RECENT APPROACHES FOR ASSESSING COGNITIVE LOAD FROM A VALIDITY PERSPECTIVE

EDITED BY: Moritz Krell, Kate M. Xu, Günter Daniel Rey and Fred Paas
PUBLISHED IN: Frontiers in Education and Frontiers in Psychology

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88974-520-3
DOI 10.3389/978-2-88974-520-3

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

RECENT APPROACHES FOR ASSESSING COGNITIVE LOAD FROM A VALIDITY PERSPECTIVE

Topic Editors:
Moritz Krell, University of Kiel, Germany
Kate M. Xu, Open University of the Netherlands, Netherlands
Günter Daniel Rey, Chemnitz University of Technology, Germany
Fred Paas, Erasmus University Rotterdam, Netherlands

Citation: Krell, M., Xu, K. M., Rey, G. D., Paas, F., eds. (2022). Recent Approaches for Assessing Cognitive Load From a Validity Perspective.
Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88974-520-3

Table of Contents

frontiers in Education

EDITORIAL
published: 24 January 2022
doi: 10.3389/feduc.2021.838422

Editorial: Recent Approaches for Assessing Cognitive Load From a Validity Perspective

Moritz Krell[1]*, *Kate M. Xu*[2], *Günter Daniel Rey*[3] and *Fred Paas*[4,5]

[1]IPN—Leibniz Institute for Science and Mathematics Education, Kiel, Germany, [2]Faculty of Educational Sciences, Open University of the Netherlands, Heerlen, Netherlands, [3]Psychology of Learning With Digital Media, Institute for Media Research, Faculty of Humanities, Chemnitz University of Technology, Chemnitz, Germany, [4]Department of Psychology, Education and Child Studies, Erasmus University Rotterdam, Rotterdam, Netherlands, [5]School of Education/Early Start, University of Wollongong, Wollongong, NSW, Australia

Keywords: cognitive load, mental load, mental effort, measurement, assessment triangle, validity, subjective and objective measures

Editorial on the Research Topic

Recent Approaches for Assessing Cognitive Load From a Validity Perspective

INTRODUCTION

"Cognitive load can be defined as a multidimensional construct representing the load that performing a particular task imposes on the learner's cognitive system" (Paas et al., 2003, p. 64). It is assumed that assessed cognitive load under various experimental conditions represents the working memory resources exerted or required during the task performance. Cognitive load is widely studied in diverse disciplines such as education, psychology, and human factors. In educational science, cognitive load is used to guide instructional designs; for instance, when developing instructional designs, overly high cognitive load should be avoided because it may hinder knowledge construction and understanding. Generally, cognitive load theory offers valuable perspectives and design principles for instruction and instructional materials (e.g., Sweller, 2005; Kirschner et al., 2006; Paas and van Merriënboer, 2020).

The present *Research Topic* invited contributions describing approaches to measure cognitive load and examining the validity of these approaches. "Measuring cognitive load is [. . .] fundamentally important to education and learning" (Chandler, 2018, p.x) and several approaches have been proposed to measure cognitive load and function as indicators of learners' working memory resources during task performance, or as input for personalized adaptive tak selection (e.g., Salden, 2006). Cognitive load measurements can also advance cognitive load theory by providing an empirical basis for testing the hypothetical effects of instructional design principles on cognitive load (Paas et al., 2003). However, Martin (2018) points out that "Finding ways to disentangle different kinds of load and successfully measure them in valid and reliable ways remains a major challenge for the research field" (p.38). The contributions of the present *Research Topic* address this challenge and, hence, contribute to the development of a fundamentally important area of cognitive load research.

In this editorial piece, we first provide a summary of recent approaches to measure cognitive load, use the *assessment triangle* (National Research Council, 2001) as a framework to systemize existing and forthcoming research in the field of cognitive load measurement, then we analyze the studies published in this *Research Topic* based on the assessment triangle.

OPEN ACCESS

Edited and reviewed by:
Gavin T. L. Brown,
The University of Auckland,
New Zealand

***Correspondence:**
Moritz Krell
krell@leibniz-ipn.de

Specialty section:
This article was submitted to Assessment, Testing and Applied Measurement, a section of the journal Frontiers in Education

Received: *17 December 2021*
Accepted: *31 December 2021*
Published: *24 January 2022*

Citation:
Krell M, Xu KM, Rey GD and Paas F (2022) Editorial: Recent Approaches for Assessing Cognitive Load From a Validity Perspective. Front. Educ. 6:838422. doi: 10.3389/feduc.2021.838422

Recent Approaches for Assessing Cognitive Load

Assessment of cognitive load has been a significant direction in cognitive load theory and research (e.g., Paas et al., 2003; Paas et al., 2008; Brünken et al., 2010). Although several classifications of cognitive load measurement methods have been proposed (e.g., analytic vs. empirical, Xie and Salvendy, 2000; direct vs. indirect, Brünken et al., 2010), the subjective-objective classification (see Paas et al., 2003) is most frequently used. Whereas subjective methods are based on self-reported data, objective methods use observations of performance, behavior, or physiological conditions. The self-reported data are preferably collected after each learning unit or test task with rating scales for perceived mental effort (e.g., Paas, 1992), perceived task difficulty (e.g., Kalyuga et al., 1999), or both (e.g., Krell, 2017; Ouwehand et al.). Subjective measures are considered less sensitive to fluctuations in cognitive load and are generally used to estimate overall cognitive load. This means that students give one rating for a whole learning unit or test task or a series of learning units or test tasks. Although it has been argued that it is not possible to differentiate between different types of cognitive load (i.e., intrinsic, extraneous, germane) with the conventional rating scales, recent research has shown that the scales can be adapted to differentiate between the different types of cognitive load successfully (e.g., Leppink et al., 2013; Leppink et al., 2014).

Although subjective measures have been found easy to use, valid, and reliable (e.g., Paas et al., 1994; Leppink et al., 2013; Krell, 2017), cognitive load researchers have continuously searched for objective measures, which are not influenced by opinions and perceptions of people. This research has resulted in a wide range of objective techniques. The first objective method used in cognitive load research was an analytical method to analyze the effectiveness of learning from conventional goal-specific problems versus nonspecific problems based on the number of statements in working memory, the number of productions, the number of cycles to a solution, and the total number of conditions matched (Sweller, 1988). Another objective technique is based on the so-called dual-task paradigm (Bijarsari), which holds that performance on a secondary task performed in parallel with a primary task is indicative of the cognitive load imposed by the primary task (for examples see: Chandler and Sweller, 1996; Van Gerven et al., 2000; Brünken et al., 2010). An objective technique that is increasingly used in cognitive load research is based on the assumption that changes in cognitive load are reflected by physiological variables. In this category of objective techniques, measures of the heart (e.g., heart-rate variability: e.g., Minkley et al., 2018; Larmuseau et al., 2020), brain (e.g., EEG: e.g., Antonenko et al., 2010; Wang et al., 2020), eye (e.g., pupil dilation: e.g., Lee et al., 2020) and skin (e.g., galvanic skin response: e.g., Nourbakhs et al., 2012; Hoogerheide et al., 2019) have been used (for a review see, Paas, and Van Merrienboer, 2020). Psychophysiological measures are sensitive to instantaneous fluctuations in cognitive load.

Independent of the specific approach, which is employed in a study to measure cognitive load (i.e., subjective or objective), it is crucial to provide evidence that the specific measure allows inferences about the amount of cognitive capacity that a learner invested for learning or solving a task and, if applicable, for which type of cognitive load (e.g., intrinsic, extraneous, germane). In the next section, a framework for the systematic development and evaluation of educational and psychological assessments is introduced: the assessment triangle.

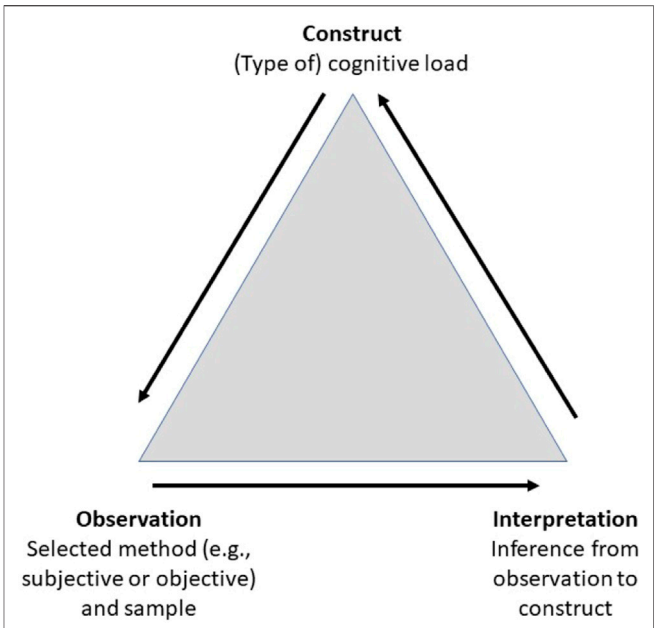

FIGURE 1 | Assessment triangle (adapted from Shavelson, 2010).

The Assessment Triangle

The assessment triangle (**Figure 1**) has been proposed by the US National Research Council (National Research Council, 2001) to emphasize that each educational assessment is a means to produce some data "that can be used to draw reasonable inferences about what students know" (p.42). While this statement refers to knowledge assessment, the framework also applies to broader contexts such as cognitive load assessment. The process of collecting evidence for supporting the specific inferences a researcher aims to draw is represented in the assessment triangle as a triad of construct (*What is measured?*), observation (*How is it measured?*), and interpretation (*Why can we infer from observation to construct?*) (Shavelson, 2010). For an assessment to be effective, each of the three elements must be considered and be in synchrony (National Research Council, 2001). Below, we explain how each dimension of the assessment triangle is reflected in the context of cognitive load research.

Construct

In the case of cognitive load assessment, the construct under consideration is cognitive load. While cognitive load is generally defined as an individual's cognitive resources used to learn or perform a task (e.g., Paas et al., 2003), further differentiations have been proposed, distinguishing between several types of cognitive load such as extraneous, intrinsic, and germane

cognitive load (e.g., Paas et al., 2003) or mental load and mental effort (e.g., Krell, 2017). For example, Choi et al. (2014) propose a model of cognitive load in which three main assessment factors of cognitive load are included: mental load, mental effort, and performance; see also (Paas and Van Merriënboer, 1994).

While the basic definition of cognitive load is shared by most researchers, the theoretical distinction between different types of cognitive load is still under debate (de Jong, 2010; Sweller et al., 2019). For example, task performance is sometimes conceptualized as being one aspect of cognitive load (e.g., Choi et al., 2014) others see it as an indicator for cognitive load (e.g., Kirschner, 2002). Furthermore, the most established conceptualization of cognitive load encompassing three types of extraneous, intrinsic, and germane cognitive load is also critically discussed (Kalyuga, 2011). For a specific cognitive load assessment, it is necessary to be clear and precise as to what one aims to measure. It is argued that an assessment approach is most effective if an explicit and clear concept of the construct of interest is used as a starting point (National Research Council, 2001).

Observation

Observation refers to the data that are collected using a specific method and aimed to be interpreted as evidence to draw inferences about the construct under investigation (National Research Council, 2001)—in this case: cognitive load. Various approaches have been suggested to measure cognitive load (Brünken et al., 2010). Subjective approaches using self-reports on rating scales are frequently used to assess cognitive load, while several objective approaches are still in an earlier stage of development and evaluation (Antonenko et al., 2010; Sweller et al., 2011; Ayres et al.). However, the validity of different approaches and the extent to which they represent cognitive load are still under debate (Solhjoo et al., 2019; Mutlu-Bayraktar et al., 2020).

Taking into consideration the different types of cognitive load suggested in the literature as well, this makes it necessary to carefully decide on which measurement method to use in a given context and for a given purpose. Furthermore, researchers should be clear about what kind of measures they decided to apply and for what reasons. Related to subjective approaches for cognitive load measurement, for instance, it is not always entirely clear which construct the items are aimed to measure. For example, many researchers use category labels related to task complexity but label them broadly as measures of cognitive load (de Jong, 2010).

Interpretation

Interpretation refers to the extent to which valid inferences can be drawn from data (e.g., self-reports on rating scales) to (the level of) an individual's invested cognitive resources. The interpretation of data as evidence for an individual's cognitive load means generalizing from a specific form of data (e.g., subjective ratings on single items) to the more global construct of cognitive load (or a specific type, such as mental effort). This is an essential step for the operationalization of the construct. However, the interpretation of measured data may be questioned, for example, if this interpretation is made only based on subjective ratings on items assessing mental effort. This is because cognitive load—as proposed by Choi et al. (2014)—is not only composed of the assessment factor of mental effort but also of mental load and performance. This demonstrates that the evaluation of the validity of the proposed interpretation of test scores is critical and complex.

In more general terms, it is proposed to consider different sources of evidence to support the claim that the proposed inferences from data to an individual's cognitive load are valid. This is why "the evidence required for validation is the evidence needed to evaluate the claims being made" (Kane, 2015, p.64).

Validity

In the *Standards for Educational and Psychological Testing*, the authors elaborate on different "sources of evidence that might be used in evaluating the validity of a proposed interpretation of test scores for a particular use" (p.13). These sources of validity evidence are: Evidence based on test content, on response processes, on internal structure, and on relations to other variables (American Educational Research Association et al., 2014). Gathering evidence based on test content hereby means analyzing the relation "between the content of a test and the construct it is intended to measure" (American Educational Research Association et al., 2014, p.14). Sources of evidence based on test content often consist of expert judgments. Concerning the assessment of cognitive load, it is necessary, for example, to ask why a specific approach (e.g., subjective) has been chosen and to what extent this decision influences the intended interpretation of the obtained data. Furthermore, quality criteria such as objectivity and reliability are necessary prerequisites for the valid interpretation of test scores (American Educational Research Association et al., 2014). The current concept of validity includes aspects of reliability and fairness in testing as part of the criteria that offer evidence of a sufficient internal structure. Gathering evidence based on response processes takes into account individuals' reasoning while answering the tasks to evaluate the extent to which the proposed inference on an individual's cognitive resources is valid. For this purpose, research methods like interviews and think-aloud protocols are typically employed. Gathering evidence based on relations to other variables means considering relevant external variables, for example, data from other assessments (e.g., convergent and discriminant evidence) or categorical variables such as different subsamples (e.g., known groups). For example, a comparison between subjective and objective measures has been proposed as a source of validity evidence for subjective measures and as a way to learn about what the different measures are measuring (Leppink et al., 2013; Korbach et al., 2018; Solhjoo et al., 2019).

Based on the considerations above, validity can be seen as an integrated evaluative judgment on the extent to which the appropriateness and quality of interpretations based on obtained data (e.g., subjective ratings or other diagnostic procedures) are supported by empirical evidence and theoretical arguments. Hence, "validity refers to the degree to

which evidence and theory support the interpretations of test scores for proposed uses of tests" (American Educational Research Association et al., 2014, p.11). Therefore, the validation of an instrument is not a routine procedure but is carried out through theory-based research, with which different interpretations of test data can be legitimized or even falsified (Hartig et al., 2008). Kane (2013) further argued that researchers have to critically demonstrate the validity of test interpretations based on a variety of evidence, especially by considering the evidence that potentially threatens the intended interpretation ("falsificationism").

The Contributions in this Research Topic

In the following, the 12 contributions of this *Research Topic* are analyzed and discussed based on the assessment triangle and the above thoughts on validity.

Minkley et al. based their study on the cognitive load framework Choi et al. (2014) and investigated relationships between causal and assessment factors of cognitive load in samples of secondary school students. The study aimed to test the assumed convergence between subjective (self-reported mental load and mental effort) and objective (heart rate) measures of cognitive load and to provide evidence for the assumed relationships between assessment factors of cognitive load (mental load and mental effort) and related causal factors in terms of learner characteristics (self-concept, interest and perceived stress). From their findings, the authors conclude that it is still unclear if objective measures can be validly interpreted as an indicator for an individual's cognitive load and in which contexts. The authors emphasize the need for a clear theoretical framework of cognitive load, including the different objective measures.

Andersen and Makransky, in their contribution, evaluated an adapted version of the widely used Cognitive Load Scale by Leppink et al. (2013) called Multidimensional Cognitive Load Scale for Physical and Online Lectures (MCLS-POL). In three studies, the authors provide validity evidence based on test content (utilizing theoretical considerations and previous studies), on internal structure (through psychometric analyses), and on relations to other variables (using group comparisons). Overall, the authors conclude that their findings provide evidence for the validity of the MCLS-POL but that some minor limitations should be considered in future studies (e.g., some subscales with only a few items).

Klepsch and Seufert evaluated items that have been formulated to measure active ("making an effort") and passive ("experiencing load") aspects of cognitive load. The authors report on two empirical studies, which are based on theoretical considerations concerning the relationship between active and passive aspects of cognitive load and intrinsic, extraneous, and germane load, as well as established load-inducing instructional design principles (e.g., the split-attention principle). Hence, the authors address validity evidence based on test content, internal structure, and relations to other variables. The findings suggest that it is possible to distinguish between active and passive aspects of load and that this can be related to the three types of cognitive load (i.e., the active load is associated with GCL, while the passive load is associated with ICL and—less strongly—with ECL). The items were not able, however, to entirely provide the expected measures of active and passive load in the different load-inducing instructional settings.

Zu et al. investigated how learner characteristics affect the validity of a subjective assessment instrument developed to assess extraneous, intrinsic, and germane cognitive load. In three experiments, the authors asked students to sort the items of the instrument and provide reasons for their groupings (experiment 1), administered the instrument alongside an electric circuit knowledge test before and after an instructional unit on electric circuits (experiment 2), and provided students with a test including different load-inducing problems and asked them to fill out the instrument subsequently (experiment 3). Overall, the findings provide validity evidence based on test content from the target population's view (experiment 1) and based on relations to other variables (i.e., known-groups comparison) from experiment 3; experiment 2, however, shows that the instrument's internal structure varied depending on the students' level of content knowledge. The authors discuss that content knowledge might moderate how students self-perceived their cognitive load. This emphasizes that learner characteristics have to be considered to draw valid inferences from self-reports on learners' cognitive load.

Ehrich et al. propose a new cognitive load index, which is derived from item response theory (IRT) estimates of relative task difficulty. The authors argue that the proposed index combines key assessment factors of cognitive load (i.e., mental load, mental effort, and performance); hence, providing theoretical arguments as validity evidence based on test content. Empirically, Ehrich et al. administered a version of "Australia's National Assessment Program—Literacy and Numeracy" test to calculate raw test scores from which relative task difficulty, that is, the proposed cognitive load index, was estimated. For this measure, they provide validity evidence based on internal structure (as the IRT model shows appropriate fit in the given context) and on relations to other variables. For the latter, the authors illustrate that students' scores on two standardized assessments (numeracy and literacy) predicted the cognitive load index as expected.

Bijarsari presents in her theoretical article current taxonomies of dual tasks for capturing cognitive load. She argues that there is a lack of standardization of dual tasks over study settings and task procedures, which—in turn—results in a lack of validity of dual-task approaches and comparability between studies. Based on a review of three dual-task taxonomies, Bijarsari proposes a "holistic taxonomy of dual-task settings," which includes parameters relevant to the design of a dual-task in a stepwise order, guiding researchers in the selection of the secondary task based on the chosen path.

Martin et al. administered the "Load Reduction Instruction Scale-Short" (LRIS-S) to students in high school science classrooms and applied multilevel latent profile analysis to identify student and classroom profiles based on students' reports on the LRIS-S and their accompanying psychological challenge and threat orientations. The authors explicitly adopted a within- and between-network construct validity approach on both the student and the classroom level. The analysis suggested

five instructional-motivational profiles (student-level within-network), which also showed differences in persistence, disengagement, and achievement (student-level between-network). At the classroom level, the authors identified three instructional-psychological profiles among classrooms (within-network) with different levels of persistence, disengagement, and achievement (between-network). Hence, the authors consider learner characteristics (motivational constructs) and environment characteristics (classrooms) and adopt a validity approach that considers evidence based on internal structure and on relations to other variables on both the student level and the classroom level.

Ayres et al., in their review, analyzed a sample of 33 experiments that used physiological measures of intrinsic cognitive load. The findings show that physiological measures related to four main categories were used in the analyzed studies (heart and lungs, eyes, skin, brain). For evaluation of the validity of the measures, the authors considered construct validity and sensitivity (i.e., the potential to detect changes in intrinsic cognitive load across tasks with different levels of complexity). The findings propose that the vast majority of physiological measures had "some level of validity" (p.13) but varied in terms of sensitivity. However, subjective measures, which were also applied in some of the studies, had the highest levels of validity. The authors conclude that a combination of physiological and subjective measures is most effective for validly and sensitively measuring intrinsic cognitive load.

Kastaun et al. examined the validity of a subjective (i.e., self-report) instrument to assess extraneous, intrinsic, and germane cognitive load during inquiry learning. Validity is evaluated by investigating relationships between causal (e.g., cognitive abilities) and assessment (e.g., eye-tracking metrics) factors about the scores on the cognitive load instrument. In two studies, secondary school students investigated a biological phenomenon and selected one of four multimedia scaffolds. Cognitive-visual and verbal abilities, reading skills, and spatial abilities were assessed as causal factors of cognitive load, and the learners indicated their representation preference by selecting one scaffold. In sum, the authors considered validity evidence based on test content and on relations to other variables, explicitly stating four validity assumptions: 1) the three scales have a sufficient internal consistency, 2) the three subjective measures detect different cognitive load levels for students in grades 9 and 11, 3) there are theoretically sound relationships between the three subjective measures and causal factors as well as 4) assessment factors. The findings consistently support assumptions 1) and 2) but only partially assumptions 3) and 4).

Thees et al. investigated the validity of two established subjective measures of cognitive load in the learning context of technology-enhanced STEM laboratory courses. Engineering students performed six experiments (presented in two different spatial arrangements) examining basic electric circuits and, immediately after the experimentation, answered both instruments. The authors analyzed various sources of validity evidence, including the instruments' internal structure and relation to other variables (i.e., group comparison). The intended three-factorial internal structure could not be found, and several subscales showed insufficient internal consistency. Only one instrument showed the expected group differences. Based on these findings, the authors suggest a combination of items from both instruments as a more valid instrument, which, however, still has low reliability in the subscale for the extraneous cognitive load.

Ouwehand et al. investigated how visual characteristics of rating scales influenced the validity of subjective cognitive load measures. They compared four rating scale measures differing in visual appearance (two numerical scales and two pictorial scales), which asked participants to rate mental load and mental effort after working on simple and complex tasks. The authors address validity evidence on test content (by asking the respondents to comment on the scales in an open-ended question) and on relations to other variables (by comparing resulting measures between scale type and task complexity). The findings show that all scales revealed expected differences in mental load and mental effort between simple and complex tasks; however, numerical scales provided expected relationships between cognitive load measures and performance on complex tasks more clearly than visual scales, while the opposite was found for simple tasks. In sum, this study hints that subtleties in measurements (i.e., item surface features such as visual appearance) can influence findings and, hence, could be a potential threat to the valid interpretation of test scores.

Schnaubert and Schneider investigated the relationship between perceived mental load and mental effort and comprehension and metacomprehension under different design conditions of multimedia material. The authors varied the design of the learning material (text-picture integrated, split attention, active integration) and tested for direct and indirect effects of mental load and mental effort on metacomprehension judgments. Beyond indirect effects via comprehension, both mental load and mental effort were directly related to metacomprehension (which differed between the multimedia design conditions). Based on their findings, the authors discuss that subjectivity (i.e., subjective experience of cognitive processes) needs to be considered more explicitly for validly assessing cognitive load with subjective methods.

SUMMARY AND DISCUSSION

To summarize, the present *Research Topic* includes two theoretical papers (i.e., literature reviews) and ten empirical studies. The theoretical papers show, for the specific areas of analysis, that there is a lack of validity in and comparability between most studies using dual tasks to capture cognitive load (Bijarsari) and that the validity and sensitivity are limited for most physiological approaches (Ayres et al.). Both studies illustrate the need for further research in terms of conceptual clarification and methods development and evaluation, respectively. Compared to other constructs, such as general cognitive abilities (e.g., Liepmann et al., 2007) or domain-specific competencies (e.g., Krüger et al., 2020), standardized representative validation studies for cognitive load assessments are highly needed but widely missing. Hence, instead of

evaluating new methods of cognitive load assessment, further systematic study of the different existing methods is needed. For example, to investigate under what conditions and for whom a specific subjective method works well and why would be more critical than just applying an existing or a new method. In addition, the measurement methods are typically considered one of several dependent variables in complex learning environments, which consist of many other elements (Choi et al., 2014). Therefore, it is challenging to disentangle the differential effects on the cognitive load measures. More fundamental research into the methods is needed to get a detailed picture of the factors affecting the obtained measures. For instance, Ouwehand et al. demonstrated that research on item surface features could provide valuable insights on the validity of subjective ratings. Such fundamental research should precede more applied research.

Of the ten empirical studies, one used cognitive load measures to investigate the proposed relationship between causal and assessment factors of cognitive load. Based on their findings, Minkley et al. specifically emphasize the need for precise conceptual integration of the various objective measures for cognitive load. The remaining nine empirical studies provide studies on the validity of newly developed cognitive load measures (e.g., active and passive load; Klepsch and Seufert) or on the validity of established measures adapted to new contexts (e.g., technology-enhanced STEM laboratory courses; Thees et al.). Studies of the latter type show that published scales should be evaluated before they can be validly used in new contexts.

Besides proposing specific cognitive load measures, the empirical studies in this *Research Topic* also provide valuable findings for cognitive load assessment in general. For example, Zu et al. show that learner characteristics should be considered to interpret subjective cognitive load measurements validly. All studies in this *Research Topic* found—to a greater or lesser extent—next to supportive evidence also evidence that potentially threatens the validity of the investigated measures. For example, Thees et al. could not find the assumed three-factorial internal structure of their data, and several of their subscales showed insufficient internal consistency. Generally, it is likely that the unsolved issue of how to conceptualize cognitive load (e.g., two vs. three types; Kalyuga, 2011) highly influences cognitive load measures and their validity—at least if scholars do not evaluate the appropriateness of the selected approach for their study context. Furthermore, it is also crucial for researchers to reflect on the consequences of new developments in cognitive load theory in the research on cognitive load measurement. One example is the recently identified possibility of working memory resource depletion, which may occur following extensive mental effort (Chen et al., 2018).

Concluding, the studies collected in this *Research Topic* illustrate the need for further research on the validity of interpretations of data as indicators for cognitive load. Such research, to be systematic and theory-guided, can be fruitfully framed within the assessment triangle (**Figure 1**).

AUTHOR CONTRIBUTIONS

Conceptualization, MK; writing—original draft preparation, MK; writing—review and editing, MK, KX, GR, FP.

REFERENCES

American Educational Research Association (2014). "American Psychological Association, & National Council on Measurement in Education," in *Standards for Educational and Psychological Testing* (Washington, DC: American Educational Research Association).

Antonenko, P., Paas, F., Grabner, R., and van Gog, T. (2010). Using Electroencephalography to Measure Cognitive Load. *Educ. Psychol. Rev.* 22, 425–438. doi:10.1007/s10648-010-9130-y

Brünken, R., Seufert, T., and Paas, F. (2010). "Cognitive Load Measurement," in *Cognitive Load Theory*. Editors J. Plass, R. Moreno, and R. Brünken (New York: Cambridge University Press), 181–202.

Chandler, P. (2018). "Foreword," in *Cognitive Load Measurement and Application*. Editor R. Zheng (New York, NY: Routledge).

Chandler, P., and Sweller, J. (1996). Cognitive Load while Learning to Use a Computer Program. *Appl. Cognit. Psychol.* 10 (2), 151–170. doi:10.1002/(sici)1099-0720(199604)10:2<151:aid-acp380>3.0.co;2-u

Chen, O., Castro-Alonso, J. C., Paas, F., and Sweller, J. (2018). Extending Cognitive Load Theory to Incorporate Working Memory Resource Depletion: Evidence from the Spacing Effect. *Educ. Psychol. Rev.* 30, 483–501. doi:10.1007/s10648-017-9426-2

Choi, H.-H., van Merriënboer, J. J. G., and Paas, F. (2014). Effects of the Physical Environment on Cognitive Load and Learning: towards a New Model of Cognitive Load. *Educ. Psychol. Rev.* 26, 225–244. doi:10.1007/s10648-014-9262-6

de Jong, T. (2010). Cognitive Load Theory, Educational Research, and Instructional Design: Some Food for Thought. *Instr. Sci.* 38, 105–134. doi:10.1007/s11251-009-9110-0

Hartig, J., Frey, A., and Jude, N. (2008). "Validität," in *Testtheorie und Fragebogenkonstruktion*. Editors H. Moosbrugger and A. Kelava (Springer), 135–163. doi:10.1007/978-3-540-71635-8_7

Hoogerheide, V., Renkl, A., Fiorella, L., Paas, F., and Van Gog, T. (2019). Enhancing Example-Based Learning: Teaching on Video Increases Arousal and Improves Problem-Solving Performance. *J. Educ. Psychol.* 111, 45–56. doi:10.1037/edu0000272

Kalyuga, S., Chandler, P., and Sweller, J. (1999). Managing Split-Attention and Redundancy in Multimedia Instruction. *Appl. Cognit. Psychol.* 13 (4), 351–371. doi:10.1002/(sici)1099-0720(199908)13:4<351:aid-acp589>3.0.co;2-6

Kalyuga, S. (2011). Cognitive Load Theory: How many Types of Load Does it Really Need. *Educ. Psychol. Rev.* 23, 1–19. doi:10.1007/s10648-010-9150-7

Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores. *J. Educ. Meas.* 50, 1–73. doi:10.1111/jedm.2013.50.issue-110.1111/jedm.12000

Kane, M. (2015). "Validation Strategies. Delineating and Validating Proposed Interpretations and Uses of Test Scores," in *Handbook of Test Development*. Editors M. Raymond, S. Lane, and T. Haladyna (New York: Routledge), 64–80.

Kirschner, P. A. (2002). Cognitive Load Theory: Implications of Cognitive Load Theory on the Design of Learning. *Learn. Instruction* 12, 1–10. doi:10.1016/s0959-4752(01)00014-7

Kirschner, P. A., Sweller, J., and Clark, R. E. (2006). Why Minimal Guidance during Instruction Does Not Work: An Analysis of the Failure of Constructivist, Discovery, Problem-Based, Experiential, and Inquiry-Based Teaching. *Educ. Psychol.* 41, 75–86. doi:10.1207/s15326985ep4102_1

Korbach, A., Brünken, R., and Park, B. (2018). Differentiating Different Types of Cognitive Load: A Comparison of Different Measures. *Educ. Psychol. Rev.* 30, 503–529. doi:10.1007/s10648-017-9404-8

Krell, M. (2017). Evaluating an Instrument to Measure Mental Load and Mental Effort Considering Different Sources of Validity Evidence. *Cogent Educ.* 4, 1280256. doi:10.1080/2331186X.2017.1280256

Krüger, D., Hartmann, S., Nordmeier, V., and Upmeier zu Belzen, A. (2020). "Measuring Scientific Reasoning Competencies," in *Student Learning in German Higher Education*. Editors O. Zlatkin-Troitschanskaia, H. Pant, M. Toepper, and C. Lautenbach (Springer), 261–280. doi:10.1007/978-3-658-27886-1_13

Larmuseau, C., Cornelis, J., Lancieri, L., Desmet, P., and Depaepe, F. (2020). Multimodal Learning Analytics to Investigate Cognitive Load during Online Problem Solving. *Br. J. Educ. Technol.* 51 (5), 1548–1562. doi:10.1111/bjet.12958

Lee, J. Y., Donkers, J., Jarodzka, H., Sellenraad, G., and van Merriënboer, J. J. G. (2020). Different Effects of Pausing on Cognitive Load in a Medical Simulation Game. *Comput. Hum. Behav.* 110, 106385. doi:10.1016/j.chb.2020.106385

Leppink, J., Paas, F., Van der Vleuten, C. P., Van Gog, T., and Van Merriënboer, J. J. (2013). Development of an Instrument for Measuring Different Types of Cognitive Load. *Behav. Res. Methods* 45, 1058–1072. doi:10.3758/s13428-013-0334-1

Leppink, J., Paas, F., Van Gog, T., Van der Vleuten, C. P. M., and van Merriënboer, J. J. G. (2014). Effects of Pairs of Problems and Examples on Task Performance and Different Types of Cognitive Load. *Learn. Instruction* 30, 32–42. doi:10.1016/j.learninstruc.2013.12.001

Liepmann, D., Beauducel, A., Brocke, B., and Amthauer, R. (2007). *Intelligenz-Struktur-Test 2000 R (I-S-T 2000 R)*. Göttingen: Hogrefe.

Martin, S. (2018). "A Critical Analysis of the Theoretical Construction and Empirical Measurement of Cognitive Load," in *Cognitive Load Measurement and Application*. Editor R. Zheng (New York, NY: Routledge), 29–44.

Minkley, N., Kärner, T., Jojart, A., Nobbe, L., and Krell, M. (2018). Students' Mental Load, Stress, and Performance when Working with Symbolic or Symbolic-Textual Molecular Representations. *J. Res. Sci. Teach.* 55, 1162–1187. doi:10.1002/tea.21446

Mutlu-Bayraktar, D., Ozel, P., Altindis, F., and Yilmaz, B. (2020). Relationship between Objective and Subjective Cognitive Load Measurements in Multimedia Learning. *Interactive Learn. Environments*, 1–13. doi:10.1080/10494820.2020.1833042

Nourbakhsh, N., Wang, Y., Chen, F., and Calvo, R. A. (2012). Using Galvanic Skin Response for Cognitive Load Measurement in Arithmetic and reading tasks. in"Proceedings of the 24th Conference on Australian Computer-Human Interaction OzCHI. New York, NY: Association for Computing Machinery.

National Research Council (2001). *Knowing what Students Know: The Science and Design of Educational Assessment*. Washington, DC: National Academy Press. doi:10.1145/2414536.2414602

Paas, F., Ayres, P., and Pachman, M. (2008). "Assessment of Cognitive Load in Multimedia Learning Environments: Theory, Methods, and Applications," in *Recent Innovations in Educational Technology that Facilitate Student Learning*. Editors D. Robinson and G. Schraw (Charlotte, NC: Information Age Publishing), 11–35.

Paas, F. G., Van Merriënboer, J. J., and Adam, J. J. (1994). Measurement of Cognitive Load in Instructional Research. *Percept Mot. Skills* 79, 419–430. doi:10.2466/pms.1994.79.1.41910.2466/pms.1994.79.1.419

Paas, F. G. W. C. (1992). Training Strategies for Attaining Transfer of Problem-Solving Skill in Statistics: A Cognitive-Load Approach. *J. Educ. Psychol.* 84, 429–434. doi:10.1037/0022-0663.84.4.429

Paas, F., Tuovinen, J. E., Tabbers, H., and Van Gerven, P. W. M. (2003). Cognitive Load Measurement as a Means to advance Cognitive Load Theory. *Educ. Psychol.* 38, 63–71. doi:10.1207/S15326985EP3801_8

Paas, F., and Van Merriënboer, J .J. G. (1994). Instructional Control of Cognitive Load in the Training of Complex Cognitive Tasks. *Educ. Psychol. Rev.* 6, 51–71. doi:10.1007/bf02213420

Paas, F., and van Merriënboer, J. J. G. (2020). Cognitive-load Theory: Methods to Manage Working Memory Load in the Learning of Complex Tasks. *Curr. Dir. Psychol. Sci.* 29, 394–398. doi:10.1177/0963721420922183

Salden, R. J. C. M., Paas, F., and van Merriënboer, J. J. G. (2006). Personalised Adaptive Task Selection in Air Traffic Control: Effects on Training Efficiency and Transfer. *Learning and Instruction* 16, 350–362.

Shavelson, R. J. (2010). On the Measurement of Competency. *Empirical Res. Voc Ed. Train.* 2, 41–63. doi:10.1007/bf03546488

Solhjoo, S., Haigney, M. C., McBee, E., van Merriënboer, J. J. G., Schuwirth, L., Artino, A. R., et al. (2019). Heart Rate and Heart Rate Variability Correlate with Clinical Reasoning Performance and Self-Reported Measures of Cognitive Load. *Sci. Rep.* 9, 14668. doi:10.1038/s41598-019-50280-3

Sweller, J., Ayres, P., and Kalyuga, S. (2011). "Measuring Cognitive Load," in *Cognitive Load Theory*. Editors J. Sweller, P. Ayres, and S. Kalyuga (New York, NY: Springer), 71–85. doi:10.1007/978-1-4419-8126-4_6

Sweller, J. (1988). Cognitive Load during Problem Solving: Effects on Learning. *Cogn. Sci.* 12 (2), 257–285. doi:10.1207/s15516709cog1202_4

Sweller, J. (2005). "Implications of Cognitive Load Theory for Multimedia Learning," in *The Cambridge Handbook of Multimedia Learning*. Editor R. E. Mayer (New York, NY: Cambridge University Press), 19–30. doi:10.1017/cbo9780511816819.003

Sweller, J., van Merriënboer, J. J. G., and Paas, F. (2019). Cognitive Architecture and Instructional Design: 20 Years Later. *Educ. Psychol. Rev.* 31, 261–292. doi:10.1007/s10648-019-81909465-5

Van Gerven, W. M., Fred, G. W. C., Paas, F., Van Merriënboer, J. J. G., and Schmidt, H. G. (2000). Cognitive Load Theory and the Acquisition of Complex Cognitive Skills in the Elderly: Towards an Integrative Framework. *Educ. Gerontol.* 26, 503–521. doi:10.1080/03601270050133874

Wang, J., Antonenko, P., Keil, A., and Dawson, K. (2020). Converging Subjective and Psychophysiological Measures of Cognitive Load to Study the Effects of Instructor-Present Video. *Mind, Brain Educ.* 14 (3), 279–291. doi:10.1111/mbe.12239

Xie, B., and Salvendy, G. (2000). Review and Reappraisal of Modelling and Predicting Mental Workload in Single- and Multi-Task Environments. *Work & Stress* 14 (1), 74–99. doi:10.1080/026783700417249

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Psychology

ORIGINAL RESEARCH
published: 18 March 2021
doi: 10.3389/fpsyg.2021.642084

The Validation and Further Development of the Multidimensional Cognitive Load Scale for Physical and Online Lectures (MCLS-POL)

Martin S. Andersen and Guido Makransky*

Department of Psychology, University of Copenhagen, Copenhagen, Denmark

Cognitive load theory (CLT) has been widely used to help understand the process of learning and to design teaching interventions. The Cognitive Load Scale (CLS) developed by Leppink and colleagues has emerged as one of the most validated and widely used self-report measures of intrinsic load (IL), extraneous load (EL), and germane load (GL). In this paper we investigated an expansion of the CLS by using a multidimensional conceptualization of the EL construct that is relevant for physical and online teaching environments. The Multidimensional Cognitive Load Scale for Physical and Online Lectures (MCLS-POL) goes beyond the CLS's operationalization of EL by expanding the EL component which originally included factors related to instructions/explanations with sub-dimensions including EL stemming from noises, and EL stemming from both media and devices within the environment. Through three studies, we investigated the reliability, and internal and external validity of the MCLS-POL using the Partial Credit Model, Confirmatory Factor Analysis, and differences between students either attending a lecture physically or online (Study 2 and 3). The results of Study 1 ($N = 250$) provide initial evidence for the validity and reliability of the MCLS-POL within a higher education sample, but also highlighted several potential improvements which could be made to the measure. These changes were made before re-evaluating the validity and reliability of the measure in a new sample of higher education psychology students ($N = 140$, Study 2), and psychological testing students ($N = 119$, Study 3). Together the studies provide evidence for a multidimensional conceptualization cognitive load and provide evidence of the validity, reliability, and sensitivity of the MCLS-POL and provide suggestions for future research directions.

Keywords: cognitive load, confirmatory factor analysis, item response theory, online lecture, Rasch measurement

OPEN ACCESS

Edited by:
Fred Paas,
Erasmus University Rotterdam, Netherlands

Reviewed by:
Matt Sibbald,
McMaster University, Canada
Tina Seufert,
University of Ulm, Germany
Christopher Lange,
Dankook University, South Korea

****Correspondence:***
Martin S. Andersen
msa@psy.ku.dk

Specialty section:
This article was submitted to Educational Psychology, a section of the journal Frontiers in Psychology

Received: *15 December 2020*
Accepted: *18 February 2021*
Published: *18 March 2021*

Citation:
Andersen MS and Makransky G (2021) The Validation and Further Development of the Multidimensional Cognitive Load Scale for Physical and Online Lectures (MCLS-POL). Front. Psychol. 12:642084. doi: 10.3389/fpsyg.2021.642084

INTRODUCTION

Cognitive load theory (CLT) posits that the strain put on working memory by the learning content plays a key role in whether or not the student succeeds in learning (Sweller et al., 2011a). A fundamental assumption is that working memory is limited in terms of capacity, but long term memory has a much greater capacity as information is stored in schemas (Chi et al., 1982). Thus, working memory becomes a form of bottleneck that requires instructors to design learning content in a way that can maximize the amount of information that is stored in the long term memory.

Originally, the CLT assumed that cognitive load was a unidimensional construct pertaining only to the total capacity of working memory (Ayres, 2018), and there is still disagreement about how to conceptualize different types of cognitive load (Kalyuga, 2011 Tindall-Ford et al., 2019). However, three distinct types of cognitive load are primarily described, including: Intrinsic Load (IL), which relates to the students perceived difficulty of the learning material, the difficulty of the learning material varies based on the materials composition and the materials' element interactivity (Sweller et al., 1998; Tindall-Ford et al., 2019). Extraneous Load (EL) which consist of non-intrinsic parts of the learning situation i.e., non-relevant information presented together with relevant information or inefficient instructional design, which will unnecessarily strain the working memory of the student (Sweller et al., 2011b). Finally, Germane Load (GL) is already existing cognitive resource which can ease the learning e.g., strategies for learning (Sweller et al., 2011b; Ayres, 2018). Some researchers have argued that GL is part of IL (Sweller, 2010; Kalyuga, 2011). Others argue that it makes sense to separate GL from IL and describe how GL is tied to actual effort that leads to a better understanding of the content (e.g., Klepsch et al., 2017), Finally, a recent article by Klepsch and Seufert argues that IL stems from a passive experience of a task, opposite GL that stems from an active experience of a task (Klepsch and Seufert, 2021).

Many attempts at measuring cognitive load have been proposed including objective tasks such as secondary tasks (Sweller et al., 2011c) and psychophysiological measures such as eye tracking (Zheng and Cook, 2012; Scharinger et al., 2020), and EEG (Antonenko et al., 2010; Makransky et al., 2019a; Baceviciute et al., 2020). Recently an article by Minkley, Xu, and Krell have compared subjective and objective factors of CL which found heart rate to be related to self-reported metal effort but not self-reported mental load, and self-reported mental effort and mental load predicted task performance better than heart rate measures (Minkley et al., 2021). However, the most common way to measure cognitive load is through self-report measures.

Previously, a single item measure by Paas (1992) has been widely used and further developed to measure several types of cognitive load (Ayres, 2006; Cierniak et al., 2009). However, single item scales have also been criticized due to several limitations including being too simplistic, making it difficult for learners to make sensible distinctions between the complexity of the material (IL) and inadequate instructions (EL; Kirschner et al., 2011). Several other self-report scales assess cognitive load with multiple items including a scale developed by Klepsch et al. (2017) which assess IL, EL, and GL, a measure to assess mental load and mental effort developed by Krell (2017), in addition to the cognitive load scale by Leppink et al. (2013), which we use in this article. We have chosen to build on the cognitive load scale by Leppink and colleagues as the items assess a broader domain such as a lecture.

In this article we aim to validate a revised version of Leppink and colleagues' Cognitive Load Scale (Leppink et al., 2013; CLS). The CLS has been widely used in educational settings, and several studies provide support for the validity and reliability of the instrument (e.g., Leppink et al., 2013; Hadie and Yusoff, 2016; Andersen and Makransky, 2020). This includes construct validity assessed through exploratory factor analysis (Leppink et al., 2013) or confirmatory factor analyses (Leppink et al., 2013; Hadie and Yusoff, 2016; Andersen and Makransky, 2020) and item response theory (Andersen and Makransky, 2020). The reliability has also been examined, typically through Cronbach's Alpha (Cronbach, 1951) or similar estimates (Leppink et al., 2013; Hadie and Yusoff, 2016; Andersen and Makransky, 2020) and furthermore the external validity has been examined by investigating how the scales are correlated to learning outcomes (Andersen and Makransky, 2020). Although there is mounting evidence of the reliability and construct validity of the CLS, there are still several gaps in this literature. A main gap in the literature are that there may be a need to revisit the content validity of the EL dimension of the CLS, and there is a need to evaluate the sensitivity of these potential dimensions of EL in physical and online lectures.

Regarding the content validity of the EL dimension, a recent study suggests that the EL may be a multidimensional construct consisting of several sub-components. Andersen and Makransky (2020) provide reliability and validity evidence that the EL dimension should be split into three subscales measuring distinct forms of EL in virtual reality environments. The subscales included: EL stemming from instructions (e.g., "The instructions and/or explanations used in the simulation were very unclear"), EL stemming from interaction (e.g., "The interaction technique used in the simulation made it harder to learn"), and EL stemming from the environment (e.g., "The virtual environment was full of irrelevant content"). This multidimensional conceptualization was theorized within immersive environments, but has not been suggested or investigated in traditional teaching environments. In this article we propose that the multidimensional conceptualization of EL is not only relevant in virtual learning environments, but rather that it is also necessary for accurately measuring cognitive load in physical and online lectures. Although, cognitive load theory does not clearly address the idea that disturbances and noises might increase EL (Sweller et al., 2011b), research suggests that multitasking using mobile devices reduce learning (Kuznekoff and Titsworth, 2013; Chen and Yan, 2016). Furthermore, research suggests that noises in learning environments can also influence learning (Ali, 2013; Servilha and Delatti, 2014), thus the idea that noises and disturbances add to EL seems straightforward. Therefore, we have devised items for three subscales to measure EL in relation to physical and online lectures addressing contemporary issues, including noises in the environment or distractions from devices such as mobile phones, which might provoke EL. In addition to the original conceptualization of EL from Leppink et al. (2013) that includes instructions and or explanations (e.g., "*The instructions and/or explanations during the activity were very unclear*"), our theoretical conceptualization of EL includes sub-dimensions stemming from noise (e.g., "*Noises in the environment made it difficult to focus on the learning content*"), and devices (e.g., "*My activities on my phone/computer made it difficult to focus on the learning content*"). Besides the newly developed EL subscales we also employed the Intrinsic Load subscale (e.g., "*The topics covered in the activity were very complex*"), and the Germane Load (GL)

TABLE 1 | Items and scales included in the Study 1.

Scale	
IL	The topics covered in the activity were very complex.
IL	The activity covered theories that I perceived as very complex.
IL	The activity covered concepts and definitions that I perceived as very complex.
EL ins	The instructions and/or explanations during the activity were very unclear.
EL ins	The instructions and/or explanations were, in terms of learning, very ineffective.
El ins	The instructions and/or explanations were full of unclear language.
EL noi	Other students talking in the classroom made it difficult to focus on the learning content.
EL noi	Students talking to me during the activity made learning ineffective.
EL noi	Other noises and distractions during the activity made it hard to learn.
EL dev	My activities on my phone/computer made it difficult to focus on the learning content.
EL dev	Messages and notifications from my phone/computer made learning unclear.
EL dev	Others' phone/computer use distracted me, making it hard to learn.
GL	The activity really enhanced my understanding of the topic(s) covered.
GL	The activity really enhanced my knowledge and understanding of cognitive load.
GL	The activity really enhanced my understanding of the theories covered.
GL	The activity really enhanced my understanding of concepts and definitions.

