeturu, S. K., & Siddiq, H. (2020). A reliability generalization meta-analysis of child oral impacts on daily performances (C-OIDP) questionnaire. *Journal of Oral Biology and Craniofacial Research, 10*, 776–781. <https://doi.org/10.1016/j.jobcr.2020.10.017>
- Peterson, R. A., & Brown, S. P. (2005). On the use of beta coefficients in meta-analysis. *Journal of Applied Psychology, 90*, 175–181. <https://doi.org/10.1037/0021-9010.90.1.175>
- Peterson, L., & Peterson, M. J. (1959). Short-term retention of individual verbal items. *Journal of Experimental Psychology, 58*, 193–198. <https://doi.org/10.1037/h0049234>
- *Petko, D., Schmid, R., & Cantieni, A. (2020). Pacing in serious games: Exploring the effects of presentation speed on cognitive load, engagement and learning gains. *Simulation & Gaming, 51*, 258–279. <https://doi.org/10.1177/1046878120902502>
- Piqueras, J. A., Martín-Vivar, M., Sandin, B., San Luis, C., & Pineda, D. (2017). The revised child anxiety and depression scale: A systematic review and reliability generalization meta-analysis. *Journal of Affective Disorders, 218*, 153–169. <https://doi.org/10.1016/j.jad.2017.04.022>
- Plass, J. L., & Kalyuga, S. (2019). Four ways of considering emotion in cognitive load theory. *Educational Psychology Review, 31*, 339–359. <https://doi.org/10.1007/s10648-019-09473-5>
- Polit, D. F., & Beck, C. T. (2006). The content validity index: Are you sure you know what's being reported? Critique and recommendations. *Research in Nursing & Health, 29*, 489–497. <https://doi.org/10.1002/nur.20147>

- Pollock, E., Chandler, P., & Sweller, J. (2002). Assimilating complex information. *Learning and Instruction, 12*, 61–86. [https://doi.org/10.1016/S0959-4752\(01\)00016-0](https://doi.org/10.1016/S0959-4752(01)00016-0)
- Rey, G. D., Beege, M., Nebel, S., Wirzberger, M., Schmitt, T. H., & Schneider, S. (2019). A meta-analysis of the segmenting effect. *Educational Psychology Review, 31*, 389–419. <https://doi.org/10.1007/s10648-018-9456-4>
- Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika, 74*, 145–154. <https://doi.org/10.1007/s11336-008-9102-z>
- Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychological Bulletin, 138*, 353–387. <https://doi.org/10.1037/a0026838>
- Schmeck, A., Opfermann, M., van Gog, T., Paas, F., & Leutner, D. (2015). Measuring cognitive load with subjective rating scales during problem solving: Differences between immediate and delayed ratings. *Instructional Science, 43*, 93–114. <https://doi.org/10.1007/s11251-014-9328-3>
- Schneider, S., Beege, M., Nebel, S., & Rey, G. D. (2018a). A meta-analysis of how signaling affects learning with media. *Educational Research Review, 23*, 1–24. <https://doi.org/10.1016/j.edurev.2017.11.001>
- *Schneider, S., Dyrna, J., Meier, L., Beege, M., & Rey, G. D. (2018b). How affective charge and text–picture connectedness moderate the impact of decorative pictures on multimedia learning. *Journal of Educational Psychology, 110*, 233–249. <https://doi.org/10.1037/edu0000209>