IL, Intrinsic Load; EL Ins, Extraneous Load Instructions; EL Noi, Extraneous Load Noises; EL Dev, Extraneous Load Devices; GL, Germane Load.

subscales (e.g., "*The activity really enhanced my understanding of the topic(s) covered*"). The new measure is labeled the Multidimensional Cognitive Load Scale for Physical and Online Lectures (MCLS-POL and can be seen in **Table 1**). Although several factors influence teaching in both online (Elkaseh et al., 2015) and offline learning (McKenzie and Schweitzer, 2001; Kappe and van der Flier, 2012) environments, we propose that these components of cognitive load are specifically relevant factors that can influence learning in offline (Klatte et al., 2013; Chen and Yan, 2016; Cerdan et al., 2018) and online lectures (Blasiman et al., 2018; Zureick et al., 2018; Costley et al., 2020). Specifically, with the global COVID-19 pandemic, the use of online teaching platforms is quickly increasing (König et al., 2020) and there is evidence that factors such as noise (Servilha and Delatti, 2014) and disturbances from devices (Chen and Yan, 2016) can create cognitive load when learning.

A related gap in the literature that we attempt to account for in this article is the limited number of studies that investigate the sensitivity of the different dimensions of cognitive load in realistic learning environments. Currently some studies have found meaningful differences of groups in cognitive load (Klepsch et al., 2017; Andersen and Makransky, 2020) and others have found predictive validity through regression analyses (Zukić et al., 2016; Andersen and Makransky, 2020).

In this paper we conduct three studies. In the first study, we validate each sub-scale of the MCLS-POL using the Partial Credit Model (PCM) from Item Response Theory (IRT), and second, we used confirmatory factor Analysis (CFA) to investigate the structural validity of the MCLS-POL. In the second and third studies we implement changes to the scale and investigate the sensitivity of the different sub-dimensions during a lecture in a higher education psychology bachelor course on the topic of educational psychology (Study 2), and a different lecture in a higher education psychology masters course on the topic of psychological testing (Study 3) which both took place in the Fall 2020 semester. Importantly, the setting of Study 2 and Study 3 took place during the COVID-19 pandemic and students were selected to either attend the lecture in person or online via Zoom (Kohnke and Moorhouse, 2020) which gave us the opportunity to investigate if the components of cognitive load differed across settings. Finally, we compared scores across Study 2 and Study 3 to investigate whether the scales would reflect the difference between the two courses. Thus, in this article our aim is to investigate whether it is possible to develop and validate questionnaires measuring cognitive load, particularly the expanded scales of extraneous cognitive load pertaining to EL from instruction, noise, and devices. In regard to comparing online with off-line learning, our research hypotheses are that there should be no differences in terms of intrinsic cognitive load or germane cognitive load between online and offline lectures as the materials and the possible germane resources should be similar. Furthermore, we don't expect any differences in relation to EL from instructions, since students in both online and off-line learning environments receive the same instructions. However, we expected to find differences across the newly developed extraneous cognitive load scales related to devices and noises because there will be differences between the online and off-line learning contexts which could influence these factors of EL.

STUDY 1

Methods Study 1

Sample

Data was collected at a European University during the fall 2019 semester. The psychology students ($N = 250$) were asked to voluntarily answer a short online survey in relation to their current course in educational psychology ($n = 120$) or psychological testing ($n = 130$). A total of 80.8% reported being females ($n = 202$), 18.4% males ($n = 46$), and 0.8% ($n = 2$) reported another gender than male or female. The mean age was 25.46 with an *SD* of 5.45.

Item Development

A team of subject matter experts consisting of an expert in educational psychology, a specialist in human computer interaction, and a psychometrician further developed the scales of Leppink and colleagues' CLS. Based on a previous study where the EL scale was conceptualized using three separate EL subscales aimed at measuring CL in virtual reality (Andersen and Makransky, 2020), we took a similar approach by conceptualizing EL as a multidimensional construct with several subscales. However, instead of being aimed at learning in virtual reality it was aimed at learning during lectures, and was based on the

literature that specified the factors that could create extraneous cognitive load within physical and online lectures. We aimed to make the items so generic that it should be possible to transfer them from one context to another without rewriting them, however in keeping with Leppink et al. (2013) formulation item 2 of the Germane Load scale specifically mentions the course subject (i.e., "The activity really enhanced my knowledge and understanding of [course subject].") and therefore has to be modified accordingly for each study (see **Table 1** for all items used in Study 1).

Statistical Analyses

In this study, we employ two methodologies to investigate the construct validity of the MCLS-POL. The first methodology is that of item response theory (IRT; Embretson and Reise, 2000) which estimates a probability function for endorsing each item of a scale in relation to the scales' total score, that allows for detailed analyses of each item. The second methodology is that of confirmatory factor analysis (CFA; Kline, 2011), in which we model the relationship between items and several latent variables called factors. Therefore, we can evaluate the fit of a model including all items and all scales as latent factors and their relation in just one model.

As the IRT approach is focused on each individual scale, it makes sense to conduct the IRT analyses first and let the knowledge from the IRT analyses inform the overall CFA model which will contain all scales. For an IRT model of the Rasch model family to be valid it must live up to five assumption (Rosenbaum, 1989). The assumptions are: (a) unidimensionality; the scale must measure one latent construct only, (b) the items must be monotonic in relation to the total scale, (c) the items must be locally independent, i.e., the items are conditionally independent after accounting for the total score, (d) the items must not show differential item functioning, e.g., students of the same ability should have equal probability of endorsing an item regardless of gender or age, (e) items must be homogenous such that that the rank order of the items of the difficulties remain the same despite differing abilities of the respondent, e.g., the most difficult item should be the most difficult item to endorse for all respondents. We will address each assumption for every scale in the analyses.

In some cases where we find deviations from assumptions of no Differential Item Functioning (DIF) (d) or no Local Dependence (LD) (c), we are still able to obtain close to optimal measurement. When DIF or LD is uniform, we can model this with a graphical log linear Rasch model (GLLRM; Kreiner and Christensen, 2004). This model can account for the differences in item functioning when DIF is present, however when using sum scores we will need to equate across DIF affected groups to make the sums scores comparable. When uniform LD is present it does not influence the sums scores, however LD dependency will inflate estimates of reliability such as Cronbach's Alpha (Cronbach, 1951) and we will instead use a Monte Carlo method to compute the estimate of reliability (Hamon and Mesbah, 2002). Factor analysis can be used to create a model where each item's relation to the scales is part of a matrix of regressions. In the confirmatory approach, we restrict the model, such that items of a given scale only load on the hypothesized factor and not any of the other factors. This allow us to not only consider the properties of the scales independently as in IRT, but also to investigate if there might be overlap between items across and other scales.

In Study 1 for IRT analyses of the polytomous items of the CL scales, we used the Partial Credit Model (PCM; Masters, 1982) in the Digram program (Kreiner and Nielsen, 2013). An overall test of DIF and homogeneity was conducted with Andersen's conditional likelihood test (Andersen, 1973). Item fit was assessed with item rest score correlations (Christensen and Kreiner, 2013). For the analyses of items-wise DIF in relation to gender, age (grouped by 1 = 0–23, and 2 = 23 and above), and course and LD we used Keldermans' likelihood ration test (Kelderman, 1984) and Goodmann and Kruskal's partial gamma correlation (Kreiner and Christensen, 2004).

For Pure PCM models in Study 1 we used Cronbach's Alpha (Cronbach, 1951) to estimate reliability. For scales with evidence of LD we used a Monte Carlo procedure to estimate the reliability since Cronbach's Alpha is prone to inflation for scales with local dependence. To account for false discovery rates due to the multiple testing we used the Benjamini and Hochberg procedure (Benjamini and Hochberg, 1995). For example, we employed the procedure to test of all possible item pairs in relation to local dependence.

To conduct the CFA we used the Lavaan package (version 0.6-5) in the R statistical programming language (version 3.6.3). To estimate the loading of the model, we used the diagonally least square method (Li, 2016), since the items were ordinal. We used the Comparative Fit Index (CFI) and the Tucker Lewis Index (TLI) with values above 0.95 to indicate acceptable fit (Hu and Bentler, 1999). Besides CFI and TLI we used the Root Mean Square Error of Approximation (RMSEA) and the Standardized Root Mean Square Residual (SRMR) were values below 0.06 and 0.08, indicate a good fit, respectively (Hu and Bentler, 1999).

Results Study 1

Results of Fit to the Partial Credit Model

The Rasch analysis of the IL scale indicated no evidence against the fit to a PCM. The overall test found no evidence of breach of homogeneity, the overall test, or the item-wise tests of DIF in relation to gender, age, or course. There was no evidence against item fit and no evidence of local dependence (see **Table 2**). The reliability of the scale measured in terms of the Cronbach's Alpha was 0.89. Therefore, we concluded that the scale provided valid and reliable measurement.

The analyses of EL instructions scale exhibited evidence of DIF in relation to course for item 1, such that it was easier for students of psychological testing to endorse the statement *"The instructions and/or explanations during the activity were very unclear"* than for students of educational psychology, despite similar levels of EL related to instructions. When the DIF was added to a graphical log linear Rasch model, then neither the overall test nor the item-wise test showed evidence of DIF. There was no evidence of breach of homogeneity, or against item fit. Finally, there was no evidence of local dependence between items and the reliability was 0.84. Therefore, we concluded that the scale provided valid and reliable measurement.

TABLE 2 | Results for the Rasch analyses of the scales in Study 1.

Scale	Overall test	Item fit	DIF gender	DIF age	DIF course	LD	*r*
IL	✓	✓	✓	✓	✓	✓	0.89
EL Ins	✓	✓	✓	✓	%	✓	0.84
EL Noi	%	✓	%	%	%	%	0.81
EL Dev	✓	✓	✓	%	✓	%	0.62
GL	✓	✓	%	✓	✓	%[6]	0.90

IL, Intrinsic Load; EL Ins, Extraneous Load Instructions; EL Noi, Extraneous Load Noises; EL Dev, Extraneous Load Devices; GL, Germane Load; DIF, Differential Item Functioning; LD, Local Dependence; r, reliability. %, unacceptable, where the check mark should be acceptable.

The Analyses of the Extraneous load scale for noise, showed evidence of both DIF and local dependence. Although, no evidence against item fit and with a reliability coefficient of 0.81, it was not possible to find a working model with LD and DIF which could converge, thus there was no fit to a PCM of a GLLRM for the EL N scale.

For the extraneous load scale in relation to devices, we found evidence of DIF in relation to age for item 2, meaning it was easier for younger students to endorse the statement: *"Messages and notifications from my phone/computer made learning unclear."* and local dependence between item 1 *"My activities on my phone/computer made it difficult to focus on the learning content,"* and item 2 "*Messages and notifications from my phone/computer made learning unclear.*" After adding these two deviations from the PCM to a GLLRM, there was no further evidence of DIF or LD, and no evidence against item fit or homogeneity, but the reliability of the scale was only 0.62. Therefore, we conclude that the scale did not fit the PCM, and that we were able to model the DIF and LD, however, the scale had low reliability.

To achieve a working model for the Germane Load scale item 2: *"The activity really enhanced my knowledge and understanding of cognitive load/psychological testing."* was omitted. Further analysis of the remaining three items showed evidence of DIF relative to gender for item 4, Such that it was easier for females to endorse: *"The activity really enhanced my understanding of concepts and definitions,"* despite having the same level of germane load. Furthermore, we found evidence of local dependence between item 3: *"The activity really enhanced my understanding of the theories covered."* and item 4: *"The activity really enhanced my understanding of concepts and definitions."* After these two instances were added to a GLLRM, there was no further evidence of DIF or LD and no evidence against item fit or against homogeneity, and the scale had a reliability of 0.90. Therefore, we concluded that the scale did not fit a pure Partial Credit Model, but could still provide close to optimal measurement after accounting for DIF and LD.

Results of Fit to the Confirmatory Factor Analysis

A Confirmatory Factor analysis was run including all items from all scales with the exception of item 2 from the GL scale because it did not fit the Partial Credit model in the previous analysis, and without the EL Noise scale which did not converge during the PCM analyses. The model grouped each scale's item so they formed a latent construct for each scale, e.g., all IL items loading only on the latent construct of IL. The model achieved acceptable fit values. The CFI was 0.999 and the TLI was 0.999. The RMSEA was <0.001 and SRMR was 0.041, thus all values indicated the model was acceptable.

Discussion Study 1

Overall the IRT and CFA analyses provided positive evidence of the construct validity of the MCLS-POL with few minor cases of LD and DIF which could be modeled. However, three major issues were identified. The first was that the EL Noise scale would not converge to a meaningful model. Adding instances of local dependence to the model led to other instances other local dependence until the program could no longer converge. Second, although the EL devices scales converged to a model after accounting for LD between two items and accounting for DIF, the scales reliability was lower than conventional cut-off for satisfactory reliability. Finally, item 2 from the GL scale had to be eliminated as it did not fit the model. These issues were dealt with in a revision of the MCLS-POL which is described in Study 2. Overall we found evidence against the validity of the EL noise scale, but we found no evidence against the validity of the other scales.

STUDY 2

Study 2 was conducted to improve the MCLS-POL based on the results of Study 1. We were interested in investigating the criterion validity of the different sub-scales within the MCLS-POL in addition to testing the reliability and validity of the measure using the PCM and CFA as in Study 1. Sensitivity was tested by using an experimental design where students experienced a lecture in educational psychology either physically or online through Zoom. An experiment was possible because restrictions due to the COVID-19 pandemic meant that approximately half of the students were assigned to a group who had to follow the lecture online instead of physically in order to increase physical distancing in the lecture hall. The students attending the lecture online followed the same lecture as the students who were physically present, while online the students could choose between seeing just the lecture slides or the teacher in front of the lecture slides, while listening to teacher speak. To examine if the uses of scales scores made sense we used the validity frame work of Kane (2013), and examined whether the scales showed meaningful differences such that online students experienced more EL than off-line students as hypothesized in the introduction.

Item Revision

Before conducting the study, we reformulated the wording of the items for the EL Noise scale so the item content became more general based on the finding that the scale did not fit the PCM in Study 1. For example, we changed the wording of item 1 from "*Other students talking in the classroom made it difficult to focus on the learning content*" to "*Noises in the environment made it difficult to focus on the learning content*" (see **Table 3** for all items used in Study 2 and Study 3). This was also useful as restrictions due to the COVID-19 pandemic meant that approximately half

TABLE 3 | Items and scales included in the Study 2 and Study 3.

Scale	
IL	The topics covered in the activity were very complex.
IL	The activity covered theories that I perceived as very complex.
IL	The activity covered concepts and definitions that I perceived as very complex.
EL Ins	The instructions and/or explanations during the activity were very unclear.
EL Ins	The instructions and/or explanations were, in terms of learning, very ineffective.
EL Ins	The instructions and/or explanations were full of unclear language.
EL Ins	Low quality audio made the instructions hard to follow.
EL Noi	Noises in the environment made it difficult to focus on the learning content.
EL Noi	Distractions in the environment made learning ineffective.
EL Noi	Unrelated events occurring in the environment made it difficult to focus.
EL Dev[a]	My activities on my phone/computer made it difficult to focus on the learning content.
EL Dev[a]	Messages and notifications from my phone/computer made learning unclear.
EL Dev[b]	Others' phone/computer use distracted me, making it hard to learn.
EL Dev	Technical issues made learning ineffective.
EL Dev	Problems with technology made it difficult to focus.
GL	The activity really enhanced my understanding of the topic(s) covered.
GL	The activity really enhanced my knowledge and understanding of [course subject].
GL	The activity really enhanced my understanding of the theories covered.
GL	The activity really enhanced my understanding of concepts and definitions.

IL, Intrinsic Load; EL Ins, Extraneous Load Instructions; EL Noi, Extraneous Load Noises; EL Dev, Extraneous Load Devices; GL, Germane Load.
[a]*Combined to make an EL Media scale.*
[b]*Omitted from final analyses.*

of the students were assigned to a group who had to follow the lecture online to increase physical distancing in the lecture hall. Given the difference between participating in a lecture physically and online we expected differences in the EL sub-scales of Noise, and Devices. We also added two more items to the EL Devices sub-scale as the results from Study 1 indicated that the scale had a low reliability.

Sample

Data were collected at a European University during the fall 2020 semester. The psychology students ($N = 140$) were asked to voluntarily answer a short online survey in relation to their current course in educational psychology. A total of 76.4% reported being females ($n = 107$), 22.9% males ($n = 32$), and 0.7% ($n = 1$) did not wish to answer the question. The mean age was 23.29 with a *SD* of 3.83. Due to the COVID-19 pandemic, the University restricted the number of students who could attend the lecture physically and students were assigned to either attend in person or online through Zoom prior to the lecture. A total of 62 students attended the lecture in person, the rest of the students attended the lecture through the Zoom online streaming service ($n = 78$). The students experienced the same lecture with the only difference being their presence in the classroom, or experiencing it online through Zoom. The MCLS-POL was administered at the end of the lecture through SurveyMonkey.

Statistical Analyses

In Study 2 for IRT analyses of the polytomous items of the CL scales, we used the Partial Credit Model (PCM; Masters, 1982) in RUMM (Andrich et al., 2003), the switch from Digram to RUMM was made as RUMM is able to handle scales based on only two items which became a necessity in Study 2. An overall test of fit to the PCM was conducted with a chi-square test, where significance indicate misfit in relation to the model (Pallant and Tennant, 2007). Item fit was deemed acceptable if the residuals of the models were within −2.5 and +2.5 (Pallant and Tennant, 2007). Local dependence was assessed by examining the residual correlations between items, where we expected the residual correlation to be close to zero. We used items residuals above 0.20 as indicative of local dependence (Christensen et al., 2017). The presence of DIF was examined through analysis of variance in items scores across age, gender, and whether the student was present physically or attended the lecture online, in cases where we tested with multiple items we corrected the *p*-values with the Bonferroni correction to adjust for false discovery rates.

Results for Study 2

Results of Fit to the Partial Credit Model

The Rasch analyses of the five scales provide almost no evidence against fit in the overall test or in relation to item fit. A minor deviation was found for the IL scale as the overall test rejected at fit ($p < 0.001$) however this might be due to item 2 which fit to the PCM was rejected at $p > 0.05$ but not $p > 0.01$ after Bonferroni correction. Furthermore, the residuals for the item fit was between −2.5 and 2.5, thus we concluded the scale fit.

The only major deviation from the model was related to the EL Devices scale where we identified strong evidence of multidimensionality. After reexamining the wording of the items it was clear that the items were assessing two separate constructs: One measuring the EL from Media with the following items (item 1, "*My activities on my phone/computer made it difficult to focus on the learning content*" and item 2, " *Messages and notifications from my phone/computer made learning unclear*") and the second measuring EL from devices with the following items (item 4, "*Technical issues made learning ineffective*" and item 5, "*Problems with technology made it difficult to focus*"). Furthermore, item 3 did not fit any of the scales and was eliminated. After the split both scales fit the model and for all scales the reliability was satisfactory (see **Table 6**).

For the other sub-scales, we only found evidence of one instance of DIF depending on whether the students attended the course physically or online in two items on the EL Instructions scale, such that is was easier for physically present students to endorse the statement in item 2 "*The instructions and/or explanations were, in terms of learning, very ineffective*" than the students attending the lecture online. Opposite of this it was easier for the online students to endorse the statement of item 4 "*Low quality audio made the instructions hard to follow,*" than for the physically present students, despite having similar levels of EL in relation to instructions. For the GL scale we again omitted item

TABLE 4 | Results for the Rasch analyses of the scales in Study 2 in RUMM.

Scale	Overall test	Item fit	DIF gender	DIF age	DIF location	LD	*r*
IL	%	✓	✓	✓	✓	✓	0.86
EL Ins	✓	✓	✓	✓	%	✓	0.73
EL Noi	✓	✓	✓	✓	✓	✓	0.85
EL Med	✓	✓	✓	✓	✓	✓	0.85
EL Dev	✓	✓	✓	✓	✓	✓	0.87
GL	✓	✓	✓	✓	✓	✓	0.88

IL, Intrinsic Load; EL Ins, Extraneous Load Instructions; EL Noi, Extraneous Load noise; EL Med, Extraneous Load Media; EL Dev, Extraneous Load devices; GL, Germane Load; DIF, Differential Item Functioning; LD, Local Dependence; r, reliability. %, unacceptable, where the check mark should be acceptable.

TABLE 5 | A comparison of the students who attended the course online and those who were physically present on the scales in the MCLS-POL in Study 2.

Scale	Online		Physical present		$t_{(138)}$	*p*	Cohen' s *d*
	M	*SD*	*M*	*SD*			
IL	2.773	0.717	2.586	0.731	−1.281	0.202	0.259
EL Ins	2.234	0.718	1.762	0.505	−4.558	<0.001	0.746
EL Noi	2.368	0.962	1.591	0.614	−5.795	<0.001	0.940
EL Med	2.820	1.165	2.226	0.904	−3.401	<0.001	0.746
EL Dev	2.532	1.070	2.016	0.784	−3.290	<0.001	0.541
GL	3.724	0.713	3.944	0.538	2.071	0.040	0.343

IL, Intrinsic Load; EL Ins, Extraneous Load instructions; EL Noi, Extraneous Load Noises; EL Med, Extraneous Load Media; EL Dev, Extraneous Load Devices; GL, Germane.

TABLE 6 | Results for the Rasch analyses of the scales in Study 3 RUMM.

Scale	Overall test	Item fit	DIF gender	DIF age	DIF zoom	LD	*R*
IL	✓	✓	✓	✓	✓	✓	0.93
EL Ins	✓	✓	✓	✓	✓	✓	0.84
EL Noi	✓	✓	✓	✓	%	✓	0.81
EL Med	✓	✓	✓	✓	✓	✓	0.85
EL Dev	✓	✓	✓	✓	✓	✓	0.90
GL	✓	✓	✓	✓	✓	✓	0.90

IL, Intrinsic Load; EL Ins, Extraneous Load Instructions; EL Noi, Extraneous Load noise; EL Med, Extraneous Load Media EL Dev, Extraneous Load Devices; GL, Germane Load; DIF, Differential Item Functioning; LD, Local Dependence; r, reliability. %, unacceptable, where the check mark should be acceptable.

2 to achieve a working model, the *p*-value for item fit of item 3 in the GL scale was 0.0163 (Bonferroni corrected cut-off was 0.016). However, as the residuals was inside −2.5 to + 2.5 we accepted it. **Table 4** illustrates how all other items fit the PCM providing evidence of the validity and reliability of the revised version of the MCLS-POL.

External Validity Results

Table 5 shows the difference between being physically present and attending the lecture online. Independent samples *t*-tests were conducted to investigate if the differences between the physical and online groups were significant. For the IL scale there was no significant difference between being physically present or attend online [$t_{(138)} = -1.281, p = 0.202$] as expected. Furthermore, EL Noise was significantly higher [$t_{(138)} = -5.795$, $p < 0.001$] for online students (M =2.368, SD = 0.962) than for physically present students (M = 1.591, SD = 0.614). EL Media was significantly higher [$t_{(138)} = -3.401$, $p < 0.001$] for online students (M =2.820, SD =1.165) than for physically present students (M = 2.226, SD = 0.904). EL Devices was significantly higher [$t_{(138)} = -3.290$, $p < 0.001$] for online student (M = 2.532, SD = 1.070) than for physically present students (M = 2.016, SD = *0.*784). All of these fit the a-priori hypotheses. However, contrary to the a-priori predictions EL Instructions was significantly higher [$t_{(138)} = -4.558$, $p < 0.001$] for online student (M = 2.234, SD = 0.718) than for physically present students (M =1.762, SD = 0.505). The online students (M = 3.724, SD = 0.713) also experienced significantly [$t_{(138)} = 2.071$, $p = 0.040$] lower GL than the physically present students (M = 3.944, SD = 0.538). These results suggest that the MCLS-POL is sensitive to differences between students learning in different environments and provides support for the external validity of the measure. However, students also reported different levels of EL related to instructions and GL which was not expected in the a-priori predictions.

Discussion Study 2

Study 2 revealed that the EL Devices scale should be split into two sub-scales in order to create valid measurement. A meaningful categorization was made by creating an EL Media subscale, and an EL Devices subscale. The EL Media sub-scale consisted of the item "*My activities on my phone/computer made it difficult to focus on the learning content*" and the item "*Messages and notifications from my phone/computer made learning unclear.*" The EL Devices sub-scale consisted of the item "*Technical issues made learning ineffective*" and the item "*Problems with technology made it difficult to focus.*" Due to the item wording (i.e., the first two items pertaining to disturbances from the devices and the second two items pertaining to technology in general) it made sense to split the scale into two distinct scales. Furthermore, comparing the students based on whether they were physically present or attending the lecture online revealed meaningful differences such that the students attending the lecture online experience significantly more EL related to instructions, noise, media, and devices, as well as significantly less GL than the students who were physically present in the lecture hall. The difference between the groups on IL was not significant. To investigate if these results would replicate in a new setting we conducted a follow-up study.

STUDY 3

Study 3 was conducted to test the validity of the MCLS-POL in a new context by replicating Study 2 in a sample of psychology master students who were participating in a lecture about psychological testing. The same statistical analyses were conducted as in Study 2.

Sample

Data was collected at a European University during the fall 2020 semester. The psychology master students (N = 119) were

TABLE 7 | A comparison of the students who attended the course online and those who were physically present on the scales in the MCLS-POL in Study 3.

Scale	Online		Physical present		$t_{(117)}$	p	Cohen' s d
	M	*SD*	*M*	*SD*			
IL	3.181	0.836	3.200	0.718	0.092	0.927	0.023
EL Ins	2.544	0.745	2.463	0.814	−0.848	0.629	0.106
EL Noi	2.370	0.950	1.741	0.669	−3.878	<0.001	0.703
EL Med	2.696	1.021	2.333	1.047	−1.608	0.111	0.354
EL Dev	2.179	1.015	1.740	0.764	−2.076	0.040	0.455
GL	3.266	0.677	3.037	0.822	−1.471	0.144	0.322

IL, Intrinsic Load; EL Ins, Extraneous Load Instructions; EL Noi, Extraneous Load Noises; EL Med, Extraneous Load Media; EL Dev, Extraneous Load Devices; GL, Germane Load.

asked to voluntarily answer a short online survey in relation to a course in psychological testing. A total of 89.1% reported being females ($n = 106$), and 10.9% males ($n = 13$). The mean age was 27.25 with an *SD* of 6.65. Similar to Study 2, students were assigned to attending the course physically or online prior to the lecture but students who were allowed to attend physically were given the option of attending online. A total of 27 students attended the lecture in person, the rest of the students attended the lecture through the Zoom online streaming service ($n = 92$). The students attending in person had to wear a mask while entering the lecture hall, which they could remove while seated and they had to sit with distance between them. The teacher also had to wear a mask when entering the lecture hall, but the teacher was allowed to remove the mask during the lecture.

Results for Study 3

Results of fit to the Partial Credit Model

The Rasch analyses of the six scales provide almost no evidence against fit in the overall test or in relation to item fit. We only found evidence DIF depending on whether the students attended the course online or by being physically present in relation to two items on the EL Noise scale, such that it was easier for physically present students to endorse "*Noises in the environment made it difficult to focus on the learning content,*" while it was easier for student attending online to endorse "*Distractions in the environment made learning ineffective,*" despite having the same level of EL related to noise. Again we split the EL device scale into two separate two-item scales. One measuring the EL from media and the second measuring EL from devices. After the split both scales fit the model and for all scales the reliability was above 0.80 and thus satisfactory (see **Table 6**).

External Validity Results

Table 7 shows the difference between being physically present and attending the lecture online. The EL noise and the EL devices scales showed significant differences across type of attendance as predicted. For the EL Noises scale scales the students who attended the lecture online ($M = 2.370$, $SD = 0.745$) experienced significantly [$t_{(117)} = -3.878$, $p < 0.001$] more EL related to noises than the students who were physically present ($M = 1.741$, $SD = 0.669$). Similarly, on the EL Devices scale the students who attended the lecture online ($M = 2.179$, $SD = 1.015$) experienced significantly [$t_{(117)} = -2.076$, $p = 0.040$] more EL related to devices than the students who were physically present ($M = 1.740$, $SD = 0.764$). However, although the online lecture group ($M = 2.696$, $SD = 1.021$) also experienced more EL related to media than the students who were physically present ($M = 2.333$, $SD = 1.047$) this difference did not reach statistical significance [$t_{(117)} = -1.608$, $p = 0.111$]. Finally, the difference between the groups on IL, EL related to instructions, and GL were not statistically significant as predicted. The results suggest that the EL noise and EL devices scales within the MCLS-POL is sensitive to differences between students learning in different environments and provides support for the external validity of the measure.

Discussion Study 3

Study 3 showed that all six scales provide valid measurement. Furthermore, comparing the students based on whether they were physically present or attending the lecture online revealed differences in a way that the students attending the lecture online experience significantly more EL related to noises and devices than the students who were physically present, other than those two scales there were no significant difference in the experienced CL. These results suggest that it is important to have a multidimensional conceptualization of EL as different components of EL can influence learning in physical and online environments differently.

It is not immediately clear why the differences in mean between students attending the lecture physically and students attending online in Study 3 are not similar to that of students in Study 2. One explanation might be that the master students attending Psychological Testing were more accustomed to lectures than the bachelor students attending Educational Psychology. Thus, we might reason that more experienced learners are less hampered by different types of extraneous load despite attending lectures online.

RESULTS OF COMBINING DATA FROM STUDY 2 AND STUDY 3

Results of the Confirmatory Factor Analysis

Before comparing the sum scores of Study 2 and Study 3 we conducted a confirmatory factor analysis with all items loading on their respective factors by combining data from Study 2 and Study 3. The fit indices were CFI = 0.99, TLI = 0.99, and both RMSEA and SRMR = 0.06 which are all satisfactory for the six factor model.

Comparing Cognitive Load Across Study 2 and Study 3

Although the samples of psychology students in Study 2 and Study 3 differed because the students in the educational psychology course were second year bachelor students and the students in the psychological testing course were master students, we were interested in comparing the cognitive load ratings across the studies. Since psychological testing is considered a

TABLE 8 | Difference between scores on the MCLS-POL in Study 2 and Study 3.

Scale/course	Educational psychology		Psychological testing		*t*	*p*	Cohen' s *d*
	M	*SD*	*M*	*SD*			
IL	2.67	0.72	3.18	0.81	−5.317	<0.001	0.67
EL Ins	2.03	0.67	2.53	0.76	−5.625	<0.001	0.70
EL Noi	2.02	0.91	2.23	0.93	−1.775	0.077	0.23
EL Med	2.56	1.09	2.61	1.04	−0.423	0.673	0.05
EL Dev	2.30	0.98	2.08	0.98	1.827	0.069	0.22
GL	3.82	0.65	3.21	0.72	7.158	<0.001	0.89

IL, Intrinsic Load; EL Ins, Extraneous Load Instructions; EL Noi, Extraneous Load Noises; EL Med, Extraneous Load Media; EL Dev, Extraneous Load Devices; GL, Germane Load.

more cognitive straining course by many students, we wanted to compare the scales across the two courses. To ensure the scales were comparable across the studies we performed a CFA which yielded satisfactory fit indices. When comparing across courses we found IL to be significantly [$t_{(257)} = -5.317$, $p < 0.001$] higher for students from psychological testing ($M = 3.18$, $SD = 0.81$) than for students of educational psychology($M = 2.67$, $SD = 0.72$). Similarly, We found EL instruction to be significantly [$t_{(257)} = -5.625$, $p < 0.001$] higher for students from psychological testing ($M = 2.53$, $SD = 0.76$) than for students of educational psychology ($M = 2.03$, $SD = 0.67$). On the other hand GL was significantly [$t_{(257)} = 7.158$, $p < 0.001$] lower for students of psychological testing($M = 3.21$, $SD = 0.65$) than for students of educational psychology ($M = 3.82$, $SD = 0.65$, see **Table 8**). The differences between the groups were not significant for the other scales.

CONCLUSION

Through three studies we describe the further development and validation of the MCLS-POL. We provide evidence of the validity and reliability of the expanded CLS which supports the multidimensional conceptualization of cognitive load. Overall there was evidence of meaningful external validity in terms of meaningful group difference between students experiencing a lecture physically and students who experience the same lecture online. However, since one of the scale was split into two separate scale with only two items each, we highly recommend that researchers wishing to use these scale enhance these two scales by adding more items to them.

The studies also reveals meaningful challenges that students as well as lectures face as more teaching is conducted online. The results from Study 2 revealed that the students attending the lectures online experienced more EL on all four EL load sub-scales, meaning that they experienced extraneous cognitive load related to instructions, noises, media, and devices, which was greater than what the students who were physically present experienced. Furthermore, the physically present students in Study 2 also reported higher GL than the students who attended the lecture online.

The students attending the lecture online in Study 3 similarly experienced more cognitive load related to EL from media and devices than the physically present students. The finding that there were no differences on IL between the students who experienced the lecture physically or online in Studies 2 and 3, but there were differences in different components of EL suggests that the MCLS-POL is sensitive at identifying different components of cognitive load.

A limitation of these studies is the lack of measurement of learning, and the subsequent hypothetical analyses of differences in learning across the students attending either online or off-line and correlations between the CL scales and learning outcome. However, as we examined the cognitive load across differing course, creating a measure of learning with similar properties across courses was difficult. Future studies might address this by examining learning in just one type of course.

The reason for providing online lectures in Studies 2 and 3 was due to the extraordinary consequences of the COVID-19 pandemic in 2020. However, many universities are aiming at providing more online lectures, due to several advantages such as accessibility issues e.g., the ability to reach more students and allow students to access high quality educational opportunities even though they are unable to be physically present (Waschull, 2001; Cascaval et al., 2008; French and Kennedy, 2017; Makransky et al., 2019b). In contrast to the benefits this article highlights some of the caveats of online teaching related to the strain it can provide for students in terms of more EL, which should be addressed when conducting online lectures.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

MA collected data conducted analyses and wrote the article. GM collected data and supervised the analyses and the writing. All authors contributed to the article and approved the submitted version.

REFERENCES

Ali, S. A. A. (2013). Study effects of school noise on learning achievement and annoyance in Assiut City, Egypt. *Appl. Acoust.* 74, 602–606. doi: 10.1016/j.apacoust.2012.10.011

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika* 38, 123–140. doi: 10.1007/BF02291180

Andersen, M. S., and Makransky, G. (2020). The validation and further development of a multidimensional cognitive load scale for virtual environments. *J. Comput. Assist. Learn.* 37, 183–196. doi: 10.1111/jcal.12478

Andrich, D., Sheridan, B. and Luo, G. (2003). *RUMM2030: Rasch Unidimensional Models for Measurement*. Perth: RUMM Laboratory.

Antonenko, P., Paas, F., Grabner, R., and van Gog, T. (2010). Using electroencephalography to measure cognitive load. *Educ. Psychol. Rev.* 22, 425–438. doi: 10.1007/s10648-010-9130-y

Ayres, P. (2006). Impact of reducing intrinsic cognitive load on learning in a mathematical domain. *Appl. Cogn. Psychol.* 20, 287–298. doi: 10.1002/acp.1245

Ayres, P. (2018). "Subjective measures of cognitive load-what can they reliably measure?" in *Cognitive Load Measurement and Application-A Theoretical Framework for Meaningful Research and Practice, 1st Edn*. eds R. Z. Zheng (New York, NY: Routledge), 9–28. doi: 10.4324/9781315296258-2

Baceviciute, S., Mottelson, A., Terkildsen, T., and Makransky, G. (2020). "Investigating representation of text and audio in educational VR using learning outcomes and EEG," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–13. doi: 10.1145/3313831.3376872

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* 57, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Blasiman, R. N., Larabee, D., and Fabry, D. (2018). Distracted students: a comparison of multiple types of distractions on learning in online lectures. *Scholarsh. Teach. Learn. Psychol.* 4, 222–230. doi: 10.1037/stl0000122

Cascaval, R. C., Fogler, K. A., Abrams, G. D., and Durham, R. L. (2008). Evaluating the benefits of providing archived online lectures to in-class math students. *J. Asynchron. Learn. Netw.* 12, 61–70. doi: 10.24059/olj.v12i3.65

Cerdan, R., Candel, C., and Leppink, J. (2018). Cognitive load and learning in the study of multiple documents. *Front. Educ.* 3:59. doi: 10.3389/feduc.2018.00059

Chen, Q., and Yan, Z. (2016). Does multitasking with mobile phones affect learning? A review. *Comput. Hum. Behav.* 54, 34–42. doi: 10.1016/j.chb.2015.07.047

Chi, M. T. H., Glaser, R., and Rees, E. (1982). "Expertise in problem solving," in *Advances in the Psychology of Human Intelligence*, Vol. 1, ed R. J. Ternberg (Lawrence Erlbaum Associates, Inc.), 7–76.

Christensen, K. B., and Kreiner, S. (2013). "Item fit statistics," in *Rasch Models in Health,* Vol. 2013, eds K. B. Christensen, S. Kreiner, and M. Mesbah (London, UK: ISTE and John Wiley and Sons, Inc.), 83–104. doi: 10.1002/9781118574454.ch5

Christensen, K. B., Makransky, G., and Horton, M. (2017). Critical values for Yen's Q3: identification of local dependence in the Rasch model using residual correlations. *Appl. Psychol. Meas.* 41, 178–194. doi: 10.1177/0146621616677520

Cierniak, G., Gerjets, P., and Scheiter, K. (2009). "Expertise reversal in multimedia learning: subjective load ratings and viewing behavior as cognitive process indicators," in *Proceedings of the Annual Meeting of the Cognitive Science Society* (Tübingen), 31.

Costley, J., Fanguy, M., Lange, C., and Baldwin, M. (2020). The effects of video lecture viewing strategies on cognitive load. *J. Comput. Higher Educ.* doi: 10.1007/s12528-020-09254-y

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika* 16, 297–334. doi: 10.1007/BF02310555

Elkaseh, A., Wong, K. W., and Fung, C. C. (2015). A review of the critical success factors of implementing e-learning in higher education. *Int. J. Technol. Learn.* 21, 1–13. doi: 10.18848/2327-0144/CGP/v22i02/49160

Embretson, S. E., and Reise, S. P. (2000). *Item Response Theory for Psychologists*. New York, NY: Lawrence Erlbaum Associates Publishers.

French, S., and Kennedy, G. (2017). Reassessing the value of University lectures. *Teach. Higher Educ.* 22, 639–654. doi: 10.1080/13562517.2016.1273213

Hadie, S. N. H., and Yusoff, M. S. B. (2016). Assessing the validity of the cognitive load scale in a problem-based learning setting. *J. Taibah Univ. Med. Sci.* 11, 194–202. doi: 10.1016/j.jtumed.2016.04.001

Hamon, A., and Mesbah, M. (2002). "Questionaires reliability under the Rasch model," in *Statistical Methods for Quality of Life Studies: Design, Measurements and Analysis*, eds M. Mesbah, B. F. Cole, and M.-L. T. Lee (Boston, MA: Kluwer Academic Publishers). 155–168. doi: 10.1007/978-1-4757-3625-0_13

Hu, L., and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Struct. Equat. Model. Multidiscipl. J.* 6, 1–55. doi: 10.1080/10705519909540118

Kalyuga, S. (2011). Cognitive load theory: how many types of load does it really need? *Educ. Psychol. Rev.* 23, 1–19. doi: 10.1007/s10648-010-9150-7

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *J. Educ. Meas.* 50, 1–73. doi: 10.1111/jedm.12000

Kappe, R., and van der Flier, H. (2012). Predicting academic success in higher education: what's more important than being smart? *Eur. J. Psychol. Educ.* 27, 605–619. doi: 10.1007/s10212-011-0099-9

Kelderman, H. (1984). Loglinear Rasch model tests. *Psychometrika* 49, 223–245. doi: 10.1007/BF02294174

Kirschner, P. A., Ayres, P., and Chandler, P. (2011). Contemporary cognitive load theory research: the good, the bad and the ugly. *Comput. Hum. Behav.* 27, 99–105. doi: 10.1016/j.chb.2010.06.025

Klatte, M., Bergstroem, K., and Lachmann, T. (2013). Does noise affect learning? A short review on noise effects on cognitive performance in children. *Front. Psychol.* 4:578. doi: 10.3389/fpsyg.2013.00578

Klepsch, M., Schmitz, F., and Seufert, T. (2017). Development and validation of two instruments measuring intrinsic, extraneous, and germane cognitive load. *Front. Psychol.* 8:1997. doi: 10.3389/fpsyg.2017.01997

Klepsch, M., and Seufert, T. (2021). Making an effort versus experiencing load. *Front. Educ.* 6. doi: 10.3389/feduc.2021.645284

Kline, R. B. (2011). *Principles and Practice of Structural Equation Modeling*. New York, NY: Guilford Publications.

Kohnke, L., and Moorhouse, B. L. (2020). Facilitating synchronous online language learning through zoom. *RELC J.* doi: 10.1177/0033688220937235

König, J., Jäger-Biela, D. J., and Glutsch, N. (2020). Adapting to online teaching during COVID-19 school closure: teacher education and teacher competence effects among early career teachers in Germany. *Eur. J. Teacher Educ.* 43, 608–622. doi: 10.1080/02619768.2020.1809650

Kreiner, S., and Christensen, K. B. (2004). Analysis of local dependence and multidimensionality in graphical loglinear Rasch models. *Commun. Stat. Theory Methods* 33, 1239–1276. doi: 10.1081/STA-120030148

Kreiner, S., and Nielsen, T. (2013). *Item Analysis in DIGRAM 3.04: Part I: Guided Tours [Research Report]*. Department of Biostastistics, University of Copenhagen.

Krell, M. (2017). Evaluating an instrument to measure mental load and mental effort considering different sources of validity evidence. *Cogent Educ.* 4:1280256. doi: 10.1080/2331186X.2017.1280256

Kuznekoff, J. H., and Titsworth, S. (2013). The impact of mobile phone usage on student learning. *Commun. Educ.* 62, 233–252. doi: 10.1080/03634523.2013.767917

Leppink, J., Paas, F., Van der Vleuten, C. P. M., Van Gog, T., and Van Merriënboer, J. J. G. (2013). Development of an instrument for measuring different types of cognitive load. *Behav. Res. Methods* 45, 1058–1072. doi: 10.3758/s13428-013-0334-1

Li, C.-H. (2016). The performance of ML, DWLS, and ULS estimation with robust corrections in structural equation models with ordinal variables. *Psychol. Methods* 21, 369–387. doi: 10.1037/met0000093

Makransky, G., Mayer, R. E., Veitch, N., Hood, M., Christensen, K. B., and Gadegaard, H. (2019a). Equivalence of using a desktop virtual reality science simulation at home and in class. *PLoS ONE* 14:e0214944. doi: 10.1371/journal.pone.0214944

Makransky, G., Terkildsen, T. S., and Mayer, R. E. (2019b). Role of subjective and objective measures of cognitive processing during learning in explaining the spatial contiguity effect. *Learn. Instruct.* 61, 23–34. doi: 10.1016/j.learninstruc.2018.12.001

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika* 47, 149–174. doi: 10.1007/BF02296272

McKenzie, K., and Schweitzer, R. (2001). Who succeeds at University? Factors predicting academic performance in first year Australian university students. *Higher Educ. Res. Dev.* 20, 21–33. doi: 10.1080/07924360120043621

Minkley, N., Xu, K., and Krell, M. (2021). Analyzing relationships between causal and assessment factors of cognitive load: associations between objective and subjective measures of cognitive load, stress, interest, and self-concept. *Front. Educ.* 6. doi: 10.3389/feduc.2021.632907

Paas, F. G. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: a cognitive-load approach. *J. Educ. Psychol.* 84, 429–434. doi: 10.1037/0022-0663.84.4.429

Pallant, J. F., and Tennant, A. (2007). An introduction to the Rasch measurement model: an example using the Hospital Anxiety and Depression Scale (HADS). *Brit. J. Clin. Psychol.* 46(Pt 1), 1–18. doi: 10.1348/014466506X96931

Rosenbaum, P. R. (1989). Criterion-related construct validity. *Psychometrika* 54, 625–633. doi: 10.1007/BF02296400

Scharinger, C., Schüler, A., and Gerjets, P. (2020). Using eye-tracking and EEG to study the mental processing demands during learning of text-picture combinations. *Int. J. Psychophysiol.* 158, 201–214. doi: 10.1016/j.ijpsycho.2020.09.014

Servilha, E. A. M., and Delatti, M. D. A. (2014). College students' perception of classroom noise and its consequences on learning quality. *Audiol. Commun. Res.* 19, 138–144. doi: 10.1590/S2317-64312014000200007

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

Sweller, J., Ayres, P., and Kalyuga, S. (2011a). "Categories of knowledge: an evolutionary approach," in *Cognitive Load Theory,* eds J. Sweller, P. Ayres, and S. Kalyuga (New York, NY: Springer), 3–14. doi: 10.1007/978-1-4419-8126-4_1

Sweller, J., Ayres, P., and Kalyuga, S. (2011b). "Intrinsic and extraneous cognitive load," in *Cognitive Load Theory,* eds J. Sweller, P. Ayres, and S. Kalyuga (New York, NY: Springer), 57–69). doi: 10.1007/978-1-4419-8126-4_5

Sweller, J., Ayres, P., and Kalyuga, S. (2011c). "Measuring cognitive load," in *Cognitive Load Theory,* eds J. Sweller, P. Ayres, and S. Kalyuga (New York, NY: Springer), 71–85. doi: 10.1007/978-1-4419-8126-4_6

Sweller, J., van Merrienboer, J. J. G., and Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. *Educ. Psychol. Review* 10, 251–296. doi: 10.1023/A:1022193728205

Tindall-Ford, S., Agostinho, S., and Sweller, J. (eds.). (2019). *Advances in Cognitive Load Theory: Rethinking Teaching* New York, NY: Routledge. doi: 10.4324/9780429283895

Waschull, S. B. (2001). The online delivery of psychology courses: attrition, performance, and evaluation. *Teach. Psychol.* 28, 143–147. doi: 10.1207/S15328023TOP2802_15

Zheng, R., and Cook, A. (2012). Solving complex problems: a convergent approach to cognitive load measurement. *Brit. J. Educ. Technol.* 43, 233–246. doi: 10.1111/j.1467-8535.2010.01169.x

Zukić, M., apo, N., and Husremović, D. (2016). Construct and predictive validity of an instrument for measuring intrinsic, extraneous and germane cognitive load. *Univ. J. Psychol.* 4, 242–248. doi: 10.13189/ujp.2016.040505

Zureick, A. H., Burk-Rafel, J., Purkiss, J. A., and Hortsch, M. (2018). The interrupted learner: how distractions during live and video lectures influence learning outcomes. *Anat. Sci. Educ.* 11, 366–376. doi: 10.1002/ase.1754

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Education

ORIGINAL RESEARCH
published: 12 April 2021
doi: 10.3389/feduc.2021.632907

Analyzing Relationships Between Causal and Assessment Factors of Cognitive Load: Associations Between Objective and Subjective Measures of Cognitive Load, Stress, Interest, and Self-Concept

Nina Minkley[1]*, *Kate M. Xu*[2] and *Moritz Krell*[3]

[1]Behavioral Biology and Biology Education, Faculty of Biology and Biotechnology, Ruhr-Universität Bochum, Bochum, Germany, [2]Faculty of Educational Sciences, Open University of the Netherlands, Heerlen, Netherlands, [3]Biology Education, Freie Universität Berlin, Berlin, Germany

OPEN ACCESS

Edited by:
Sedat Sen,
Harran University, Turkey

Reviewed by:
Mary Roduta Roberts,
University of Alberta, Canada
H. Cigdem Bulut,
Çukurova University, Turkey

***Correspondence:**
Nina Minkley
nina.minkley@rub.de

Specialty section:
This article was submitted to Assessment, Testing and Applied Measurement, a section of the journal Frontiers in Education

Received: *24 November 2020*
Accepted: *08 February 2021*
Published: *12 April 2021*

Citation:
Minkley N, Xu KM and Krell M (2021) Analyzing Relationships Between Causal and Assessment Factors of Cognitive Load: Associations Between Objective and Subjective Measures of Cognitive Load, Stress, Interest, and Self-Concept. Front. Educ. 6:632907. doi: 10.3389/feduc.2021.632907

The present study is based on a theoretical framework of cognitive load that distinguishes causal factors (learner characteristics affecting cognitive load e.g., self-concept; interest; perceived stress) and assessment factors (indicators of cognitive load e.g., mental load; mental effort; task performance) of cognitive load. Various assessment approaches have been used in empirical research to measure cognitive load during task performance. The most common methods are subjective self-reported questionnaires; only occasionally objective physiological measures such as heart rates are used. However, the convergence of subjective and objective approaches has not been extensively investigated yet, leaving unclear the meaning of each kind of measure and its validity. This study adds to this body of research by analyzing the relationship between these causal and assessment (subjective and objective) factors of cognitive load. The data come from three comparable studies in which high school students (N = 309) participated in a one-day out of school molecular biology project and completed different tasks about molecular biology structures and procedures. Heart rate variability (objective cognitive load) was measured via a chest belt. Subjective cognitive load (i.e., mental load and mental effort) and causal factors including self-concept, interest, and perceived stress were self-reported by participants on questionnaires. The findings show that a) objective heart rate measures of cognitive load are related to subjective measures of self-reported mental effort but not of mental load; b) self-reported mental effort and mental load are better predictors of task performance than objective heart rate measures of cognitive load; c) self-concept, interest and perceived stress are associated with self-reported measures of mental load and mental effort, and self-concept is associated with one of the objective heart rate measures. The findings are discussed based on the theoretical framework of cognitive load and implications for the validity of each measure are proposed.

Keywords: cognitive load, causal and assessment factors, stress, interest, self-concept, heart rate, mental load, mental effort

INTRODUCTION

Cognitive load can be broadly defined as a psychological construct representing an individual's cognitive resources used to learn or perform a task. As such, cognitive load is an established construct in education and psychology, often used as a guidance to optimize instructional designs (e.g., Paas and van Merriënboer, 1994; Kirschner et al., 2006) and is considered as a control variable in assessment contexts (e.g. Minkley et al., 2018; Nehring et al., 2012). It is assumed that measures of cognitive load under various experimental conditions represent the working memory resources exerted or required during task performance. Within assessment contexts but traditionally also within instructional design contexts, cognitive load has been conceptualized in terms of the perceived complexity of tasks (mental load) and the invested mental effort while working on the tasks (e.g., Paas and van Merriënboer, 1994; Choi et al., 2014; Krell, 2017; Skuballa et al., 2019). Mental load refers to the amount of cognitive resources required to solve the problem, whereas mental effort refers to the cognitive sources that are actually invested during problem solving. This theoretical distinction is powerful because it allows to separate internal and external dimensions of cognitive load and can guide further research (Paas and van Merriënboer, 1994; Choi et al., 2014). However, mental load and mental effort are typically assessed using subjective self-reports on questionnaires (e.g., Krell, 2017), which assumes that the respondents are aware of their actual amount of cognitive load, which they invested to solve a task (Solhjoo et al., 2019). Furthermore, such subjective measures have been critically discussed due to issues of validity (de Jong, 2010; Kirschner et al., 2011; van Gog and Paas, 2008). Hence, some studies use objective, physiological measures as indicators for cognitive load (e.g., various heart rate or pupillometric measures; Solhjoo et al., 2019; Zheng and Cooke, 2012). However, it is not clear to which extent objective measures converge with subjective measures as indicators for an individual's cognitive load in the corresponding contexts. The convergence of assessment methods provides evidence for validity of these measures.

Validity is an integrated evaluative judgment on the extent to which the appropriateness and quality of interpretations and measures based on test scores (or other diagnostic procedures) are supported by empirical evidence and theoretical arguments (Messick, 1995; Kane, 2013). According to the argument-based approach to validation (Kane, 2013), validation depends on the intended interpretation and use of test scores and requires to provide argumentative evidence that an intended test score interpretation is legitimate. Hence, the validation of an instrument is not a routine procedure, but is carried out through theory-based research, with which different interpretations of a test score can be legitimized or even falsified (Hartig et al., 2012). In the *Standards for Educational and Psychological Testing*, it is emphasized that "validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (AERA et al., 2014, p. 11). The authors further elaborate on different "sources of evidence that might be used in evaluating the validity of a proposed interpretation of test scores for a particular use" (p. 13), including validity evidence based on relations to other variables. Hence, what has been named the external aspect of construct validity (Messick, 1995) are conceptualized as one source of validity evidence in the argument-based approach to validation (Kane, 2013). Specifically, comparison studies with subjective and objective measures of cognitive load "may lead to new insights on convergence between biological [i.e., objective] and subjective measures and on what these different types of measures are measuring" (Leppink et al., 2013, p. 1070). Hence, comparison between subjective and objective measures have been proposed as a source of validity evidence for subjective measures (Solhjoo et al., 2019).

Clearly more research about subjective and objective cognitive load measures, their interrelationships, and association with theoretically important variables such as emotion and motivation are needed to contribute to a comprehensive understanding of cognitive load and its assessment approaches. The present study contributes to this body of research by examining the convergence of subjective and objective measures of cognitive load and the relationship between causal and assessment factors of cognitive load.

Causal Factors and Assessment Factors of Cognitive Load

In a theoretical framework of how cognitive load might be conceptualized in the context of problem solving, Paas and van Merriënboer (1994) distinguished between causal and assessment factors of cognitive load. Causal factors include learner characteristics (e.g., prior knowledge, cognitive capabilities, motivation, and affect) as well as task environment (e.g., task complexity, time pressure). In a more recent revision of the framework, the dimension of task environment in the *causal factors* has been subdivided into learning task and the physical learning environment (Choi et al., 2014). Furthermore, two- and three-way-interactions illustrate the fact that each causal dimension may affect cognitive load depending on characteristics of the other dimensions. For example, the amount of cognitive load a learner invests to solve a given task might depend on task complexity, the specific context or goal (e.g., just for fun vs. high stakes test), and the learner's interest related to the given problem. In terms of *assessment factors*, the authors distinguished between a task-relevant cognitive load dimension of mental load, a person-relevant cognitive load dimension of mental effort and task performance (Paas and van Merriënboer, 1994). Mental load is based on characteristics of the task, representing the cognitive capacity needed to process a task. In contrast, mental effort reflects an individual's invested cognitive capacity while working on a task. Assessments of mental effort are thought to provide information about the amount of controlled processing a person is engaged in (Paas and van Merriënboer, 1994).

Some of the assessment factors are hypothesized to be affected by the causal factors during problem solving. In particular Paas

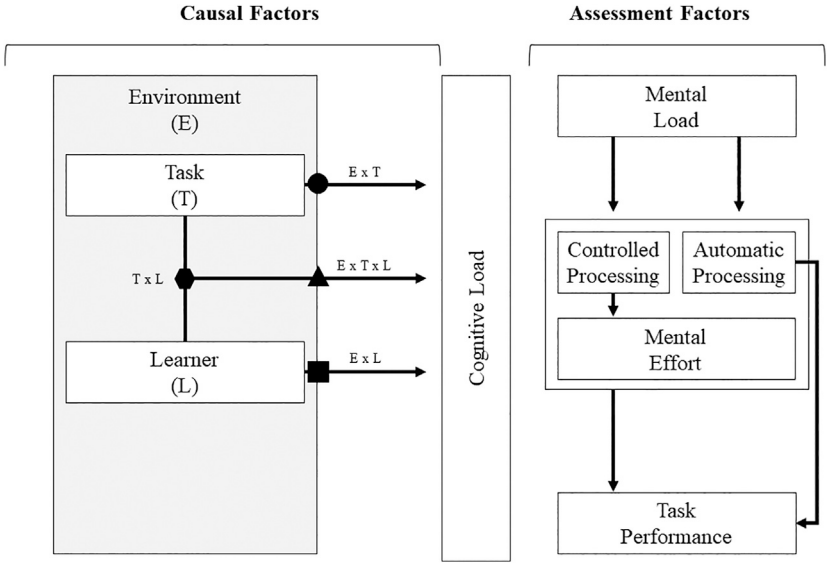

FIGURE 1 | Cognitive load as reflected by causal factors and assessment factors, including two- and three-way interactions between the causal factors (adapted from Choi et al., 2014, p. 229).

and van Merriënboer (1994) conceptualized mental load as independent from person characteristics and, thus, as being constant for a given task (e.g., in terms of cognitive capacity necessary to process the number of elements in a given task). Whereas mental effort and task performance are affected by all three causal dimension factors. Sweller et al. (2011) propose mental load and mental effort as being two different but, in most cases, positively correlated constructs, with the former being the hypothetically required and the latter being the occurring cognitive resources in relation to the learning task. However, the relationship between mental load and mental effort, as well as the relationship between mental load, mental effort, and task performance, might not necessarily be positive. For example, both high or low mental load could result in rather low mental effort due to the moderating role of person characteristics such as motivational variables and persons may reach the same number of correct answers on a test but need to work with different amounts of mental effort (Paas et al., 2003). Relatedly, Moreno (2010) suggested conceptualizing cognitive load within a cognitive-affective theory of learning and emphasizes 'that cognitive capacity is a parameter that students bring to the learning task whereas motivation determines the actual amount of cognitive resources invested in the learning task' (p. 137).

This framework provides a powerful tool for researchers as it allows to further investigate cognitive load by narrowing down into its constituent dimensions and provides venues for further research. For example, relating measures of mental effort and task performance allows to investigate the cognitive capacity needed for reaching a specific level of performance (Paas and van Merriënboer, 1994). Further research questions that can be derived from the framework are related to the specific relationships between learner characteristics, such as emotion and motivation, and individual's invested mental effort (Moreno, 2010; Hawthorne et al., 2019; Skuballa et al., 2019).

The present study focuses on this framework of cognitive load (**Figure 1**) and is, therefore, related to the assessment and causal factors of cognitive load, assuming that the actual cognitive capacity that needs to be investigated to process a given task (i.e., mental load) is necessarily intertwined with the causal factors such as learner characteristics (especially relatively stable characteristics such as prior knowledge) and, thus, is likely to vary between individuals. Thus, it focuses on the associations between the assessment factors and causal factors, in particular the learner characteristic aspects in terms of motivation and affect. Furthermore, we examined the level of convergence between cognitive load measures obtained via objective and subjective approaches. Below we first elaborate on recent research on subjective and objective measures of assessment factors of cognitive load, then we review literature on learner characteristic related variables including self-concept, interest, and perceived stress as causal factors of cognitive load.

Subjective Measures of Cognitive Load

Subjective measures of cognitive load ask respondents to self-report the amount of cognitive load after working on a task (Sweller et al., 2011) and this has been the primary approach in research practice (e.g., Paas 1992; Nehring et al., 2012). One basic assumption of subjective measures is that individuals are aware and can quantify and report on their cognitive load (Solhjoo et al., 2019). One of the first scales that have been proposed for subjective measurement of cognitive load was developed by Paas (1992), who introduced a single-item nine-point mental effort rating scale. This scale asks respondents to rate their invested mental effort, ranging from "very, very low mental effort" to "very, very high mental effort".

However, subjective measurement of cognitive load 'has become highly problematic' (Kirschner et al., 2011, p. 104). Krell (2015) summarized several reasons for this:

(1) Many studies adapt the scale initially developed by Paas (1992) and change the wording or number of category labels without re-evaluating its psychometric properties (Paas et al., 2003; van Gog and Paas 2008).
(2) Often, only a single item is used to measure cognitive load, although the use of several items would increase measurement precision (Leppink et al., 2013).
(3) Sometimes, it is not entirely clear which construct the items are aimed to measure. Whereas Paas (1992) focused on mental effort, many researchers use category labels related to task complexity such as 'difficulty' and still consider it as measures of cognitive load (van Gog and Paas 2008; de Jong 2010).
(4) Finally, van Gog and Paas (2008) emphasize that most measures target cognitive load as a whole and not specific dimensions of it, such as mental effort or mental load.

In response to such criticisms, new subjective measures of cognitive load have been proposed. For example, some authors (Klepsch et al., 2017; Leppink et al., 2013) developed instruments to assess three dimensions of cognitive load: intrinsic load, extraneous load, and germane load (cf. Paas et al., 2003). These dimensions reflect the cognitive capacity caused by the complexity of a problem (intrinsic cognitive load), the design of learning material (extraneous cognitive load), and the effort invested to solve a given task (germane cognitive load). Hence, related to the framework used in the present study (**Figure 1**), germane and extraneous cognitive load are related to ME, whereas intrinsic cognitive load is related to mental load (Choi et al., 2014). For both instruments, validity evidence based on various sources (e.g., internal structure, test content) have been provided (cf. AERA et al., 2014) and the authors conclude that their instrument is useful, feasible, and reliable (Klepsch et al., 2017) or that it could be used for research purposes in various knowledge domains (Leppink et al., 2013), respectively.

Related to the framework of cognitive load, which is used in the present study (**Figure 1**), Krell (2015, 2017) proposed the Students' Mental Load and Mental Effort in Biology Education-Questionnaire ("StuMMBE-Q"). This instrument was designed to measure students' mental load and mental effort on 12 rating scale items. Besides its development in the context of student assessment in biology education, it has been widely applied in various other contexts (e.g., Nebel et al., 2016; Knigge et al., 2019). Evidence for the valid interpretation of the ratings as indicators for students' mental load and mental effort have been provided based on test content, internal structure, and relation to other variables (AERA et al., 2014). For example, psychometric analyses confirmed a two-dimensional data structure, representing measures of mental load and mental effort (Krell, 2015). Furthermore, students' self-reported mental load and mental effort significantly increased with increasing task-complexity (Krell, 2017). In preceding studies using the StuMMBE-Q, students' task performance was significantly negatively correlated with their self-reported mental load but not associated with their self-reported mental effort (Krell, 2015, 2017).

In sum, subjective measurement has been the primary approach in assessment of cognitive load. However, this approach is subject to influence from causal factors, in particular individual differences such as prior knowledge, interest and motivation. Thus, subjectively perceived cognitive load might not accurately reflect the characteristics and demands of the learning task. Objective measures, on the other hand, may provide more accurate reflection on task complexity presented to the learner.

Objective Measures of Cognitive Load

Besides the measurement of cognitive load via self-reports on questionnaires (e.g., Paas, 1992; Leppink et al., 2013; Krell, 2017), other approaches suggest or use more objective, physiological measures as indicators for respondents' cognitive load (cf. Sweller et al., 2011); such as eye movements (Ikehara and Crosby, 2005; Zu et al., 2019), degree of pupil dilation (Huh et al., 2019), or physiological stress parameters such as heart rate and cortisol secretion (Veltman and Gaillard, 1993; Kennedy and Scholey, 2000; Cranford et al., 2014). Previous research has attempted to triangulate both objective and subjective cognitive load measures (Kahneman and Peavler., 1969; Antonenko et al., 2010; Zu et al., 2019). Physiological measures have also been proposed as sources for validity evidence for self-reports, if positive correlations between both measures can be shown (e.g., Solhjoo et al., 2019). However, some studies were able to show such positive relations (e.g. Solhjoo et al., 2019), while others found differences between self-reports and physiological measures, suggesting that both might assess separate aspects of cognitive load (e.g., Zheng and Cooke, 2012 for pupillometric measures).

Based on the Biopsychosocial Model of Challenge and Threat (Blascovich and Mendes, 2000; Blascovich, 2008) individuals can perceive performance tasks as 'challenge' or 'threat', depending on how they assess task demands and their own resources. 'Challenge' occurs when resources exceed demands (comparable with low mental load) and 'threat' if demands exceed resources (comparable with high mental load). In this model, the individual relevance of the performance goal is also important, as task engagement (comparable to mental effort) is necessary for challenge and threat states (Blascovich and Mendes, 2000). Challenge and threat states are marked by different cardiovascular patterns. Challenge states are indexed by increases in heart rate, dilation of arteries and increased blood pump, whereas threat states result in little or small increase in heart rate, constriction, and less bloodstream (Seery, 2011; Seery, 2013).

Besides the measurement of the heart rate itself (heart rate, e.g., Solhjoo et al., 2019), some heart rate variability measures were also used in educational and psychological research, including the ratio between high frequency and low frequency components (LF/HF ratio, e.g., Minkley et al., 2018), and the time-related root mean square of successive differences (RMSSD,

e.g., Minkley et al., 2018). If a person is confronted with a complex problem or task, where demands exceed their resources, the RMSSD typically decreases (e.g., Malik et al., 1996), as it represents parasympathetic activity which in turn characterizes relaxation (Laborde et al., 2017). In contrast the LF/HF ratio as well as the heart rate increases (e.g., Malik et al., 1996; Isowa et al., 2006). Both parameters basically represent (at least the proportion of; LF/HF ratio) sympathetic activity, which in turn increases under stress (Laborde et al., 2017).

Thus, it is plausible to assume a relationship between subjective measures and heart rate- or hormone- based objective measures of cognitive load, which is rarely investigated yet. One study of the authors found a significant positive correlation between self-reported mental load (i.e., perceived task complexity) and perceived stress but no or rather small correlations between self-reported mental load and physiological stress responses (cortisol concentration, heart rate variability; Minkley et al., 2018). Veltman and Gaillard (1993) report similar findings and conclude that cortisol concentration might not be a valid indicator for mental activity and that heart rate variability might indicate cognitive load only in high demanding tasks. The study of Kennedy and Scholey (2000) points in a similar direction: Although a correlation between heart rate and demanding (mental arithmetic, word retrieval) vs. non demanding (control) tasks could be observed, this correlation was not present when the mental arithmetic tasks were further differentiated regarding their difficulty (serial subtraction of three vs. seven). In contrast, Cranford et al. (2014) observed a higher heart rate in chemical tasks designed to achieve high cognitive load compared to the heart rate in low cognitive load tasks. Unfortunately, the authors did not assess any self-report measure of cognitive load from their participants. However, a recent study by Solhjoo et al. (2019) found positive correlations between self-report measures of cognitive load and heart rate variability for a small sample of ten medical students. In contrast, Woody et al. (2018) found no association between cognitive load and cortisol concentration or heart rate. Hence, it is still vague regarding to what extent and under which circumstances objective measurements capture cognitive load and to what extent they converge with subjective measures of cognitive load.

Motivation and Emotion as Causal Factors of Cognitive Load: Interest, Self-Concept, and Stress Perception

As shown in **Figure 1**, in the theoretical framework proposed by Paas and van Merriënboer (1994), causal factors of cognitive load consist of both task characteristics and learner characteristics. To date much research of causal factors have focused on varying task characteristics such as worked example vs problem solving. Although there has been research on the roles of learner characteristics in terms of learner prior knowledge (Kalyuga et al., 2003) and age (van Gerven et al., 2002) on cognitive load, less is known about motivation and emotion and how they affect cognitive load. The related cognitive theory of multimedia learning (Moreno, 2010) also suggested that learner's motivational and affective factors play an important role in the cognitive processes that are captured by the germane cognitive load, which in part is reflected by the learner's invested mental effort. Motivated learners are more likely to be engaged during the learning process, thus their cognitive resources are more likely directed at schema constructions important for knowledge acquisition. In a similar way, empirical research also confirmed that factors such as learner's emotion, enjoyment, or stress, can influence available working memory resources (Plass and Kalyuga, 2019) thus affecting the learner's experienced cognitive load, whereas other constructs such as academic self-concept are less studied in the current context but may also affect cognitive load (Seufert, 2020). In the present study, we look specifically at constructs representing interest, stress related emotion, and academic self-concept of ability (ASCA) and how they may relate to cognitive load measurements. Below we review relevant literature in these motivational (interest and ASCA) and affective (emotion - stress) causal factors and how they affect or could theoretically affect cognitive load. Additionally, to account for the relevance of objective and subjective differentiation of cognitive load measures in the present study, we also review a relatively small pool of literature that studied the relationships between motivation, emotion and objectively measured cognitive load.

Interest

According to Ainsley et al. (2002), p.545, interest is an individual's "predisposition to attend to certain objects and events and to engage in certain activities". Since an interested learner is more likely to attend to the learning tasks, their working memory resources are more effectively directed at processing new information afforded by the learning task (i.e., more germane processing and thus higher mental effort). On the other hand, the focused attention and positive affect associated with interest (or "flow"; see Csikszentmihalyi, 2013) could also mean the learner may experience less perceived task difficulty (i.e., less mental load). Learner's interest is often found inversely associated with perceived cognitive load during task performance and higher invested effort. For example, learners with higher learning interest viewed the same tasks to be less difficult (i.e., mental load) even when they invested more effort (Milyavskaya et al., 2018). Skuballa et al. (2019) also found that learner's topic interest was associated with decreased perceived task difficulty. It seems that the more motivated the learner is, the less difficulty they perceive the task, and the more mental effort the learner will invest in the learning task. This body of literature largely supports the role of interest as a causal factor of cognitive load.

Academic Self-Concept

The ASCA represents one's self-evaluation of their competencies in academic domains (Marsh et al., 2012). Although to our knowledge no previous study has investigated the association between ASCA and cognitive load, ASCA may affect the perceived cognitive load in several possible ways. ASCA has been shown to have reciprocal relationships with both the learner's achievement and their interest (e.g., Marsh et al., 2005). The bi-directional relationships with achievement and interest implies that a learner with a higher ASCA may also

have a higher level of prior knowledge (i.e., achievement), and is also more interested in the topic. For a given task, a higher prior knowledge is likely to be associated with less cognitive load, because the learner already has consolidated knowledge "chunks" and therefore the number of elements to be processed for performing the task is reduced (Sweller et al., 2019). Consequently, the learner is likely to experience relatively less cognitive load, in particular perceived difficulty (i.e., mental load). Similarly, since ASCA is reciprocally related to interest, a learner with high ASCA is likely to be interested in the task, perceive less task difficulty (i.e., mental load), and invest more effort (i.e., mental effort). Based on this theoretical conjecture, ASCA is a causal factor of cognitive load, and this relationship is likely supported by empirical research.

Emotion (Stress)

Emotions, in particular negative valence emotions, have been suggested to be a source of increased extraneous load (Sweller et al., 2019). Plass and Kalyuga (2019) elaborated on how emotion might affect cognitive load and they suggested that negative emotion such as stress may indeed increase the burden of neural circuitry involved in learning thus increasing cognitive load and reducing task processing efficiency. Moreover, Blascovich (2008) claims that individuals perceive tasks as 'challenge' or 'threat', depending on their assessment of task difficulty and possibilities of dealing with them. As a result, the individual may pursue or disengage with the task. Empirical research has yielded somewhat inconsistent results regarding this proposition (e.g., Fraser et al., 2012; Knörzer et al., 2016). Based on a sample of undergraduate medical students who worked on a simulation training, Fraser et al. (2012) found that stress-related invigorating emotions are associated with higher perceived cognitive load. In their experimental study, Knörzer et al. (2016) induced negative emotions prior to a learning task. Although participants in the experimental group rated higher perceived task difficulty (i.e., mental load) than those in the control group (neutral emotion), this effect was not statistically significant. The ratings on mental effort also did not differ between the two groups. However, both studies were based on relatively small sample sizes ($n < 100$) thus further evidence is needed in larger, multiple samples to provide stronger statistical power. In sum, although there is a theoretical basis for a learner's emotion to be a causal factor of cognitive load, further empirical research is necessary to confirm this hypothesis.

Motivation, Emotion and Objectively Measured Cognitive Load

Most of the literature relating motivation and affect to cognitive load is based on cognitive load reported from questionnaire measurements. There is limited empirical research examining whether objectively assessed cognitive load is also associated with motivation and affect (typically measured by questionnaire-based method). In an early experimental study, Kahneman and Peavler (1969) demonstrated the potential effect of motivation by showing that individuals performed better on the test items associated with higher monetary values. Furthermore, while working on task items with higher incentive, participants also showed larger pupillary dilation - an objective indicator of mental effort (Goldinger and Papesh, 2012). As discussed by Paas and van Merriënboer (1994), a higher performance could indicate higher mental effort, thus, the link between causal and assessment factors is supported by an association between motivation (prompt by monetary value) and cognitive load (illustrated by performance). Kahneman and Peavler (1969)'s research showed that there may be an association between motivation and objectively measured cognitive load. Given the scarcity of the existing research demonstrating the relationship between motivation, affect and objectively measured cognitive load, the present study aims to contribute to the literature by investigating the relationship between interest, ASCA, stress and cognitive load measures assessed both subjectively via questionnaire as well as objectively through heart rate.

Aims of the Study, Research Questions and Hypotheses

The aims of this study are to evaluate the convergence between subjective (self-reports on mental load and mental effort) and objective (heart rate measures LF/HF ratio, RMSSD and heart rate) measures of cognitive load (Leppink et al., 2013; Solhjoo et al., 2019) and to provide evidence for the assumed relationships between assessment factors of cognitive load, that is mental effort and mental load (Paas and van Merriënboer, 1994) and conceptually related causal factors in terms of positive and negative motivation and affect, that is self-concept, interest and perceived stress (cf. Moreno, 2010; Minkley et al., 2014; Minkley et al., 2018; Solhjoo et al., 2019).

Research Questions

(1) How do cognitive load measures converge via subjective and objective measures?

H1: There is a linear relationship between subjective and objective measures of cognitive load (positive for LF/HF ratio and heart rate, negative for RMSSD), with still a medium to large variance component specific to each measure, because subjective and objective measures are likely to capture separate aspects of cognitive load (Zheng and Cooke, 2012; Solhjoo et al., 2019).

(2) How do subjective and objective measures of cognitive load predict task performance?

H2: It is expected that (1) subjective (i.e., mental load and mental effort) and (2) objective (i.e., RMSSD, LF/HF ratio, and heart rate) measures of cognitive load contribute to predict students' task performance because (1) a higher level of mental load indicates more challenging tasks (Krell, 2017) whereas higher mental effort indicates a more intense engagement with tasks (Paas et al., 2003). (2) Regarding objective measures, RMSSD typically decreases, whereas the LF/HF ratio and heart rate increase with increasing task difficulty (e.g., Malik et al., 1996; Isowa et al., 2006).

(3) How do students' self-concept, interest and stress perception predict their subjectively (mental effort, mental load) and objectively measured cognitive load?

H3a: It is expected that students' (1) self-concept and interest, (2) but not their stress perception, predict their self-reported level of mental effort, because (1) higher amounts of self-concept and interest might elicit a more intense engagement with tasks (Milyavskaya et al., 2018; Skuballa et al., 2019). (2) In contrast, the effect of personal classification of perceived stress (as challenge or threat) is less consistent and can lead to either increased ambition or give up (Lazarus and Folkman, 1984).

H3b: It is expected that students' (1) self-concept and interest (2) and also their stress perception predict their self-reported level of mental load, because (1) higher amounts of self-concept and interest might elicit a perception of less complex tasks (Milyavskaya et al., 2018; Skuballa et al., 2019) and (2) demanding tasks, beyond what is manageable according to the self-concept, should be perceived as threatening, leading to a stress response (Dickerson and Kemeny, 2004) thus a higher perceived mental load.

H3c: It is expected that students' self-concept and interest, (2) but not their stress perception, predict objective measures of cognitive load (i.e., RMSSD, LF/HF ratio, and heart rate), because (1) higher self-concept and interest contribute to perceive a task as less challenging (Milyavskaya et al., 2018; Skuballa et al., 2019). (2) In contrast, most previous studies did not find systematic associations between stress perception and cognitive load measures, assuming inter alia that both measures differ between individuals and are affected by the time of measurement in varying degrees (Campbell and Ehlert, 2012).

MATERIALS AND METHODS

Context and Design of the Studies

The data come from three earlier studies, one of which (study 1) has been published elsewhere (Minkley et al., 2018), and which are secondarily analyzed for the purpose of this study. In each of the three studies, high school students participated in a one-day out of school learning project about molecular biology techniques with instructional and practical hands-on phases, aimed at identifying genetically modified food in the teaching and learning laboratory at Ruhr-Universität Bochum. The procedure and design of the three studies were the same and also the structure of the tasks' and the measurements. The differences between the three studies lie in the fact that the content of the tasks differed in each study, although they all dealt with topics in molecular biology (see sections "**Procedure**" and "**Tasks**").

Participants

The participants (N = 309, from three studies; **Table 2**) were upper-level students with a mean age of 17.48 years (SD = 1.07; **Table 1**), from 22 high schools in North Rhine-Westphalia, Germany. Complete heart rate variability measurements were obtained from 227 participants. Among the 82 participants which were excluded from further analysis, some did not complete the study tasks. For those who did complete the task, there were technical errors; e.g., a lack of coupling between the measurement sensor and the storage device, which leads to the fact that the measurement data could not be read out, or too many measurement artifacts. For the latter, according to the procedure of Laborde et al. (2017), the threshold value for data being considered as artifacts, was set to 0.45 s (= very low) difference between a single heartbeat interval from the local average. That is, participants with regular fluctuations above a range of 0.45 s were excluded from the data set. To avoid possible confounding effects, 18 participants with a body mass index >30 kg/m^2, serious medical conditions or frequent smoking were excluded from all heart rate variability analyses (Piestrzeniewicz et al., 2008; Thayer et al., 2010). Thus, we examined heart rate variability parameters from 209 students.

TABLE 1 | Demographic characteristics and measurement values.

	M	*SD*	*n*
Demographic variables			
Sex, % female	54.4		309
Age, years	17.48	1.07	309
Measured variables			
Mental load	4.03	1.16	305
Mental effort	4.35	1.08	305
PS, mm	29.80	21.70	260
LF/HF, ratio	2.63	2.23	209
RMSSD, ms	51.05	35.52	209

PS, perceived stress; LF/HF ratio, ratio between low and high frequencies; RMSSD, root mean square of successive differences.

Procedure

In all three studies the procedure was comparable in terms of learning environment and tasks. After arrival, the students', respectively their parents' written informed consent was collected and the participants were informed about the procedure of the following studies, which were conducted in accordance with the Declaration of Helsinki and approved by the Ethics Commission of the local medical school. Thereafter all students participated in the same molecular biology project, which was not associated with the present studies. The actual study started after about one half of the molecular biology project. Before the students worked on the test tasks, they were equipped with a chest belt and moved to another room. Then each of them was placed in front of a laptop, separated from each other. The students also filled in a demographic questionnaire (sex, age) including medical information (weight, height, chronic diseases, medication intake), and two scales measuring the ASCA and interest regarding biology (Sparfeldt et al., 2004; Rost et al., 2007; Wilde et al., 2009; Minkley et al., 2014). All participants were randomly seated in front of a laptop, on which the test booklets were installed. The laptops were separated from each other, so that the participants could not see the tasks of their schoolmates. The students got 10 up to 20 min to work on the tasks (depending on the study). After

completion of the tasks, the students filled in a questionnaire to self-report their mental load and mental effort during working on the tasks (Krell, 2015, 2017) and indicated their perceived stress on a Visual Analogue Scale (Luria, 1975; cf. **Table 2**). Heart rate variability was measured continuously via a chest belt and a storage device (Polar V800).

Tasks

In each of the three studies, participants completed several tasks that had the same structure and differed only slightly in content, but all related to molecular structures and phenomena covered in biology classes (e.g., DNA structure, diffusion). In order to achieve a test design that is as close to school reality as possible, we have checked the tasks for alignment with the course content and the biology curriculum in high school. In each task, there was a representation of a molecular structure or process. The task complexity was varied by considering different cognitive demands in all three studies. In the simpler tasks (about one half) the participants had to recognize and select the correct name or a true statement about a depicted molecular structure or process from several answering options (i.e., single-best answer format). In the more complex tasks, the participants had to explain or compare the depicted molecular structure or process in a short written text (i.e., constructed response format). The tasks also differed regarding the types of representation (one half of the tasks included purely symbolic representations, the other half combined symbolic–textual representations; Minkley et al., 2018). The distribution between simple and complex tasks as well as the type of representation was systematically distributed in all three studies. The tasks have been discussed in several rounds by the first and the last author of this article to evaluate their content validity. In the present study, however, task complexity is not a focus, given the abundance of prior research.

Assessments

The instruments used in the present investigation are comparable across the studies included. **Table 2** lists the assessments in the three studies. Further explanations are provided in the text below. All assessed data have been z-standardized for further analyses.

Mental Load and Mental Effort

The StuMMBE-Q instrument was applied in the present study to achieve indicators for the students' mental load and mental effort (Krell, 2015, 2017). For both dimensions of cognitive load, that is mental load (e.g., "The tasks were challenging.") and mental effort (e.g., "I have tried hard to answer the tasks correctly."), six rating scale items for self-reporting are included in the StuMMBE-Q. For each item, a seven-point rating scale ranging from 'not at all' (=1) to 'totally' (=7) was provided. Measures for mental load and mental effort have been computed by calculating the mean score of the six respective items. For the total sample, mean scores are $M_{\text{Mental load}} = 4.09$ ($SD = 1.19$) and $M_{\text{Mental effort}} = 4.39$ ($SD = 1.09$), respectively. The internal consistency of the questionnaire is satisfying, with Cronbach's alpha for mental load = 0.85 and for mental effort = 0.81.

Academic Self-Concept of Ability (ASCA)

In order to assess students' biology-specific ASCA in study 1 and 2, we used a modification (Minkley et al., 2014) of the DISC-Grid (Rost et al., 2007). There, the participants have to rate eight items (e.g., "For me it is easy to solve biology problems." or "I have a good feeling when thinking about my achievements concerning biology.") on a 6-point rating scale, ranging from 1 ("does not apply to me at all") to 6 ("fully applies to me"). Measures of ASCA have been computed by calculating the sum score of the eight items. The mean score is $M_{\text{ASCA}} = 32.34$ ($SD = 9.74$). The internal consistency of the questionnaire is excellent, with Cronbach's alpha for ASCA = 0.92.

Interest

In study 1, the students' interest regarding biology was assessed by rating 8 items on a 6-point rating scale ranging from "does not apply to me at all" (=1) to "fully applies to me" (=6) (e.g., "I am interested in Biology." or "I enjoy working on tasks in biology.") adapted from Sparfeldt et al. (2004). In study 2, three items with a 7-point rating scale ranging from "does not apply to me at all" (=1) to "fully applies to me" (=7) have been used to assess the same construct based on Wilde et al. (2009) (e.g.," Biology lessons are interesting." or "I enjoy biology lessons."). Measures of interest have been computed by calculating the mean score of the items. For the total sample, mean scores are $M_{\text{Interest}} = 3.89$ ($SD = 1.04$) for study 1 and $M_{\text{Interest}} = 5.29$ ($SD = 1.26$) for study 2, respectively. The internal consistency of the questionnaire is excellent, with Cronbach's alpha = 0.91 (for study 1) and = 0.91 (for study 2).

TABLE 2 | Assessments in the three studies.

	Subjective measures of cognitive load	Objective measures of cognitive load (heart rate measures)	Measures of interest, stress perception and self-concept
Study 1	Mental load	Heart rate	Self-concept
(n = 93)	Mental effort	LF/HF ratio	Interest
		RMSSD	Perceived stress
Study 2 (n = 145)	Mental load	–	Self-concept
	Mental effort	LF/HF ratio	Interest
		RMSSD	Perceived stress
Study 3 (n = 133)	Mental load	Heart rate	–
	Mental effort	LF/HF ratio	–
		RMSSD	Perceived stress

LF/HF ratio, ratio between low and high frequencies; RMSSD, root mean square of successive differences.

Perceived Stress

The perceived stress was assessed using a Visual Analogue Scale (Luria, 1975) in all three studies. The Visual Analogue Scale consists of a 100mm-long line, with the label 'no stress' on the left end and 'maximum stress' on the right end. The participants placed a cross on this line at that point which expresses how stressed they felt at that moment. Afterward the distance between the left end of the line and the participants' cross was measured; therefore, the possible score range is from 0 to 100. For the total sample, the mean score is $M_{\text{Stress}} = 29.84$ ($SD = 21.82$).

Heart rate Variability Measures

The heart rate variability of the participants was measured continuously via a chest belt with an integrated ECG-sensor (V800, Polar). After the measurement, the data was transmitted to a software (Kubios) to calculate time and frequency domain measures. We calculated the root mean square of successive differences (RMSSD) as a common time domain measure of heart rate variability, reflecting vagal tone (Laborde et al., 2017), or rather parasympathetic activity (Malik et al., 1996; Hjortskov et al., 2004). Additionally, we calculated LF/HF ratio as a frequency domain measure reflecting the ratio between parasympathetic (high frequency components; 0.15–0.4 Hz) and sympathetic (low frequency components; 0.04–0.15 Hz) nervous system activity. For the total sample, mean scores are $M_{\text{Heart rate}} = 90.28$ ($SD = 17.65$), $M_{\text{RMSSD}} = 51.64$ ($SD = 36.11$), and $M_{\text{LF/HF ratio}} = 2.63$ ($SD = 2.29$), respectively.

Task Performance

To assess task performance, performance expectations were prepared for all tasks and points were awarded as follows. For the tasks in single-best answer format, respondents received a full score for the correct answer (=1) or no points if they selected a wrong answering option. For the constructed response items, scoring was done according to a predetermined scoring scheme, which also allowed for partially correct answers. This performance expectation was discussed and revised in several rounds as part of the task development (see **section 2.4**). To calculate a task performance score, the percentage of achieved points relative to maximum points was used; therefore, the possible score range is from 0 to 100. For the total sample, the mean score is $M_{\text{Performance}} = 41.83$ ($SD = 24.03$).

Data Analysis

The Software IBM SPSS Statistics was Used for Data Analysis.

Measurements of almost all scales had skewness and kurtosis statistics between -2 and 2 (except kurtosis of LF/HF ratio = 3.55 ± 0.32), indicating approximately normal distribution (Gravetter and Wallnau, 2012).

For the analysis of the convergence of cognitive load measures via subjective and objective measures (RQ 1), correlational analyses were performed with mental load or mental effort and the objective measures (heart rate, LF/HF ratio, RMSSD).

For the analysis of how subjective and objective measures of cognitive load predict task performance (RQ 2), first a basic analysis of the correlations between mental load, mental effort, objective measures, and task performance was carried out. Subsequently, a joint test of the variables was performed in the form of a regression analysis with objective and subjective measures of cognitive load as predictor variables for task performance.

For the analysis of how students' self-concept, interest and stress perception predict their subjectively and objectively measured cognitive load (RQ 3), also basic analyses of the correlations between subjective and objective measures of cognitive load, self-concept, interest, and stress perception were carried out. Subsequently, linear regression analyses were performed with subjective and objective measures of cognitive load as dependent variables and self-concept, interest, and perceived stress as predictor variables.

For all linear regression analyses, no serious violations of assumptions could be found; the Durban-Watson-statistic is in the range of 1 and 3, indicating that there is no considerable autocorrelation, and VIF is < 10 for all items, indicating no serious multicollinearity (Field, 2009). (See regression tables for the exact values.)

RESULTS

In the following, findings are presented for the total sample ($N = 309$). The separate findings for the three individual studies ($n_1 = 93$, $n_2 = 145$, $n_3 = 133$; **Table 2**) can be found in the **Supplementary Material**.

Related to our research question on how cognitive load measures converge via subjective and objective measures (RQ1), the findings (**Table 3**) show no statistically significant correlation between self-reported mental load and the objective measures; for mental effort, there are statistically significant correlations between mental effort and LF/HF ($p < 0.05$) and mental effort and heart rate ($p < 0.001$). For RMSSD, the p-values indicate marginally

TABLE 3 | Pearson correlation coefficients r between objective and subjective measures of cognitive load for the total sample.

		Mental effort	LF/HF ratio	RMSSD	Heart rate
Mental load	*r*	–0.05	0.03	–0.05	–0.04
	p	0.38	0.67	0.42	0.63
	n	367	224	223	134
Mental effort	*r*		0.15	–0.11	–0.27
	p		**0.03**	0.11	**0.00**
	n		224	223	134
LF/HF ratio	*r*			–0.52	–0.04
	p			**0.00**	0.62
	n			226	135
RMSSD	*r*				0.23
	p				**0.01**
	n				134

LF/HF ratio, ratio between low and high frequencies; RMSSD, root mean square of successive differences; significant values are bold.

significant correlations. However, even statistically significant correlations are mostly small and indicate shared variance (R^2) of less than 10%.

Related to our research question on how subjective and objective measures of cognitive load predict task performance (RQ2), initial basic correlational analyses findings show a significant negative correlation between mental load and task performance and a significant positive correlation between mental effort and task performance; the correlation coefficients are small to medium (**Table 4**). There is a positive correlation toward LF/HF and a negative toward heart rate, both with small correlation coefficients.

In the subsequent linear regression analysis with objective and subjective measures of cognitive load as predictor variables for task performance, both mental effort and mental load contribute significantly to predict task performance, while LF/HF contributes marginally significantly (**Table 5**).

Related to our research question on how students' self-concept, interest and stress perception predict their subjectively and objectively measured cognitive load (RQ3), the initial basic correlational analyses show significant correlations between mental load and mental effort and the motivation variables (i.e., ASCA and interest). While mental effort is positively related to ASCA (r = 0.23) and interest (r = 0.20), the opposite could be found for mental load (ASCA: r = −0.29; interest: r = −0.18). Perceived stress was found to be not related to mental effort, but significantly related to mental load (r = 0.14). The effect sizes are mostly small (**Table 6**).

Subsequent linear regression analyses with mental effort (**Table 7**), mental load (**Table 8**), LF/HF ratio (**Table 9**), RMSSD (**Table 10**), and heart rate (**Table 11**) as dependent variables and ASCA, interest and perceived stress as predictor variables show more specific results. While mental effort is not significantly explained by any of the three variables (H3a), ASCA (negatively) and perceived stress (positively) contribute to explain mental load (H3b). Related to the objective cognitive load measures (H3c), the same two variables contribute to explain LF/HF ratio (ASCA positively, perceived stress negatively), while this is only the case for one of these two predictor variables for RMSSD (perceived stress contributes negatively) and heart rate (ASCA contributes positively).

TABLE 4 | Pearson correlation coefficients *r* between objective and subjective measures of cognitive load and task performance for the total sample.

		Mental load	Mental effort	LF/HF ratio	RMSSD	Heart rate
Task performance	*r*	−0.38	0.26	0.16	−0.04	−0.21
	p	**0.00**	**0.00**	**0.018**	0.511	**0.015**
	n	363	363	225	224	135

LF/HF ratio, ratio between low and high frequencies; RMSSD, root mean square of successive differences; significant values are bold.

TABLE 5 | Linear regression analysis with objective and subjective measures of cognitive load as predictor variables for **task performance** (for the total sample).

Coefficient	*B*	*SE*(*B*)	BETA	*p*	VIF
(Constant)	0.337	0.065	–	**<0.001**	–
Mental load	−0.305	0.070	−0.326	**<0.001**	1.015
Mental effort	0.268	0.068	0.309	**<0.001**	1.143
LF/HF ratio	0.140	0.071	0.174	**0.051**	1.435
RMSSD	−0.064	0.067	−0.084	0.344	1.442
Heart rate	−0.091	0.067	−0.107	0.177	1.144

LF/HF ratio, ratio between low and high frequencies; RMSSD, root mean square of successive differences; adjusted R^2 = 0.280; Durbin–Watson–Statistic = 1.466; significant values are bold.

TABLE 6 | Pearson correlation coefficients r between ASCA, interest, perceived stress, and subjective and objective measures of cognitive load for the total sample.

		Mental load	Mental effort	LF/HF ratio	RMSSD	Heart rate
ASCA	*r*	−0.29	0.23	0.19	−0.07	−0.22
	p	**0.00**	**0.00**	**0.02**	0.37	0.08
	n	234	234	153	152	61
Interest	*r*	−0.18	0.20	0.01	0.08	0.12
	p	**0.01**	**0.00**	0.91	0.32	0.35
	n	234	234	153	153	61
Perceived stress	*r*	0.14	0.05	−0.08	0.10	0.13
	p	**0.01**	0.37	0.24	0.15	0.13
	n	316	316	227	226	135

ASCA, Academic self-concept of ability; LF/HF ratio, ratio between low and high frequencies; RMSSD, root mean square of successive differences; significant values are bold.

TABLE 7 | Linear regression analysis with variables ASCA, interest and perceived stress as predictor variables for **mental effort** (for the total sample).

Coefficient	*B*	*SE*(*B*)	BETA	*p*	VIF
(Constant)	−0.254	0.070		**<0.001**	
ASCA	0.094	0.082	0.092	0.252	1.207
Interest	0.133	0.077	0.138	0.087	1.206
Perceived stress	0.114	0.074	0.114	0.124	1.011

ASCA, Academic self-concept of ability; adjusted R^2 = 0.035; Durbin–Watson–Statistic = 1.664; significant values are bold.

TABLE 8 | Linear regression analysis with variables ASCA, interest and perceived stress as predictor variables for **mental load** (for the total sample).

Coefficient	*B*	SE(B)	BETA	*p*	VIF
(Constant)	−0.198	0.060		**0.001**	
ASCA	−0.177	0.070	−0.199	**0.012**	1.207
Interest	−0.049	0.066	−0.059	0.458	1.206
Perceived stress	0.149	0.063	0.170	**0.020**	1.011

ASCA, Academic self-concept of ability; adjusted R^2 = 0.069; Durbin–Watson–Statistic = 1.807; significant values are bold.

TABLE 9 | Linear regression analysis with variables ASCA, interest and perceived stress as predictor variables for **LF/HF ratio** (for the total sample).

Coefficient	*B*	*SE*(*B*)	BETA	*p*	VIF
(Constant)	–0.183	0.065		**0.006**	
ASCA	0.177	0.076	0.206	**0.021**	1.242
Interest	–0.052	0.075	–0.061	0.492	1.258
Perceived stress	–0.127	0.067	–0.153	0.058	1.019

ASCA, Academic self-concept of ability; adjusted R^2 = 0.043; Durbin–Watson–Statistic = 1.495; significant values are bold.

TABLE 10 | Linear regression analysis with variables ASCA, interest and perceived stress as predictor variables for **RMSSD** (for the total sample).

Coefficient	*B*	*SE*(*B*)	BETA	*p*	VIF
(Constant)	0.069	0.082		0.398	
ASCA	–0.127	0.095	–0.119	0.185	1.242
Interest	0.117	0.094	0.113	0.214	1.258
Perceived stress	0.166	0.084	0.161	**0.048**	1.019

ASCA, Academic self-concept of ability; adjusted R^2 = 0.027; Durbin–Watson–Statistic = 1.083; significant values are bold.

TABLE 11 | Linear regression analysis with variables ASCA, interest and perceived stress as predictor variables for **heart rate** (for the total sample).

Coefficient	*B*	*SE*(*B*)	BETA	*p*	VIF
(Constant)	0.400	0.150		**0.010**	
ASCA	–0.416	0.189	–0.297	**0.032**	1.128
Interest	0.231	0.150	0.214	0.127	1.185
Perceived stress	0.044	0.143	0.040	0.760	1.076

ASCA, Academic self-concept of ability; adjusted R^2 = 0.046; Durbin–Watson–Statistic = 1.252; significant values are bold.

DISCUSSION

The aims of this study were to evaluate the convergence between subjective (self-reports on mental load and mental effort) and objective (heart rate measures: LF/HF ratio, RMSSD and heart rate) measures of cognitive load and to provide evidence for the assumed relationships between cognitive load and conceptually related constructs of positive and negative motivation and affect (self-concept, interest, and perceived stress). Related to the convergence of cognitive load measures via subjective and objective measures (RQ1), we only found significant correlations between self-reported ME and two of the heart rate variability measures (positive for LF/HF ratio and negative for heart rate) and no correlation between self-reported mental load and any heart rate measure. Hence, the assumption of a significant linear relationship between subjective and objective measures of cognitive load (positive for LF/HF ratio and heart rate, negative for RMSSD; H1), can only be confirmed for the specific relationship between mental effort and LF/HF ratio. Even for this relationship, the shared variance is rather small. Therefore, the present findings support the assumption that subjective and objective measures are likely to capture separate aspects of cognitive load (Zheng and Cooke, 2012; Solhjoo et al., 2019).

Related to the question how subjective and objective measures of cognitive load predict task performance (RQ2), we found that the subjective measures contributed significantly to explain variance in students' task performance and one of the objective measures (LF/HF ratio) contributed marginally significantly. Based on these findings, the assumption of a contribution of subjective and objective measures of cognitive load to predict students' task performance can be confirmed for mental effort and mental load and also for LF/HF ratio. As suggested in prior research, higher mental load indicates the perception of a more challenging task and is, thus, associated with lower achievement, whereas higher mental effort indicates students' engagement (Krell, 2017). However, in the basic correlational analysis the coefficients are rather small, indicating that different relationships may also exist. For example, for some students, high mental effort might also indicate high load because of an overly demanding task (Paas et al., 2003).