- *Schneider, S., Krieglstein, F., Beege, M., & Rey, G. D. (2021). How organization highlighting through signaling, spatial contiguity and segmenting can influence learning with concept maps. *Computers and Education Open, 2*, 100040. <https://doi.org/10.1016/j.caeo.2021.100040>
- *Schneider, S., Nebel, S., Beege, M., & Rey, G. D. (2018c). Anthropomorphism in decorative pictures: Benefit or harm for learning? *Journal of Educational Psychology, 110*, 218–232. <https://doi.org/10.1037/edu0000207>
- *Schneider, S., Nebel, S., Beege, M., & Rey, G. D. (2018d). The autonomy-enhancing effects of choice on cognitive load, motivation and learning with digital media. *Learning and Instruction, 58*, 161–172. <https://doi.org/10.1016/j.learninstruc.2018d.06.006>
- *Schneider, S., Nebel, S., Pradel, S., & Rey, G. D. (2015). Mind your Ps and Qs! How polite instructions affect learning with multimedia. *Computers in Human Behavior, 51*, 546–555. <https://doi.org/10.1016/j.chb.2015.05.025>
- *Schneider, S., Häbler, A., Habermeyer, T., Beege, M., & Rey, G. D. (2019a). The more human, the higher the performance? Examining the effects of anthropomorphism on learning with media. *Journal of Educational Psychology, 111*, 57–72. <https://doi.org/10.1037/edu0000273>
- *Schneider, S., Wirzberger, M., & Rey, G. D. (2019b). The moderating role of arousal on the seductive detail effect in a multimedia learning setting. *Applied Cognitive Psychology, 33*, 71–84. <https://doi.org/10.1002/acp.3473>
- *Schrader, C., Seufert, T., & Zander, S. (2021). Learning from instructional videos: Learner gender does matter; speaker gender does not. *Frontiers in Psychology, 12*, 1593. <https://doi.org/10.3389/fpsyg.2021.655720>
- Schroeder, N. L., & Ceneci, A. T. (2018). Spatial contiguity and spatial split-attention effects in multimedia learning environments: A meta-analysis. *Educational Psychology Review, 30*, 679–701. <https://doi.org/10.1007/s10648-018-9435-9>
- Schuman, H., Presser, S., & Ludwig, J. (1981). Context effects on survey responses to questions about abortion. *Public Opinion Quarterly, 45*, 216–223. <https://doi.org/10.1086/268652>
- Schuurman, N. K., & Hamaker, E. L. (2019). Measurement error and person-specific reliability in multilevel autoregressive modeling. *Psychological Methods, 24*, 70–91. <https://psycnet.apa.org/doi/https://doi.org/10.1037/met0000188>
- Schweppe, J., & Rummel, R. (2014). Attention, working memory, and long-term memory in multimedia learning: An integrated perspective based on process models of working memory. *Educational Psychology Review, 26*, 285–306. <https://doi.org/10.1007/s10648-013-9242-2>
- Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*, 107–120. <https://doi.org/10.1007/S11336-008-9101-0>
- Silver, N. C., & Dunlap, W. P. (1987). Averaging correlation coefficients: Should Fisher's z transformation be used? *Journal of Applied Psychology, 72*, 146–148. <https://doi.org/10.1037/0021-9010.72.1.146>