Regarding the question, how students' self-concept, interest and stress perception predict their subjectively and objectively measured cognitive load (RQ3), the assumption of a positive linear relationship between learner characteristics (self-concept and interest) and self-reported level of mental effort has to be rejected and the missing of a linear relationship between stress perception and students' self-reported level of mental effort (H3a) can be confirmed, because none of the variables self-concept, interest and perceived stress significantly explained self-reported mental effort. Hence, unlike it was found in preceding studies (Milyavskaya et al., 2018; Skuballa et al., 2019), the students' self-concept and interest did not elicit a more intense engagement with the tasks. Related to mental load, a negative linear relationship to measures of self-concept and interest can be partly confirmed (for ASCA) and also a positive linear relationship to students' stress perception (H3b) can be confirmed. This indicates that the students' self-concept and perceived stress significantly contributed to how cognitively challenging the tasks were perceived. Hence, unlike originally proposed by Paas and van Merriënboer (1994), the present measures of mental load are not independent from person characteristics: although it is possible to objectively define a given task's complexity (e.g., in terms of interacting elements to be processed or level of cognitive demands; Krell, 2018; Minkley et al., 2018), the perception of task complexity and amount of cognitive resources, which are necessary to process it, is intertwined with person characteristics, such as self-concept.

Finally, related to the objective measures and based on the regression analyses, the assumption of a linear relationship between objective measures of cognitive load and measures of self-concept and interest and of no linear relationship to perceived stress (H3c) can only be partly confirmed: ASCA contributes to heart rate (positively) and LF/HF ratio (positively), while perceived stress contributes to LF/HF ratio (negatively) and RMSSD (negatively). The former refuses earlier findings where higher ASCA values are associated with less stress (Minkley et al., 2014), as we found higher LF/HF and heart rate values in students with a higher self-concept compared to those

with a lower self-concept. This contrasts with the Biopsychosocial Model of Challenge and Threat (Blascovich and Mendes, 2000; Blascovich, 2008), according to which threat and stress arises when task demands exceed resources. For students with a high self-concept—which corresponds to high resources—tasks should be more a challenge than a threat, as they should assess their own resources to exceed task demands. But perhaps in our study the threat increases also for the students with a high self-concept because they want to maintain their own high ASCA and fear not to meet it in the test, which seems to be quite difficult indicated by the relatively high mental load reported by the participants and the rather low test performance.

The assumed and (at least partly) confirmed lack of a relationship between objective stress measurements and perceived stress might result from the differences between physiological and psychological stress responses, which—like mental load—corresponds with the complexity of a task (e.g., Kahneman and Peavler, 1969; Veltman and Gaillard, 1993; Hjortskov et al., 2004; Minkley and Kirchner, 2012): According to the transactional stress model of Lazarus and Folkman (1984) and the Biopsychosocial Model of Challenge and Threat (Blascovich and Mendes, 2000; Blascovich, 2008), stress is the result of the interplay between situational demands and individual resources. Thus, the perception of high demands and low resources creates the feeling of being stressed and raises various physiological responses (e.g., decreased heart rate variability, increased cortisol secretion; Rensing et al., 2006) which are often not related to the respondents' perceived stress (for a review see Campbell and Ehlert, 2012). Several studies which already investigated the association between physiological (primarily cortisol secretion) and psychological (perceived stress or anxiety) stress responses reported heterogeneous results (for a review see Campbell and Ehlert, 2012). In most studies, there was no systematic relationship between these parameters (Buchanan et al., 1999; Weekes et al., 2006; Campbell and Ehlert, 2012; Minkley et al., 2014; Kärner et al., 2018; Ringeisen et al., 2019). Only a few studies found low or moderate associations between increased cortisol concentration and perceived stress (e.g., Spangler et al., 2002; Lindahl et al., 2005). Campbell and Ehlert (2012) discuss various factors as possible reasons for this disassociation (e.g., differing assessment protocols, mediating factors and interindividual differences in the degree of psychophysiological correspondence). Beyond these formal reasons—which are partly founded in the matter itself—it is also conceivable that physiological and psychological stress responses represent different aspects of a person's reaction toward a stress situation.

Comparing the findings for the total sample presented above with the findings for the three individual studies (**Supplementary Material**), it becomes evident that - in most cases - there is the same trend in the individual studies and the total sample. Caused by the smaller sample sizes, several coefficients do not reach the 5% p-level in the individual studies but then do in the total sample analysis. For example, while there are four statistically significant correlation coefficients in **Table 4**, only two (studies 1 and 2) or one (study 3) of them reaches the 5% level in the individual studies (**Supplementary Material**); however, there have been significant results in some of the single studies that could not be detected in the total sample. Most notably, related to the convergence of cognitive load measures via subjective and objective measures (**Table 3**), no significant association between LF/HF ratio and heart rate was found for the total sample but for study 1 and study 3. Albeit not being related to one of the research questions and hypotheses addressed here, the opposed findings from two studies indicate the challenging nature of assessing physiological measures of cognitive load. One challenge here could be that our tests did not take place under real conditions at school and therefore may only have had a low personal relevance for the participants. This can lead to a generally low physiological stress response, since the factor "personal relevance", which contributes to something becoming a stressor, is not or only slightly pronounced. With these rather low stress responses, even small differences (e.g., regarding the content of the tasks) could have different effects on the different heart rate variability measures.

It has been suggested that comparison studies with subjective and objective measures of cognitive load may lead to new insights on what these two types of measures are in fact measuring, hence advancing our understanding of the construct of cognitive load in terms of its subjective and objective measurement (Leppink et al., 2013). Other scholars proposed such comparison studies as a source of validity evidence for subjective measures (Solhjoo et al., 2019). As several objective (i.e., physiological) measures of cognitive load have been suggested (e.g., various heart rate or pupillometric measures; Solhjoo et al., 2019; Zheng and Cooke, 2012; cf. Sweller et al., 2011), it remains unclear as to which objective measure can be validly interpreted as an indicator for an individual's cognitive load and in which contexts. This is additionally highlighted in the present study, with specific findings for the three heart rate measures (e.g., **Table 3**). Nevertheless, taken that each cognitive load measure may capture specific aspects of cognitive load (Zheng and Cooke, 2012), a systematic association between subjective and objective measures of cognitive load, with still a medium to large variance component specific to each measure, might be seen as providing evidence for the measures in fact indicating cognitive load. As this study revealed, objective heart rate measures of cognitive load are significantly related to self-reported mental effort but not to mental load. Such objective measures might primarily be used in future studies to indicate the person-relevant cognitive load dimension of mental effort and not the task-relevant cognitive load dimension of mental load (Paas and van Merriënboer, 1994). Similarly, the relationships between subjective measures of mental load and mental effort and further variables of task performance and learner characteristics are generally in line with what has been expected based on the framework presented in **Figure 1**. This can be seen as further validity evidence based on relations to other variables for the subjective measures of mental load and mental effort (Krell, 2015, 2017). Opposed to this, the findings related to the objective heart rate measures are less clear. This further illustrates the challenges associated with establishing objective cognitive load measures; for example, "physiological measures have proved insufficiently sensitive

to indicate the differences in cognitive load generated by the instructional designs used by cognitive load theory" (Sweller et al., 2011, p.81). In sum, as validity refers to the extent to which evidence and theory support the intended interpretation of test scores (Kane, 2013; AERA et al., 2014), a clear theoretical framework of cognitive load including the different objective measures would be needed to derive clear hypotheses about the specific relationships between measures of cognitive load and, hence, allow to interpret related findings as validity evidence for single measures. From this point of view, the present study rather adds to our understanding of the construct of cognitive load in terms of its subjective and objective measurement (Leppink et al., 2013), than to provide strong validity evidence for subjective measures (Solhjoo et al., 2019).

Limitations

Naturally, this study has several limitations. First, self-reported measures are vulnerable to several biases (e.g., social desirability, cultural background), especially in the case of measuring "interest", where we have used different instruments. Second, we assumed, tested, and found linear relationships between some of the included variables; however, for some variables, non-linear relationships might be considered as well. For example, Paas et al. (2003) discuss a non-linear relationship between task performance and ME. Third, we did not systematically investigate the causal factors of learning task and learning environment (Choi et al., 2014) and also did not consider the two- and three-way-interactions proposed in the framework for cognitive load (**Figure 1**). As a minor limitation, one of the measurements (LF/HF ratio) shows kurtosis above 2, which is outside normal distribution. This could be due to sample fluctuation. Hence, future studies should investigate whether the specific relationships between person characteristics and dimensions of cognitive load, which were found in this study, might be specific for the present tasks and environment (i.e., tasks dealing with molecular structures and phenomena, presented digitally on laptops, and solved during an out-of-school experience). Also, future study could replicate the findings through other objective measures such as pupillary measures and skin conductance measures.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available upon request to the first author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethik Kommission der Medizinischen Fakultät der Ruhr-Universität Bochum, Gesundheitscampus 33, 44801 Bochum, Germany. Written informed consent to participate in this study was obtained from participants and their legal guardians if they were under the age of majority.

AUTHOR CONTRIBUTIONS

NM conceived the project, designed the study, conducted the experiment, analyzed the data, and wrote the manuscript. MK conceived the project, designed the study, analyzed the data, and wrote the manuscript. KX specified the theoretical framework and wrote the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2021.632907/full#supplementary-material.

REFERENCES

AERA, APA, and NCME. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Ainley, M., Hidi, S., and Berndorff, D. (2002). Interest, learning, and the psychological processes that mediate their relationship. *J. Educ. Psychol.* 94 (3), 545. doi:10.1037/0022-0663.94.3.545

Antonenko, P., Paas, F., Grabner, R., and van Gog, T. (2010). Using electroencephalography to measure cognitive load. *Educ. Psychol. Rev.* 22 (4), 425–438. doi:10.1007/s10648-010-9130-y

Blascovich, J. (2008). "Challenge, threat, and health," in *Handbook of motivation science*. Editors W. L. Gardner and W. L. Gardner (New York: The Guilford Press), 481–493.

Blascovich, J., and Mendes, W. B. (2000). "Challenge and threat appraisals: the role of affective cues," in *Studies in emotion and social interaction, second series. Feeling and thinking: the role of affect in social cognition*. Editor J. P. Forgas (New York: Cambridge University Press), 59–82.

Buchanan, T. W., al'Absi, M., and Lovallo, W. R. (1999). Cortisol fluctuates with increases and decreases in negative affect. *Psychoneuroendocrinology* 24 (2), 227–241.doi:10.1016/s0306-4530(98)00078-x

Campbell, J., and Ehlert, U. (2012). Acute psychosocial stress: does the emotional stress response correspond with physiological responses? *Psychoneuroendocrinology* 37 (8), 1111–1134. doi:10.1016/j.psyneuen.2011.12.010

Choi, H.-H., van Merriënboer, J. J. G., and Paas, F. (2014). Effects of the physical environment on cognitive load and learning: towards a new model of cognitive load. *Educ. Psychol. Rev.* 26, 225–244. doi:10.1007/s10648-014-9262-6

Cranford, K. N., Tiettmeyer, J. M., Chuprinko, B. C., Jordan, S., and Grove, N. P. (2014). Measuring load on working memory: the use of heart rate as a means of measuring chemistry students' cognitive load. *J. Chem. Educ.* 91 (5), 641–647. doi:10.1021/ed400576n

Csikszentmihalyi, M. (2013). *Flow: the psychology of happiness*. London: Random House.

de Jong, T. (2010). Cognitive load theory, educational research, and instructional design: some food for thought. *Instr. Sci.* 38, 105–134. doi:10.1007/s11251-009-9110-0

Dickerson, S. S., and Kemeny, M. E. (2004). Acute stressors and cortisol responses: a theoretical integration and synthesis of laboratory research. *Psychol. Bull.* 130 (3), 355–391. doi:10.1037/0033-2909.130.3.355

Fraser, K., Ma, I., Teteris, E., Baxter, H., Wright, B., and McLaughlin, K. (2012). Emotion, cognitive load and learning outcomes during simulation training. *Med. Educ.* 46 (11), 1055–1062. doi:10.1111/j.1365-2923.2012.04355.x

Goldinger, S. D., and Papesh, M. H. (2012). Pupil dilation reflects the creation and retrieval of memories. *Curr. Dir. Psychol. Sci.* 21 (2), 90–95.doi:10.1177/0963721412436811

Gravetter, F. J., and Wallnau, L. B. (2012). *Statistics for the behavioral sciences*. Belmont, CA: Wadsworth Cengage Learning.

Hartig, J., Frey, A., and Jude, N. (2012). "Validität," in *Testtheorie und Fragebogenkonstruktion*. Editors H. Moosbrugger and A. Kelava (Berlin, Heidelberg: Springer). doi:10.1007/978-3-642-20072-4_7

Hawthorne, B. S., Vella-Brodrick, D. A., and Hattie, J. (2019). Well-being as a cognitive load reducing agent: a review of the literature. *Front. Educ.* 4, 121. doi:10.3389/feduc.2019.00121

Hjortskov, N., Rissn, D., Blangsted, A. K., Fallentin, N., Lundberg, U., and Sgaard, K. (2004). The effect of mental stress on heart rate variability and blood pressure during computer work. *Eur. J. Appl. Physiol.* 92 (1-2), 84–89. doi:10.1007/s00421-004-1055-z

Huh, D., Kim, J. H., and Jo, I. H. (2019). A novel method to monitoring changes in cognitive load in video-based learning. *J. Comput. Assist. Learn.* 35, 721–730. doi:10.1111/jcal.12378

Ikehara, C. S., and Crosby, M. E. (2005). "Assessing cognitive load with physiological sensors," in Proceedings of the 38th Annual Hawaii international conference on system sciences, 1–9.

Isowa, T., Ohira, H., and Murashima, S. (2006). Immune, endocrine and cardiovascular responses to controllable and uncontrollable acute stress. *Biol. Psychol.* 71 (2), 202–213. doi:10.1016/j.biopsycho.2005.04.002

Kahneman, D., and Peavler, W. S. (1969). Incentive effects and pupillary changes in association learning. *J. Exp. Psychol.* 79 (2Pt.1), 312–318. doi:10.1037/h0026912

Kalyuga, S., Ayres, P., Chandler, P., and Sweller, J. (2003). The expertise reversal effect. *Educ. Psychol.* 38, 23–31. doi:10.1207/S15326985ep3801_4

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *J. Educ. Meas.* 50 (1), 1–73. doi:10.2307/23353796doi:10.1111/jedm.12000

Kärner, T., Minkley, N., Rausch, A., Schley, T., and Sembill, D. (2018). Stress and resources in vocational problem solving. *Vocations Learn.* 11 (2), 365–398. doi:10.1007/s12186-017-9193-8

Kennedy, D. O., and Scholey, A. B. (2000). Glucose administration, heart rate and cognitive performance: effects of increasing mental effort. *Psychopharmacology* 149, 63–71. doi:10.1007/s002139900335

Kirschner, P. A., Ayres, P., and Chandler, P. (2011). Contemporary cognitive load theory research: The good, the bad and the ugly. *Comput. Human Behav.* 27 (1), 99–105. doi:10.1016/j.chb.2010.06.025

Kirschner, P. A., Sweller, J., and Clark, R. E. (2006). Why minimal guidance during instruction does not work: an analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educ. Psychol.* 41, 75–86. doi:10.1207/s15326985ep4102_1

Klepsch, M., Schmitz, F., and Seufert, T. (2017). Development and validation of two instruments measuring intrinsic, extraneous, and germane cognitive load. *Front. Psychol.* 8, 1997. doi:10.3389/fpsyg.2017.01997

Knigge, M., Krauskopf, K., and Wagner, S. (2019). Improving socio-emotional competencies using a staged video-based learning program? Results of two experimental studies. *Front. Educ.* 4, 142. doi:10.3389/feduc.2019.00142

Knörzer, L., Brünken, R., and Park, B. (2016). Facilitators or suppressors: effects of experimentally induced emotions on multimedia learning. *Learn. Instruct.* 44, 97–107. doi:10.1016/j.learninstruc.2016.04.0

Krell, M. (2017). Evaluating an instrument to measure mental load and mental effort considering different sources of validity evidence. *Cogent Educ.* 4, 1280256. doi:10.1080/2331186X.2017.1280256

Krell, M. (2018). Schwierigkeitserzeugende Aufgabenmerkmale bei Multiple-Choice-Aufgaben zur Experimentierkompetenz im Biologieunterricht: eine Replikationsstudie. *ZfDN* 24, 1–15. doi:10.1007/s40573-017-0069-0

Krell, M. (2015). Evaluating an instrument to measure mental load and mental effort using item response theory. *Sci. Educ. Rev. Lett. Res. Lett.* 2015, 1–6. doi:10.18452/8212

Laborde, S., Mosley, E., and Thayer, J. F. (2017). Heart rate variability and cardiac vagal tone in psychophysiological research - recommendations for experiment planning, data analysis, and data reporting. *Front. Psychol.* 08, 1–18. doi:10.3389/fpsyg.2017.00213

Lazarus, R. S., and Folkman, S. (1984). *Stress, appraisal, and coping*. New York, NY: Springer.

Leppink, J., Paas, F., van der Vleuten, C. P. M., van Gog, T., and Van Merriënboer, J. J. G. (2013). Development of an instrument for measuring different types of cognitive load. *Behav. Res.* 45 (4), 1058–1072. doi:10.3758/s13428-013-0334-1

Lindahl, M., Theorell, T., and Lindblad, F. (2005). Test performance and self-esteem in relation to experienced stress in Swedish sixth and ninth graders--saliva cortisol levels and psychological reactions to demands. *Acta Paediatr.* 94, 489–95.doi:10.1111/j.1651-2227.2005.tb01922.x

Luria, R. E. (1975). The validity and reliability of the visual analogue mood scale. *J. Psychiatr. Res.* 12 (1), 51–57. doi:10.1016/0022-3956(75)90020-5

Malik, M., Bigger, J. T., Camm, A. J., Kleiger, R. E., Malliani, A., Moss, A. J., et al. (1996). Heart rate variability: Standards of measurement, physiological interpretation, and clinical use. *Eur. Heart J.* 17 (3), 354–381. doi:10.1093/oxfordjournals.eurheartj.a014868

Marsh, H. W., Trautwein, U., Lüdtke, O., Köller, O., and Baumert, J. (2005). Academic self-concept, interest, grades, and standardized test scores: reciprocal effects models of causal ordering. *Child. Dev.* 76 (2), 397–416. doi:10.1111/j.1467-8624.2005.00853.x

Marsh, H. W., Xu, M., and Martin, A. J. (2012). "Self-concept: a synergy of theory, method, and application," in *Theories, constructs, and critical issues. APA educational psychology handbook*. (New York, NY: American Psychological Association), 427–458.

Messick, S. (1995). Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *Am. Psychol.* 50, 741–749. doi:10.1037/0003-066x.50.9.741

Milyavskaya, M., Galla, B., Inzlicht, M., and Duckworth, A. (2018). *More effort, less fatigue: how interest increases effort and reduces mental fatigue*. PsyArXiv http://psyarxiv.com/8npfx/

Minkley, N., Kärner, T., Jojart, A., Nobbe, L., and Krell, M. (2018). Students' mental load, stress, and performance when working with symbolic or symbolic-textual molecular representations. *J. Res. Sci. Teach.* 55, 1162–1187. doi:10.1002/tea.21446

Minkley, N., and Kirchner, W. H. (2012). Influence of test tasks with different cognitive demands on salivary cortisol concentrations in school students. *Int. J. Psychophysiology* 86 (3), 245–250. doi:10.1016/j.ijpsycho.2012.09.015

Minkley, N., Westerholt, D. M., and Kirchner, W. H. (2014). Academic self-concept of ability and cortisol reactivity. *Anxiety Stress Coping* 27, 303–316. doi:10.1080/10615806.2013.848273

Moreno, R. (2010). Cognitive load theory: more food for thought. *Instr. Sci.* 38, 135–141. doi:10.1007/s11251-009-9122-9

Nebel, S., Beege, M., Schneider, S., and Rey, G. D. (2016). The higher the score, the higher the learning outcome? Heterogeneous impacts of leaderboards and choice within educational videogames. *Comput. Hum. Behav.* 65, 391–401. doi:10.1016/j.chb.2016.08.042

Nehring, A., Nowak, K., Upmeier zu Belzen, A., and Tiemann, R. (2012). Doing inquiry in chemistry and biology: the context's influence on the students' cognitive load. *La Chimica nella Scuola* IV, 253–258. doi:10.4324/9781315748047

Paas, F. G. W. C. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: a cognitive-load approach. *J. Educ. Psychol.* 84, 429–434. doi:10.1037/0022-0663.84.4.429

Paas, F. G. W. C., and Van Merriënboer, J. J. G. (1994). Instructional control of cognitive load in the training of complex cognitive tasks. *Educ. Psychol. Rev.* 6 (4), 351–371. doi:10.1007/BF02213420

Paas, F., Tuovinen, J. E., Tabbers, H., and van Gerven, P. W. M. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educ. Psychol.* 38, 63–71. doi:10.1207/S15326985EP3801_8

Piestrzeniewicz, K., Łuczak, K., Lelonek, M., Wranicz, J. K., and Goch, J. H. (2008). Obesity and heart rate variability in men with myocardial infarction. *Cardiol. J.* 15 (1), 43–9. doi:10.1207/s15326985ep3801_8

Plass, J. L., and Kalyuga, S. (2019). Four ways of considering emotion in cognitive load theory. *Educ. Psychol. Rev.* 31 (2), 339–359. doi:10.1007/s10648-019-09473-5

Rensing, L., Koch, M., Rippe, B., and Rippe, V. (2006). *Mensch im Stress. Psyche, Körper, Moleküle. [Man under stress]*, München: Spektrum,

Ringeisen, T., Lichtenfeld, S., Becker, S., and Minkley, N. (2019). Stress experience and performance during an oral exam: the role of self-efficacy, threat appraisals, anxiety, and cortisol. *Anxiety Stress Coping* 32 (1), 50–66. doi:10.1080/10615806.2018.1528528

Rost, D., Sparfeldt, J., and Schilling, S. (2007). *DISK-GITTER mit SKSLF-8. Differentielles schulisches Selbstkonzept-Gitter mit Skala zur Erfassung des Selbstkonzepts schulischer Leistungen und Fähigkeiten [The DISC-grid with SKSLF-8. Differential self-concept grid with a scale for capturing the self-concept of school relevant accomplishments and abilities]*. Göttingen: Hogrefe.

Seery, M. D. (2011). Challenge or threat? Cardiovascular indexes of resilience and vulnerability to potential stress in humans. *Neurosci. Biobehav. Rev.* 35 (7), 1603–1610. doi:10.1016/j.neubiorev.2011.03.003

Seery, M. D. (2013). The biopsychosocial model of challenge and threat: using the heart to measure the mind. *Social Personal. Psychol. Compass* 7 (9), 637–653. doi:10.1111/spc3.12052

Seufert, T. (2020). Building bridges between self-regulation and cognitive load-an invitation for a broad and differentiated attempt. *Educ. Psychol. Rev.* 32, 1151–1162. doi:10.1007/s10648-020-09574-6

Skuballa, I. T., Xu, K. M., and Jarodzka, H. (2019). The impact of co-actors on cognitive load: when the mere presence of others makes learning more difficult. *Comput. Hum. Behav.* 101, 30–41. doi:10.1016/j.chb.2019.06.016

Solhjoo, S., Haigney, M. C., McBee, E., van Merriënboer, J. J. G., Schuwirth, L., Artino, A. R., et al. (2019). Heart rate and heart rate variability correlate with clinical reasoning performance and self-reported measures of cognitive load. *Sci. Rep.* 9, 14668. doi:10.1038/s41598-019-50280-3

Spangler, G., Pekrun, R., Kramer, K., and Hofmann, H. (2002). Students' emotions, physiological reactions, and coping in academic exams. *Anxiety Stress Coping* 15 (4), 413. doi:10.1080/1061580021000056555

Sparfeldt, J. R., Rost, D. H., and Schilling, S. R. (2004). Schulfachspezifische interessen—ökonomisch gemessen. *Psychol. Erziehung Unterricht* 51, 213–220. doi:10.1207/s15326985ep3801_8

Sweller, J., Ayres, P., and Kalyuga, S. (2011). "Measuring cognitive load," in *Cognitive load theory*. Editors J. Sweller, P. Ayres, and S. Kalyuga (New York: Springer), 71–85.

Sweller, J., van Merriënboer, J. J. G., and Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educ. Psychol. Rev.* 31 (2), 261–292. doi:10.1007/s10648-019-09465-5

Thayer, J. F., Yamamoto, S. S., and Brosschot, J. F. (2010). The relationship of autonomic imbalance, heart rate variability and cardiovascular disease risk factors. *Int. J. Cardiol.* 141, 122–131. doi:10.1016/j.ijcard.2009.09.543

van Gerven, P. W. M., Paas, F. G. W. C., Van Merriënboer, J. J. G., and Schmidt, H. G. (2002). Cognitive load theory and aging: effects of worked examples on training efficiency. *Learn. Instruction* 12, 87–105. doi:10.1016/s0959-4752(01)00017-2

van Gog, T., and Paas, F. (2008). Instructional efficiency: revisiting the original construct in educational research. *Educ. Psychol.* 43 (1), 16–26. doi:10.1080/00461520701756248

Veltman, J. A., and Gaillard, A. W. K. (1993). Indices of mental workload in a complex task environment. *Neuropsychobiology* 28, 72–75. doi:10.1159/000119003

Weekes, N., Lewis, R., Patel, F., Garrison-Jakel, J., Berger, D. E., and Lupien, S. J. (2006). Examination stress as an ecological inducer of cortisol and psychological responses to stress in undergraduate students. *Stress* 9 (4), 199–206. doi:10.1080/10253890601029751

Wilde, M., Bätz, K., Kovaleva, A., and Urhahne, U. (2009). Überprüfung einer Kurzskala intrinsischer Motivation (KIM). *Z. für Didaktik der Naturwissenschaften* 15, 31–45. doi:10.1007/978-3-658-02539-7_7

Woody, A., Hooker, E. D., Zoccola, P. M., and Dickerson, S. S. (2018). Social-evaluative threat, cognitive load, and the cortisol and cardiovascular stress response. *Psychoneuroendocrinology* 97, 149–155. doi:10.1016/j.psyneuen.2018.07.009

Zheng, R., and Cook, A. (2012). Solving complex problems: a convergent approach to cognitive load measurement. *Br. J. Educ. Technol.* 43, 233–246. doi:10.1111/j.1467-8535.2010.01169.x

Zu, T., Hutson, J., Loschky, L. C., and Rebello, N. S. (2019). Using Eye movements to measure intrinsic, extraneous, and germane load in a multimedia learning environment. *J. Educ. Psychol.*, 112, 1338. doi:10.1037/edu0000441

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Education

ORIGINAL RESEARCH
published: 22 April 2021
doi: 10.3389/feduc.2021.648324

An Item Response Modeling Approach to Cognitive Load Measurement

John Fitzgerald Ehrich[1], Steven J. Howard[2], Sahar Bokosmaty[2] and Stuart Woodcock[3]*

[1] Faculty of Arts, Macquarie University, Sydney, NSW, Australia, [2] School of Education, Faculty of Social Sciences, University of Wollongong, Wollongong, NSW, Australia, [3] School of Education and Professional Studies, Faculty of Arts, Education and Law, Griffith University, Southport, QLD, Australia

The accurate measurement of the cognitive load a learner encounters in a given task is critical to the understanding and application of Cognitive Load Theory (CLT). However, as a covert psychological construct, cognitive load represents a challenging measurement issue. To date, this challenge has been met mostly by subjective self-reports of cognitive load experienced in a learning situation. In this paper, we find that a valid and reliable index of cognitive load can be obtained through item response modeling of student performance. Specifically, estimates derived from item response modeling of relative difficulty (i.e., the difference between item difficulty and person ability locations) can function as a linear measure that combines the key components of cognitive load (i.e., mental load, mental effort, and performance). This index of cognitive load (*relative* difficulty) was tested for criterion (concurrent) validity in Year 2 learners (N = 91) performance on standardized educational numeracy and literacy assessments. Learners' working memory (WM) capacity significantly predicted our proposed cognitive load (relative difficulty) index across both numeracy and literacy domains. That is, higher levels of WM were related to lower levels of cognitive load (relative difficulty), in line with fundamental predictions of CLT. These results illustrate the validity, utility and potential of this objective item response modeling approach to capturing individual differences in cognitive load across discrete learning tasks.

Keywords: cognitive load, item response theory, mental effort, working memory, standardized test

OPEN ACCESS

Edited by:
Kate M. Xu,
Open University of the Netherlands, Netherlands

Reviewed by:
Andrew J. Martin,
University of New South Wales, Australia
Melina Klepsch,
Abt. Lehr-Lernforschung, Universität Ulm, Germany

****Correspondence:***
John Fitzgerald Ehrich
john.ehrich@mq.edu.au

Specialty section:
This article was submitted to Assessment, Testing and Applied Measurement, a section of the journal Frontiers in Education

Received: *31 December 2020*
Accepted: *22 March 2021*
Published: *22 April 2021*

Citation:
Ehrich JF, Howard SJ, Bokosmaty S and Woodcock S (2021) An Item Response Modeling Approach to Cognitive Load Measurement. Front. Educ. 6:648324. doi: 10.3389/feduc.2021.648324

INTRODUCTION

The core goal of cognitive load theory (CLT) is the creation of learning environments that make optimal use of learners' cognitive resources and reduce any demands extraneous to learning in order to optimize learning success (Paas et al., 2003, 2004). In addition to the inherent complexity of information that is to be learned, the method of presenting information to learners also affects the cognitive load learners experience when acquiring knowledge and skills. However, the understanding and application of CLT requires methods to appraise cognitive load, which could be expected to differ across tasks, contexts and learners. To-date, this has been indexed mostly by subjective self-reports of cognitive load experienced in a learning situation. In this study, we evaluated a more objective and sensitive approach to indexing cognitive load experienced by learners.

Cognitive Load: Definition, Sources and Measurement

Cognitive load is considered to be a complex multidimensional construct that consists of: (1) causal factors relating to the task, the learner and their interactive components; and (2) assessment factors such as mental load (ML), mental effort (ME), and performance (e.g., Paas and van Merriënboer, 1994). The cognitive resources needed for a certain task comprise ML, which is a result of a task's content, presentation, structure, complexity and difficulty (Paas, 1992). On the other hand, the cognitive resources that are devoted to a task comprise ME (Paas, 1992; Paas et al., 2003). ME is intrinsic to the learner, and constitutes the degree to which cognitive resources are mobilized to enable processing and completion in complex tasks (Paas and van Merriënboer, 1994). The causal factor of cognitive load relates to aspects such as the novelty of the task and environmental conditions, while factors relating to the learner involve aspects like working memory (WM) capacity and expertise. These task and learner factors interact to further influence performance through their influence on, for example, motivation.

Cognitive load can be understood within three broad categories – intrinsic, extraneous, and germane (Sweller et al., 2019). Intrinsic cognitive load has to do with the complexity of the information which is being processed and subsumes the idea of "element interactivity" (Sweller et al., 2019). Element interactivity depends on the nature of the information and the prior knowledge of the learner processing the information. For example, complex tasks which require the processing of multiple interconnected elements are considered to have high element interactivity. By contrast, extraneous cognitive load has to do with how information is presented and the instructional procedures involved in the task. Manipulations of the presentation of instructional procedures can affect the level of element interactivity. Finally, germane cognitive load refers "[. . .] to the WM resources available to deal with the element interactivity associated with intrinsic cognitive load" (Sweller, 2010, p. 126). Therefore, germane cognitive load is both linked to intrinsic and extraneous cognitive load. Germane cognitive load resources can only be utilized if extraneous cognitive load is not depleting WM resources. Moreover, germane cognitive load can redistribute WM resources to process complex tasks with high element interactivity (Sweller et al., 2019).

As a covert psychological construct, which can be expected to vary across tasks, contexts and learners, cognitive load constitutes a serious challenge in terms of its accurate measurement. Without precision in its capture, application of CLT is limited to the identification of conditions under which learning is superior or inferior, without the ability to accurately tailor these principles to the specific tasks, conditions and learners involved in a particular learning situation. For instance, the split attention effect would suggest that when learners are novice, essential information should be well integrated; however, this might not be expected at higher levels of expertise. Application of this principle to optimize learning outcomes amongst diverse tasks (e.g., in reading, numeracy, and science), diverse learners (e.g., in expertise and WM capacity), and in different contexts to which the research was conducted, is complicated without the ability to carefully appraise changes in cognitive load as conditions change.

When cognitive load is measured it is most often done through the use of a subjective ranking using a Likert scale asking for invested ME (e.g., Marcus et al., 1996; Tindall-Ford et al., 1997; Salden et al., 2004; Halabi et al., 2005). A primary reason is that this method is straightforward, simple to apply, shows evidence of reliability, construct validity, and does not interfere with learning (Paas et al., 1994; Sweller et al., 1998). For instance, Paas (1992) used a one-dimensional 9-point symmetrical category rating scale (Likert-type scale) for assessing learners' ME in different phases of learning and performance. The scale ranged from 1 (very low mental effort) to 9 (very high mental effort), on which learners rank their ME during a learning and performance task. Paas et al. (1994) tested this subjective scale for its measurement properties and found that it had good reliability (e.g., Cronbach $\alpha = 0.82$) and was sensitive to variation in small levels of cognitive load. Such evidence is taken to suggest that learners are capable of introspecting their cognitive processes and use this to quantify their ME.

However, this scale has been interpreted by some cognitive load researchers by substituting "mental effort" with "task difficulty" (e.g., Ayres, 2006; Cierniak et al., 2009). By itself, asking learners to rank difficulty of learning tasks as a measure of ME is problematic. While ME and task difficulty are no doubt related, as a consequence of factors such as prior knowledge, they are not identical (van Gog and Paas, 2008). For instance, when tasks are very difficult for learners, research shows they are often not stimulated to put in the required ME (Wright, 1984; Wright et al., 1986) and, as a result, may not be reflective of the task's cognitive load. Despite this, Sweller et al. (2011, p. 74) state that the subjective ME scale has "[. . .] been shown to be the most sensitive measure available to differentiate the cognitive load imposed by different instructional procedures."

From these scales, ME (cognitive load) is indexed through a combination of the learning result and learners' ME. That is, a learning experience is considered more optimal if it has a higher average performance than an alternative condition. Yet when two instructional conditions record the same average performance the learning condition that requires less ME has higher instructional efficiency. Accordingly, the learning condition that needs more ME is considered to be less efficient than the one that requires learners to exert less ME. Using a cognitive load framework, Paas and van Merriënboer (1993) suggested a method for quantifying this instructional efficiency. Their formula, $E = \frac{(P-R)}{\sqrt{2}}$, reconciles: (E), the relative efficiency of the instructional condition; (P), the standardized z-scores for test performance scores; and (R), the standardized z-scores for the ratings of cognitive load related to the task. Based on this formula, a learning condition would be more efficient when lower subjective ratings of cognitive load correspond with higher performance scores. These scores are calculated per learner and per task, and interpreted relative to an ideal slope of 1, where instructional efficiency = 0 (or performance is equal to ME). Proximity above or below this slope denotes high or low mental efficiency, respectively. This mental efficiency model

has since been expanded to include factors such as motivation (Hummel et al., 2004).

However, Hoffman and Schraw (2010) have pointed out fundamental measurement concerns with Paas and van Merriënboer's (1993) cognitive load efficiency model beyond the well-documented issues of using self-report measures, such as measurement error arising from rater bias and over-confidence (e.g., Stone, 2000; Burson et al., 2006). Hoffman and Schraw note that task performance scores and ME scores are not commensurable and do not share a common unit of measurement. Calculations derived from incommensurable variables are problematic for interpretative and computational reasons (see Hoffman and Schraw, 2010).

Recently, studies have attempted to measure the different aspects of cognitive load (e.g., Leppink et al., 2013; Klepsch et al., 2017; Krell, 2017). For example, Krell (2017) developed a seven-point Likert scale to measure self-reported levels of cognitive load. In this study, Krell used an item response theory (IRT) approach to test the linear functioning of the self-report scale. This scale consists of 12 items, half of which measure ML (i.e., the cognitive capacity to process tasks) and the other half to measure ME (the investment of cognitive capacity by persons to process tasks). Krell tested the scale on a large sample of high school students on the performance of a standardized science test. Krell found evidence that ML and ME were different dimensions and some evidence which suggest a causal role between ML and performance but not ME and performance.

Whereas the majority of cognitive load researchers have used subjective self-report, a range of objective cognitive load measurement techniques have also been explored by cognitive load researchers (for overviews see Paas et al., 2003; Paas et al., 2008). Whereas subjective techniques are normally used to get an estimate of overall cognitive load, that is, experienced load based on the whole task procedure, continuous objective techniques can be used to determine the dynamics of cognitive load through fluctuations in cognitive load from the beginning to the end of the task (Xie and Salvendy, 2000; Paas et al., 2003). Such approaches include neuroscience (e.g., Antonenko et al., 2010; Howard et al., 2015), physiological measurements such as heart rate (e.g., Paas and van Merriënboer, 1994), pupil dilation (van Gerven et al., 2004), and blood glucose levels (e.g., Scholey et al., 2001).

Other objective cognitive load measurement techniques involve the use of secondary tasks. Secondary-task techniques are based on the assumption that performance on a secondary task can be used to reflect the level of cognitive load imposed by a primary task, and have been used successfully by several cognitive load researchers (e.g., Chandler and Sweller, 1996; Marcus et al., 1996). A recent and promising example of this technique is the rhythm method (Park and Brünken, 2015; Korbach et al., 2018). With this technique participants have to execute a previously practiced rhythm continuously by foot tapping (secondary task) while learning (primary task). Eye-tracking analysis is another objective technique to measure cognitive load. These studies investigate fixation time and number of fixations on visual stimuli as indications of ME and cognitive load (see Korbach et al., 2017; Krejtz et al., 2018).

In summary, cognitive load has been measured primarily through the use of subjective self-report scales. Less common objective measures of cognitive load have been attained through brain imaging, the monitoring of physiological processes, the use of secondary tasks, and eye tracking. While such studies (e.g., neuroscientific (fMRI) approaches to cognitive load measurement) have shown great potential (Whelan, 2007) they are cumbersome, intrusive, require considerable technical expertise beyond the capability of most CLT researchers, and are unclear about which type of cognitive load is being measured. Moreover, such measurement approaches lack ecological validity and occur within laboratory settings outside of the typical classroom learning environment. An ideal measure of cognitive load would be objective, unobtrusive, and measurable within a typical classroom environment.

A Measure of Cognitive Load Through Rasch Modeling

Self-report Likert scale ratings do not constitute measures in so far as, technically, they are *observations* and, as such, do not meet the basic requirements of measurement (Wright, 1997). Likert scale raw scores provide ordinal data, which means that: (1) the scale is finite or limited to a small number of observations (e.g., 5-, 7-, or 9-point); and (2) that differences between observations (i.e., ratings) are not equidistant from each other, as in an interval or ratio level scale. For a scale to qualify as a linear measure it needs to be boundless, or not limited to a finite set of observations and, critically, needs to consist of equally divisible units. Hence, a serious problem of measurement error arises when Likert scales are used as substitute measures in parametric analyses, such as analyses of variance (ANOVA) (Wright, 1997). Ideally, a behavioral measure of ME would be derived through an objective procedure that fulfills the measurement principles of a linear continuum with interval-level units. Item response modeling presents such an opportunity, while using some of the same data (e.g., performance) as CLT efficiency indices.

The Rasch Model

The Rasch (1960) model, or the one parameter logistic model (1PL), is a commonly used model in IRT. The Rasch model is a mathematical model of probability predicated on a hierarchy of item difficulties. This hierarchy of item difficulty is determined by conformity to a Guttman scalar pattern. The model depicts the probability of getting an item correct/incorrect as a logistic function of the distance between a person's location (ability) and an item's location (difficulty). These location estimates are situated on the same linear scale (i.e., logit scale). This relationship is expressed below in mathematical form for dichotomous data (e.g., correct/incorrect test answers):

$$P\{X_{ni} = x\} = \frac{e^{x(\beta_n - \delta_i)}}{1 + e^{(\beta_n - \delta_i)}}$$

Where P = probability of X at person n for item i and where x represents either a correct ($x = 1$) or an incorrect ($x = 0$) response. Person locations are denoted as β_n and item locations as δ_i (Andrich et al., 2010).

According to this model, an item's difficulty is defined as being equal to the level of ability at which 50% of persons respond successfully to that item. When the difficulty of any given item exceeds the ability of any given group of persons, a smaller percentage of persons respond successfully. A major strength of this model is that an analysis on raw data provide reliable and valid independent (stand-alone) measures of a person's ability and the difficulty of items. These reliable person ability and item difficulty parameters, attained through a person-item interaction, potentiates an objective measure of cognitive load. That is, the difference between item difficulty δ and person ability β or (δ –β) provides an objective and performance-derived estimate of *relative* difficulty (or cognitive load experiences by the learner as a function of the learning task). The more the difficulty of an item exceeds the ability of the person, the greater the relative difficulty of *that* item for *that* person and, hence the greater cognitive load involved in correctly solving the item.

This approach reflects the interaction between measurable elements of cognitive load (i.e., ML, ME, and performance) and calibrates them within a single scalable trait/dimension. ML is captured through the transformation of raw performance data into reliable estimates of item/task difficulty. ME is estimated through the transformation of raw performance data into ability measures (and degree to which variation occurs with respect to difficulty estimates). This relative difficulty of items is analogous to ME as *a measure of the amount of cognitive load* involved in correctly responding to the task/item. This provides a summary interval level measurement of cognitive load derived by an objective mathematical procedure. It is important to note that this proposed cognitive load measure involves intrinsic cognitive load only and does *not encompass extraneous cognitive load*. The proposed measure deals solely with the complexity of the tasks and or difficulty of the test questions (element interactivity) and the background knowledge of the learners (e.g., their numeracy and literacy abilities).

By contrast with Paas and van Merriënboer's (1993) efficiency model, which stem from calculations involving incommensurable variables, this IRT approach provides a psychometrically sound alternative. For example, Paas and van Merriënboer's (1993) efficiency model *uses two distinct scales* to derive a measure of cognitive efficiency/cognitive load and calculates the difference between *z* score performance and z score effort as an efficiency measure. By contrast, a probabilistic IRT analysis transforms the raw data of a *single* performance measure and derives item difficulty and person ability parameters from this measurement scale (i.e., test or task scores). IRT probabilistic transformation of raw performance scores into these two parameter estimates are located on a single logit scale in interval level units. Hence, the subtraction of the ability estimates from the difficulty estimates per person item interaction is psychometrically sound as these estimates share a common logit scale.

The Present Study

We understand the concept of test validity as defined by Kane (2013) who presents an argument-based approach. In this approach "...to validate an interpretation or use of test scores is to evaluate the plausibility of the claims based on the test scores" (p. 1). This validity framework consists of (1) stating the proposed interpretation and use of the test scores and (2) evaluating the plausibility of such proposals (Kane, 2013).

In the current study, and following from Kane's (2013) argument-based approach, we specifically propose that IRT derived statistics from standardized numeracy and literacy test scores can provide proxy measures to determine variance in learners' intrinsic cognitive load. In order to evaluate the plausibility of this proposal we demonstrate two types of validity evidence: construct validity and concurrent criterion validity. Evidence of construct validity is demonstrated through an IRT analysis on the National Assessment Program – Literacy and Numeracy (NAPLAN) standardized test data (e.g., correct item functioning, reliability testing, and fit to the Rasch model). Moreover, we evaluate the plausibility of this proposal by attaining concurrent criterion validity evidence. Our hypothesis (H1) for criterion validity was that WM should inversely predict the relative difficulty/cognitive load requirement of learners. That is, concordant with CLT theory, higher WM capacity would decrease the experience of cognitive load and give preliminary support for the utility of this index to measure learners' cognitive load.

MATERIALS AND METHODS

Participants

Ninety-one primary school primary school-aged learners in Grade 2 (aged 7–8 years) participated in this study. Learners were recruited across three regional (n = 29) and two metropolitan schools (n = 62), with a balanced gender ratio of boys (n = 42), and girls (n = 49). All learners spoke English as their first language and had no known developmental delay or disorder.

Measures

Learning Assessment

An out-of-circulation version of Australia's National Assessment Program – Literacy and Numeracy (NAPLAN) test was administered as the learning task (ACARA, 2011). Specifically, a numeracy test (35 multiple-choice questions) and a language conventions test which consists of a spelling subtest (25 multiple-choice questions) and a grammar subtest (25 multiple-choice questions) of NAPLAN were selected to provide raw performance data. These assessments were administered in a group setting within the students' classrooms, which followed the protocols of the NAPLAN test.

Working Memory

Phonological and visual-spatial WM was measured by respective "Not This" and "Mr Ant" tasks from the Early Years Toolbox (EYT; Howard and Melhuish, 2017). These tasks are administered via iPad to collect scores and timing measures.

Phonological WM

The iPad-based EYT "Not This" task (Howard and Melhuish, 2017) involves the presentation of an auditory instruction, against a blank screen, to find a stimulus that does not have certain

characteristics of color, shape, or size (or a combination of these; e.g., 'Point to a shape that is not red and not a circle). After a brief retention interval, participants are then shown a stimulus array from which to identify a stimulus that satisfies the auditory instruction. The task increases in complexity from level 1 (one feature to recall) to level eight (eight features to recall). Each level consists of five trials and at least three successful responses are required to proceed to the next level. The task ends if participants fail to achieve three or more successful trials within a level, or the completion of level eight. WM capacity is estimated using a point score, calculated as: one point for each successive level, starting at the first, in which at least three trials are performed correctly and then 1/5 of a point for each successful trial thereafter.

Visual-Spatial WM

The iPad-based EYT "Mr Ant" task (Howard and Melhuish, 2017) involves recall of an increasing number of stickers placed on various locations of a cartoon ant. The task increases in complexity from level one (recalling the placement of one sticker) to level eight (recalling the placement of eight stickers). The task consists of three trials per level and failure on all trials at a given level (or completion of level eight) ends the task. In test trials, a cartoon ant with sticker/s is presented for 5 s, followed by a blank screen for 4 s, before the return of the cartoon ant without any stickers. Participants respond by tapping on the location of the missing sticker/s. WM capacity is estimated by a point score, calculated as: 1 point for each successive level, starting at the first, in which at least two trials are performed correctly and then 1/3 of a point for each successful trial thereafter.

Procedure

NAPAN tests were administered in two group sessions within students' classrooms, across 2 days, starting with language conventions. This order and spacing is consistent with NAPLAN administration (Board of Studies Teaching and Educational Standards NSW (BOSTESNSW), 2015). Absent students completed the missed test on the day of their return to school. After completion of the NAPLAN assessments, the WM tasks were administered in a single session individually and in a quiet room. The tasks were administered in a fixed random order, as follows: RSPM; Mr Ant; and Not This. The classroom teacher was present throughout the testing phase and was on hand to assist students who had questions.

RESULTS

Rasch Analyses

The proposed indices of cognitive load were derived from Rasch modeling analyses of the NAPLAN test performances (numeracy and language conventions). These data were analyzed using the dichotomous Rasch model, run on Rasch Unidimensional Measurement Modeling (RUMM) 2030 software (Andrich et al., 2010; for a complete interpretation of Rasch analysis, see Tennant and Conaghan, 2007). Overall fit of the data to the Rasch model indicated good model fit for both tests (chi-square all $p > 0.05$) (see **Table 1** for summary of fit statistics). The Person Separation Index (PSI), a reliability index on the transformed logistic data, indicated very good reliability for all three tests (0.85–0.86), as did the Cronbach alpha reliability indices (0.86–0.94).

The individual fit of items to the Rasch model are identified by fit residuals outside the acceptable ranges ($\leq$2.50 and $>$2.50). Residuals constitute the difference between the observed values and the theoretical Rasch estimates. Individual item misfit can also be detected by significant chi-square and F statistics, where an insignificant p value (>0.05) indicates good fit to the Rasch model. Misfit can also be detected by examination of an item's item characteristic curve (ICC). ICCs plot the observed values against the theoretical Rasch-derived estimates represented as an s-shaped curve; the closer the proximity between the observed values and the theoretical curve the better the fit and vice versa.

One item in the language conventions test (item 48) was found to misfit the model ($\chi^2 = 0.72$, $p < 0.001$) at Bonferroni adjusted alpha = 0.001 and was removed from the analysis. Also, Item 25 in the language conventions test had an extreme score (defined as all responses correct or incorrect) and was not used in the analysis. Otherwise, individual item fit was acceptable for all items of each test. Overall, all tests showed evidence of good reliability and construct validity (as good fit to the unidimensional Rasch model and correct functioning of items). The spread of items relative to the ability of the learners in the numeracy and language conventions tests are depicted in **Figures 1**, **2**, respectively.

The high reliability indices and well-functioning of items according to the Rasch model constitutes significant evidence of the precision of the test score data which we will use to formulate our proposed intrinsic cognitive load measure. Following Kane's (2013) validity argument approach, such evidence of the precision of our test score data will support the plausibility and generalizability of our proposed measure.

Relative Difficulty/Cognitive Load Measures

Essentially, our proposed cognitive load index is a measure of the relative difficulty of test items. This relative difficulty measure was calculated from the subsequent IRT analysis on the NAPLAN numeracy and language conventions test data. These relative difficulty/cognitive load measures were calculated for each test dimension by subtracting the IRT derived person ability estimates from the item difficulty estimates for each person-item interaction. The descriptives for these measures are depicted in **Table 2** as logits and depict the mean relative difficulty/cognitive load for each person-item interaction across the two test domains.

TABLE 1 | Rasch analysis summary statistics of the NAPLAN numeracy and language conventions tests.

	Item trait Interaction		PSI	α
Test type	**Value (df)**	***p***		
Numeracy	088.3 (70)	0.07	0.85	0.86
Language conventions	105.8 (96)	0.23	0.86	0.94

**ps < 0.05 are statistically significant. PSI, person separation index.*

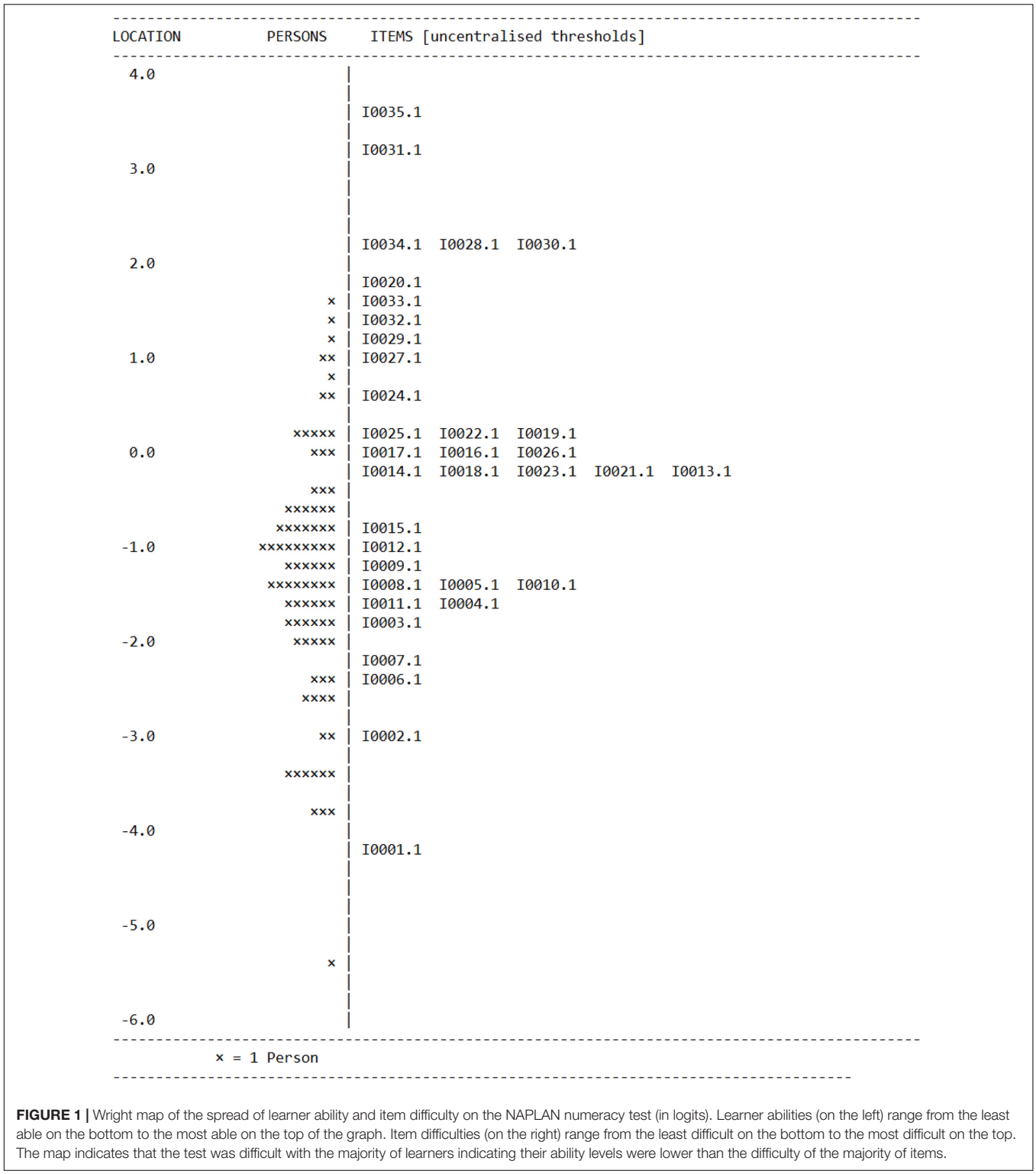

FIGURE 1 | Wright map of the spread of learner ability and item difficulty on the NAPLAN numeracy test (in logits). Learner abilities (on the left) range from the least able on the bottom to the most able on the top of the graph. Item difficulties (on the right) range from the least difficult on the bottom to the most difficult on the top. The map indicates that the test was difficult with the majority of learners indicating their ability levels were lower than the difficulty of the majority of items.

Multiple Regression Analyses

The results of the multiple regression for Model 1 (numeracy relative difficulty/cognitive load) indicated that the two WM predictors significantly explained 20% of the variance [$R^2 = 0.20$, $F(2,87) = 10.63$, $p < 0.001$]. Phonological WM made the strongest contribution to explaining numeracy relative difficulty/cognitive load and accounted for 9% unique variance while visual-spatial WM was found to contribute 6% unique variance. It was

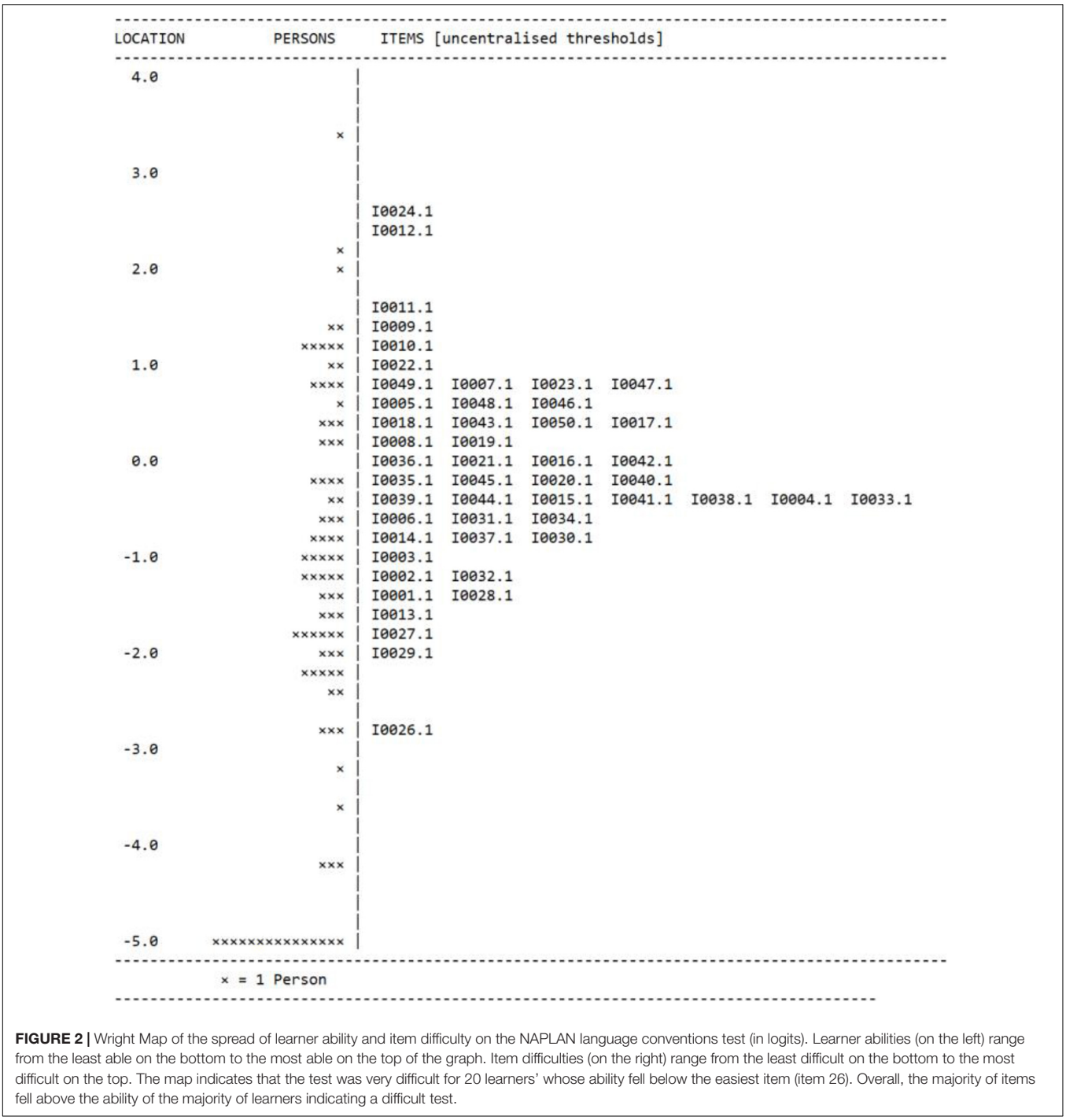

FIGURE 2 | Wright Map of the spread of learner ability and item difficulty on the NAPLAN language conventions test (in logits). Learner abilities (on the left) range from the least able on the bottom to the most able on the top of the graph. Item difficulties (on the right) range from the least difficult on the bottom to the most difficult on the top. The map indicates that the test was very difficult for 20 learners' whose ability fell below the easiest item (item 26). Overall, the majority of items fell above the ability of the majority of learners indicating a difficult test.

found that as phonological WM increased by one standard deviation the relative difficulty/cognitive load index decreased by 0.31 standard deviations ($\beta = -0.31$, $p < 0.01$), as did visual spatial WM, which decreased by 0.26 standard deviations ($\beta = -0.26$, $p < 0.05$). Model 2 (language conventions relative difficulty/cognitive load) indicated that the predictors explained 7% of the variance ($R^2 = 0.07$, $F(2,87) = 0\ 3.18$, $p < 0.05$). However, only phonological WM significantly contributed to unique variance (6%). As phonological WM increased by 1 standard deviation the relative difficulty/cognitive load index decreased by 0.26 standard deviations ($\beta = -0.26$, $p < 0.05$). Correlations of these variables are listed in **Table 3** and results of the regression models are summarized in **Table 4**.

TABLE 2 | Descriptive statistics for item response derived measures of relative difficulty/cognitive load for the numeracy and language conventions tests.

Relative difficulty	Mean	SD	Skewness	Kurtosis
Numeracy	3.31	1.29	0.20 (0.25)	0.54 (0.50)
Language conventions	4.20	2.11	0.33 (0.25)	−1.23 (0.50)

These metrics are denoted in logit values and indicate the average amount of cognitive load capacity utilized to complete the full tests (numeracy and language conventions) per test taker. SD = standard deviation. Standard errors are denoted in parentheses.

TABLE 3 | Summary of intercorrelations.

Measure	1	2	3	4
1. Numeracy (relative difficulty)	–	−0.539***	−0.369***	−0.338**
2. Language conventions (relative difficulty)	–	–	−0.262*	−0.095
3. Phonological working memory	–	–	–	−0.221*
4. Visual spatial working memory	–	–	–	–

**ps < 0.05; **ps < 0.01; ***ps < 0.001.*

TABLE 4 | Multiple regression results for working memory predicting relative difficulty/cognitive load measures.

	B	SE B	β	*t*
Model 1				
Numeracy				
Constant	5.846	0.573		10.201***
Phonological WM	−0.529	0.169	−0.311	−03.130**
Visual spatial WM	−0.289	0.111	−0.260	−02.609*
Model 2				
Language conventions				
Constant	6.426	1.009		6.372***
Phonological WM	−0.707	0.297	−0.255	−2.375*
Visual spatial WM	−0.058	0.195	−0.032	0.767

**ps < 0.05; **ps < 0.01; ***ps < 0.001 are statistically significant.*

DISCUSSION

The aim of the current study was to evaluate the potential of item response modeling to generate an objective measure of intrinsic cognitive load. Results indicated that valid and reliable indices of intrinsic cognitive load can be attained by item response modeling of raw test data (or other series of complex tasks/problems within a single domain) at an interval scale level. The interaction of the two parameter estimates (item difficulty and person ability) combine into a single scalable measure, in logits, subsuming critical elements of the measurable aspects of cognitive load: ML (i.e., task difficulty) and ME (performance measures transposed into ability logits). In support of our hypothesis (H1), resulting relative difficulty indices–that is, subtraction of the person ability estimates from the item difficulty estimates–were related to cognitive resources, in the expected direction, functions as an estimate of cognitive load. This IRT approach to estimating intrinsic cognitive load is superior to subjective self-report measures as it meets the requirements of objective measurement (Andrich, 2004).

Our findings provide clear validity evidence for the plausibility of our interpretations and utility of our IRT-based measure to indicate a learner's intrinsic cognitive load capacity. This evidence was demonstrated through a concurrent criterion validity approach in that a learner's WM capacity was found to significantly predict our proposed cognitive load index within both numeracy and literacy domains. We found both phonological and visual spatial WM scores significantly accounted for 20% of the variance of cognitive load in the numeracy domain. This finding is consistent with prior research which has found that phonological and visual spatial WM are important predictors of numeracy processing (Alloway and Alloway, 2010; Alloway and Passolunghi, 2011). While phonological WM significantly captured 7% of the variance of our novel cognitive load index in the language conventions domain (combined spelling and grammar tasks), visual-spatial WM played no significant role.

A possible explanation for these results, that is, the small amount of variance captured by phonological WM and lack of predictive role of visual-spatial WM on our cognitive load measure may have to do with the nature of the language convention spelling and grammar tasks. In the language conventions sections of the NAPLAN tests, the spelling items consist of identification of misspelt words. The mental resources needed for this type of processing do not require deliberate thought and essentially require retrieval from long-term memory if the word is known and guessing in the case of an unknown word (though in some cases the application of spelling rules may apply). Similarly, in the grammatical section of the language conventions test the format consists of short cloze activities where a sentence is presented, and students choose the correct missing grammatical form. Here, knowledge of the correct conjugation or form of the verb or auxiliary is all that is needed to successfully complete the task. The degree to which deliberate thought is needed to control the processing of information is minimal and hence the ME and WM capacities on these tasks would not be optimal. According to Paas and van Merriënboer's (1994) cognitive load model, the automatic processing of information bypasses the requirement of drawing on ME resources and feeds directly into performance. Hence, this type of automatic processing may have sufficiently limited the cognitive capacity requirements in the language conventions domain.

Our findings may also simply be reflective of the reduced role of visual spatial WM in language processing. For example, it is well established that visual spatial WM is important for early numeracy processing (McKenzie et al., 2003; Bull et al., 2008). Moreover, in the year three NAPLAN numeracy tests many questions comprise visual "patterns" (or similar) and consequently involve visual processing along the lines of what was assessed by the visual spatial WM tasks. By contrast, such item types requiring visual processing were not present in the language conventions test used. Therefore, this may explain the lesser role of visual spatial WM processing as a predictor of our proposed cognitive load index.

Overall, however, our findings indicated that higher levels of cognitive resources were related to lower levels of cognitive load

requirements and vice versa. This is consistent with fundamental underpinnings of CLT (Sweller et al., 2019), which suggest that: cognitive load and WM capacity share an inverse relationship, such that deficiency in one aspect can be rectified by reduction in the other; and that a reduction in cognitive load can facilitate learning and performance.

Our proposed IRT modeling approach to cognitive load measurement provides a relatively simple and straightforward procedure to attain reliable and valid estimates of intrinsic cognitive load. While IRT modeling and Rasch analysis has been available to social scientists and psychologists for many decades now few have taken advantage of its superior measurement capabilities. Moreover, the creative potential of IRT modeling and its applications to cognitive load research, as well as educational and psychological research in general, has yet to be actualized.

As we have shown in this study, IRT modeling can provide an objective measure of intrinsic cognitive load outside of subjective self-report. This is particularly pertinent given the difficulty in attaining reliable self-report measures on cognitive processing of younger children (i.e., less than 7 years) (Conjin et al., 2020). The ability to ascertain reliable and valid measures of intrinsic cognitive load through a performance-based objective mathematical procedure is highly beneficial, especially for cognitive load researchers interested in measuring younger learners' cognitive load. Moreover, this objective IRT modeling approach has ecological validity in that the performance data (i.e., tasks, problems, and questions) are collected within the classroom learning environment and are unobtrusive. The innovation of IRT and Rasch modeling into the cognitive load research paradigm offers exciting measurement opportunities beyond subjective self-report approaches.

Limitations

We wish to acknowledge several limitations of this study. First, while our study has demonstrated the utility and validity of IRT modeling to quantify intrinsic cognitive load it is important to note that IRT analysis requires large sample sizes. In the case of the current study sample size was not such an issue because we used standardized tests which have already been validated with large (nationwide) samples using IRT analyses (ACARA, 2020). Normally, a reliable IRT analysis requires (N = 200) or so (Linacre, 1994). Hence, IRT analysis may be beyond the scope of typical smaller experimental classroom-based cognitive load investigations. Second, our sample of learners were younger than the target age of the tests and this was reflected somewhat in the IRT analysis, in that many learners found the test difficult.

Future Directions

The current study has shown that our relative difficulty/cognitive load index varies with WM in relation to intrinsic cognitive load. Further validation of this measure would benefit from evaluation of the index to determine whether it varies according to the learner task following CLT principles (e.g., extraneous and germane load) and through construct (i.e., convergent) validity testing to establish the measure's relationship with other cognitive load scales (e.g., Paas, 1992; Leppink et al., 2013; Krell, 2017). Such research is needed to show that our proposed cognitive load index varies with theoretical variations in cognitive load. Additionally, it would be desirable to investigate the performance of our proposed cognitive load index with learners at varying stages of age and development. Finally, our proposed cognitive load index may be a useful measure for those undertaking intervention research where the index can be used to assess shifts in relative difficulty (cognitive load) scores across stages of learner development.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by University of Wollongong. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

REFERENCES

ACARA (2011). *NAPLAN. Australian Curriculum Assessment and Reporting Authority (ACARA)*. Available online at: http://www.nap.edu.au/naplan/naplan.html (accessed January 2, 2020).

ACARA (2020). Reliability and Validity of NAPLAN. Australian Curriculum Assessment and Reporting Authority. Available online at: https://www.nap.edu.au/resources (accessed January 2, 2020).

Alloway, T. P., and Alloway, R. G. (2010). Investigating the predictive roles of working memory and IQ in academic attainment. *J. Exp. Child Psychol.* 106, 20–29. doi: 10.1016/j.jecp.2009.11.003

Alloway, T. P., and Passolunghi, M. C. (2011). The relationship between working memory, IQ, and mathematical skills in children. *Learn. Individ. Dif.* 21, 133–137. doi: 10.1016/j.lindif.2010.09.013

Andrich, D. (2004). Controversy and the Rasch model: a characteristic of incompatible paradigms? *Med. Care* 42(Suppl. 1), 1–7. doi: 10.1097/01.mlr.0000103528.48582.7c

Andrich, D., Sheridan, B., and Luo, G. (2010). *RUMM2030: A Windows Program for the Rasch Unidimensional Measurement Model (User Manual: Part 1 Dichotomous Data)*. Perth, WA: RUMM Laboratory.

Antonenko, P., Paas, F., Grabner, R., and van Gog, T. (2010). Using electroencephalography to measure cognitive load. *Educ. Psychol. Rev.* 22, 425–438. doi: 10.1007/s10648-010-9130-y

Ayres, P. (2006). Impact of reducing intrinsic cognitive load on learning in a mathematical domain. *Appl. Cogn. Psychol.* 20, 287–298. doi: 10.1002/acp.1245

Board of Studies Teaching and Educational Standards NSW (BOSTESNSW) (2015). *NAPLAN*. Available online at: http://www.boardofstudies.nsw.edu.au/naplan/ (accessed May 31, 2016).

Bull, R., Espy, K. A., and Wiebe, S. A. (2008). Short-term memory, working memory, and executive functioning in preschoolers: longitudinal predictors of mathematical achievement at age 7 years. *Dev. Neuropsychol.* 33, 205–228. doi: 10.1080/87565640801982312

Burson, K. A., Larrick, R. P., and Klayman, J. (2006). Skilled or unskilled, but still unaware of it: perceptions of difficulty drive miscalibration in relative comparisons. *J. Pers. Soc. Psychol.* 90, 60–77. doi: 10.1037/0022-3514.90.1.60

Chandler, P., and Sweller, J. (1996). Cognitive load while learning to use a computer program. *Appl. Cogn. Psychol.* 10, 151–170. doi: 10.1002/(sici)1099-0720(199604)10:2<151::aid-acp380>3.0.co;2-u

Cierniak, G., Scheiter, K., and Gerjets, P. (2009). Explaining the split attention effect: is the reduction of extraneous cognitive load accompanied by an increase in germane cognitive load? *Comput. Hum. Behav.* 25, 315–324. doi: 10.1016/j.chb.2008.12.020

Conjin, J. M., Smits, N., and Hartman, E. E. (2020). Determining at what age children provide sound self-reports: an illustration of the validity-index approach. *Assessment* 27, 1604–1618. doi: 10.1177/1073191119832655

Halabi, A. K., Tuovinen, J. E., and Farley, A. A. (2005). Empirical evidence on the relative efficiency of worked examples versus problem-solving exercises in accounting principles instruction. *Issues Account. Educ.* 20, 21–32. doi: 10.2308/iace.2005.20.1.21

Hoffman, B., and Schraw, G. (2010). Conceptions of efficiency: applications in learning and problem solving. *Educ. Psychol.* 45, 1–14. doi: 10.1080/00461520903213618

Howard, S., Burianova, H., Ehrich, J., Kervin, L., Calleia, A., Barkus, E., et al. (2015). Behavioural and fMRI evidence of the differing cognitive load of domain-specific assessments. *Neuroscience* 297, 38–46. doi: 10.1016/j.neuroscience.2015.03.047

Howard, S. J., and Melhuish, E. C. (2017). An early years toolbox (EYT) for assessing early executive function, language, self-regulation, and social development: validity, reliability, and preliminary norms. *J. Psychoeduc. Assess.* 35, 255–275. doi: 10.1177/0734282916633009

Hummel, H. G. K., Paas, F., and Koper, E. J. R. (2004). Cueing for transfer in multimedia programmes: process worksheets vs. worked-out examples. *J. Comput. Assist. Learn.* 20, 387–397. doi: 10.1111/j.1365-2729.2004.00098.x

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *J. Educ. Meas.* 50, 1–73. doi: 10.1111/jedm.12000

Klepsch, M., Schmitz, F., and Seufert, T. (2017). Development and validation of two instruments measuring intrinsic, extraneous, and germane cognitive load. *Front. Psychol.* 8:1997. doi: 10.3389/fpsyg.2017.01997

Korbach, A., Brünken, R., and Park, B. (2017). Measurement of cognitive load in multimedia learning: a comparison of different objective measures. *Instr. Sci.* 45, 515–536. doi: 10.1007/s11251-017-9413-5

Korbach, A., Brünken, R., and Park, B. (2018). Differentiating different types of cognitive load: a comparison of different measures. *Educ. Psychol. Rev.* 30, 503–529. doi: 10.1007/s10648-017-9404-8

Krejtz, K., Duchowski, A. T., Niedzielska, A., Biele, C., and Krejtz, I. (2018). Eye tracking cognitive load using pupil diameter and microsaccades with fixed gaze. *PLoS One* 13:e0203629. doi: 10.1371/journal.pone.0203629

Krell, M. (2017). Evaluating an instrument to measure mental load and mental effort considering different sources of validity evidence. *Cogent Educ.* 4:1280256. doi: 10.1080/2331186X.2017.1280256

Leppink, J., Paas, F., Vander Vleuten, C. P. M., van Gog, T., and van Merriënboer, J. J. G. (2013). Development of an instrument for measuring different types of cognitive load. *Behav. Res. Methods* 45, 1058–1072. doi: 10.3758/s13428-013-0334-1

Linacre, J. M. (1994). Sample size and item calibrations stability. *Rasch Meas. Trans.* 7:328.

Marcus, N., Cooper, M., and Sweller, J. (1996). Understanding instructions. *J. Educ. Psychol.* 88, 49–63.

McKenzie, B., Bull, R., and Gray, C. (2003). The effects of phonological and visual-spatial interference on children's arithmetical performance. *Educ. Child Psychol.* 20, 93–108.

Paas, F. G. W. C. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: a cognitive-load approach. *J. Educ. Psychol.* 84, 429–434. doi: 10.1037/0022-0663.84.4.429

Paas, F. G. W. C., Ayres, P., and Pachman, M. (2008). "Assessment of cognitive load in multimedia learning environments: theory, methods, and applications," in *Recent Innovations in Educational Technology that Facilitate Student Learning*, eds D. H. Robinson, and G. J. Schraw (Charlotte, NC: Information Age), 11–35.

Paas, F. G. W. C., Renkl, A., and Sweller, J. (2004). Cognitive load theory: instructional implications of the interaction between information structures and cognitive architecture. *Instr. Sci.* 32, 1–8. doi: 10.1023/b:truc.0000021806.17516.d0

Paas, F. G. W. C., Tuovinen, J., Tabbers, H., and van Gerven, P. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educ. Psychol.* 38, 63–71. doi: 10.1207/s15326985ep3801_8

Paas, F. G. W. C., and van Merriënboer, J. J. G. (1993). The efficiency of instructional conditions: an approach to combine mental effort and performance measures. *Hum. Factors* 35, 737–743. doi: 10.1177/001872089303500412

Paas, F. G. W. C., and van Merriënboer, J. J. G. (1994). Instructional control of cognitive load in the training of complex cognitive tasks. *Educ. Psychol. Rev.* 6, 351–371. doi: 10.1007/bf02213420

Paas, F. G. W. C., van Merriënboer, J. J. G., and Adam, J. J. (1994). Measurement of cognitive load in instructional research. *Percept. Mot. Skills* 79, 419–430. doi: 10.2466/pms.1994.79.1.419

Park, B., and Brünken, R. (2015). The rhythm method: a new method for measuring cognitive load—an experimental dual–task study. *Appl. Cogn. Psychol.* 29, 232–243. doi: 10.1002/acp.3100

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.

Salden, R. J. C. M., Paas, F. G. W. C., Broers, N. J., and van Merriënboer, J. J. G. (2004). Mental effort and performance as determinants for the dynamic selection of learning tasks in air traffic control training. *Instr. Sci.* 32, 153–172. doi: 10.1023/b:truc.0000021814.03996.ff

Scholey, A. B., Harper, S., and Kennedy, D. O. (2001). Cognitive demand and blood glucose. *Physiol. Behav.* 73, 585–592. doi: 10.1016/s0031-9384(01)00476-0

Stone, N. J. (2000). Exploring the relationship between calibration and self-regulated learning. *Educ. Psychol. Rev.* 12, 437–476.

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

Sweller, J., Ayres, P., and Kalyuga, S. (2011). *Cognitive Load Theory*. London: Springer.

Sweller, J., van Merriënboer, J. J. G., and Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. *Educ. Psychol. Rev.* 10, 251–296.

Sweller, J., van Merriënboer, J. J. G., and Paas, F. G. W. C. (2019). Cognitive architecture and instructional design: 20 years later. *Educ. Psychol. Rev.* 31, 261–292. doi: 10.1007/s10648-019-09465-5

Tennant, A., and Conaghan, P. G. (2007). The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis Care Res.* 57, 1358–1362. doi: 10.1002/art.23108

Tindall-Ford, S., Chandler, P., and Sweller, J. (1997). When two sensory modes are better than one. *J. Exp. Psychol. Appl.* 3, 257–287. doi: 10.1037/1076-898x.3.4.257

van Gerven, P. W. M., Paas, F. G. W. C., van Merriënboer, J. J. G., and Schmidt, H. G. (2004). Memory load and the cognitive pupillary

response in aging. *Psychophysiology* 41, 167–174. doi: 10.1111/j.1469-8986.2003.00148.x

van Gog, T., and Paas, F. G. W. C. (2008). Instructional efficiency: revisiting the original construct in educational research. *Educ. Psychol.* 43, 16–26. doi: 10.1080/00461520701756248

Whelan, R. R. (2007). Neuroimaging of cognitive load in instructional multimedia. *Educ. Res. Rev.* 2, 1–12. doi: 10.1016/j.edurev.2006.11.001

Wright, B. (1997). A history of social science measurement. *Educ. Meas. Issues Pract.* 16, 36–52.

Wright, R. (1984). Motivation, anxiety, and the difficulty of avoidance control. *J. Pers. Soc. Psychol.* 46, 1376–1388. doi: 10.1037/0022-3514.46.6.1376

Wright, R., Contrada, R., and Patane, M. (1986). Task difficulty, cardiovascular response, and the magnitude of goal valence. *J. Pers. Soc. Psychol.* 51, 837–843. doi: 10.1037/0022-3514.51.4.837

Xie, B., and Salvendy, G. (2000). Prediction of mental workload in single and multiple task environments. *Int. J. Cogn. Ergon.* 4, 213–242. doi: 10.1207/s15327566ijce0403_3

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Education

ORIGINAL RESEARCH
published: 28 April 2021
doi: 10.3389/feduc.2021.645284

Making an Effort Versus Experiencing Load

Melina Klepsch and Tina Seufert*

Department Learning and Instruction, Institute of Psychology and Education, Ulm University, Ulm, Germany

In cognitive load theory (CLT), the role of different types of cognitive load is still under debate. Intrinsic cognitive load (ICL) and germane cognitive load (GCL) are assumed to be highly interlinked but provide different perspectives. While ICL mirrors the externally given task affordances which learners experience passively, germane resources are invested by the learner actively. Extraneous affordances (ECL) are also experienced passively. The distinction of passively experienced load and actively invested resources was inspired by an investigation where we found differential effects of a learning strategy training, which in fact resulted in reduced passive load and increased actively invested effort. This distinction is also mirrored in the active and passive forms for effort in German language: "es war anstrengend" (it has been strenuous) vs. "ich habe mich angestrengt" (I exerted myself). In two studies, we analyzed whether we could distinguish between these active and passive aspects of load by using these phrases and how this distinction relates to the three-partite concept of CLT. In two instructional design studies, we included the active and passive items into a differentiated cognitive load questionnaire. We found the factor structure to be stable, with the passive item loading on the ICL factor and the active item loading on the GCL factor. We conclude that it is possible to distinguish between active and passive aspects of load and that further research on this topic could be constructive, especially for learning tasks where learners act in a more self-regulated way and learner characteristics are taken into account.

Keywords: cognitive load theory (CLT), differentiated measurement, self-regulated learning (SRL), assessing cognitive load, active vs. passive cognitive load, measurement, validity

OPEN ACCESS

Edited by:
Fred Paas,
Erasmus University Rotterdam, Netherlands

Reviewed by:
Slava Kalyuga,
University of New South Wales, Australia
Jeroen Van Merrienboer,
Maastricht University, Netherlands

****Correspondence:***
Melina Klepsch
melina.klepsch@uni-ulm.de

Specialty section:
This article was submitted to Educational Psychology, a section of the journal Frontiers in Education

Received: *22 December 2020*
Accepted: *11 February 2021*
Published: *28 April 2021*

Citation:
Klepsch M and Seufert T (2021) Making an Effort Versus Experiencing Load. Front. Educ. 6:645284. doi: 10.3389/feduc.2021.645284

INTRODUCTION AND THEORETICAL BACKGROUND

During the last two decades, the concept of cognitive load has been widely discussed. The basic assumption that cognitive processes of schema construction require mental resources is noncontroversial. However, the aspects that contribute to these requirements while learning, and particularly, the role of different types of load, are still under debate. In its original account, cognitive load theory (CLT) differentiated between three different types of cognitive load (Sweller et al., 1998). Two of them are widely acknowledged: 1) intrinsic cognitive load (ICL) represents the complexity of the learning task itself, determined by the number of elements and their interrelations. Intrinsic load can only be altered by either changing the learning task or when learners can use their prior knowledge. Existing schemata would ease to relate the given elements and would thus reduce the intrinsic affordances. 2) Extraneous cognitive load (ECL) arises from an inappropriate design and poses load onto the learner, which is not task-relevant (e.g., search processes between different representations). This type of load was addressed in a vast amount of studies on instructional design (e.g., Paas et al., 2003). Particularly, in multimedia learning contexts, various design guidelines have

been established which recommend design features which would reduce the extraneous affordances of the learning environment and thus reduce extraneous load (e.g., Mayer, 2014). The third type, that is, 3) germane cognitive load (GCL), is controversial. GCL arises when learners actively allocate resources to deal with the task, for example, to build cognitive schemata. In other words, learners use resources, which are germane while dealing with the intrinsically given the requirements of the task. Germane load therefore was also labeled as germane resources (Kalyuga, 2011). This description makes it obvious that intrinsic load and germane load are highly interlinked.

Nevertheless, if self-regulation is taken into account, that is, learners actively regulate their learning processes, ICL and GCL provide different perspectives. While ICL mirrors the externally given task affordances which learners experience rather passively, germane resources are invested by the learner actively (Seufert, 2018). The distinction of passively experienced load vs. actively invested resources could be an interesting approach to better understand the effects of how learners deal with or use their effort. This idea of different perspectives on cognitive load is actually not new and has already been reflected by Paas and van Merriënboer (1994). Besides the causal effects which determine cognitive load, they differentiate between different assessment factors which mirror the experienced load for each individual learner, namely, mental load, mental effort, and performance. The first two factors match the abovementioned distinction of actively vs. passively experienced load. While mental load reflects the task-centered dimension, which is determined externally by the task, mental effort is human-centered, and thus determined internally and dependent on learner's decisions and characteristics. Not only the location of the determination is crucial but also the possibility to control these aspects. While mental load is not controllable and due to Paas and van Merriënboer "constant for a given task in a given environment," mental effort is the "amount of controlled processing the individual is engaged in" (1994). Scheiter et al. (2020) referred to this distinction in their review on how to measure cognitive load and link it to an analog distinction which is made in the metacognitive literature. When learners rate their mental effort while metacognitively monitoring their learning process, they can also take these two perspectives as outlined by Koriat et al. (2006). When monitoring their effort, learners can on the one hand take into account the effort which is intrinsic to the task and thus externally triggered, which would be a data-driven appraisal of the experienced load in accordance with the concept of mental load. On the other hand, they could refer to their deliberately invested engagement based on their own goals, which is thus called a goal-driven appraisal of the experienced load in accordance with the concept of mental effort.

Not only contemporary approaches of linking research on cognitive load with research on self-regulation refer to the distinction of invested effort vs. experienced load (see also Schnaubert and Schneider, 2020) but also current research on measurement issues of cognitive load. Krell (2017), for example, developed a scale to measure mental effort and mental load. In a recent study, it could be shown that the distinction between load and effort could be even linked to objective measures of load, like the heart rate, which was only related to self-reported mental effort, but not to mental load (Minkley et al., 2021).

Experiencing Load Versus Investing Effort: Passive and Active Load

In the context of metacognitive monitoring, it becomes obvious that learners are in agency of their effort experience when it comes to effort appraisals, while they are not with load appraisals. This is what led us to the terms active and passive (or re-active as the activation is triggered externally) load. The use of the terms active and passive load—instead of mental effort and mental load—is especially motivated by the passive and active use of the word effort in German language: "es war anstrengend" (it has been strenuous) vs. "ich habe mich angestrengt" (I exerted myself). As there are already different items in use to measure mental effort and mental load (see, e.g., Paas, 1992; Krell, 2017), we refer to the name active and passive load as long as there is no comparison with the given items. The passive and active phrases of effort in the German language could be applied to measure the active and passive aspects of cognitive load, which we did in a study on the effects of learning strategies' training for children (Taxis et al., 2010). We compared a group with training to a control group without training regarding their strategy use, learning outcomes, and cognitive load. As children usually have problems in understanding the item to measure load—"My mental effort was. . ."—we used the abovementioned German phrases in active and passive forms for measuring load. The training comprised step-by-step instructions on reading strategies and metacognitive strategies. Besides the expected effects of the training on strategy use and learning outcomes, we found evidence for differentiated effects of the two load items. While the children experienced a reduced passive load as the training provided explicit guidelines for learning, they reported an increase in their actively invested effort. This study provided a first hint that the activation of cognitive processes can enhance and reduce different aspects of cognitive load at the same time.

We started our discussion of different perspectives on cognitive load with the classical three-partite description of intrinsic, extraneous, and germane load. Thus, we now must ask how the two-fold perspective of active and passive load would align to these three load types. As argued above, ICL and GCL are closely connected as they refer to the same affordances given by a task. While ICL is data-driven, externally determined, and thus passively experienced, GCL is goal-driven, internally determined, and thus actively invested. But how about extraneous affordances of a task (ECL), like intricate navigation in an online-learning environment or other unnecessary, that is, task-irrelevant search processes? As they are also determined externally and are usually not under control of the learner, we would expect them to also relate to passive aspects of cognitive load. However, it seems also plausible that learners may also manage this kind of extraneous load by adapting their strategies, and thus are no longer passive. This self-management of load is described in an article of Eitel et al. (2020), who also link the concept of cognitive load with self-regulation in learning. They describe that learners can enhance

their effort in selecting relevant aspects when confronted with irrelevant seductive details.

How to Elicit and Validate Active and Passive Load

Based on Kane's (2013) argument-based approach for validity, we formulate the following interpretation/use argument (IUA): We plan to measure cognitive load after learning, using an item on active as well as on passive load to assess learners' perceived invested effort and experienced load during learning: all for the purposes of discerning the effects differences in learning settings or through learner characteristics. This IUA results in different inferences (scoring, generalization, and extrapolation) that are described in the following paragraphs and will be discussed in the end.

In order to analyze the relations between the classical three-partite load types and active and passive load, we refer to well-researched design principles or instructional design effects. With their specific theoretical assumptions on how they would affect cognitive load and performance, we can deduce specific assumptions on the three classical types of cognitive load. This leads to a scoring inference which will be operationalized as hypotheses below: If active load is associated with GCL and passive load with ICL and eventually also with ECL, then the effects on all those measures should be in accordance. As we aim at analyzing the active and passive aspects of load, we use established variations of tasks, their design, and learners' aptitudes to elicit ICL, ECL, and GCL as a starting point. If we are using already established multimedia principles within classical experimental designs—as typical for studies on cognitive load—as well as often examined learner characteristics, then this enables us to assume some generalization inference for the found results. Given that if, for example, we can find an effect in one worked-example study, then we are hopefully able to find it in another, as there is already broad evidence for the worked-example effect.

To elicit *intrinsic load effects*, one could alter the complexity of the task (element interactivity, e.g., Sweller et al., 1998). Alternatively, one could use the learner-based variation of existing schemata, which would also affect the perceived ICL. Learners with higher expertise would rate a complex task less intrinsically loading than a learner with low expertise (Artino, 2008). Therefore, a scoring inference is as follows: Whenever ICL increases, the measure of passive load should also increase. Moreover, if ICL increases in a way that it exceeds learners' working memory capacity, this should also be indicated by a decrease in learning performance.

For inducing effects on ECL, well-researched design principles could be used, like the multimedia principle, the modality principle, the split-attention principle, the worked-example principle, or the redundance principle (for an overview, see Mayer, 2014). The split-attention principle, which is also called the spatial contiguity principle, states that instructional material which needs to be integrated should be placed nearby or even integrated, like words printed into a picture instead of separately in a text (Ayres and Sweller, 2014). Thus, learners do not need extraneous resources for searching for corresponding elements, and with the freed-up resources, learners can engage in mental integration which would be germane to the task. The same ECL-reducing effect can be assumed when implementing worked examples (for an overview, see Renkl, 2014). Learners are provided with a problem and an additional example on how to solve this problem before starting to solve the same or comparable problems on their own. With the guidance of the worked-example, learners do not need to spend extraneous resources on inadequate problem-solving strategies (Van Gog et al., 2008) and improve their learning outcomes. Van Gog et al. (2008) also argue that with the freed-up resources, learners could then invest more mental effort which would be germane to the task. Thus, worked examples may trigger both, a decreased ECL and an increased GCL. Such interdependences of the three load types have been also examined by Park (2010) in her study program on the additivity assumption of load. She could demonstrate that the enhancing effects of different instructions for learning or cognitive load cannot simply be added but that they might interact. This complex interplay has also been revealed by Wirzberger et al. (2016). So, the next scoring assumption is as follows: If ECL decreases by matters of design, then the score on the passive item should also decrease.

Regarding GCL, there are also several techniques proposed, which intend to foster generative activities. Moreno and Mayer (2010) analyzed the effects of personalization: guided activities like prompts to elaborate on the learning content, feedback, or self-explanation prompts. Their review provides strong evidence for improved learning outcomes when these techniques were used. But can these techniques enhance germane cognitive load? In their review on the effects of journal writing for learning, Nückles et al. (2020) provided evidence for a germane load–enhancing effect but they also emphasized that such an enhancing effect would depend on the learner's ability to apply the requested generative activity. This is also confirmed in a study of Park et al. (2016), where prompting learners to mentally animate a complex scenario only led to higher mental effort when learners had the necessary spatial abilities to perform these mental animations. Despite triggering generative activities and thus germane load externally, one could also analyze whether learners with task-specific abilities and particularly with strategic skills would invest more germane resources. Thus, the impulse to invest germane resources would be elicited internally. However, none of these studies used a specific measure for germane load but only deduced their interpretation by combining measures of mental effort with learning outcome measures. Only in two studies reported by Klepsch and Seufert (2020), effects of germane load–inducing instructions were revealed by using the germane cognitive load scale of the differentiated questionnaire by Klepsch et al. (2017). This leads to the assumption that if GCL increases, then the scores on the active load item should also increase, which gives us another scoring inference.

Overall, there are several instructional means with which the different types of load can be elicited. Besides the theoretical arguments, there are also empirical studies confirming these effects, but only a few measured the effects with differentiated measures of load (e.g., Klepsch and Seufert, 2020).

Present Studies

The goal of the present studies is to investigate the concept of passive load, that is, experiencing load and active load, that is, investing effort in relation to the classical three-partite types of load based on CLT. From a theoretical point of view, one would assume that the passively experienced load should be linked to those aspects of load that come along with the task affordances and its presentation, that is, ICL and ECL. The active load on the other hand would link to the actively invested germane resources. The purpose of the studies is to examine whether this theoretically assumed mapping of active and passive load with the concepts of intrinsic, extraneous, and germane load can be substantiated empirically. In addition, we use a simple measure for active and passive load by using the German active and passive linguistic forms of experiencing load. With such an easy-to-apply—and even for younger kids easy-to-understand—measure, the concepts of active and passive load could be investigated easily instead of or in addition to other differentiated measures (e.g.,Leppink et al., 2013; Klepsch et al., 2017; Krell, 2017). If the scores on the active and the passive load item are given, then the scores of other cognitive load measurements could be predicted, as well as learning outcome, which would allow an extrapolation inference.

In two experimental studies, we address the following questions: 1) Can we distinguish between these active and passive aspects of load and 2) relates this distinction to the three-partite concept of CLT. With classical instructional design variations, we analyze 3) the prognostic validity of these measures based on the abovementioned theories and studies on how to elicit the different types of load.

In the first study, we used worked examples in comparison to problem-solving as an external factor to elicit extraneous load and combined it with self-explanation prompts as an external factor to elicit germane load. In the second study, we used a split-source format in comparison to an integrated format to externally elicit extraneous load and learners' cognitive style to either process information in a holistic or serialistic way, which can elicit germane processes internally. We did not vary ICL as a between-subjects factor but considered learners' prior knowledge as an internal variation of ICL in both studies. In both studies, learners must rate their perceived intrinsic, extraneous, and germane load with the differentiated questionnaire for cognitive load by Klepsch et al. (2017) and their active and passive load with the additional items "ich habe mich angestrengt" (I exerted myself) and "es war anstrengend" (it was strenuous). With a factor analysis and additional correlations, the structural interrelation between the two lines of concepts is analyzed. Moreover, the expected main effects and interactions for the instructional variations were analyzed for validating the respective measures. Regarding the abovementioned research questions, we have the following hypotheses.

TABLE 1 | Number of participants in each experimental group of study 1.

Experimental groups	*N*
Problem solving * without prompt	20
Problem solving * with prompt	15
Worked examples * without prompt	20
Worked examples * with prompt	18

(H1) We expect that making an effort (active) in an explorative factor analysis is loading on the GCL factor and that experiencing load (passive) is loading either on the ICL or ECL factor.
(H2) We expect to find correlations between making an effort (active) and GCL and between experiencing load (passive) and ICL and/or ECL.
(H3) In terms of a validation of both measures, we expect making an effort (active) and germane cognitive load to increase when generative learning activities are triggered by the design (H3_active). We expect ECL to decrease when the design is optimized regarding unnecessary processes, and we expect no group differences for ICL, as we do not vary complexity of the learning content in our studies. Experiencing load (passive) should be either more in line with ICL or ECL (H3_passive). Additionally, learning outcome between groups will differ (H3_learning_outcome).

To examine these hypotheses, we conducted two studies, where we analyzed in which way the concept of active and passive load, measured by the active and passive German forms of effort (**Supplementary Appendix A**), relates to the three types of cognitive load measured by a differentiated measurement instrument (**Supplementary Appendix A**; Klepsch et al., 2017).

STUDY 1

In order to elicit different levels of different aspects of cognitive load, study 1 used the worked-example effect and the prompting principle. Learners had to deal with two mathematical topics, that is, extremum problems and Taylor polynomial tasks.

Based on theoretical assumptions and empirical findings, we assume for study 1 effects of worked examples and prompts on ICL, ECL, GCL, and the passive and the active item. While worked examples are meant to reduce the perceived ECL, prompts should enhance GCL. As we did not change the content to be learned, there should not be any difference in perceived ICL. The assumptions for the active and passive load should correspond with the effects for either ICL and/or ECL (passive) or for GCL (active). We additionally analyze the effects on learning outcomes of both instructional variations.

Method

In **Table 1**, the experimental conditions of study 1 and number of participants in these conditions are listed.

Participants

We collected data from 73 learners. Participants were at average 22.34 (*SD* = 4.27) years old, and 13.70% were male. They were students from a German university.

Design

We conducted a 2 × 2 between-subject study with four experimental conditions (**Table 1**). Independent variables are based on the worked-example effect (problem-solving vs. worked examples) and the prompting principle (no prompts vs. self-explanation prompts). As dependent variables, we assessed cognitive load in a differentiated way, the active and passive parts of load and learning outcome. As control variables, age, sex, and prior knowledge have been assessed.

Procedure

At the beginning, learners were informed about the procedure of the study they participated in and signed an informed consent form. All participants were aware that they could withdraw their data at any point in the study without having any disadvantage. Then, each participant was randomly assigned to one of the experimental groups. As a next step, each participant filled out a questionnaire asking for demographic data. Then, prior knowledge was assessed (described in Material). Afterward, participants had to deal with the learning material either with or without worked examples (described in Material). If they were in one of the experimental conditions with self-explanation prompts, learners then were presented the prompts (described in Material). After learning, the cognitive load questionnaire (**Supplementary Appendix A**; Klepsch et al., 2017) had to be answered, as well as the items to assess the active and passive parts of load (**Supplementary Appendix A**). In the end, learning outcome was measured; for each topic, two tasks which are structurally similar to those in the learning material and one transfer task with an unknown structure had to be solved.