- Simms, L. J., Zelazny, K., Williams, T. F., & Bernstein, L. (2019). Does the number of response options matter? Psychometric perspectives using personality questionnaire data. *Psychological Assessment, 31*, 557–566. <https://doi.org/10.1037/pas0000648>
- *Skulmowski, A., Pradel, S., Kühnert, T., Brunnett, G., & Rey, G. D. (2016). Embodied learning using a tangible user interface: The effects of haptic perception and selective pointing on a spatial learning task. *Computers & Education, 92*, 64–75. <https://doi.org/10.1016/j.compedu.2015.10.011>
- *Skulmowski, A., & Rey, G. D. (2018). Realistic details in visualizations require color cues to foster retention. *Computers & Education, 122*, 23–31. <https://doi.org/10.1016/j.compedu.2018.03.012>
- *Skulmowski, A., & Rey, G. D. (2020a). Subjective cognitive load surveys lead to divergent results for interactive learning media. *Human Behavior and Emerging Technologies, 2*, 149–157. <https://doi.org/10.1002/hbe2.184>
- *Skulmowski, A., & Rey, G. D. (2020b). The realism paradox: Realism can act as a form of signaling despite being associated with cognitive load. *Human Behavior and Emerging Technologies, 2*, 251–258. <https://doi.org/10.1002/hbe2.190>
- *Stark, L., Malkmus, E., Stark, R., Brünken, R., & Park, B. (2018). Learning-related emotions in multimedia learning: An application of control-value theory. *Learning and Instruction, 58*, 42–52. <https://doi.org/10.1016/j.learninstruc.2018.05.003>
- *Stárková, T., Lukavský, J., Javora, O., & Brom, C. (2019). Anthropomorphisms in multimedia learning: Attract attention but do not enhance learning? *Journal of Computer Assisted Learning, 35*, 555–568. <https://doi.org/10.1111/jcal.12359>
- Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment, 80*, 99–103. https://doi.org/10.1207/S15327752JPA8001_18
- Sundararajan, N., & Adesope, O. (2020). Keep it coherent: A meta-analysis of the seductive details effect. *Educational Psychology Review, 32*, 707–734. <https://doi.org/10.1007/s10648-020-09522-4>
- Sungur, S. (2007). Modeling the relationships among students' motivational beliefs, metacognitive strategy use, and effort regulation. *Scandinavian Journal of Educational Research, 51*, 315–326. <https://doi.org/10.1080/00313830701356166>
- Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*, 275–285. [https://doi.org/10.1016/0364-0213\(88\)90023-7](https://doi.org/10.1016/0364-0213(88)90023-7)
- Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational Psychology Review, 22*, 123–138. <https://doi.org/10.1007/s10648-010-9128-5>
- Sweller, J. (2016). Working memory, long-term memory, and instructional design. *Journal of Applied Research in Memory and Cognition, 5*, 360–367. <https://doi.org/10.1016/j.jarmac.2015.12.002>
- Sweller, J. (2018). Measuring cognitive load. *Perspectives on Medical Education, 7*, 1–2. <https://doi.org/10.1007/s40037-017-0395-4>
- Sweller, J. (2020). Cognitive load theory and educational technology. *Educational Technology Research and Development, 68*, 1–16. <https://doi.org/10.1007/s11423-019-09701-3>
- Sweller, J. (2021). The role of evolutionary psychology in our understanding of human cognition: Consequences for cognitive load theory and instructional procedures. *Educational Psychology Review, 1*–13. <https://doi.org/10.1007/s10648-021-09647-0>
- Sweller, J., Ayres, P., & Kalyuga, S. (2011). Measuring cognitive load. In J. Sweller, P. Ayres, & S. Kalyuga (Eds.), *Cognitive load theory* (pp. 71–85). Springer. https://doi.org/10.1007/978-1-4419-8126-4_6
- Sweller, J., & Chandler, P. (1994). Why some material is difficult to learn. *Cognition and Instruction, 12*, 185–233. https://doi.org/10.1207/s1532690xci1203_1
- Sweller, J., Van Merriënboer, J. J., & Paas, F. G. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*, 251–296. <https://doi.org/10.1023/A:1022193728205>
- Sweller, J., van Merriënboer, J. J., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review, 31*, 261–292. <https://doi.org/10.1007/s10648-019-09465-5>
- Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education, 48*, 1273–1296. <https://doi.org/10.1007/s11165-016-9602-2>
- *Tang, M., Ginns, P., & Jacobson, M. J. (2019). Tracing enhances recall and transfer of knowledge of the water cycle. *Educational Psychology Review, 31*, 439–455. <https://doi.org/10.1007/s10648-019-09466-4>
- Tavakol, M., & Dennick, R. (2011a). Making sense of Cronbach's alpha. *International Journal of Medical Education, 2*, 53–55. <https://doi.org/10.5116/ijme.4dfb.8df8>