Material

Each learner had to deal with two types of mathematical learning material: 1) extremum problems and 2) Taylor polynomial tasks. Prior knowledge was assessed after participants received an example task for each type of material: in two questions, they had to answer on a 7-point Likert-type scale, whether they were familiar with such tasks, and in a second step, they had to write down how they would solve the task. Afterward, the learning phase started with a short introduction on the topic, followed by two example tasks. To realize differences in ECL as an independent variable, problem-solving or worked examples were included. In the problem-solving group, the correct solution was presented, but no information on how the solution could be calculated was given. In the worked-example group, the solution steps for calculating the correct answer for the example tasks were provided. Participants had 5 min time for each domain to learn the content and another 15 min to conduct the post test. Performance was measured by a posttest for each domain containing two analog tasks (similar in structure to the example tasks in the learning material), and one complex transfer task with an unknown problem structure. For each task, six points could be reached, and therefore, as a maximum, 36 points could be reached in the posttest. Task performance was calculated by summing up points given for correct answers.

The self-explanation prompts were presented after working through the learning material. We asked participants to reflect what is new for them in the learning material. They should write down in own words how they would solve problems like those presented in the learning material and reflect if there is something, they do not yet understand.

Cognitive load was assessed through a differentiated questionnaire (**Supplementary Appendix A**; Klepsch et al., 2017). For the active and passive aspects of load, two items (**Supplementary Appendix A**) were used: "es war anstrengend" ("It has been strenuous," passive) vs. "ich habe mich angestrengt" ("I exerted myself," active).

Data Analysis

For all three scales of the differentiated questionnaire, reliability was estimated by using McDonald's ω. A principal component analysis (PCA) was conducted to provide evidence that the items are forming the intended factors. Then, the items on the active and passive aspects of load were included into the factor analysis, to show that the active items fit into the GCL factor and the passive item fits into the ICL or ECL factor. Then, correlations were conducted to again provide evidence if especially the passive item correlates with load resulting from the material (ICL and/or ECL), and the active item correlates with the GCL scale.

Additionally, we analyzed group differences regarding the control variables, prior knowledge, and age to check for differences despite the randomization. Also, correlations between the control variables and the dependent variables have been calculated. Whenever a significant difference between groups or a significant correlation could be found, the affected variable was included in the following analyses of variance as a covariate. To identify group differences, AN(C)OVAS were conducted with ICL, ECL, GCL, the active or the passive item, or learning outcomes as dependent variables.

Results

Relation of Intrinsic, Extraneous, and Germane Load With Active and Passive Load

We used McDonalds ω to assess the reliability of the three scales of the differentiated cognitive load questionnaire. For ICL and GCL, two items, and for ECL, three items were included, and we could find sufficient levels of McDonalds ω (ICL: 0.80, ECL: 0.85, and GCL: 0.77).

The principal component analysis (PCA) was conducted with the items of the differentiated questionnaire with orthogonal rotation (varimax). The Kaiser–Meyer–Olkin measure verified the sampling adequacy for the analysis (*KMO* = 0.59). Barlett's test of sphericity (χ^2 (21) = 206.34, $p < 0.001$) indicated that correlations between items were sufficiently large for PCA. An initial analysis was run to obtain eigenvalues for each component

TABLE 2 | Rotated PCA (a) with items from Klepsch et al. (2017) and (b) including the passive and active items for study 1.

(a)	Component			(b)	Component		
Item	1	2	3	Item	1	2	3
ICL 1	–0.015	0.918	0.088	ICL 1	0.842	–0.076	0.127
ICL 2	0.237	0.867	0.069	ICL 2	0.892	0.179	0.120
ECL 1	0.780	0.394	–0.067	ECL 1	0.450	0.750	–0.033
ECL 2	0.895	–0.115	0.209	ECL 2	–0.072	0.904	0.171
ECL 3	0.909	0.130	–0.118	ECL 3	0.173	0.896	–0.107
GCL 1	0.089	0.235	0.844	GCL 1	0.138	0.084	0.813
GCL 2	–0.067	–0.065	0.888	GCL 2	–0.107	–0.053	0.884
				Passive	0.789	0.328	–0.032
				Active	0.175	0.001	0.885

in the data. Three components had eigenvalues over Kaiser's criterion of 1 and in combination explained 81.76% of the variance. **Table 2** (a) shows the factor loadings after rotation. The items that cluster on the same components suggest that component 1 represents ECL, component 2 ICL, and component 3 GCL.

Including the active and passive items into the PCA resulted in the following outcome: The Kaiser–Meyer–Olkin measure verified the sampling adequacy for the analysis (KMO = 0.64). Barlett's test of sphericity (χ^2 (36) = 349.47, p < 0.001) indicated that correlations between items were sufficiently large for PCA. An initial analysis was run to obtain eigenvalues for each component in the data. Three components had eigenvalues over Kaiser's criterion of 1 and in combination explained 78.48% of the variance. **Table 2** (b) shows the factor loadings after rotation. The items that cluster on the same components suggest that component 1 represents ICL including the passive item, component 2 ECL, and component 3 GCL including the active item.

We found significant relationships between the active item and the original two-item GCL scale (r = 0.77, p (one-tailed) < 0.001) as well as the original two-item ICL scale (r = 0.21, p (one-tailed) = 0.03), and between the passive item and the original two-item ICL scale (r = 0.63, p (one-tailed) <. 001) as well as the original three-item ECL scale (r = 0.44, p (one-tailed) < 0.001). All correlations can be found in **Supplementary Appendix B**.

Validating the Measures Regarding the Implemented Design

Concerning the control variables, we could not find any group differences for the variables age (F < 1, n.s.) and prior knowledge (F (1,69) = 2.47, p = 0.12, η^2 = 0.04). Correlations between these variables and dependent variables showed that age is significantly correlated with learning outcome (r = −0.41, p < 0.001) and prior knowledge is significantly correlated with learning outcome (r = 0.32, p < 0.01), ECL (r = −0.24, p = 0.04), and the passive item (r = −0.29, p = 0.01), and were thus integrated as covariates in the respective analyses. All correlations can be found in **Supplementary Appendix C**.

All means and standard deviations of the dependent variables can be found in **Table 3**, and effects are visualized in **Figure 1**.

For ICL, we found no significant main effect of worked examples (F < 1, n.s.) or prompts (F < 1, n.s.), neither an interaction (F (1,69) = 1.39, p = 0.24, η^2 = 0.02): ICL was reported to be equal in the experimental groups.

For ECL, we found a significant main effect of worked examples (F (1,69) = 13.57, p < 0.001, η^2 = 0.17), no main effect of prompts (F > 1, n.s.), and no interaction (F > 1, n.s.): ECL was reported to be lower in the groups with worked examples.

For GCL, we found no main effect of worked examples (F (1,69) = 1.99, p = 0.16, η^2 = 0.03), a significant main effect of prompts (F (1,69) = 6.26, p = 0.02, η^2 = 0.08), and no interaction (F > 1, n.s.): GCL was reported to be higher in the groups without prompts.

For the passive item, we found no main effect of worked examples (F < 1, n.s.), no main effect of prompts (F (1,68) = 1.58, p = 0.21, η^2 = 0.02), but an interaction (F (1,68) = 4.49, p = 0.04, η^2 = 0.06): Simple main effects show that in the worked-example group, there is a significant difference if a learner gets prompts or gets no prompts (F (1,68) = 6.25, p = 0.02, η^2 = 0.08), with getting a prompt resulting in higher passive load. Thus, this result pattern does not fully correspond to those of either ICL or ECL.

For the active item, we found no significant main effect of worked examples (F < 1, n.s.) or prompts (F < 1, n.s.), neither an interaction (F < 1, n.s.): The active item was reported to be equal in the experimental groups. Thus, this result pattern does also not fully correspond to that of GCL.

For learning outcome, we found a significant main effect of worked examples (F (1,67) = 10.137, p < 0.01, η^2 = 0.13), no main effect of prompts (F (1,67) = 1.38, p = 0.25, η^2 = 0.02), and no

TABLE 3 | Means and standard deviation of all dependent variables in study 1.

	Problem-solving *without prompt	Problem-solving *with prompt	Worked examples *without prompt	Worked examples *with prompt
	M (SD)	M (SD)	M (SD)	M (SD)
ICL	4.39 (0.93)	3.97 (0.82)	4.28 (1.41)	4.48 (1.22)
ECL[1]	3.78 (0.22)	3.63 (0.26)	2.72 (0.22)	2.99 (0.23)
GCL	5.53 (0.73)	5.02 (1.46)	5.28 (0.77)	4.59 (1.04)
Passive[1]	4.72 (0.26)	4.49 (0.29)	4.15 (0.25)	5.05 (0.26)
Active	4.98 (0.99)	4.90 (1.50)	4.93 (1.04)	4.89 (1.35)
Learning outcome[1]	15.64 (1.85)	21.12 (2.16)	24.84 (1.80)	24.03 (1.89)

[1]*Corrected, if covariates are included.*

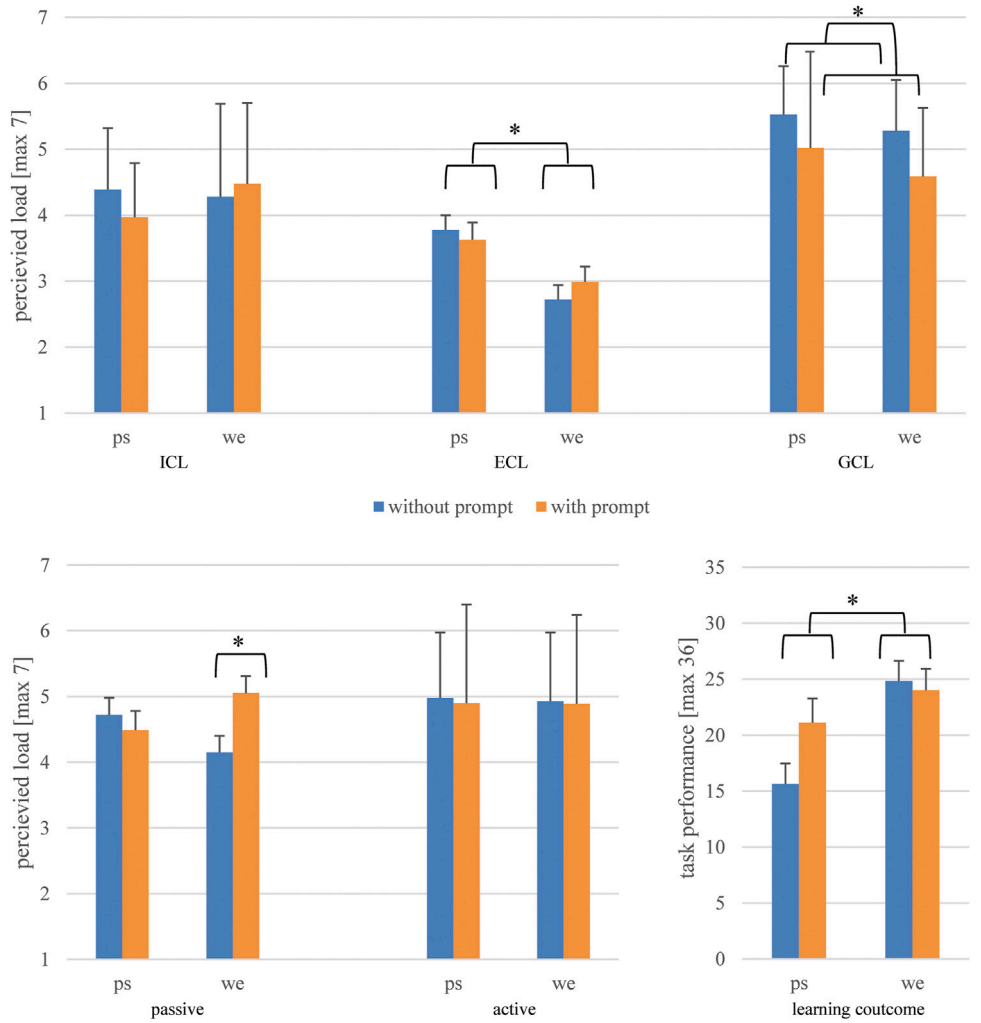

FIGURE 1 | Effects on the dependent variables for study 1. Note: ps = problem-solving, we = worked examples; * (simple) main effect with $p \leq 0.05$.

interaction (F (1,67) = 2.63, p = 0.11, η^2 = 0.04): Learning outcome was higher, when worked examples were given.

Discussion

Regarding the relation of the active and passive load measures with the three classical types of cognitive load, the PCA revealed that the active item is loading on the GCL factor and the passive item is loading on the ICL factor (H1). We also found the expected correlations between the active item and the original two-item GCL scale and between the passive-item and the original two-item ICL scale (H2). These results will be discussed more in depth in the general discussion.

Regarding the validation of the different load measures in the different instructional settings, we could not find any effects on ICL, which we expected, as we did not change the learning content between conditions. But we found a main effect of worked examples on ECL. In H3, we assumed that the passive item would correspond with either the ICL or the ECL scale, but the passive item is not aligned with the patterns for either ICL or ECL. Instead, we found a significant interaction effect, which is defined by the difference of learners with or without prompts in the worked-example group: Learners report higher passive load when provided with a prompt. The combination of worked examples and prompts seems to result in additional affordances that are necessary to link the worked examples with the prompts, which is in fact an alteration and increase in task affordances, that is, more interactive elements to deal with. The increase of passive load reported by this group thus indicates in our view that the passive item would be more closely linked to the concept of ICL than to ECL (H3_passive).

In this first study, we used prompts to elicit GCL, but the GCL items and the active item did not reveal corresponding results: We assumed a main effect of prompting on GCL and active load, which we could only find on GCL but not on active load. However, the main effect of prompts on GCL revealed unexpected results with higher levels of GCL for learners without prompts. This may be explained with the fact that they deliberately invested more mental effort as a human-

TABLE 4 | Number of participants in each experimental group of study 2.

Experimental groups	*n*
Split-source format * serialist	20
Split-source format * holist	16
Integrated format * serialist	21
Integrated format * holist	15

centered dimension, as they have not been guided with hints. Assuming that self-regulation also results in cognitive load (Seufert, 2018), giving prompts might have reduced the amount of needed self-regulation, and therefore can result in lower GCL. These arguments would be in line with seeing the missing prompts as a desirable difficulty (Bjork, 2017). However, as the learning outcomes did not increase for learners without prompts, we rather appraise missing prompts as a difficulty, which the participants in our study were willing to compensate with increased germane load. In the worked-example condition where they were relieved from additional load, this investment paid off, and learning outcomes were comparable to those of learners with prompts. For the active load measure, we could not find such effects.

To complete, we could find an effect of worked examples on learning outcome, which was better when worked examples were provided (H3_learning_outcome). For learning outcome, we could neither find an effect of prompts nor an interaction of worked examples and prompts, which was unexpected, but is in line with the found results on GCL and the active load item, as already mentioned, and will be discussed in the general discussion.

STUDY 2

As already mentioned, study 2 deals with the split-attention effect and Pask's learning styles (Pask, 1976). Learners had to deal with a biological topic about the structure and functions of the human kidney.

Based on theoretical assumptions and empirical findings, we assume effects of split-attention and learning style on ICL, ECL, GCL, and passive load and active load for study 2. While the integrated format in contrast to the split format should reduce the perceived ECL, learners' style of processing learning material in an either serialistic or holistic way should enhance GCL when it matches the presentation format. Learners with a holistic learning style tend to grasp the overall picture of the learning content and to integrate the most relevant parts while serialistic learners focus on specific details without trying to integrate them into an overall picture (Pask, 1976). Thus, no main effect but an interaction effect is hypothesized with more germane load for holistic learners in the integrated format, which helps them to get an overview, and for serialistic learners in the split format, which helps them to concentrate on details. As we did not change the learning content, there should not be any difference in perceived ICL. The assumptions for the active and passive load should correspond with the effects for either ICL and/or ECL (passive) or for GCL (active). We again analyzed the effects on learning outcomes: We assume that serialists reach higher levels of learning outcome when provided with the split-source material, whereas holists reach higher levels of learning outcome when provided with the integrated material

Method

In **Table 4**, the experimental conditions of study 2 and the number of participants in these conditions are listed.

Participants

We collected data from 72 learners. Participants were at average 22.99 (*SD* =3.98) years old, and 16.70 % were male. They were students from a German university.

Design

We conducted a 2 × 2 between-subject aptitude–treatment interaction study with four experimental conditions. Independent variables are related to the split-attention principle (split-source format vs. integrated format) and participants' learning style (serialist vs. holist). As dependent variables, we assessed cognitive load in a differentiated way, the active and passive parts of load and learning outcome. As control variables, age, sex, and prior knowledge have been assessed.

Procedure

As in study 1, at the beginning, learners were informed about the procedure of the study they participated in and signed an informed consent form. All participants were aware that they could withdraw their data at any point in the study without having any disadvantage. As a next step, each participant filled out a questionnaire asking for demographic data. Then, they filled out the questionnaire for assessing their learning type (described in Material). Also, in this study, prior knowledge was assessed (described in Material). Afterward, participants had to deal with the learning material either with a split-source format or an integrated format (described in Material), which was randomly assigned. After learning, the cognitive load questionnaire (**Supplementary Appendix A**; Klepsch et al., 2017) had to be answered, as well as the items to assess the active and passive parts of load (**Supplementary Appendix A**). In the end, the knowledge test had to be answered (described in Material).

Material

Prior knowledge about the human kidney was assessed through 10 open questions, covering content of the following learning material. The learning material itself was about the structure and functions of the kidney. It consisted of four pictures and seven corresponding texts. As a dependent variable, ECL was varied through either split-source material or integrated material. In the split-source material, the text was on the left side of the paper and the four pictures on the right side. References connecting the pictures with the text were included through corresponding numbers in the text and the pictures. In the integrated format, the connections between the four pictures were made clear by integrating the four pictures into one overall picture, giving the frame with the four detailed pictures like a zoom-in in different parts of the kidney. In addition, each text was placed near the corresponding part in the pictures. The aim of this approach was to minimize search processes between text and pictures. Each

TABLE 5 | Rotated PCA (a) with items from Klepsch et al. (2017) and (b) including the passive and active items for study 2.

(a)	Component			(b)	Component		
Item	1	2	3	Item	1	2	3
ICL 1	−0.155	0.083	0.883	ICL 1	−0.231	0.028	0.837
ICL 2	0.207	−0.008	0.857	ICL 2	0.154	0.007	0.849
ECL 1	0.818	−0.077	0.153	ECL 1	0.799	−0.058	0.193
ECL 2	0.843	−0.164	−0.131	ECL 2	0.856	−0.116	−0.083
ECL 3	0.880	−0.074	0.027	ECL 3	0.867	−0.037	0.045
GCL 1	−0.021	0.926	−0.036	GCL 1	−0.097	0.878	−0.071
GCL 2	−0.239	0.870	0.122	GCL 2	−0.317	0.808	0.089
				Passive	0.315	0.334	0.750
				Active	0.148	0.813	0.278

participant had 30 min to learn the content. To assess learning performance, a posttest had to be filled in consisting of 13 tasks. Altogether, 35 points could be reached.

Pask's learning styles (Pask, 1976) were assessed with the questionnaire of Ford (1985). Based on the calculated value, learners were either categorized as serialists or holists.

Cognitive load was again assessed with a differentiated questionnaire (**Supplementary Appendix A**; Klepsch et al., 2017). For the active and passive aspects of load, the same two items (**Supplementary Appendix A**) were used as in study 1, namely, "es war anstrengend" ("It has been strenuous," passive) vs. "ich habe mich angestrengt" ("I exerted myself," active).

Data Analysis

For all three scales of the differentiated questionnaire, reliability was estimated by using McDonald's ω. A principal component analysis (PCA) was conducted to provide evidence that the items are forming the intended factors. Then, the items on the active and passive aspects of load were included into the factor analysis, to show that the active items fit into the GCL factor and the passive item fits into the ICL or ECL factor. Additionally, correlation analysis was conducted to provide evidence again if especially the passive item correlates with load resulting from the material (ICL and/or ECL), and the active item correlates with the GCL scale.

Again, group differences were analyzed regarding the control variables such as prior knowledge and age. Also, correlations between these variables and the dependent variables have been calculated. Whenever a significant difference between groups or a significant correlation could be found, the affected variable was included in the following analyses of variance as a covariate. To identify group differences, AN(C)OVAS were conducted with ICL, ECL, GCL, passive load or active load, and learning outcome as dependent variables.

Results

Relation of Intrinsic, Extraneous, and Germane Load With Active and Passive Load

We used McDonalds ω to assess the reliability of the three scales of the differentiated cognitive load questionnaire. For ICL and GCL, two items, and for ECL, three items were included, and we could find satisfying levels of McDonalds ω (ICL: 0.72, ECL: 0.75, and GCL: 0.76).

The PCA was conducted with the items of the differentiated questionnaire with orthogonal rotation (varimax). The Kaiser–Meyer–Olkin measure verified the sampling adequacy for the analysis ($KMO = 0.63$). Barlett's test of sphericity (χ^2 (21) = 157.54, $p < 0.001$) indicated that correlations between items were sufficiently large for PCA. An initial analysis was run to obtain eigenvalues for each component in the data. Three components had eigenvalues over Kaiser's criterion of 1 and in combination explained 78.71% of the variance. **Table 5** (a) shows the factor loadings after rotation. The items that cluster on the same components suggest that component 1 represents ECL, component 2 GCL, and component 3 ICL.

Including the active and passive items into the PCA resulted in the following outcome: The Kaiser–Meyer–Olkin measure verified the sampling adequacy for the analysis ($KMO = 0.70$). Barlett's test of sphericity (χ^2 (36) = 258.17, $p < 0.001$) indicated that correlations between items were sufficiently large for PCA. An initial analysis was run to obtain eigenvalues for each component in the data. Three components had eigenvalues over Kaiser's criterion of 1 and in combination explained 75.21% of the variance. **Table 5** shows the factor loadings after rotation. The items that cluster on the same components suggest that component 1 represents ECL, component 2 GCL including the active item, and component 3 ICL including the passive item.

We found significant relationships between the active item and the original two-item GCL scale ($r = 0.59$, p (one-tailed) < 0.001) as well as the original two-item ICL scale ($r = 0.23$, p (one-tailed) = 0.03), and between the passive item and the original two-item ICL scale ($r = 0.60$, p (one-tailed) < 0.001) and the original three-item ECL scale ($r = 0.27$, p (one-tailed) = 0.01). All correlations can be found in **Supplementary Appendix B**.

Validating the Measures Regarding the Implemented Design

Concerning the control variables, we could not find any group differences for the variables such as age ($F < 1$, n.s.) and prior knowledge (F (1,68) = 1.14, $p = 0.29$, $\eta^2 = 0.02$). Correlations between these variables and dependent variables showed that prior knowledge was significantly correlated with learning outcome ($r = 0.40$, $p < 0.01$), ICL ($r = -0.26$, $p = 0.03$) and the passive item ($r = -0.25$, $p = 0.04$), and therefore included as a covariate in the respective analyses. All correlations can be found in **Supplementary Appendix C**.

All means and standard deviations of the dependent variables can be found in **Table 6**, and effects are visualized in **Figure 2**.

For ICL, we found no main effect of learning style ($F < 1$, n.s.) or material ($F < 1$, n.s.), neither an interaction ($F < 1$, n.s.): ICL was reported to be equal in the experimental groups.

For ECL, we found no main effect of learning style ($F < 1$, n.s.), a significant main effect of material (F (1,68) = 5.55, $p = 0.02$, $\eta^2 = 0.08$), and no interaction ($F > 1$, n.s.): ECL was reported to be higher with split-source material than integrated material.

For GCL, we found no main effect of learning style ($F < 1$, n.s.), or material (F (1,68) = 3.34, $p = 0.07$, $\eta^2 = 0.05$), but a significant interaction (F (1,68) = 5.82, $p = 0.02$, $\eta^2 = 0.05$): Simple main effects showed that with split-source material, serialists reported

TABLE 6 | Means and standard deviation of all dependent variables in study 2.

	Split-source material * serialist	Split-source material * holist	Integrated material * serialist	Integrated material * holist
	M (SD)	*M (SD)*	*M (SD)*	*M (SD)*
ICL[1]	5.57 (0.21)	5.51 (0.24)	5.44 (0.21)	5.82 (0.21)
ECL	4.74 (1.10)	4.78 (1.22)	4.25 (1.39)	3.92 (1.04)
GCL	5.40 (0.65)	4.59 (1.58)	5.24 (1.34)	5.77 (0.88)
Passive[1]	6.06 (0.27)	5.57 (0.26)	5.34 (0.30)	5.82 (0.32)
Active	6.03 (0.57)	4.94 (1.60)	5.64 (1.01)	5.73 (1.19)
Learning outcome[1]	15.40 (1.45)	10.22 (1.62)	12.44 (1.40)	14.06 (1.70)

[1]*Corrected, if covariates are included.*

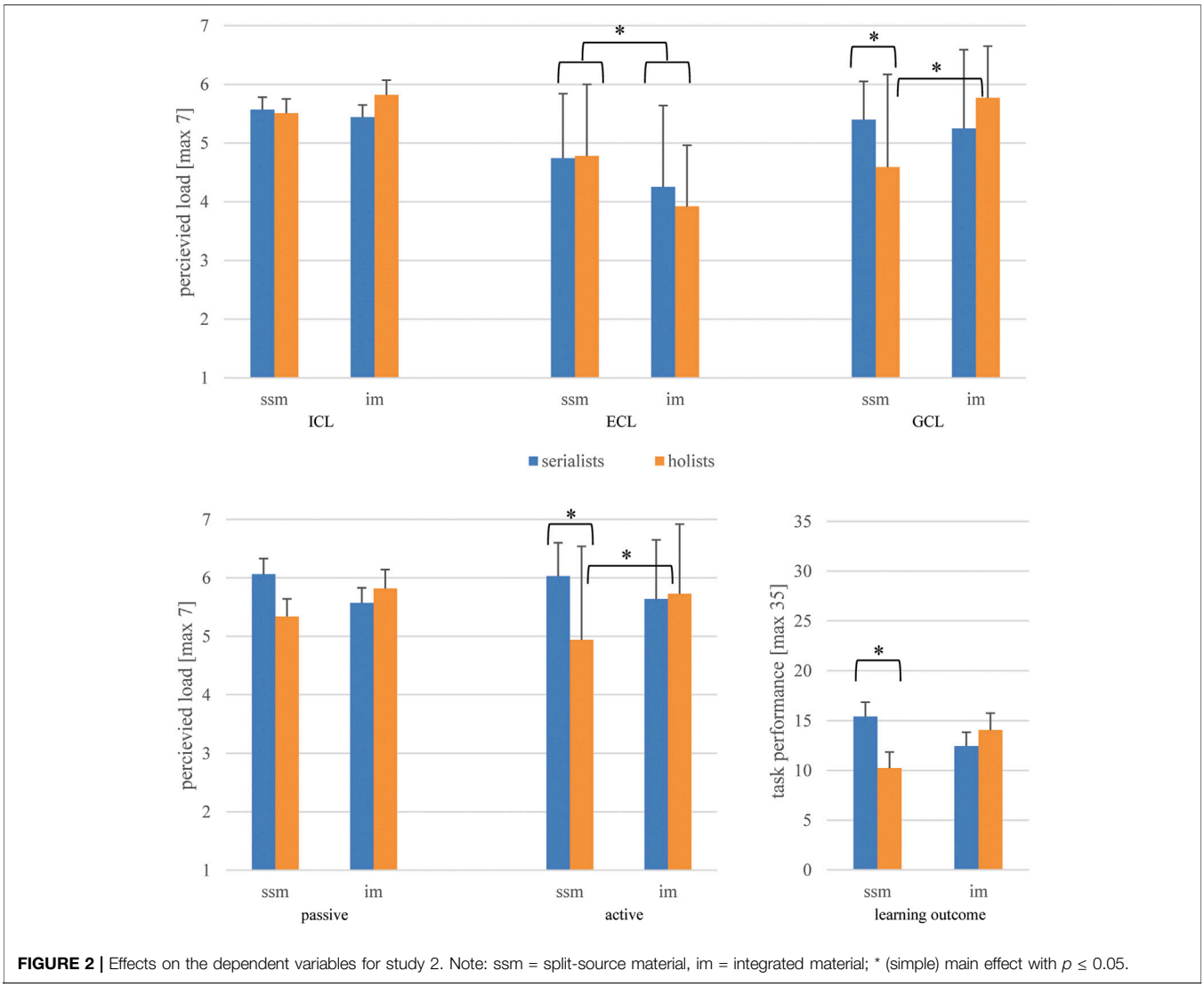

FIGURE 2 | Effects on the dependent variables for study 2. Note: ssm = split-source material, im = integrated material; * (simple) main effect with $p \leq 0.05$.

higher GCL than holists (F (1,68) = 4.28, p = 0.04, η^2 = 0.06), and holists reported higher GCL with integrated material than split-source material (F (1,68) = 7.89, p < 0.01, η^2 = 0.10).

For the passive item, we found no main effect of learning style (F < 1, n.s.), material (F < 1, n.s.), or interaction (F (1,68) = 2.89, p = 0.09, η^2= 0.04): The responses to the passive item were equal in the experimental groups. Thus, this result pattern corresponds to that of ICL.

For the active item, we found no main effect of learning style (F (1,68) = 3.50, p = 0.07, η^2 = 0.05), or material (F < 1, n.s.), but a

significant interaction (F (1,68) = 4.89, p = 0.03, η^2 = 0.07): Simple main effects showed that with split-source material, serialists reported higher active investment than holists (F (1,68) = 8.39, p < 0.01, η^2 = 0.11), and holists reported higher GCL with integrated material than split-source material (F (1,68) = 3.91, p = 0.05, η^2 = 0.05). Thus, this result pattern corresponds to that of GCL.

For the overall learning outcome, we found no main effect of learning style (F (1,67) = 1.34, p = 0.25, η^2 = 0.02) or of material (F < 1, .s.), but the expected significant interaction (F (1,67) = 4.86, p = 0.03, η^2 = 0.07): Simple main effects showed that with split-source material, serialists reached higher levels of overall learning outcome than holists (F (1,67) = 5.79, p = 0.02, η^2 = 0.08).

Discussion

Also, for study 2 with the PCA, we could show that the active item is loading on the GCL factor and the passive item is loading on the ICL factor (H1). We also found the expected correlations between the active item and the original two-item GCL scale and between the passive item and the original two-item ICL scale (H2). These results will be discussed more in depth in the general discussion.

The results of the second study show evidence for H3: We found the same effects on GCL and active load, where we categorized learners based on their learning style. A significant interaction of learner's aptitude and the treatment shows that learners with different prerequisites benefit from different treatments. This effect could be found for the GCL scale as well as for the active item (H3_active).

In line with our hypotheses, we could not find any effects, neither of the independent variables or their interaction with ICL. For the ECL measure, however, we were able to find a main effect of split attention. For the passive item, we cloud not find any main effects and no interaction, which indicates that the passive item is rather related to ICL and not to ECL (H3_passive).

GENERAL DISCUSSION

The goal of the present studies was to investigate the concept of passive load, that is, experiencing load, and active load, that is, investing effort, in relation to the classical three-partite types of load based on CLT. Therefore, in a first step, we discuss the results of the two studies based on hypotheses and address the question if reported active and passive load can be connected to reported ICL, ECL, and GCL measured with the differentiated questionnaire of Klepsch et al. (2017). Next, strengths and weaknesses of the studies as well as of the concept of active and passive load in general are discussed, and further directions are addressed. Finally, we discuss the usefulness of the concepts of active and passive load, especially the theoretical and practical implications, and the need for further research. The stated inferences (scoring, generalizations, and extrapolation) will also be discussed to provide broader evidence for validity.

Active and Passive Aspects of Load and Their Connection to ICL, ECL, and GCL

Based on our findings, we conclude that it is possible to distinguish between active and passive aspects of load and that the concept can be related to the three types of cognitive load, ICL, ECL, and GCL measured with the differentiated questionnaire by Klepsch et al. (2017).

We could find sufficient evidence to underline our first two hypotheses (H1 and H2) and therefore also for the scoring inference that active load is associated strongly with GCL and passive load strongly with ICL and weakly with ECL. First, we could show that in both studies, PCA of the items for the three original load types results in the same factor structure as reported by Klepsch et al. (2017). Thus, reliability of the scales is given, which is a necessary but not sufficient condition for validity. Including the active and passive items into the PCA showed that the active item extends the GCL scale and the passive item extends the ICL scale. This was the case in both studies. In a second step, the correlation analysis showed that in both studies, the active item has a moderate-to-strong positive relationship with the GCL scale and that the passive item has a moderate-to-strong positive relationship with the ICL scale. Thus, we can state that these links between the active item with GCL and the passive item with ICL are rather stable and replicable in two independent studies. However, the active item also correlated weakly with ICL in both studies, which underpins the idea that ICL and GCL might be highly interlinked (Kalyuga, 2011). But as the strengths of the correlations highly differ—the active item is much stronger related to GCL than to ICL—we can state that active investment is more clearly linked to germane load. However, learners can and must only invest additional effort if the complexity of the task is sufficiently high, that is, when ICL is high. The role of ECL is not that clearly linked to the active–passive construct. We found a moderate correlation of the passive item with ECL in both studies. Thus, the extraneous affordances do play a role for learners' appraisal of passive load, but in our studies, it was less considerable than ICL. However, whether this would also be the case in settings where ECL is remarkably higher overall and would be thus more striking and disrupting, the impact of ECL for passive load appraisals might increase in relation to ICL.

Overall, we can state that experienced load can be differentiated with the two items for measuring active and passive load. This distinction becomes important as soon as learners act in a self-regulated way and decide to invest effort in processing learning content.

Validating the Results in Instructional Design Studies

In the third hypothesis, we assumed to find corresponding result patterns for the active load item and GCL whenever generative learning activities are triggered by the design (i.e., all measures should increase) (H3_active), and also corresponding result patterns for the passive load item with ICL or ECL (H3_passive).

In study 2, we could show that generative learning activities result in differences in GCL and the active load item, but we could not show the same pattern for study 1.

But why did learners rate the active load item in accordance with the GCL items in the second study but not in the first? Perhaps prompts, which aims at triggering germane load externally, and

internally available task-specific abilities—like strategy skills to process information in a holistic or serialistic way—have different effects on perceived active investment. Applying internally available abilities in learning settings that matches these abilities, for example, providing holists with integrated material, is directly experienced as active investment. Externally triggered investment, in our case prompting, might not be experienced as one's own active investment. Whether this argument is generalizable has to be proven in future studies with measures of active load in situations with more successful triggers for generative activities. Nevertheless, the role of learners' aptitudes in interaction with different instructional settings as an internal trigger of germane investment seems to be promising and should be analyzed in a study with a direct comparison to an externally triggered germane investment. With reference to the postulated scoring inference on GCL (H3_active; where we stated that if GCL increases, then also the scores on the active load items should increase), we can state that we could find evidence for that.

At least one of the stated scoring inferences on ICL and ECL cannot be assured, as we could not find a stable link between ECL and the passive item. In contrast, in both studies, we could show that the passive item is more in line with ICL (H3_passive). But we are not able to fully substantiate this, as our studies do not include variations in ICL. Therefore, at the moment, we can only state that we have found a promising link between ICL and the passive item through PCA and the conducted correlation analyses. Based on the given design of the studies, we are not able to provide evidence whether rated passive load would increase if ICL of the task increases. This should be investigated in further studies.

Overall, interesting to mention is that all the design effects, which are reflected in ECL, cannot be found for either the passive or active load measure. To complete the picture of the instructional design effects, we found the expected effect of worked examples with better learning outcomes when worked examples were provided. Unfortunately, we could not confirm positive effects on learning outcomes when prompts were provided. As we already discussed, this might be due to the additional perceived passive load when prompts and worked examples were given because learners might have felt overwhelmed in self-explaining. For the split-attention effect in the second study, we found the expected interaction with learners' style to deal with either holistic (integrated) or serially (split) presented material. Overall, the implemented design effects can be largely seen as valid means to trigger the expected effects. Thus, these results substantiate the validity of the result patterns of the load measures and the correspondence of the passive load measure with the ICL scale and the active load measure with the GCL scale.

Strengths and Weaknesses

Overall, we can state that both learning materials in the two studies have been rather complex and difficult. This is also reflected in learning outcome. In study 1, learners did hardly reach two-thirds of the points, whereas in study 2, most learners did not even reach half of the points in the posttest. In study 2, the internal complexity is also reflected in reported high ICL and passive load. In study 1, reported ICL and passive load are not that pronounced, but still high. Also, important to mention, both studies are classical experimental studies conducted with students at a German university, and both include many more female than male participants, which were fortunately randomized equal over experimental groups. Thus, despite the evidence of the studies which allows some generalization, the generalization inference is not fully proven yet. Results should be replicated with broader and more balanced samples. Moreover, the samples in both studies have been rather small. We had at least 15 participants in each experimental condition, but we are aware that this results in power restrictions. Therefore, it would be of interest to conduct studies with larger samples and in a more realistic learning setting, for example, in school. Using a school setting would additionally allow to analyze whether younger kids and teens would be able to understand the items for active and passive load, as we would assume, which would even provide more proof of generalization.

The concept of active and passive load comes with some positive but also some negative aspects that should be considered. On the one hand, active and passive load in their operationalization should be easy to understand even for younger kids and teens. They can be answered rather quickly, and therefore can be used for repeated online measures of cognitive load during a learning phase, without interfering with learning. On the other hand, we could not find any evidence that design aspects in the learning material are reflected in passive or active load. Therefore, these items seem not to be useful if one is especially interested in differences in the design effects of learning materials. However, as soon as learner characteristics are considered, the concept of active and passive load seems to become important, which is addressed in the next section.

Future Directions of Research

As already mentioned, ICL mirrors the externally given task affordances, and GCL can be seen as the actively invested resources by the learner (Seufert, 2018). Our results show that this can be linked to passive and active load. Passive load shows in study 2 a similar pattern like ICL. In study 1, the overall pattern is also similar, but we found a simple main effect of prompts in the worked-example group. As the prompt was new for the learners, it might have resulted in passive rather than active load—as intended by us—because the prompt belongs to the task-centered dimension (Paas and van Merriënboer, 1994) or can be seen as data-driven (Scheiter et al., 2020). Further studies should have a closer look at prompts and other active load–enhancing techniques, like asking learners to use special learning strategies. Based on the presented results, we would assume that all these active load–enhancing techniques at the beginning result in more passive load and over time—when learners are used to the techniques and can use them more automatically—the techniques should result in higher active load instead of passive load (see also mathemathantic effect; Clarck, 1990). Whether prompts for generative activities can be fruitful for learners and are thus experienced as active and germane rather than hindering and passive, highly depends on learners' prerequisite. Learners' prior knowledge or their strategy skills should be taken into account (see, e.g., Nückles et al., 2020; Seufert, 2020). Study 2 also provides evidence that learner

characteristics, which are internally given (even per se or through training) in interaction with the instructional design, can result in differences in active load. Only when the external affordances match the internal abilities germane resources are invested, that is, learners report germane load as well as active load investment. In our case, in study 2, split-source material would fit more to learners with a serialistic approach of learning, whereas integrated material would fit more to learners with a more holistic approach of learning. The fact that this fit is also crucial for being successful with this engagement can be seen with the same effects mirrored in learning outcomes. In future studies, the complex interplay of learner characteristics, tasks affordances, and design aspects should be addressed.

As already mentioned, the two items for active and passive load can be easily used to assess load online, by asking learners repeatedly during learning to rate their perceived active and passive load. This would be particularly relevant when intrinsic load is changing over time, for example, when learning content gets more complex with each step, a learner moves forward or when learners gain expertise. Especially an observation over longer periods of time, when, for example, a training on learning strategies is implemented, would be of interest to provide evidence if the use of a new learning strategy shifts from passively experienced load to active investment as soon as the strategy is internalized and therefore more easily used automatically.

Generally, the two items that measure for active and passive load could be used to either substitute or extend other measures of cognitive load. It would be particularly interesting to analyze the items in combination with other differentiated measures of cognitive load by Leppink et al. (2013) or by Krell (2017) to substantiate its validity. Doing so would also help to provide more evidence for the stated extrapolation inference to get evidence if one measure can predict another and especially if learning outcomes in further tasks could be predicted. In the two studies, we analyzed the relations between the active and passive items and the cognitive load questionnaire of Klepsch et al. (2017). If measurement methods of cognitive load have predictive validity, it would be of interest to have a closer look on different methods of measurement and their mutual prediction of each other. Also, the extrapolation of learning outcome should be surveyed more closely. Using the items repeatedly during a study could, for example, be used to predict if the next task will be performed correctly or with errors.

CONCLUSION

We found interesting evidence for the concept of active and passive load and its connection to the classic three-partite concept of cognitive load. However, there is still research needed to investigate active and passive load more in depth. Based on the interplay between the two approaches, the active–passive aspects and the original three load types, further studies should be conducted and analyzed, to discuss the concept of active and passive load on a theoretical level. This is especially necessary when considering learner's active role in self-regulatory learning (Seufert, 2018; Seufert, 2020) and trying to connect the concept of active and passive load with analogous concepts like the human- or task-centered dimension (Paas and van Merriënboer, 1994) or the concept of data-driven and goal-driven appraisals of cognitive load as discussed at the crossroad between CLT and self-regulated learning research (De Bruin et al., 2020).

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

Ethical review and approval was not required for the studies on human participants in accordance with the local legislation, institutional requirements and due to the recommendations of the German Research Association: No subject was in risk of physical or emotional pressure, we fully informed all subjects in all studies about the goals and processes of the study they participated in, and none of the subjects were patients, minors, or persons with disabilities. In all studies, participation was voluntary, and all subjects signed a written informed consent and were aware that they had the chance to withdraw their data at any point of the study.

AUTHOR CONTRIBUTIONS

MK and TS contributed to the conception and design of the studies. MK developed the used material and questionnaires and led data collection for studies. MK analyzed and interpreted the data. TS and MK drafted the work. All authors provided approval of the final submitted version of the manuscript and agreed to be accountable for all aspects of the work by ensuring that questions related to the accuracy or integrity of any part of the work were appropriately investigated and resolved.

ACKNOWLEDGMENTS

We like to thank our former students involved in one of the studies. They contributed to this research as part of their curriculum, in form of a bachelor or master thesis, or as student research assistants.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2021.645284/full#supplementary-material.

REFERENCES

Artino, A. R., (2008). Cognitive load theory and the role of learner experience. An abbreviated review for educational practitioners. *AACE J.* 16, 425–439.

Ayres, P., and Sweller, J. (2014). in *"The split-attention principle in multimedia learning," in* the Cambridge handbook of multimedia learning. Editor R. E. Mayer. Second edn (New York, NY: Cambridge University Press), 206–226.

Bjork, R. A. (2017). "Creating desirable difficulties to enhance learning," in *In progress*. Editors I. Wallace, L. Kirkman, and L. Vergne (Bethel, CT: Crown House Publishing)), 81–89.

Clarck, R. E. (1990). "When teaching kills learning: research on mathemathantics," in analysis of complex skills and complex knowledge domains: learning and instruction," in *European research in an international context*. Editors H. Mandl, E. De Corte, N. Bennett, and H. F. Friedrich (Oxford, UK: Pergamon Press), 1–22.

De Bruin, A. B. H., Roelle, J., Roelle, J., Carpenter, S. K., and Baars, M. (2020). Synthesizing cognitive load and self-regulation theory: a theoretical framework and research agenda. *Educ. Psychol. Rev.* 32, 903–915. doi:10.1007/s10648-020-09576-4

Eitel, A., Endres, T., and Renkl, A. (2020). Self-management as a bridge between cognitive load and self-regulated learning: the illustrative case of seductive details. *Educ. Psychol. Rev.* 32, 1073–1087. doi:10.1007/s10648-020-09559-5

Ford, N. (1985). Learning styles and strategies of postgraduate students. *Br. J. Educ. Technol.* 16, 65–77. doi:10.1111/j.1467-8535.1985.tb00483.x

Kalyuga, S. (2011). Cognitive load theory: how many types of load does it really need? *Educ. Psychol. Rev.* 23, 1–19. doi:10.1007/s10648-010-9150-7

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *J. Educ. Meas.* 50, 1–73. doi:10.1111/jedm.12000

Klepsch, M., Schmitz, F., and Seufert, T. (2017). Development and validation of two instruments measuring intrinsic, extraneous, and germane cognitive load. *Front. Psychol.* 8, 1997. doi:10.3389/fpsyg.2017.01997

Klepsch, M., and Seufert, T. (2020). Understanding instructional design effects by differentiated measurement of intrinsic, extraneous, and germane cognitive load. *Instr. Sci.* 48, 45–77. doi:10.1007/s11251-020-09502-9

Koriat, A., Ma'ayan, H., and Nussinson, R. (2006). The intricate relationships between monitoring and control in metacognition: lessons for the cause-and-effect relation between subjective experience and behavior. *J. Exp. Psychol. Gen.* 135, 36–69. doi:10.1037/0096-3445.135.1.36

Krell, M. (2017). Evaluating an instrument to measure mental load and mental effort considering different sources of validity evidence. *Cogent Edu.* 4, 1280256. doi:10.1080/2331186X.2017.1280256

Leppink, J., Paas, F., van der Vleuten, C. P. M., van Gog, T., and Van Merriënboer, J. J. G. (2013). Development of an instrument for measuring different types of cognitive load. *Behav. Res.* 45, 1058–1072. doi:10.3758/s13428-013-0334-1

Moreno, R., and Mayer, R. E. (2010). "Techniques That Increase Generative Processing in Multimedia Learning: Open Questions for Cognitive Load Research," in *Cognitive Load Theory*. Editors J. L. Plass, R. Moreno, and R. Brünken (Cambridge, New York: Cambridge University Press), 153–177.

Minkley, N., Xu, K. M., and Krell, M. (2021). Analyzing Relationships Between Causal and Assessment Factors of Cognitive Load: Associations Between Objective and Subjective Measures of Cognitive Load, Stress, Interest, and Self-Concept. *Front. Educ.* 6, 632907. doi:10.3389/feduc.2021.632907

Minkley, N., Xu, K., and Krell, M. (2021). Analyzing relationships between causal and assessment factors of cognitive load: associations between objective and subjective measures of cognitive load, stress, interest, and self-concept. *Front. Educ.* doi:10.3389/feduc.2021.632907

Nückles, M., Roelle, J., Glogger-Frey, I., Waldeyer, J., and Renkl, A. (2020). The self-regulation-view in writing-to-learn: using journal writing to optimize cognitive load in self-regulated learning. *Educ. Psychol. Rev.* 32, 1089–1126. doi:10.1007/s10648-020-09541-1

Paas, F. G. W. C. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: a cognitive-load approach. *J. Educ. Psychol.* 84, 429–434. doi:10.1037/0022-0663.84.4.429

Paas, F. G. W. C., and Van Merriënboer, J. J. G. (1994). Instructional control of cognitive load in the training of complex cognitive tasks. *Educ. Psychol. Rev.* 6, 351–371. doi:10.1007/BF02213420

Paas, F., Renkl, A., and Sweller, J. (2003). Cognitive load theory and instructional design: recent developments. *Educ. Psychol.* 38, 1–4. doi:10.1207/S15326985EP3801_1

Park, B., Münzer, S., Seufert, T., and Brünken, R. (2016). The role of spatial ability when fostering mental animation in multimedia learning: an ATI-study. *Comput. Hum. Behav.* 64, 497–506. doi:10.1016/j.chb.2016.07.022

Park, B. (2010). "Testing the additivity hypothesis of cognitive load theory,". M.S Dissertation, Saarbrücken, (Germany): Universität Des Saarlandes.

Pask, G. (1976). Styles and strategies of learning. *Br. J. Educ. Psychol.* 46, 128–148. doi:10.1111/j.2044-8279.1976.tb02305.x

R. E. Mayer (Editors) (2014). *The Cambridge handbook of multimedia learning*. Second edition (New York, NY: Cambridge University Press).

Renkl, A. (2014). "The worked examples principle in multimedia learning,"in *the Cambridge handbook of multimedia learning*. Editor R. E. Mayer. Second edn (New York, NY: Cambridge University Press), 391–412.

Scheiter, K., Ackerman, R., and Hoogerheide, V. (2020). Looking at mental effort appraisals through a metacognitive lens: are they biased? *Educ. Psychol. Rev.* 32, 1003–1027. doi:10.1007/s10648-020-09555-9

Schnaubert, L., and Schneider, S. (2020). How learners use mental load and mental effort as indicators for metacomprehension under different load-inducing conditions of multimedia design.

Seufert, T. (2018). The interplay between self-regulation in learning and cognitive load. *Educ. Res. Rev.* 24, 116–129. doi:10.1016/j.edurev.2018.03.004

Seufert, T. (2020). Building bridges between self-regulation and cognitive load-an invitation for a broad and differentiated attempt. *Educ. Psychol. Rev.* 32, 1151–1162. doi:10.1007/s10648-020-09574-6

Sweller, J., Van Merriënboer, J. J. G., and Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. *Educ. Psychol. Rev.* 10, 251–296. doi:10.1023/a:1022193728205

Taxis, S.-S., Gutmann, C., Herzmann, P., and Seufert, T. (2010). "Effects of a learning strategy training for children," in *"Instructional Design for motivated and competent learning in a digital world": program Book*. Editors M. Hopp and F. Wagner, (Ulm, Germany: Ulm University), 44–46.

Van Gog, T., Paas, F., and Van Merriënboer, J. J. G. (2008). Effects of studying sequences of process-oriented and product-oriented worked examples on troubleshooting transfer efficiency. *Learn. Instr.* 18, 211–222. doi:10.1016/j.learninstruc.2007.03.003

Wirzberger, M., Beege, M., Schneider, S., Nebel, S., and Rey, G. D. (2016). One for all?! simultaneous examination of load-inducing factors for advancing media-related instructional research. *Comput. Educ.* 100, 18–31. doi:10.1016/j.compedu.2016.04.010

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Education

ORIGINAL RESEARCH
published: 10 May 2021
doi: 10.3389/feduc.2021.647097

Subjective Measure of Cognitive Load Depends on Participants' Content Knowledge Level

Tianlong Zu[1], Jeremy Munsell[2] and N. Sanjay Rebello[2,3]*

[1] Department of Physics, Lawrence University, Appleton, WI, United States, [2] Department of Physics and Astronomy, Purdue University, West Lafayette, IN, United States, [3] Department of Curriculum and Instruction, Purdue University, West Lafayette, IN, United States

Cognitive load theory (CLT) posits the classic view that cognitive load (CL) has three-components: intrinsic, extraneous and germane. Prior research has shown that subjective ratings are valid measures of different CL subtypes. To a lesser degree, how the validity of these subjective ratings depends on learner characteristics has not been studied. In this research, we explored the extent to which the validity of a specific set of subjective measures depends upon learners' prior knowledge. Specifically, we developed an eight-item survey to measure the three aforementioned subtypes of CL perceived by participants in a testing environment. In the first experiment (N = 45) participants categorized the eight items into different groups based on similarity of themes. Most of the participants sorted the items consistent with a threefold construct of the CLT. Interviews with a subgroup (N = 13) of participants provided verbal evidence corroborating their understanding of the items that was consistent with the classic view of the CLT. In the second experiment (N = 139) participants completed the survey twice after taking a conceptual test in a pre/post setting. A principal component analysis (PCA) revealed a two-component structure for the survey when the content knowledge level of the participants was initially lower, but a three-component structure when the content knowledge of the participants was improved to a higher level. The results seem to suggest that low prior knowledge participants failed to differentiate the items targeting the intrinsic load from those measuring the extraneous load. In the third experiment (N = 40) participants completed the CL survey after taking a test consisting of problems imposing different levels of intrinsic and extraneous load. The results reveals that how participants rated on the CL survey was consistent with how each CL subtype was manipulated. Thus, the CL survey developed is decently effective measuring different types of CL. We suggest instructors to use this instrument after participants have established certain level of relevant knowledge.

Keywords: cognitive load, subjective measure, validity, test, content knowledge

OPEN ACCESS

Edited by:
Fred Paas,
Erasmus University Rotterdam, Netherlands

Reviewed by:
John Sweller,
University of New South Wales, Australia
Patricia O'Sullivan,
University of California, San Francisco, United States

****Correspondence:***
Tianlong Zu
tianlong.zu@lawrence.edu

Specialty section:
This article was submitted to Assessment, Testing and Applied Measurement, a section of the journal Frontiers in Education

Received: *29 December 2020*
Accepted: *12 April 2021*
Published: *10 May 2021*

Citation:
Zu T, Munsell J and Rebello NS (2021) Subjective Measure of Cognitive Load Depends on Participants' Content Knowledge Level. Front. Educ. 6:647097. doi: 10.3389/feduc.2021.647097

INTRODUCTION

Cognitive load theory (CLT) attends to the limited working memory capacity (Cowan, 2001) for instruction and learning. It posits that optimal design of instruction and learning should not overload learners' working memory capacity (Sweller, 1988, 1994, 2010; Sweller et al., 1998). This is because novel information and previously learned information from long-term memory need

to be consciously processed in working memory. However, working memory is limited in capacity and duration when processing novel information, especially without deliberate rehearsal (Baddeley, 1992; Cowan, 2001). This is in contrast to the long-term memory, which is an unlimited, permanent repository for organized knowledge that governs our cognitive processes (Sweller, 2010).

Cognitive load is defined as the working memory load experienced when performing a specific task (Kalyuga, 2011; Sweller et al., 2011; van Merriënboer and Sweller, 2005). This places a requirement on instruction to avoid overloading the working memory during learning. Cognitive load is a multifaceted construct. Historically, extraneous cognitive load (ECL) was the first CL subtype introduced by Sweller (1988). ECL refers to the working memory resources allocated to unproductive cognitive processes. The level of ECL is related to the presentation format (e.g., visual, audio, and text), spatial and temporal organization of various information, etc. For example, high ECL can be caused if the same information is presented simultaneously in both text and audio modalities, since the redundancy would cause cognitive resources to be wasted which could potentially hinder learning. The second CL subtype is the intrinsic cognitive load (ICL) which is related to the working memory resources allocated to dealing with the learning objectives (Sweller, 1994). Finally, a third kind of CL, germane cognitive load (GCL), refers to the working memory resources used for constructing, chunking and automating schemas (Sweller et al., 1998). From a common theoretical perspective, John Sweller suggested that all three subtypes of CL can be defined in terms of the core concept of element interactivity (Sweller, 2010). Under this theoretical formalism, the load is extraneous if the element interactivity can be reduced without altering the learning objective. The load is intrinsic if reducing the element interactivity alters the learning objective. Germane load simply refers to the working memory resources used for processing the intrinsic load, such as through chunking, schema generation and automation, and therefore is also tied to element interactivity. Kalyuga (2011) also argues germane load is not an independent load type, since there is no theoretical argument for any difference between GCL and ICL. Thus, it has been suggested that GCL can be readily incorporated into the definition of ICL by redefining the cognitive processes involving GCL as pertaining to the learning goals related to ICL. Thus, there are only two independent components in CLT: ECL and ICL, which are additive. Scientists would favor a two-component model if a two-component construct of CL has equal or more explanatory power than a three-component model. This comes from Occam's razor argument which says the simpler model is usually the right one. Jiang and Kalyuga (2020) have provided evidence supporting a two-component model over a three-component model using subjectively rated CL surveys.

The aforementioned latest development in CLT suggests any measurement of CL should focus on differentiating ICL and ECL. Monitoring the levels of different types of CL perceived by students could help maximize the learning outcomes. However, Sweller (2010) has also argued that learners' content knowledge level could affect their ability in discerning ICL from ECL. Learners with low content knowledge level may have difficulty differentiating irrelevant information from relevant information, or productive learning process from unproductive learning process. Therefore, the knowledge level moderates students' capability to discern ICL from ECL. This certainly places even further challenges for educators since measuring different CL subtypes is already an ongoing challenge in the CLT research community (Kirschner et al., 2011). Thus, it begs researchers to design appropriate measurement tools and identify the condition for the appropriate use of the tools.

There are two widely used approaches toward measuring CL: subjective (e.g., self-report) and objective (e.g., tests and physiological measures). Subjective measures of Sweller's three subtypes of CL have been more extensively explored (Hart and Staveland, 1988; Paas, 1992; Kalyuga et al., 1998; Gerjets et al., 2004, 2006; Ayres, 2006; Leppink et al., 2013). Subjective measures typically require participants to evaluate their own cognitive processes during a learning task. Thus, they rely on the participant's ability to introspect on their learning experience. A great deal of work has gone into developing such Likert scale style subjective measures of CL. In these studies, researchers generally manipulated ICL in terms of adjusting the amount of information presented to students, such as modular vs. molar solutions comparison (Gerjets et al., 2004, 2006), changing the number of arithmetic operations (Ayres, 2006), or adjusting the learning material complexity (Windell and Wieber, 2007). These studies have shown that subjective measures could discern different levels of ICL using a difficulty rating (Windell and Wieber, 2007; Cierniak et al., 2009), a mental effort rating (Ayres, 2006), sub-items of NASA-TLX, such as stress, devoted effort, task demands (Gerjets et al., 2004, 2006), and mental demands (Windell and Wieber, 2007). The authors manipulated ECL in terms of a split-attention effect (Kalyuga et al., 1998; Windell and Wieber, 2007), and a modality effect (Windell and Wieber, 2007). They showed that ECL can be measured using a weighted workload of NASA-TLX (Windell and Wieber, 2007), a difficulty rating (Kalyuga et al., 1998), and a rating of difficulty of interaction with the material (Cierniak et al., 2009). According to Sweller (2010), GCL is affected only by the learner's motivation. Some researchers manipulated GCL/motivation through, instructional format (Gerjets et al., 2006). These studies showed GCL can be measured using, sub-items of NASA-TLX, such as task demands, effort, and navigational demands (Gerjets et al., 2006), or multiple survey items evaluating learning performance (Leppink et al., 2013).

In this work, we developed and validated a subjective survey for assessing the CL experienced by learners taking a conceptual physics test. The survey was adapted from the CL survey developed by Leppink et al. (2013). The motivation for developing this survey is that the survey developed by Leppink et al. (2013) and many previous subjective surveys were designed to measure the CL during instructional activities. In educational psychology, it has been suggested that quizzes and tests can be used as learning practice for learners (Roediger and Karpicke, 2006; Karpicke et al., 2014). This is contradictory to the traditional view that tests can only be used as summative evaluation of learner performance. Many different pedagogical

methods require instructors to create problems of their own (Mazur, 2013). Instructors need to create different testing tasks if they want to use frequent testing as learning practice for students to construct knowledge. Thus, it is important to make sure the tasks on the test are optimally designed. For example, the task should not use confusing language in the statements. On the other hand, many different problem tasks purposefully provide more information than needed to solve the problem, such as context rich problem (Ogilvie, 2009). It is possible students will process the unnecessary information if they do not have the relevant knowledge, or they will report the statement of a problem task is confusing even if the statement is perfectly clear to an expert. A CL survey that can be used to measure the three types of CL could inform instructors if the tasks created used clear language and provide instructors feedback about how students process unnecessary information.

In this study, we adopt an argument-based approach to describe the validation of the CL survey we developed (Kane, 2013). During the development and validation of the CL survey, we payed attention to both reliability and validity. Reliability refers to the consistency of the items designed to measure the same theoretical attribute, and validity refers to the appropriateness of interpreting the subjective ratings on the survey.

We conducted three experiments to validate the CL survey. Together these three experiments contributed to establishing the validity of the CL survey and the condition under which it should be administered. The studies involving human participants were reviewed and approved by the IRB office at Purdue University. The participants provided their written informed consent to participate in this study.

EXPERIMENTS

Experiment One

Materials and Procedure

We adapted the first six items on the survey used by Leppink et al. (2013), and adapted two other items targeting GCL based on previous literature (Paas, 1992; Salomon, 1984; see **Table 1**). Each item was rated on a Likert scale from 1 (not at all the case) to 9 (completely the case).

TABLE 1 | A mapping between all items of the cognitive load survey and what they are constructed to measure.

ICL	1. The topics covered on the physics test were very complex.
	2. The physics test covered formulas that I perceived to be very complex.
	3. The test covered concepts and definitions that I perceived as very complex.
ECL	4. The questions on the physics test had confusing language that was not clear to me.
	5. It was very hard to identify what information is relevant to answering the questions on the physics test.
	6. There was a lot of distracting information in the question statements on the physics test.
GCL	7. I concentrated a lot as I answered the questions on the physics test.
	8. I devoted a lot of mental effort in finding and applying the relevant concepts needed to answer the questions on the physics test.

In experiment one our goal was to establish the construct validity of the survey to verify whether the survey items indeed measure the three different cognitive sub-loads that they were intended to measure. We asked a different group of participants (N = 45) to categorize the items on the CL survey into three groups based on the common theme. A subgroup of the participants (N = 13) were interviewed about how they perceived the items on the survey. Follow-up questions were asked during the interview process. Participants were asked to provide their reasoning for the way in which they grouped the items and were asked to express the similarities and differences among the items grouped together.

Results of Experiment One

In experiment one, we asked N = 45 participants to sort the eight items on the CL survey into three groups based on a common theme. Participants were offered extra credit equal to 1% of their total course grade for their participation. The items were presented to the participants in a randomized order. Twenty-nine of 45 participants (64%) sorted the items as expected (Group A: items 1, 2, 3; Group B: items 4, 5, 6; Group C: items 7, 8). Nine of 45 participants (20%) misplaced one or two items different from the expected grouping. Seven of 45 participants (16%) grouped the items following no apparent pattern.

Thirteen of the 29 participants were randomly selected for a follow-up interview after they completed the sorting task. Each participant was asked to provide the similarities and differences between the items of each group that they had created. Participants' responses were audio recorded and coded. Ten (10) of the 13 participants grouped the survey statements in a way that was consistent with the CLT model. These interviews allowed us to detect participants' perception of the meaning of these statements. The discussion with participants' and their descriptions about each group are described below. The second author (JM) conducted the interview for experiment one since he was not associated with the course in any way at the time.

Discussion

Participants commonly responded that they grouped ICL items together because they were dealing with complexity. When confronted with questions about what makes a problem complex, participants often responded that a problem with several different elements was complex. One student remarked "circuit problems incorporate a lot of different elements. There is an element of knowing how bulbs in series function, but there also bulbs in parallel, and there is also a switch." It was also commonly reported that not having a familiarity with the ideas contained in a question makes it complex. To that end, one student said, "For this question about the power delivered to these circuits you have to know the definition of power and resistors and understand how circuits work." Here students' argument clearly aligns well with how CLT defines ICL in terms of element interactivity which is reflected in the complexity (Sweller, 2010).

When participants were asked what differentiates these statements in this group from each other, they often remarked

that the statements spoke to differing levels of organization. One student said "some of the items are about topics and the others are about sub-topics." Statement one referred to topics, while statements two and three mentioned formulas and concepts, respectively. Participants generally agreed that topics are more general than concepts which is broader than formulas.

The common theme expressed by the participants about the ECL items was confusion. "These statements have to do with ambiguity, distraction, and confusion while taking a test and those things go together." Some participants also related these items to the statement of the questions in the test, saying things like "These statements deal with the language of the question rather than the content," while another said "These were within the question itself. Like language and extra information, the wording of the test." Students' understanding of the items seem to be consistent with what CLT describes ECL in terms of its detrimental impact for learning since they are mostly distracting to learning (Sweller, 1988, 2010).

When participants were asked what differentiates the statements in this group from each other, they replied that the statements were talking about different ways in which things can be confusing. For example, "These statements are different because (5) is about finding the relevant information (6) is about having too much information, and (4) is about the question being asked in a confusing way." Participants were asked what makes confusing language different from distracting information. To which one replied "confusing information is related to things I don't understand, where distracting information is related to having more information than I need to solve the problem," while another observed "Language has a lot to do with the words you are using, distracting information deals with words and sentences that have no purpose." Further probing, many respondents were asked what makes a test question confusing. Participants commented that questions they didn't know how to answer were confusing. For example, "This question about electric field is confusing because I don't know much about electric field." This last response seems to indicate the possibility of conflating the reasons for grouping items as ICL or ECL.

Participants grouped the GCL items on the survey as thematically similar often reflected that they were related to one's subjective experience of the test instead of difficult content or confusing language. One student stated, "These items covered what you felt toward the exam, not objective difficulty but more like your personal experience" while another indicated "These items dealt less with the test itself and more with the test taker, more about concentration and mental effort, having to put more thought into answering the questions instead of difficult concepts or the wording of questions." Another participant said "Both items had to do with thinking. This one (7) was concentrating a lot, and this one (8) was exerting a lot of mental effort to figure out what was needed." Here what the students reported were consistent with the definition of GCL as proposed by Sweller that GCL has to do with the personal experience of students, especially how much cognitive resources were devoted to learning as reflected by the words: "concentration" and "mental effort" (Salomon, 1984; Paas, 1992; Sweller, 2010).

When asked about what individuated these statements, participants often designated the difference in the concepts "mental effort" and "concentration." In a series of subsequent questions most participants were unable to definitively make a distinction between the two constructs, with responses such as, "Concentration is more like focus. Mental effort is more like thinking about the information you already know to find the answer." In order to further probe their interpretation of these statements the interviewer asked participants to reflect on "concentration" and "mental effort" in terms of the steps to for solving a physics problem. Participants generally stated that reading and taking in the information of the problem statement required concentration while building a model of the problem and formulating a solution required mental effort.

In summary, results of experiment one indicate that participants were remarkably astute not just in grouping the statements, but also in articulating their criteria for the groupings. Further they were able to describe with clarity how the various statements within each group were similar to and different from each other. These results support the face validity and construct validity of the items on the survey.

Experiment Two

Materials

In experiment two, our goal was to determine whether the validity of survey i.e., its alignment with the three-component model of CLT was conditional to the level of content knowledge of the participants. We know from literature that knowledge level can influence the perceived CL since low knowledge learners may not differentiate ICL from ECL, while high knowledge learners can distinguish between these two constructs (Sweller, 1994, 2010). Based on this, we hypothesize that the items used for assessing the ICL and ECL will load to the same component when they have low knowledge, and the items used for assessing ICL and ECL will load to two separate components when they have high knowledge level. We conducted a principal component analysis to confirm if different items on the survey aligned with the three different CL subtypes.

Procedure

N = 139 participants enrolled in a physics class for elementary education majors participated in the study. In a pre/post-test design, we asked participants to complete the DIRECT (Determining and Interpreting Resistive Electric Circuit Concepts Test) assessment at the beginning and at the end of an instructional unit on DC electric circuits. DIRECT was developed by Engelhardt and Beichner (2004) for assessing conceptual understanding of circuits. DIRECT has 29 multiple choice items on it. There is only one correct answer to each item. It usually took students ∼30 min to complete. After each test (pre- and post-test), we administered this CL survey individually. Presumably, participants had low knowledge level of the relevant concepts at the beginning of the instructional unit (as confirmed by their performance on the test); and they had higher knowledge level of the relevant material by the end of the instructional unit (again, confirmed by their performance on the test).

To validate the CL survey that has three underlying components, a principal component analysis (PCA) was conducted for the two sets (pre and post) of data using IBM SPSS version 24 software. PCA analysis was used since we are exploring how many components the 8 items of the CL survey would load on to. As such, PCA is a proper analysis. The results of the PCA are shown in **Table 2** and **Table 3**.

Results of Experiment Two

For the CL survey results collected after participants had completed the DIRECT pre-test, there were no outliers or extreme skewness or kurtosis, as well as sufficient inter-item correlation; KMO (Kaiser-Meyer-Olkin test) = 0.839, Bartlett's $\chi^2_{(28)} = 481.779$, $p < 0.001$. A high value of KMO indicates the data is suitable for a factor analysis; Bartlett's test of sphericity indicates the data is suitable for a factor analysis when the test achieves significance level (0.05). Both KMO and Bartlett tests indicated the data collected were fit for a PCA analysis. KMO is measure of sampling adequacy. Given the small sample size, a PCA was conducted. Varimax rotation was performed to investigate the correlational nature of the underlying components. When, eigenvalue one was used as criteria for determining the number of underlying components, a two-component model emerged with 68% of the variation explained by the two components. Items 1–6 are loaded to the first component and items 7–8 are loaded to the second component. This seems to support the idea that when participants do not have high knowledge level, they could not differentiate ICL from ECL as suggested by CLT (Sweller, 2010). Reliability analysis for the six components loaded to the same construct revealed Cronbach's alpha values of 0.874 for Items 1, 2, 3, 4, 5, 6 (1, 2, 3 are expected to measure ICL; 4, 5, 6 are expected to measure ECL); and 0.782 for Items 7, 8 (expected to measure GCL).

TABLE 2 | Means (SD), skewness, kurtosis, and components loadings for study one (pre-test).

Item	Mean (SD)	Skewness	Kurtosis	Component loading	
				C1	C2
Component one					
Item 1	5.67(1.77)	–0.292	–0.507	0.755	0.288
Item 2	4.92(2.19)	–0.153	–0.873	0.766	0.279
Item 3	5.79(1.88)	–0.338	–0.789	0.804	0.275
Item 4	5.14(2.02)	–0.172	–0.565	0.761	0.076
Item 5	4.89(2.16)	–0.140	–0.901	0.786	0.120
Item 6	3.46(1.81)	0.477	–0.528	0.748	–0.149
Component two					
Item 7	6.41(1.74)	–0.567	–0.324	0.076	0.909
Item 8	5.50(1.96)	–0.315	–0.644	0.182	0.856

TABLE 3 | Means (SD), skewness, kurtosis, and components loadings for study one (post-test).

Item	Mean (SD)	Skewness	Kurtosis	Component loading		
				C1	C2	C3
Component one						
Item 1	4.35(1.82)	0.176	–0.657	0.643	0.343	0.386
Item 2	3.45(1.70)	0.346	–0.776	0.842	0.282	0.041
Item 3	3.39(1.70)	0.772	0.206	0.874	0.037	0.125
Component two						
Item 4	3.44(1.71)	0.849	0.478	0.514	0.541	0.148
Item 5	4.26(2.16)	0.405	–0.751	0.008	0.910	0.116
Item 6	3.40(1.96)	0.851	0.010	0.321	0.819	0.023
Component three						
Item 7	5.98(1.89)	–0.436	–0.777	–0.008	0.208	0.910
Item 8	4.82(2.07)	–0.106	–1.117	0.497	–0.092	0.742

For the CL survey results collected after participants had completed the DIRECT post-test, there were no outliers or extreme skewness or kurtosis, as well as sufficient inter-item correlation; KMO = 0.708, Bartlett's $\chi^2_{(28)} = 496.201$, $p < 0.001$. Both KMO and Bartlett tests indicated the data collected were fit for a PCA analysis. Given the small sample size, a PCA was conducted. Varimax rotation was performed to investigate the correlational nature of the underlying components. When, eigenvalue 1 was used as criteria for determining the number of underlying components, a three-component model emerged with 77% of the variation explained by the three-components. Items 1, 2, and 3 are loaded to the first component; items 4, 5, and 6 are loaded to the second component; items 7, and 8 are loaded to the third component. This provides evidence that when participants have high knowledge level (as on the post-test), they could differentiate ICL from ECL as suggested by CLT (Sweller, 2010). Reliability analysis for the three-components revealed Cronbach's alpha values of 0.816 for Items 1, 2, 3 (1, 2, 3 are expected to measure ICL); and 0.763 for 4, 5, 6 (4, 5, 6 are expected to measure ECL); and 0.687 for Items 7, 8 (expected to measure GCL).

Discussion

In summary, results of experiment two indicate that, in a testing environment, the CL survey we developed could capture the three kinds of CL subtypes. Results confirmed that students' capability of differentiating ICL from ECL depends on their content knowledge level. Thus, when students have relatively low content knowledge, they fail to differentiate ICL from ECL, which suggests even information related to learning can be confusing to students. This is consistent with what Sweller (2010) has suggested that content knowledge level could moderate how students self-perceive their CL. In addition, the items on our CL survey seem to be able to capture GCL regardless of the content knowledge level of students, which suggests students could introspect their level of effort for making sense and applying knowledge.

Experiment Three

Materials

In this experiment, we developed two pairs of tasks with clear manipulations using a 2 (High/Low ECL) × 2 (High/Low ICL) design (see **Figure 1**) based on Sweller (2010) and the redundancy effect of multimedia learning theory (Mayer, 2014).

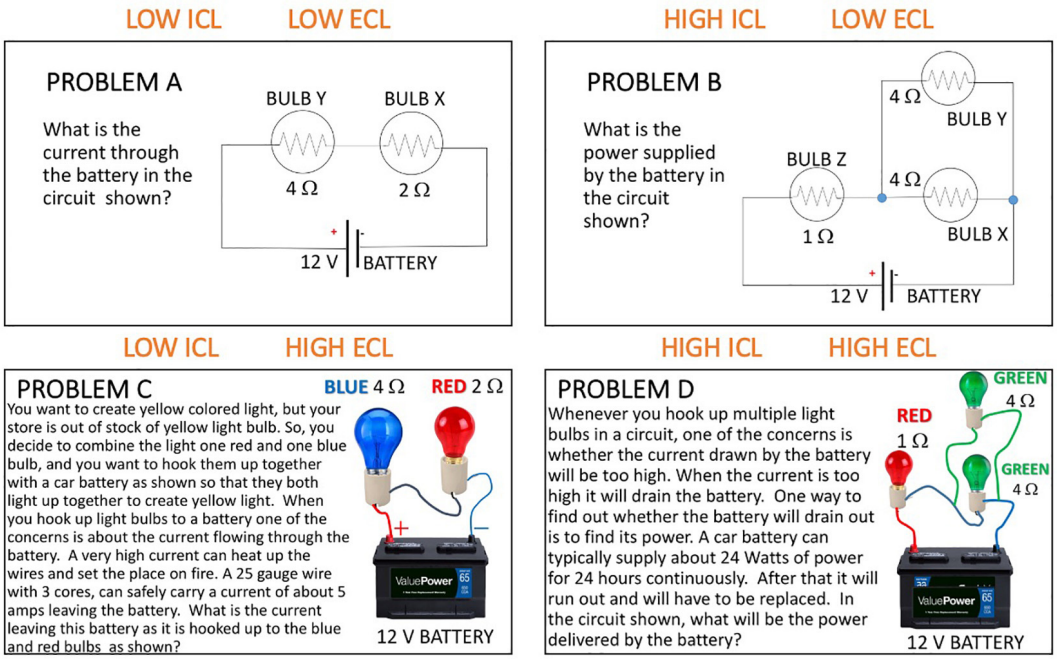

FIGURE 1 | Problems solved and compared by participants and their respective ICL and ECL levels.

The ICL level is the same (low) for both problems A and C because the underlying principle is series circuit law. Problem C has a higher ECL compared to A since it also presents redundant textual information. Problems B and D have higher ICL compared to problems A and C, since the underlying physics principle for both these problems (B and D) is a combination of series and parallel circuit laws. Compared to problem B, problem D has a higher ECL since it presents redundant textual information as well.

Procedure

A group (different from Experiments 1 and 2) of $N = 40$ elementary education majors participated in this study. We first asked participants to solve the four problem tasks on a short physics quiz. After participants completed the quiz, they were shown three pairs of problems, that they had just solved juxtaposed with each other: Pair A–B, where both problems were manipulated to impose low ECL but different ICL (A < B); Pair A–C, where both problems were manipulated to impose low ICL, but different ECL (A < C); and Pair A–D, where the problems were manipulated to impose different ICL (A < D) and ECL (A < D).

Although there were six potential problem pairs, we chose these three pairs because they would allow us to probe the extent to which participants were able to discern differences in both ICL and ECL when only one of those two had been manipulated to be different (as in pairs A–B and A–C), and one in which both had been manipulated to be different (pair A–D).

Each of the eight items on our CL survey were presented to the participants, and then they were asked to answer a question like: "If you are asked to rate: "The topics covered on the physics question were very complex" on a scale from 1 (Not at all the case) to 9 (Completely the case)", select from one of three options: (i) I would rate A is higher than C on the scale; (ii) I would rate A is the same as C on the scale; (iii) I would rate A is lower than C on the scale". An example is shown in **Figure 2**.

Results

We collapsed the answers to the items on the survey targeting the same underlying construct and calculated the percentage of participants selecting each of the three options. The results can be found in **Tables 4–6**.

For problem pair (A, B) comparison, 56% of participants rated A and B as having the same ICL. Eighty-six (86%) of participants rate A and B have the same ECL, and 76% of participants rated that they devoted the same total energy when solving A and B. This means that participants did not perceive the combination of series and parallel circuits (Problem B) as more complicated than the simple series circuit (Problem A). They perceived no distracting information for both A and B. They perceived the same level of germane load for solving these two problems. Notice the percentage for GCL is not close to the percentage of ICL suggesting that GCL and ICL can be differentiated by the measurement.

For problem pair (A, C) comparison, 54% of participants rated A and C have the same ICL. 74% of participants rated A had less ECL than B. Sixty-three (63%) of participants rated the same level of germane load when solving A and C. This means that participants did realize that the A and C are investigating the same underlying principles. They also perceived more distracting information for A than C. Again, the percentage for GCL is not

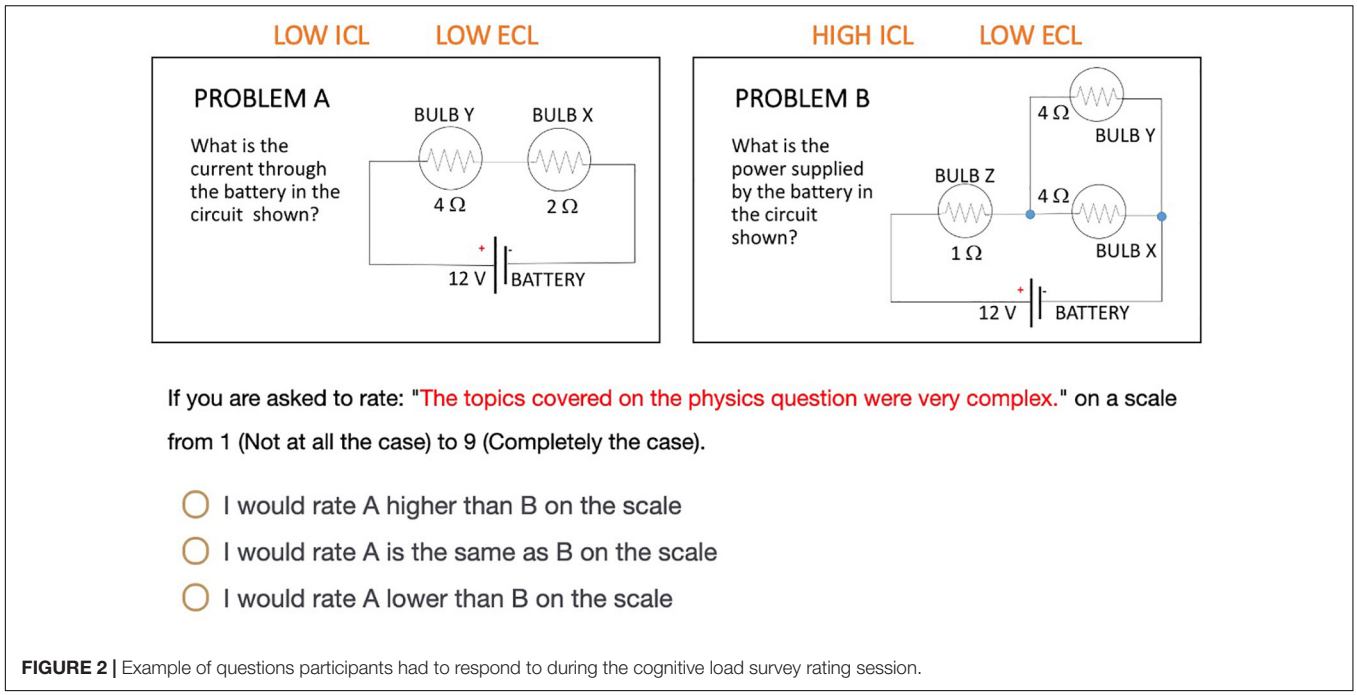

FIGURE 2 | Example of questions participants had to respond to during the cognitive load survey rating session.

the close to the percentage of ICL suggesting that GCL and ICL can be differentiated by measurement.

For problem pair (A,D) comparison, most (56%) of participants rated A had lower ICL than D. Most (70%) of participants rated A had lower ECL than D, and most (50%) of participants rated the same germane load when solving A and D. This means that participants did realize that combination of series circuit and parallel law is more complicated than the simple series circuit law. They also perceived more distracting information for A than D. Again, the ratings for "A is the same as D" is different for items corresponding to ICL than items corresponding to GCL suggesting that ICL and GCL might be differentiated on a subjective survey.

TABLE 4 | Percentage of each option selected by participants for each aspect of cognitive load for (A,B) comparison.

	% (ICL)	% (ECL)	% (GCL)
A is lower than B	39%	12%	21%
A is the same as B	56%	86%	76%
A is higher than B	5%	2%	3%

TABLE 5 | Percentage of each option selected by participants for each aspect of cognitive load for (A,C) comparison.

	% (ICL)	% (ECL)	% (GCL)
A is lower than C	44%	74%	37%
A is the same as C	54%	23%	63%
A is higher than C	2%	3%	0%

TABLE 6 | Percentage of each option selected by participants for each aspect of cognitive load for (A,D) comparison.

	% (ICL)	% (ECL)	% (GCL)
A is lower than D	56%	70%	42%
A is the same as D	39%	28%	50%
A is higher than D	5%	2%	8%

Discussion

In general, experiment three results are consistent the three-component model of CLT (Sweller, 2010). Specifically, we found that for all problem pair comparisons, participants rated the GCL differently than the ICL, which is consistent with the notion that GCL and ICL should be independent constructs, even though GCL may not provide an independent source of CL as suggested by Sweller (2010) and Kalyuga (2011). In problem pairs with the same ECL level (e.g., A–B), indeed most participants (86%), as expected rated the ECL of the two problems to be the same. In problem pairs with different ECL levels (e.g., A–C and A–D), most participants rated C (74%) and D (70%) as imposing a higher ECL than A as expected. In problem pair (e.g., A–B), where both problems were manipulated to impose low ECL but different ICL levels (e.g., A < B), most participants (56%) rated the ICL level of A and B to be the same. However, in a problem pair (e.g., A–D) where problems were manipulated to impose different ECL (A < D) as well as different ICL (A < D), participants indeed rated A as imposing lower ICL than D. This result demonstrates participants' perceived differences in ICL might depend on the levels of ECL. When ECL is high, they might have confused ECL with ICL as in the A–D comparison case. Given that these participants were students in an introductory physics class for elementary education majors and were unfamiliar with material before being exposed to it in the class, we can assume that they were low prior knowledge

students. Therefore, this result seems to be consistent with the notion that participants may not be able to differentiate ICL from ECL when they have low prior knowledge. This result is consistent with the results for experiment two in this study as well as the theoretical formalism from Sweller (2010), and certainly calls for more research to further understand low prior knowledge learners' ability to distinguish between changes in ICL and ECL.

GENERAL DISCUSSION

In this study, we developed an eight-item cognitive load (CL) survey measuring intrinsic load, extraneous load, and germane load of participants while taking a multiple-choice conceptual physics test. We conducted three experiments to validate the survey.

In the first experiment participants were asked to sort the items into groups according to the common theme. A vast majority of the participants sorted the items consistent with the CLT formalism. Namely, participants grouped items relevant to ICL together, items relevant to ECL together, and items relevant to GCL together. In follow-up interviews we found evidence that participants understand the items on the survey consistent with how ICL, ECL, and GCL were theorized in CLT (Sweller, 2010).

In the second experiment, we administered the CL survey both at the beginning and at the end of an instructional unit on electric circuits in a conceptual physics class for elementary education majors. A PCA revealed a two-component model when knowledge level was low (beginning of the unit) confirming what Sweller (2010) has proposed that low knowledge level participants might not differentiate relevant information from irrelevant information. All the items relevant to ICL and ECL loaded onto one component and all the items relevant to GCL loaded onto another component. A PCA revealed a three-component model when knowledge level was high (end of the teaching unit) confirming what Sweller (2010) has proposed that participants can differentiate relevant and irrelevant information at a high knowledge level. All the items relevant to ICL loaded onto one component; all the items relevant to ECL were loaded onto a second component; and all items relevant to GCL were loaded onto yet another component. This seems to indicate that this survey is better able to distinguish between ICL and ECL on a post-test rather than on a pretest.

In the third experiment, we asked participants to solve four physics problems of varying levels of ICL and ECL. Two of the four problems had the same ICL (low level), the other pair had the same ICL (high level). For each pair of problems with same level of ICL, one had low ECL, and the other had high ECL. After having solved the four problems, we asked participants to compare how they would rate the eight items on the CL survey differently when comparing selected pairs of problems. The results showed that most participants selected the option that they devoted the same amount of GCL to both problems in each of the compared pairs of problems. However, their ratings were not the same as on the items corresponding to ICL suggesting GCL and ICL can be measured separately, contrary to what the theoretical construct suggests (Kalyuga, 2011; Jiang and Kalyuga, 2020). When they were asked to compare a problem of high ECL with a problem of low ECL, they rated ECL as expected. When they were asked to compare a pair of problems of the same ICL, they rated the ICL as expected. When they were asked to compare two problems—one of high ICL with one of low ICL—when both problems had a low ECL, they rated the problems having the same ICL. However, when asked to compare two problems—one of high ICL and ECL with another of low ICL and ECL, they rated the problems as having different levels of ICL (A < D) which might indicate how participants perceive ICL depends on the existence of ECL. When ECL is high, they might have confused ECL with ICL. This seems to be consistent with the idea that participants may not be able to differentiate ICL from ECL when they have low prior knowledge as shown by experiment one in this study as well as the theoretical formalism from Sweller (2010). Overall, the results of the three experiments taken together provide clear evidence supporting the classic theoretical construct of CLT, i.e., a three-component construct (Sweller, 2010). These results also provide clear validation of the CL survey items.

Cognitive load theory proposes a multi-faceted construct of CL. Given the significance of CL in learning and instruction, the measurement of sub aspects of the load is important. This work will be beneficial to assessment designers who are interested in attending to the issues of CL in the design of assessment instruments. Our work adds to the existing literature by developing and adapting a subjective survey for measuring three aspects of CL.

Prior studies have not looked at if their CL survey/items were stable over the progression of students' learning. This is an overlooked area in the CLT community. This work offers evidence supporting what Sweller (2010) has argued for a long time that students' capability of differentiating ICL from ECL depends on their knowledge level. This is a challenge for CLT community if we want to measure the three types of CL reliably, we have to take the knowledge level of students into consideration. As for the proper use of the CL survey developed in this work, we suggest using it when students have developed certain level of knowledge. In terms of instruction, when instructors design questions, it usually happens post-instruction when students have already constructed a certain level of knowledge which is a good time for using the survey.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by IRB office at Purdue University. The

patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

TZ conceptualized and drafted the initial article and incorporated edits from coauthors. JM conducted interviews for experiment two and drafted the section in the initial article. JM and NR offered suggestions for terms and assisted with editing. NR reviewed and edited the manuscript. All authors read and approved the final manuscript for submission.

FUNDING

This work was supported in part by the U.S. National Science Foundation Grant No. 1348857.

REFERENCES

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

Baddeley, A. (1992). Working memory. *Science* 255, 556–559. doi: 10.1126/science.1736359

Cierniak, G., Scheiter, K., and Gerjets, P. (2009). Explaining the split-attention effect: is the reduction of extraneous cognitive load accompanied by an increase in germane cognitive load? *Comput. Hum. Behav.* 25, 315–324. doi: 10.1016/j.chb.2008.12.020

Cowan, N. (2001). Metatheory of storage capacity limits. *Behav. Brain Sci.* 24, 154–176.

Engelhardt, P. V., and Beichner, R. J. (2004). Participants' understanding of direct current resistive electrical circuits. *Am. J. Phys.* 72, 98–115. doi: 10.1119/1.1614813

Gerjets, P., Scheiter, K., and Catrambone, R. (2004). Designing instructional examples to reduce intrinsic cognitive load: molar versus modular presentation of solution procedures. *Instr. Sci.* 32, 33–58. doi: 10.1023/B:TRUC.0000021809.10236.71

Gerjets, P., Scheiter, K., and Catrambone, R. (2006). Can learning from molar and modular worked examples be enhanced by providing instructional explanations and prompting self-explanations? *Learn. Instr.* 16, 104–121. doi: 10.1016/j.learninstruc.2006.02.007

Hart, S. G., and Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. *Adv. Psychol.* 52, 139–183. doi: 10.1016/s0166-4115(08)62386-9

Jiang, D., and Kalyuga, S. (2020). Confirmatory factor analysis of cognitive load ratings supports a two-factor model. *Tutor. Quant. Methods Psychol.* 16, 216–225. doi: 10.20982/tqmp.16.3.p216

Kalyuga, S. (2011). Cognitive load theory: how many types of load does it really need? *Educ. Psychol. Rev.* 23, 1–19. doi: 10.1007/s10648-010-9150-7

Kalyuga, S., Chandler, P., and Sweller, J. (1998). Levels of expertise and instructional design. *Hum. Factors* 40, 1–17. doi: 10.1518/001872098779480587

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *J. Educ. Meas.* 50, 1–73. doi: 10.1111/jedm.12000

Karpicke, J. D., Lehman, M., and Aue, W. R. (2014). "Retrieval-based learning: an episodic context account", in *Psychology of Learning and Motivation* Vol. 61, ed. B. H. Ross (San Diego, CA: Academic Press), 237–284.

Kirschner, P. A., Ayres, P., and Chandler, P. (2011). Contemporary cogni- tive load theory research: the good, the bad and the ugly. *Comput. Hum. Behav.* 27, 99–105. doi: 10.1016/j.chb.2010.06.025

Leppink, J., Paas, F., Van der Vleuten, C. P., Van Gog, T., and Van Merriënboer, J. J. (2013). Development of an instrument for measuring different types of cognitive load. *Behav. Res. Methods* 45, 1058–1072. doi: 10.3758/s13428-013-0334-1

Mayer, R. E. (2014). Incorporating motivation into multimedia learning. *Learn. Instr.* 29, 171–173. doi: 10.1016/j.learninstruc.2013.04.003

Mazur, E. (2013). *Peer Instruction*. London: Pearson Education.

Paas, F. G. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: a cognitive-load approach. *J. Educ. Psychol.* 84:429. doi: 10.1037/0022-0663.84.4.429

Ogilvie, C. A. (2009). Changes in students' problem-solving strategies in a course that includes context-rich, multifaceted problems. *Phys. Rev. Phys. Educ. Res.* 5:020102. doi: 10.1103/PhysRevSTPER.5.020102

Roediger III, H. L., and Karpicke, J. D. (2006). Test-enhanced learning: taking memory tests improves long-term retention. *Psychol. Sci.* 17, 249–255. doi: 10.1111/j.1467-9280.2006.01693.x

Salomon, G. (1984). Television is "easy" and print is "tough": the differential investment of mental effort in learning as a function of perceptions and attributions. *J. Educ. Psychol.* 76, 647–658. doi: 10.1037/0022-0663.76.4.647

Sweller, J. (1988). Cognitive load during problem solving: effects on learning. *Cogn. Sci.* 12, 257–285. doi: 10.1207/s15516709cog1202_4

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learn. Instr.* 4, 295–312. doi: 10.1016/0959-4752(94)90003-5

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

Sweller, J., Ayres, P., and Kalyuga, S. (2011). *Cognitive Load Theory*. New York, NY: Springer.

Sweller, J., Van Merrienboer, J. J., and Paas, F. G. (1998). Cognitive architecture and instructional design. *Educ. Psychol. Rev.* 10, 251–296.

van Merriënboer, J., and Sweller, J. (2005). Cognitive load theory and complex learning: recent development and future directions. *Educ. Psychol. Rev.* 17, 147–177. doi: 10.2307/23363899

Windell, D., and Wieber, E. N. (2007). "Measur-ing cognitive load in multimedia instruction: A comparison of two instruments," in *Paper Presented at American Educational Research Association Annual Conference* (Chicago, IL).

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

PERSPECTIVE
published: 20 May 2021
doi: 10.3389/fpsyg.2021.648586

A Current View on Dual-Task Paradigms and Their Limitations to Capture Cognitive Load

*Shirin Esmaeili Bijarsari**

Department of Educational Psychology, Institute of Psychology, Goethe University Frankfurt, Frankfurt, Germany

Dual-task paradigms encompass a broad range of approaches to measure cognitive load in instructional settings. As a common characteristic, an additional task is implemented alongside a learning task to capture the individual's unengaged cognitive capacities during the learning process. Measures to determine these capacities are, for instance, reaction times and interval errors on the additional task, while the performance on the learning task is to be maintained. Opposite to retrospectively applied subjective ratings, the continuous assessment within a dual-task paradigm allows to simultaneously monitor changes in the performance related to previously defined tasks. Following the Cognitive Load Theory, these changes in performance correspond to cognitive changes related to the establishment of permanently existing knowledge structures. Yet the current state of research indicates a clear lack of standardization of dual-task paradigms over study settings and task procedures. Typically, dual-task designs are adapted uniquely for each study, albeit with some similarities across different settings and task procedures. These similarities range from the type of modality to the frequency used for the additional task. This results in a lack of validity and comparability between studies due to arbitrarily chosen patterns of frequency without a sound scientific base, potentially confounding variables, or undecided adaptation potentials for future studies. In this paper, the lack of validity and comparability between dual-task settings will be presented, the current taxonomies compared and the future steps for a better standardization and implementation discussed.

Keywords: cognitive load, dual task, secondary task, measurement, validity, comparability, cognitive load measurement, taxonomy

OPEN ACCESS

Edited by:
Günter Daniel Rey,
Technische Universität Chemnitz, Germany

Reviewed by:
Paul Ayres,
University of New South Wales, Australia
Maik Beege,
Technische Universität Chemnitz, Germany

***Correspondence:**
Shirin Esmaeili Bijarsari
s.esma@psych.uni-frankfurt.de

Specialty section:
This article was submitted to Educational Psychology, a section of the journal Frontiers in Psychology

Received: *31 December 2020*
Accepted: *28 April 2021*
Published: *20 May 2021*

Citation:
Esmaeili Bijarsari S (2021) A Current View on Dual-Task Paradigms and Their Limitations to Capture Cognitive Load. Front. Psychol. 12:648586. doi: 10.3389/fpsyg.2021.648586

INTRODUCTION

Empirical studies in educational research are often accompanied by the term cognitive load and its measurement. As a construct based on the Cognitive Load Theory (Sweller et al., 1998), it is depicted to reflect the utilization of mental resources, in particular the working memory of an individual, *via* their level of exhaustion. It is assumed to vary between a higher or lower state, depending on the tasks performed, for instance, writing an essay versus reciting simple vocabulary. By identifying the parameters exhausting the mental resources, instructional settings can be adapted for a higher learning outcome. For this purpose, different methods to measure cognitive load have been developed over the years. Brünken et al. (2003) classify these methods based on their objectivity and causal relationship into four categories:

subjective-direct, subjective-indirect, objective-direct, and objective-indirect methods.