- Tavakol, M., & Dennick, R. (2011b). Post-examination analysis of objective tests. *Medical Teacher*, *33*, 447–458. <https://doi.org/10.3109/0142159X.2011.564682>
- *Thees, M., Kapp, S., Altmeyer, K., Malone, S., Brünken, R., & Kuhn, J. (2021). Comparing two subjective rating scales assessing cognitive load during technology-enhanced STEM laboratory courses. *Frontiers in Education*, *6*, 705551. <https://doi.org/10.3389/educ.2021.705551>
- *Thees, M., Kapp, S., Strzys, M. P., Beil, F., Lukowicz, P., & Kuhn, J. (2020). Effects of augmented reality on learning and cognitive load in university physics laboratory courses. *Computers in Human Behavior*, *108*, 106316. <https://doi.org/10.1016/j.chb.2020.106316>
- Thompson, S. G. (1994). Systematic Review: Why sources of heterogeneity in meta-analysis should be investigated. *BMJ*, *309*, 1351–1355. <https://doi.org/10.1136/bmj.309.6965.1351>
- Thompson, B. L., Green, S. B., & Yang, Y. (2010). Assessment of the maximal split-half coefficient to estimate reliability. *Educational and Psychological Measurement*, *70*, 232–251. <https://doi.org/10.1177/0013164409355688>
- Thompson, B., & Vacha-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. *Educational and Psychological Measurement*, *60*, 174–195. <https://doi.org/10.1177/0013164400602002>
- Vacha-Haase, T. (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement*, *58*, 6–20. <https://doi.org/10.1177/0013164498058001002>
- van der Stel, M., & Veenman, M. V. (2010). Development of metacognitive skillfulness: A longitudinal study. *Learning and Individual Differences*, *20*, 220–224. <https://doi.org/10.1016/j.lindif.2009.11.005>
- van Gog, T., & Paas, F. (2008). Instructional efficiency: Revisiting the original construct in educational research. *Educational Psychologist*, *43*, 16–26. <https://doi.org/10.1080/00461520701756248>
- Vaske, J. J., Beaman, J., & Sponarski, C. C. (2017). Rethinking internal consistency in Cronbach's alpha. *Leisure Sciences*, *39*, 163–173. <https://doi.org/10.1080/01490400.2015.1127189>
- Wakita, T., Ueshima, N., & Noguchi, H. (2012). Psychological distance between categories in the Likert scale: Comparing different numbers of options. *Educational and Psychological Measurement*, *72*, 533–546. <https://doi.org/10.1177/0013164411431162>
- *Wang, Z., Ardasheva, Y., Carbonneau, K., & Liu, Q. (2021a). Testing the seductive details effect: Does the format or the amount of seductive details matter? *Applied Cognitive Psychology*, *35*, 761–774. <https://doi.org/10.1002/acp.3801>
- *Wang, B., Ginns, P., & Mockler, N. (2021b). Sequencing tracing with imagination. *Educational Psychology Review*, 1–29. <https://doi.org/10.1007/s10648-021-09625-6>
- Warrens M. J. (2015). On Cronbach's alpha as the mean of all split-half reliabilities. In R. Millsap, D. Bolt, L. Ark van der, W.C. Wang (Eds.), *Quantitative psychology research* (pp. 293–300). Springer. https://doi.org/10.1007/978-3-319-07503-7_18
- Weidenmann, B. (2002). Multicodierung und Multimodalität im Lernprozess [Multicoding and multimodality in the learning process]. In L. J. Issing & P. Klimsa (Eds.), *Information und Lernen mit Multimedia* (3rd edition, pp. 45–62). Beltz PVU.
- Westen, D., & Rosenthal, R. (2003). Quantifying construct validity: Two simple measures. *Journal of Personality and Social Psychology*, *84*, 608–618. <https://doi.org/10.1037/0022-3514.84.3.608>
- *Xiong, J. (2017). The impact of control belief and learning disorientation on cognitive load: The mediating effect of academic emotions in two types of hypermedia learning environments. *TOJET: The Turkish Online Journal of Educational Technology*, *16*, 177–189.
- Zavgorodniaia, A., Duran, R., Hellas, A., Seppala, O., & Sorva, J. (2020, September). Measuring the cognitive load of learning to program: A replication study. In J. Maguire, & Q. Cutts (Eds.), *United Kingdom & Ireland Computing Education Research Conference* (pp. 3–9). <https://doi.org/10.1145/3416465.34164>
- Zu, T., Hutson, J., Loschky, L. C., & Rebello, N. S. (2020). Using eye movements to measure intrinsic, extraneous, and germane load in a multimedia learning environment. *Journal of Educational Psychology*, *112*, 1338–1352. <https://doi.org/10.1037/edu0000441>
- Zu, T., Munsell, J., & Rebello, N. S. (2021). Subjective measure of cognitive load depends on participants' content knowledge level. *Frontiers in Education*, *6*, 647097. <https://doi.org/10.3389/educ.2021.647097>

Authors and Affiliations

Felix Krieglstein¹  · **Maik Beege**² · **Günter Daniel Rey**¹ · **Paul Ginns**³ · **Moritz Krell**⁴ · **Sascha Schneider**⁵

¹ Psychology of Learning with Digital Media, Institute for Media Research, Faculty of Humanities, Chemnitz University of Technology, Chemnitz, Germany

² Digital Media in Education, Department of Psychology, University of Education, Freiburg, Germany

³ Sydney School of Education and Social Work, The University of Sydney, Sydney, Australia

⁴ Department Biology Education, IPN - Leibniz Institute for Science and Mathematics Education, Kiel, Germany

⁵ Educational Technology, Institute of Education, Faculty of Arts and Social Sciences, University of Zurich, Zurich, Switzerland