Subjective measurements can be summarized as self-reports like questionnaires (Leppink et al., 2013) to assess the perceived mental effort. It is not a method best used for continuous assessment as it is executed retrospectively (Brünken et al., 2003) and seems to be influenced in the sensitivity and accuracy of its results by the timing and frequency of its use (Chen et al., 2011; van Gog et al., 2012). Nonetheless, it is so far the only method to attempt to identify the cognitive load distinguished by its three dimensions intrinsic, extraneous, and germane load (Brünken et al., 2010; Leppink et al., 2013; Klepsch et al., 2017). In contrast, objective measurements assess the performance of the individual simultaneously to the task and vary from physiological methods like electroencephalography (Antonenko et al., 2010) or fMRI (Whelan, 2007) to dual tasks (Park and Brünken, 2018). Chen et al. (2011) found the objective measurements more lacking compared to subjective measurements, because of their lower sensitivity toward small changes in the cognitive load during a task. Brünken et al. (2003), however, emphasized the difference in accuracy between indirect and direct measurements based on the causal relation of mental effort and experienced cognitive load. In that regard, indirect measurements tend to be unreliable in their interpretation as other factors might have influenced the reported responses (Brünken et al., 2010). Objective-direct measurements like neuroimaging and dual tasks, however, relate directly to the experienced cognitive load (Brünken et al., 2003). And while neuroimaging methods like fMRI seem promising, some limitations arise by the intrusiveness of the technical device. Dual tasks, often also referred to as secondary tasks, present an objective-direct measurement in which two tasks are to be performed simultaneously to observe performance drops in either task. There are two ways to conduct dual tasks, either to induce or to assess cognitive load (Brünken et al., 2002; Klepsch et al., 2017). To induce cognitive load, the secondary task is designed to demand the mental resources needed for the primary task, for instance, by tapping or humming a melody (Park and Brünken 2015; Sun and Shea, 2016). Therefore, the performance of the primary task is affected. In contrast, the cognitive load can also be assessed by simple decision-making tasks like mathematical tasks (Lee et al., 2015; Tang et al., 2015), to observe the performance of the secondary task without influencing the primary task.

Due to these differences in objectivity and causal relation, dual tasks might be seen as an adequate alternative to assess cognitive load as a simultaneous, objective-direct measurement. However, the current state of research showcases a broad variety and heterogeneity of dual-task methods that lack standardization and continuity in their implementation. This in turn hinders the validity and comparability between studies as well as an accurate depiction of the cognitive load throughout the learning process. To further expand on this discrepancy between intent and implementation of dual tasks, this paper will discern the underlying cause of the lack of validity and comparability and present the current state on the taxonomy of dual tasks.

THE LACK OF VALIDITY AND COMPARABILITY IN DUAL-TASK SETTINGS

For a better understanding of the proclaimed issues, the validation as formulated by Kane (2013) should be consulted. He states in his argument-based approach that two steps have to be executed to ensure validity: specifying the proposed interpretation or use of the test and evaluating these claims based on appropriate evidence. The evidence is collected through four inferences that build up from a single observation in a test setting, for instance, a multiple-choice question, to the implementation of the target score as a reflection of the real-life performance. In the dual-task setting, it is comparable to question who and what the task is going to assess, which parameters encompass the proposed interpretation and use and if the determined parameters result in its successful accomplishment. However, aside a few exceptions, there is a lack of empirical investigation of secondary tasks, not only regarding their psychometric properties but also in relation to their respective dual-task settings (Watter et al., 2001; Jaeggi et al., 2010). Contrary to the assumption of validity being universal for every setting of its respective test (Kane, 2013), validity has to be examined for each new proposed interpretation and use. A similar sentiment can be found in the study of Jaeggi et al. (2010), where one of the more common secondary tasks, the n-back task, was examined on its validity. The mixed results showed not only difficulty in confirming its validity but also a further need for implementation and examination in different settings.

Another issue arises in the form of lacking comparability between the different dual-task studies. Currently, most dual tasks are custom-made for their specific instructional setting, without any reference to an evaluated and standardized method. Most often, the decision behind the choice of a dual-task method is not further discussed, which in turn might hinder future researchers in continuing or implementing these studies. The different types of dual task not only lack a framework by which a fitting task can be chosen but they also ignore natural limitations in combining different tasks, for instance, a primary motoric task of walking and a secondary task of typing on a phone. This setting would result in a reduced performance of the primary task as the secondary task is naturally intrusive by limiting the field of vision (Lamberg and Muratori, 2012). Nor do they focus as much on the aspect that experience in multitasking can increase the ability to dual task (Strobach et al., 2015) or that dual tasks are great to measure progress in novices but not experts (Haji et al., 2015). Similarly, to the topic of experts, there can be confounding variables, for instance, response automatization (van Nuland and Rogers, 2016) and age, in particular dementia, influencing the participants (Toosizadeh et al., 2016; Sawami et al., 2017).

THE CURRENT TAXONOMY OF DUAL TASKS

Despite the broad heterogeneity of dual-task methods in instructional settings, one common denominator can be found.

A dual-task setting consists of two tasks: the primary task that the researcher wants to observe and the secondary task that has no connection to it beyond its competitive nature. The participant has to perform both tasks concurrently. Apart from that, most attempts at creating a systematic approach toward the variety of dual-task methods have been few and far between and lacking a holistic view.

One of the earlier taxonomies by Brown (1978) postulated four design factors to determine differences between dual-task methods: the information processing demand, the prioritized task performance, the temporal structure and the locus of interference. The first design factor focused on the demand the chosen secondary task puts onto the information processing – either by stimuli with constant or variable demands, for example, changing between easy and complex tasks, or by continuously variable and continuously constant demands not bound to specific stimuli. Another role played the priority given to the secondary task, which could be either primary, secondary, or of equal importance to the primary task. It could be compared to the priorly mentioned ways of inducing or assessing cognitive load (Brünken et al., 2002; Klepsch et al., 2017). van Nuland and Rogers (2016) further recommended the task priority to be explicitly stated in the participants' instructions, as there otherwise might be a task performance trade-off. The third design factor by Brown (1978) focused on the temporal structure of the secondary task, which was either force-paced by the experimental setting, self-paced by the participant or force-paced by the experimental setting within a specific time interval. Lastly, the locus of interference between both tasks could either be at the sensory input or motor output, within the process of the tasks or a combination of all three. He argued though that both sensory input and motor output should not be used as a locus of interference as the dual-task method intends to focus on the mental resources and therefore needs to be used during the process of the mental activity.

Another attempt at categorizing and standardizing dual tasks from a physician's viewpoint has been made by McIsaac et al. (2015). Three main categories were stated: tasks by action, task complexity, and task novelty. The category of tasks by action distinguishes between dual tasks consisting of both cognitive, both motor, and cognitive-motor or motor-cognitive primary and secondary task combinations. Therefore, the selection of the proper dual-task method does not only focus on finding a fitting secondary task contentwise but also on its execution in combination with the primary task. The second category, task complexity, is in general a relevant factor but not easy to standardize. The complexity of a task might be felt differently for someone that has never done it versus an experienced user. In this case, task novelty also plays a role as the experience influences the complexity and therefore also the measurement results (Strobach et al., 2015).

Lastly, the recent taxonomy by Wollesen et al. (2019) focused on the different task types. They distinguished between reaction time tasks, controlled processing tasks, visuospatial tasks, mental tracking tasks, working memory tasks, and discrimination tasks. The reaction time tasks were defined as tasks that rely on the reaction time between the sensory stimulus and the behavioral response, for example, pressing a button whenever a light goes on. The controlled processing task expands the reaction time task by the addition of a decision-making process, for example, pressing a button only when a specific symbol appears. The visuospatial task focuses on detecting or processing visual information, for example, finding a symbol in a rotated position. The mental tracking tasks require the memorization of information and are split into two subcategories: the arithmetic tests, for example, counting backward in 3 s (n-back tasks), and the verbal fluency, for example, naming words starting with the same letter. The working memory tasks are a simpler form of the mental tracking tasks as they only require holding information but not processing it, for example, memorizing a picture that has to be found again afterward. Lastly, the discrimination tasks focus on the selective attention toward a specific stimulus, for example, the Go/NoGo tasks in which participants have to either provide or withhold a response depending on the stimulus (Verbruggen and Logan, 2008).

Expanding on the visuospatial tasks presented by Wollesen et al. (2019), a few more modality-related classifications can be found. The method of tapping or humming melodies (Park and Brünken 2015; Sun and Shea, 2016), mathematical tasks (Lee et al., 2015; Tang et al., 2015), and visual tasks like reading text or symbols (Scerbo et al., 2017; Wirzberger et al., 2018) showcase that the modality between primary and secondary task can differ between auditory/vocally, visually, and motoric tasks. Furthermore, as mentioned by Brown (1978) and Wollesen et al. (2019), there can be differences in the frequency of the dual task, from event- or interval-based tasks that appear, for example, every 3, 5, or 7 s to continuous tasks that constantly request the participants' attention. Yet, there is not really a study to be found that uses dual tasks continuously. Most rely on either interval- or event-based frequency.

OUTLINING A HOLISTIC TAXONOMY

The three taxonomies presented lack a holistic view of the dual-task setting and tend to either simplify or strongly limit the classification. For instance, McIsaac et al. (2015) categorizes tasks by action into cognitive or motor tasks even though the description of detecting a cognitive action outside of an fMRI setting seems contradictory. The participant needs to either act motoric or verbally to respond. In contrast, the taxonomy of Wollesen et al. (2019) expands on the task action by displaying a broader variety of secondary tasks but stays limited to only this one parameter. Furthermore, simply the difference between the two dual-task types of inducing and assessing cognitive load needs to be included in a taxonomy as it changes the intent and therefore the use of it. For this purpose, an attempt at a holistic taxonomy was made (**Figure 1**).

Parameters relevant to the design of the dual-task setting were included in a stepwise order, ultimately resulting in the selection of the secondary task based on the chosen path. Most of the options are not unique at that, for instance, middle complex tasks can be event-based too. Following the yellow-colored path as an example, after selecting to induce the cognitive load, the stimulus modality and task action modality of the

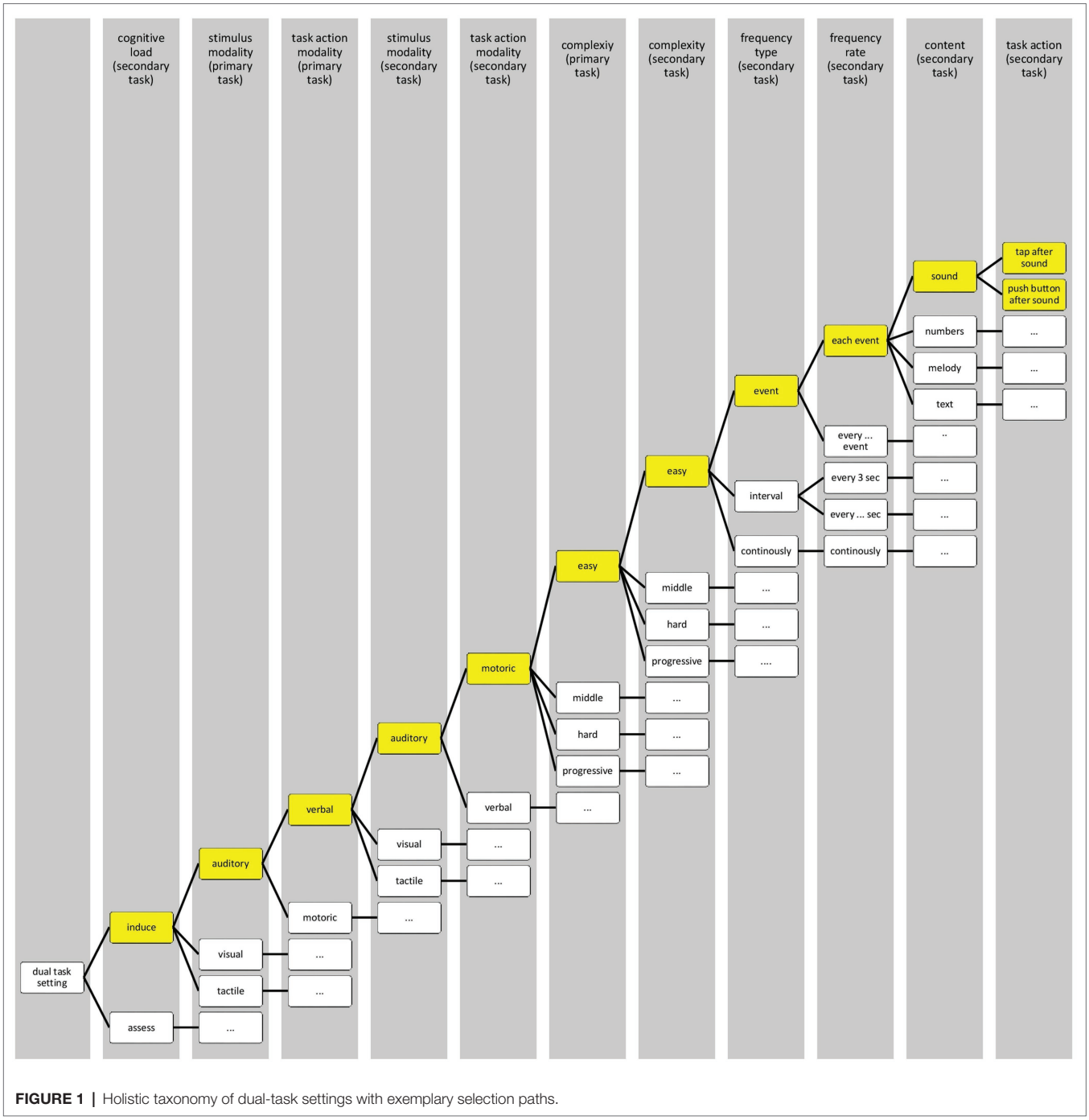

FIGURE 1 | Holistic taxonomy of dual-task settings with exemplary selection paths.

primary task have to be regarded. For instance, choosing a verbal primary task would in turn either hinder a verbal secondary task or restrict the option of higher frequency types in the subsequent parameters. These selections are followed by the complexity of both tasks, and lastly the possible frequency types, frequency rate, and content of the secondary task. Lastly, the task action should show the possible options regarding the prior selections, in this case to either tap or push a button after the sound event, as the secondary task was intended to be auditory in its stimulus but motoric in its action. However, it should be noted that the taxonomy needs to be standardized to be usable as a guide or framework in designing a dual-task setting. The variations of the parameters need to be tested and validated, which, aside from a few exceptions, has yet to be done.

DISCUSSION

So far, the classifications of the current dual-task paradigms show a mix of different factors without a theoretical framework.

Most studies lack a detailed explanation of the reasoning behind the implementation or adaptation of a secondary task, aside the general assumption of using a fitting cognitive load measurement. The presented taxonomies show a broad range of parameters but do not find a common ground. While McIsaac et al. (2015) summarize the different tasks by their action of cognitive versus motoric tasks, the complexity and the novelty of the task, Wollesen et al. (2019) go a bit further and categorize dual tasks by their execution, but with no regards to other parameters. In addition, both taxonomies need to be further specified for a profound framework, especially regarding the different modalities and frequency of dual tasks (Brown, 1978). According to the dual-coding theory (Paivio, 1971, 1991), both verbal information and nonverbal/visual information interact for a better recall, but their information is processed differently in their own channel. Therefore, there should be a higher regard toward the selection of the task modalities and their influence on the cognitive load measurement. Using the same modalities in primary and secondary tasks might contribute to a higher cognitive load measurement because the information is not already distinguished simply by its sensory input. Further influences might be found in the different temporal structure of dual tasks, in particular the frequency in which the secondary task should be used. So far, even empirical studies that describe their task as continuous, end up being high-interval tasks or tasks that cannot be done over a longer time frame because of physical exhaustion, for instance, constant humming or tapping (Park and Brünken 2015; Sun and Shea, 2016). This bears the question on how to change the lack of continuous dual tasks as this particular ability makes it a noteworthy measurement for the cognitive load. Furthermore, it not only needs to be usable over a longer period but also have more variations to be applicable in different settings. For this, it is advisable to look back at the modalities and the restrictions they contain as the physical strain and execution interfere with a continuous dual task. For example, humming a melody might influence an emotional reaction (Schellenberg et al., 2013), but also simply put a physical strain over a longer period. Visual dual tasks would be hard to be kept up in a continuous setting as it would be hard to split the focus of the eyes toward two different tasks, see split-attention effect (Ayres and Cierniak, 2012). A solution might be the use of eye-tracking to adapt the secondary task into a less intrusive method, for example, by changing colors and symbols in the background of the instructional setting to observe the eye movement. In motoric tasks, primary tasks usually cannot be physical as it tends to disturb the secondary task and heightens the physical strain. An exception can be created with physical tasks that work disconnected from each other, for example, tapping on a pedal while sitting and repairing machinery.

Conclusively, future research in relation to dual-task paradigms should take a step back in creating or expanding the different methods of dual tasks and firstly focus on creating a profound and universal taxonomy. Furthermore, the currently existing methods should be evaluated and adapted to create a standardized and reliable use. This of course needs an extensive analysis of the instructional settings and the possibilities to implement dual tasks based on pre-defined variables so that in the future researchers can more easily choose the fitting dual-task paradigms. Dual tasks should furthermore work more toward creating truly continuous tasks to ensure the direct measurement of cognitive load that it proclaims to be (Brünken et al., 2003).

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, and further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

REFERENCES

Antonenko, P., Paas, F., Grabner, R., and van Gog, T. (2010). Using electroencephalography to measure cognitive load. *Educ. Psychol. Rev.* 22, 425–438. doi: 10.1007/s10648-010-9130-y

Ayres, P., and Cierniak, G. (2012). "Split-attention effect" in *Springer Reference. Encyclopedia of the Sciences of Learning*. ed. N. M. Seel (Boston, MA: Springer US), 3172–3175.

Brown, I. D. (1978). Dual task methods of assessing work-load. *Ergonomics* 21, 221–224. doi: 10.1080/00140137808931716

Brünken, R., Plass, J. L., and Leutner, D. (2003). Direct measurement of cognitive load in multimedia learning. *Educ. Psychol.* 38, 53–61. doi: 10.1207/S15326985EP3801_7

Brünken, R., Seufert, T., and Paas, F. (2010). "Measuring cognitive load" in *Cognitive Load Theory*. eds. J. L. Plass, R. Moreno and R. Brünken (Cambridge, New York: Cambridge University Press), 181–202.

Brünken, R., Steinbacher, S., Plass, J. L., and Leutner, D. (2002). Assessment of cognitive load in multimedia learning using dual-task methodology. *Exp. Psychol.* 49, 109–119. doi: 10.1027//1618-3169.49.2.109

Chen, S., Epps, J., and Chen, F. (2011). "A comparison of four methods for cognitive load measurement," in *Proceedings of the 23rd Australian Computer-Human Interaction Conference (OzCHI 2011)*. eds. N. Colineau, C. Paris, and D. Stevenson; Australian National University, Canberra; ACM SIGCHI; November 28–December 2, 2011; ACM, 76–79.

Haji, F. A., Khan, R., Regehr, G., Drake, J., de Ribaupierre, S., and Dubrowski, A. (2015). Measuring cognitive load during simulation-based psychomotor skills training: sensitivity of secondary-task performance and subjective ratings. *Adv. Health Sci. Educ. Theory Pract.* 20, 1237–1253. doi:10.1007/s10459-015-9599-8

Jaeggi, S. M., Buschkuehl, M., Perrig, W. J., and Meier, B. (2010). The concurrent validity of the N-back task as a working memory measure. *Memory* 18, 394–412. doi: 10.1080/09658211003702171

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *J. Educ. Meas.* 50, 1–73. doi: 10.1111/jedm.12000

Klepsch, M., Schmitz, F., and Seufert, T. (2017). Development and validation of two instruments measuring intrinsic, extraneous, and germane cognitive load. *Front. Psychol.* 8:1997. doi: 10.3389/fpsyg.2017.01997

Lamberg, E. M., and Muratori, L. M. (2012). Cell phones change the way we walk. *Gait Posture* 35, 688–690. doi: 10.1016/j.gaitpost.2011.12.005

Lee, H.-I., Park, S., Lim, J., Chang, S. H., Ji, J.-H., Lee, S., et al. (2015). Influence of driver's career and secondary cognitive task on visual search behavior in driving: a dual-task paradigm. *Adv. Phys. Educ.* 5, 245–254. doi: 10.4236/ape.2015.54029

Leppink, J., Paas, F., van der Vleuten, C. P. M., van Gog, T., and van Merriënboer, J. J. G. (2013). Development of an instrument for measuring different types of cognitive load. *Behav. Res. Methods* 45, 1058–1072. doi: 10.3758/s13428-013-0334-1

McIsaac, T. L., Lamberg, E. M., and Muratori, L. M. (2015). Building a framework for a dual task taxonomy. *Biomed. Res. Int.* 2015:591475. doi: 10.1155/2015/591475

Paivio, A. (1971). *Imaginary and verbal processes*. New York: Holt, Rinehart and Winston, Inc.

Paivio, A. (1991). Dual coding theory: retrospect and current status. *Can. J. Psychol./ Revue Canadienne De Psychologie* 45, 255–287. doi: 10.1037/h0084295

Park, B., and Brünken, R. (2015). The rhythm method: A new method for measuring cognitive load-An experimental dual-task study. *Appl. Cogn. Psychol.* 29, 232–243. doi: 10.1002/acp.3100

Park, B., and Brünken, R. (2018). "Secondary task as a measure of cognitive load" in *Cognitive Load Measurement and Application: A Theoretical Framework for Meaningful Research and Practice*. ed. R. Zheng (New York, NY, US: Routledge/Taylor & Francis Group), 75–92.

Sawami, K., Katahata, Y., Suishu, C., Kamiyoshikawa, T., Fujita, E., Uraoka, M., et al. (2017). Examination on brain training method: effects of n-back task and dual-task. *F1000Research* 6:116. doi: 10.12688/f1000research.10584.1

Scerbo, M. W., Britt, R. C., and Stefanidis, D. (2017). Differences in mental workload between traditional and single-incision laparoscopic procedures measured with a secondary task. *Am. J. Surg.* 213, 244–248. doi: 10.1016/j.amjsurg.2016.09.056

Schellenberg, E. G., Weiss, W. M., (2013). "Music and cognitive abilities" in *The Psychology of Music*. *3rd Edn*. ed. D. Deutsch (London: Academic Press), 499–550.

Strobach, T., Becker, M., Schubert, T., and Kühn, S. (2015). Better dual-task processing in simultaneous interpreters. *Front. Psychol.* 6:1590. doi: 10.3389/fpsyg.2015.01590

Sun, R., and Shea, J. B. (2016). Probing attention prioritization during dual-task step initiation: a novel method. *Exp. Brain Res.* 234, 1047–1056. doi: 10.1007/s00221-015-4534-z

Sweller, J., van Merrienboer, J. J. G., and Paas, F. G. W. C. (1998). *Educ. Psychol. Rev.* 10, 251–296. doi: 10.1023/A:1022193728205

Tang, P.-F., Yang, H.-J., Peng, Y.-C., and Chen, H.-Y. (2015). Motor dual-task timed up & go test better identifies prefrailty individuals than single-task timed up & go test. *Geriatr. Gerontol. Int.* 15, 204–210. doi: 10.1111/ggi.12258

Toosizadeh, N., Najafi, B., Reiman, E. M., Mager, R. M., Veldhuizen, J. K., O'Connor, K., et al. (2016). Upper-extremity dual-task function: an innovative method to assess cognitive impairment in older adults. *Front. Aging Neurosci.* 8:167. doi: 10.3389/fnagi.2016.00167

van Gog, T., Kirschner, F., Kester, L., and Paas, F. (2012). Timing and frequency of mental effort measurement: evidence in favour of repeated measures. *Appl. Cogn. Psychol.* 26, 833–839. doi: 10.1002/acp.2883

van Nuland, S. E., and Rogers, K. A. (2016). E-learning, dual-task, and cognitive load: the anatomy of a failed experiment. *Anat. Sci. Educ.* 9, 186–196. doi: 10.1002/ase.1576

Verbruggen, F., and Logan, G. D. (2008). Response inhibition in the stop-signal paradigm. *Trends Cogn. Sci.* 12, 418–424. doi: 10.1016/j.tics.2008.07.005

Whelan, R. R. (2007). Neuroimaging of cognitive load in instructional multimedia. *Educ. Res. Rev.* 2, 1–12. doi: 10.1016/j.edurev.2006.11.001

Watter, S., Geffen, G. M., and Geffen, L. B. (2001). The n-back as a dual-task: P300 morphology under divided attention. *Psychophysiology* 38, 998–1003. doi: 10.1111/1469-8986.3860998

Wirzberger, M., Herms, R., Bijarsari, S. E., Eibl, M., and Rey, G. D. (2018). Schema-related cognitive load influences performance, speech, and physiology in a dual-task setting: a continuous multi-measure approach. *Cognit. Res.* 3:46. doi: 10.1186/s41235-018-0138-z

Wollesen, B., Wanstrath, M., van Schooten, K. S., and Delbaere, K. (2019). A taxonomy of cognitive tasks to evaluate cognitive-motor interference on spatiotemoporal gait parameters in older people: a systematic review and meta-analysis. *Eur. Rev. Aging Phys. Act.* 16:12. doi: 10.1186/s11556-019-0218-1

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 01 July 2021
doi: 10.3389/fpsyg.2021.656994

Assessing Instructional Cognitive Load in the Context of Students' Psychological Challenge and Threat Orientations: A Multi-Level Latent Profile Analysis of Students and Classrooms

Andrew J. Martin[1]*, *Paul Ginns*[2], *Emma C. Burns*[3], *Roger Kennett*[1], *Vera Munro-Smith*[4], *Rebecca J. Collie*[1] and *Joel Pearson*[1]

[1] University of New South Wales, Kensington, NSW, Australia, [2] The University of Sydney, Sydney, NSW, Australia, [3] Macquarie University, Sydney, NSW, Australia, [4] The King's School, North Parramatta, NSW, Australia

OPEN ACCESS

Edited by:
Kate M. Xu,
Open University of the Netherlands, Netherlands

Reviewed by:
Alexandre Morin,
Concordia University, Canada
André Tricot,
Epsylon Laboratory EA 4556, France

***Correspondence:**
Andrew J. Martin
andrew.martin@unsw.edu.au

Specialty section:
This article was submitted to Educational Psychology, a section of the journal Frontiers in Psychology

Received: 21 January 2021
Accepted: 27 May 2021
Published: 01 July 2021

Citation:
Martin AJ, Ginns P, Burns EC, Kennett R, Munro-Smith V, Collie RJ and Pearson J (2021) Assessing Instructional Cognitive Load in the Context of Students' Psychological Challenge and Threat Orientations: A Multi-Level Latent Profile Analysis of Students and Classrooms. Front. Psychol. 12:656994. doi: 10.3389/fpsyg.2021.656994

To better understand instructional cognitive load, it is important to operationalize and assess it in novel ways that can reveal how different students perceive and experience this load as either challenging or threatening. The present study administered a recently developed instruction assessment tool—the Load Reduction Instruction Scale-Short (LRIS-S)—to $N = 2{,}071$ students in 188 high school science classrooms. Multilevel latent profile analysis (LPA) was used to identify student and classroom profiles based on students' reports of instructional cognitive load (load reduction instruction, LRI; using the LRIS-S) and their accompanying psychological challenge orientations (self-efficacy and growth goals), and psychological threat orientations (anxiety and failure avoidance goals). In phase 1 of analyses (investigating students; Level 1), we identified 5 instructional-psychological student profiles that represented different presentations of instructional load, challenge orientation, and threat orientation, ranging from the most maladaptive profile (the Instructionally-Overburdened & Psychologically-Resigned profile) to the most adaptive profile (Instructionally-Optimized & Psychologically-Self-Assured profile). The derived profiles revealed that similar levels of perceived instructional load can be accompanied by different levels of perceived challenge and threat. For example, we identified two profiles that were both instructionally-supported but who varied in their accompanying psychological orientations. Findings also identified profiles where students were dually motivated by both challenge and threat. In turn, these profiles (and their component scores) were validated through their significant associations with persistence, disengagement, and achievement. In phase 2 of analyses (investigating students and classrooms; Levels 1 and 2), we identified 3 instructional-psychological classroom profiles that varied in instructional cognitive load, challenge orientations, and threat orientations: Striving classrooms, Thriving classrooms, and Struggling classrooms. These three classroom

profiles (and their component scores) were also validated through their significant associations with classroom-average persistence, disengagement, and achievement—with Struggling classrooms reflecting the most maladaptive outcomes and Thriving classrooms reflecting the most adaptive outcomes. Taken together, findings show that considering instructional cognitive load (and new approaches to empirically assessing it) in the context of students' accompanying psychological orientations can reveal unique insights about students' learning experiences and about important differences between classrooms in terms of the instructional load that is present.

Keywords: cognitive load, load reduction instruction, cognitive appraisal, engagement, achievement, latent profile analysis, multilevel, construct validity

INTRODUCTION

In this study, we propose that instructional cognitive load is likely to be perceived and experienced in different ways by different students. Following cognitive appraisal theory (e.g., Lazarus and Folkman, 1984; for reviews, see Roseman and Smith, 2001; Moors et al., 2013), we suggest that some students will perceive cognitive load in an approach- and challenge-oriented way, while other students will perceive cognitive load in an avoidant- and threat-oriented way. This being the case, we suggest that we can better understand instructional cognitive load when we consider it in the context of students' accompanying psychological challenge and threat orientations. The present study does so from a person-centered perspective (using latent profile analysis; LPA) based on students' reports of instructional load (load reduction instruction, LRI; using the Load Reduction Instruction Scale-Short, LRIS-S) and their accompanying psychological challenge orientations (self-efficacy and growth goals) and psychological threat orientations (anxiety and failure avoidance goals).

In addressing these issues, we adopt a *construct validation* approach. Researchers in educational psychology have long emphasized the importance of evaluating instruments within a construct validation framework (e.g., Marsh, 2002; Martin, 2007, 2009). Construct validation can be classified in terms of *within-network* and *between-network* approaches. Within-network approaches explore the internal structure of a construct or network and between-network approaches typically seek to establish a logical, theoretically meaningful pattern of relations between constructs and networks. Both approaches tend to employ variable-centered analyses such as reliability and factor analysis (for within-network validity) and correlation and regression (for between-network validity). Indeed, this study's central measurement tool (the Load Reduction Instruction Scale; LRIS) has been validated using these variable-centered within- and between-network approaches (Martin and Evans, 2018).

However, as discussed below, variable-centered approaches to construct validation can mask important phenomena among subpopulations of the wider population. Person-centered approaches, on the other hand, are well-placed for validation at a more granular subpopulation level. Therefore, the present investigation applies a person-centered approach to the assessment of within- and between-network validity. Using multilevel LPA and integrating and synthesizing cognitive load theory (Sweller, 2012) with cognitive appraisal theory (Lazarus and Folkman, 1984), we test within-network validity by identifying a network of theoretically plausible student and classroom profiles that are based on student reports of instructional load (via the LRIS) and students' accompanying psychological challenge and threat orientations. We then test between-network validity by exploring the links between the network of student and classroom profiles and a theoretically plausible network of outcome variables (persistence, disengagement, achievement). In essence, then, the present study contributes a novel multilevel construct validity approach to person-centered analysis in a bid to understand the nomological network of instructional cognitive load. **Figure 1** shows the hypothesized model.

This approach to construct validity draws on guidelines advanced by methodologists in the measurement space—very much inspired by the seminal work of Campbell and Fiske (1959). Psychological research (including educational psychology research) typically involves hypothetical constructs that are unobservable, conceptual, or theoretical abstractions (Marsh et al., 2006). Constructs are often inferred indirectly via observable indicators. A vital question, then, is how well these indicators represent the hypothetical construct, including the extent to which the construct is: well-represented by derived scores, is well-defined, and is related to variables to which it should be theoretically connected (Marsh, 2002; Marsh et al., 2006). In light of these critical questions, it is recommended that construct validity research comprises multiple perspectives based on multiple methods. According to Marsh et al. (2006), this involves the use of "multiple indicators of each construct, multiple constructs and tests of their a priori relations, multiple outcome measures, multiple independent/manipulated variables, multiple methodological approaches … the multiple perspectives provide a foundation for evaluating construct validity based on appropriate patterns of convergence and divergence and for refining measurement instruments, hypotheses, theory, and research agendas" (p. 442). Accordingly, the present construct validation study includes latent factors (comprised of multiple indicators), multiple factors and tests of their relations, numerous outcome measures, and integration of two analytical methods

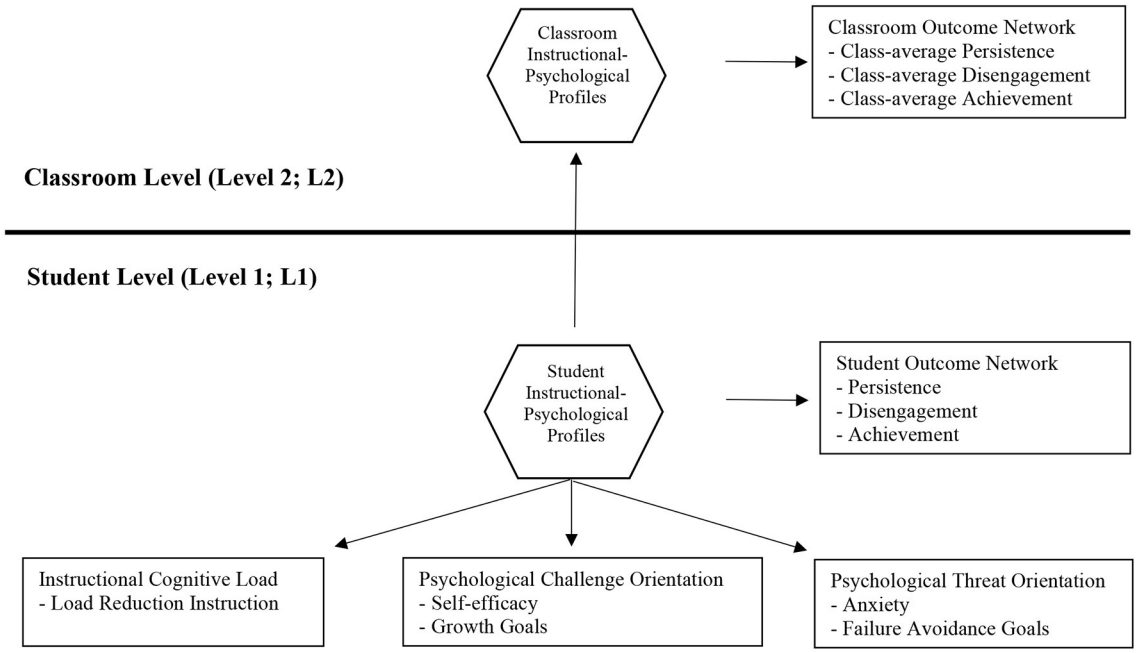

FIGURE 1 | Hypothesized models tested in the study at the student- and classroom-level.

by way of LPA and multi-level modeling. In these ways, we claim to provide a unique perspective on multilevel construct validity.

COGNITIVE LOAD THEORY

Cognitive load theory has identified principles of instruction that are aimed at easing the cognitive burden on students as they learn (Sweller et al., 2011; Sweller, 2012). According to cognitive load theory, there are two kinds of cognitive load that can be imposed during instruction and that impede learning: intrinsic and extraneous load (Sweller et al., 2011). Intrinsic cognitive load is a function of the inherent complexity of instructional activity and material, given the learner's prior knowledge. Teachers can manage intrinsic cognitive load by presenting instructional material that is appropriate to the level of knowledge of students (Sweller et al., 2011). Extraneous cognitive load emanates from the way that instructional material is structured and presented (Sweller et al., 2011). Teachers can manage this latter form of load by presenting instructional material sequentially, clearly, and explicitly to students; in doing so, students are guided through learning in a structured and linear fashion, leading to low extraneous load. However, extraneous load is high when instructional material is presented such that students need to figure out the informational structure, have to decide between a range of potential solutions, and/or apply information about which they have low prior knowledge (Sweller et al., 2011). Extraneous cognitive load is identified as an unnecessary burden on students as it does not contribute to learning (Sweller et al., 2011).

To date, the bulk of research into cognitive load theory has been experimental, with cognitive load induced through presentation and manipulation of instructional/learning material to elicit conditions of low or high cognitive load (see Sweller, 2012 for a review). There is significant research assessing cognitive load through self-reports of cognitive burden (e.g., Leppink et al., 2013; Krell, 2017), and other approaches such as through electrodermal activity, neurological activity, eye tracking, blood flow, physical pressure exerted on a computer mouse, etc. (e.g., Paas et al., 1994; Ikehara and Crosby, 2005; Wang et al., 2014; Howard et al., 2015; Ghaderyan et al., 2018). This research has assessed the presence of cognitive load, as well as the instructional techniques aimed at managing or reducing cognitive load. Much of this research has taken place under experimental conditions. This experimental work has significant internal validity and has been critical to rigorously measuring cognitive load and identifying some major effects that are now well-established in the cognitive load tradition (e.g., expertise reversal effect, modality effect, split attention effect, etc.). Alongside this experimental work it is also important to extend assessment to attend to other matters of validity—such as conducting classroom-based assessment research that has the potential to inform the ecological validity of cognitive load assessment. As one significant step in this direction, recent work has articulated a multi-factor instructional framework—load reduction instruction (LRI)—aimed at (a) reducing instructional cognitive load on learners, (b) developing a multi-factor survey instrument to assess this instructional framework, and (c)

administering and assessing this instrument in the classroom context (Martin, 2016; Martin and Evans, 2018, 2019; Martin et al., 2020a).

LOAD REDUCTION INSTRUCTION (LRI) AND ITS MEASUREMENT

Harnessing key cognitive load theory principles, Martin (2016; see also Martin and Evans, 2018, 2019) proposed LRI as an instructional means to manage the cognitive burden students can experience as they learn. According to cognitive load theory, there is a need for instruction that accommodates the reality of the limits of working memory and helps students transfer knowledge between working and long-term memory (Paas et al., 2003; Sweller, 2004). A key means by which this can occur is by developing students' fluency and automaticity in skill and knowledge. This frees up working memory resources, reduces cognitive burden, and better enables students to transfer novel information into long-term memory (Rosenshine, 2009). Based on the Martin (2016) LRI framework, fluency and automaticity are developed through its first four principles: (principle #1) reducing the difficulty of instruction in the initial stages of learning, as appropriate to the learner's level of prior knowledge (see also Pollock et al., 2002; Mayer and Moreno, 2010); (principle #2) providing appropriate support and scaffolding to learn the relevant skill and knowledge (see also Renkl and Atkinson, 2010; Renkl, 2014); (principle #3) allowing sufficient opportunity for practice (see also Purdie and Ellis, 2005; Rosenshine, 2009; Nandagopal and Ericsson, 2012); and (principle #4) providing appropriate feedback-feedforward (combination of corrective information and specific improvement-oriented guidance) as needed (see also Shute, 2008; Hattie, 2009; Mayer and Moreno, 2010). Through these four principles, students develop fluency and automaticity and are then well-positioned to apply their skill and knowledge more independently—including through novel tasks, higher order reasoning and thinking, problem solving, and guided discovery (Martin, 2016; Martin and Evans, 2019)—which is important to guard against potential expertise reversal effects (Kalyuga et al., 2012; Chen et al., 2017). This represents principle #5: guided independent learning.

Having articulated the five key principles of LRI (Martin, 2016), Martin and Evans (2018) developed a novel tool, the Load Reduction Instruction Scale (LRIS), to administer to students to report on the instructional practices of their teacher. The LRIS comprises five factors to assess the five LRI principles (difficulty reduction, support and scaffolding, practice, feedback-feedforward, guided independence). Each factor is composed of five items (yielding a 25-item instrument). In the first empirical study of the LRIS, students in 40 high school mathematics classrooms completed the instrument (Martin and Evans, 2018). Findings revealed the scores of each factor to be reliable, the factor structure to be sound, and significant bivariate correlations with intrinsic and extraneous cognitive load in predicted directions. In a follow-up investigation that linked the Martin and Evans (2018) data with a previous survey, results showed that LRI was associated with gains in mathematics motivation, engagement, and achievement (Evans and Martin, Submitted). In another study, Martin et al. (2020a) introduced and explored a brief form of the LRIS (the LRIS-Short; LRIS-S) that was designed to capture a unidimensional latent factor in keeping with the higher order LRIS factor in the Martin and Evans (2018) research. Martin et al. (2020a) used this LRIS-S in a multilevel study of more than 180 science classrooms, finding that the link between LRI in science and science achievement (at student- and classroom-levels) was mediated by class participation, future aspirations, and enjoyment in science. However, all these studies (including their construct validity aspects) have been variable-centered. As is now discussed, to even better understand the nomological network of LRI, a construct validity approach from a multilevel person-centered perspective has much to offer.

PERSON-CENTERED AND MULTILEVEL ASSESSMENT

Person-Centered Assessment

As noted above, the bulk of research assessing cognitive load in students' academic lives (including LRI research) has been variable-centered. Variable-centered approaches aim to assess relations between factors across individuals (Collie et al., 2020). This has been important for conducting classic construct validity work with cognitive load factors and for identifying what cognitive load factors predict or are predicted by other factors (e.g., Leppink et al., 2013; Klepsch and Seufert, 2020). However, variable-centered approaches can mask important phenomena among subpopulations of the wider population. In contrast, person-centered research utilizes the factors in a study to identify distinct subpopulations (or profiles) of individuals (Bauer and Curran, 2004; Collie et al., 2015; Morin et al., 2017). For example, different subpopulations of students may reflect different patterns in how LRI manifests and relates to other educational factors in their classroom and academic experience. Thus, LRI may not function in a similar way for all students. To the extent this is the case, it is important to assess LRI to capture distinct subpopulations of students, if such subpopulations exist. This will generate practical yields in identifying specific student profiles that teachers can attend to in their pedagogy.

Person-centered analyses (in this investigation, latent profile analysis—LPA) are ideal for addressing these issues. Specifically, by revealing the way LRI co-occurs with other factors among subpopulations of students, person-centered assessment helps identify the different types of students that reside within the classroom and the distinct ways LRI manifests in these students' academic lives. As we describe below, we seek to further assess the construct validity of LRI from a person-centered perspective by including psychological challenge and threat orientation measures alongside LRI measures to better understand the nomological network of instructional-psychological profiles. LPA enables us to see whether there might be subpopulations of students with similar LRI levels, but who differ on psychological factors (and vice versa). As we describe below, it is theoretically plausible that students who have different psychological challenge and threat orientations will experience

instructional load in different ways and our person-centered construct validity approach seeks to confirm this. The LPA represents the within-network validity aspect of our study by way of its assessment of a target network of instructional-psychological profiles.

Multilevel Assessment

It is also the case that the bulk of cognitive load research is almost exclusively conducted at the student-level, without appropriate regard for the classrooms to which the students belong (however, LRI research is an exception—described below). There is now broad recognition of the importance of analyzing hierarchical data in appropriate ways (Marsh et al., 2012), especially when the variables of interest include references to class-wide phenomena, such as instruction (as LRI does). In single-level research designs there are statistical biases (e.g., within-group dependencies; confounding of variables within and between groups) and multilevel analyses seek to resolve biases like these (see Raudenbush and Bryk, 2002; Goldstein, 2003; Marsh et al., 2009). Multilevel analysts have identified the reciprocity of group and individual dynamics: the group can affect the individuals and individuals can affect the group (Raudenbush and Bryk, 2002; Goldstein, 2003; Marsh et al., 2009). To date, most LRI research has recognized this and conducted multilevel analyses at student- and classroom-levels—indeed, demonstrating the validity of LRI at student- and classroom-levels (Martin and Evans, 2018, Evans and Martin, Submitted; Martin et al., 2020a). But, as noted, these studies have involved variable-centered analyses, which may mask differences between subpopulations of students and classrooms. The present study extends this research by conducting multilevel LPA that not only identifies student instructional-psychological profiles, but also classroom instructional-psychological profiles. Given LRI is about classroom instruction, its construct validity must be reflected in theoretically logical classroom profiles.

LRI AND ACCOMPANYING PSYCHOLOGICAL ORIENTATIONS

As described earlier, it is plausible that cognitive load will be perceived and experienced by students in different ways. Theories of cognitive appraisal (e.g., Lazarus and Folkman, 1984; for reviews, see Roseman and Smith, 2001; Moors et al., 2013) suggest that when presented with a task, an individual subjectively appraises its demands and their capacity to meet the demands. The individual perceives challenge when they believe they can meet the demands; the individual perceives threat when they believe they cannot meet the demands (see also Putwain and Symes, 2014, 2016, Uphill et al., 2019). Thus, when cognitive load is unacceptably high there may be a greater likelihood that anxiety and fear are present, and when cognitive load is effectively managed, more positive appraisals occur. At the same time, it is also the case that different students can appraise the same stimuli in different ways. For example, following cognitive appraisal theory (Lazarus and Folkman, 1984), some students will perceive cognitive load in an approach- and challenge-oriented way, while other students will perceive cognitive load in an avoidant- and threat-oriented way. Indeed, recent reviews of challenge and threat suggest that there may even be the dual presence of both challenge and threat (Uphill et al., 2019; see also Rogat and Linnenbrink-Garcia, 2019 for dual goals under approach-avoidance goal frameworks)—for example, in the event of cognitive load there may be students who perceive it as a challenge and opportunity for learning, but who also fear failing and see it as a potential threat. This idea had been previously raised by Covington (2000) and Martin and Marsh (2003) in the form of "overstrivers" who are students investing effort (in a challenge-like way), but driven in large part by a fear of poor performance. This brings into consideration the extent to which there exist student profiles based on different combinations of LRI and psychological challenge and threat orientations. LPA is designed to explore such possibilities and thus represents an important (and novel) means of assessing the study's contended network of instructional-psychological profiles (i.e., the within-network validity aspect of the study).

Consistent with Martin et al. (2021), recent challenge-threat frameworks (e.g., Putwain and Symes, 2014, 2016; Putwain et al., 2015; Uphill et al., 2019), and the latest approach-avoidance perspectives that have introduced growth goals (e.g., Elliot et al., 2011, 2015), we propose psychological challenge orientation can be inferred through students' self-efficacy and growth goals. Self-efficacy refers to a belief in one's capacity to meet or exceed a task demand or task challenge (Bandura, 1997; Schunk and DiBenedetto, 2014). This being the case, self-efficacy has been inferred as an analog of perceived challenge (Uphill et al., 2019) and operationalized as an indicator of perceived challenge in LPA (Martin et al., 2021). Perceived challenge has also been linked to approach orientations in motivational psychology. For example, Elliot defines approach motivation as the "energization of behavior by, or the direction of behavior toward, positive stimuli (objects, events, possibilities)" (2006, p. 112). According to Elliot, goals are a major means by which individuals' approach (and avoidance) orientations are manifested. Recent research has identified growth goals (i.e., self-improvement, or personal best goals) as one example of an approach motivation orientation (Martin and Liem, 2010; Elliot et al., 2011, 2015; Martin and Elliot, 2016a,b; Burns et al., 2018, 2019, 2020b). We recognize mastery goals are also approach-oriented, but it has previously been argued that growth goals represent a particularly ambitious and challenge-oriented goal striving (in keeping with our intent to capture challenge orientation) and are shown to explain variance in engagement beyond the effects of mastery goals (Yu and Martin, 2014; Martin and Elliot, 2016a). Growth goals are thus inferred as having an underlying psychological challenge dimension to them, along the lines of Uphill et al.'s (2019) review.

Also following Martin et al. (2021) and recent challenge-threat and approach-avoidance frameworks (e.g., Elliot et al., 2011, 2015; Putwain and Symes, 2014, 2016; Putwain et al., 2015; Uphill et al., 2019), we propose psychological threat orientation can be inferred through students' anxiety and failure avoidance goals. Anxiety reflects an individual's perception that a task demand exceeds their personal resources and capacity and poses

a self-relevant threat to them in some way (Britton et al., 2011). Anxiety has thus been associated with threat appraisals (e.g., see Putwain et al., 2015, 2017; see also Uphill et al., 2019) and has been used as an indicator of perceived threat in LPA (Martin et al., 2021). Perceived threat has also been linked to avoidance orientations in motivational psychology. Elliot defines avoidance motivation as the "energization of behavior by, or the direction of behavior away from, negative stimuli (objects, events, possibilities)" (2006, p. 112). In goal theory research, avoidance goals are a salient example of avoidance motivation (Elliot, 2006; Van Yperen et al., 2015). Avoidance goals are those where the individual strives to avoid poor performance and failure or the implications of poor performance (Covington, 2000; Elliot, 2006; Martin, 2007, 2009), and being an element of an avoidance orientation, are suggested as analogs of an inherent psychological threat orientation, along the lines of Uphill et al.'s (2019) review.

To summarize, the present study assesses students' self-efficacy and growth goals (to infer challenge) and anxiety and failure avoidance goals (to infer threat) as key psychological orientations to include as indicators alongside LRI in LPA. Importantly, however, although we conceptually categorize these as challenge and threat indicators, they are modeled as separate indicators so that these conceptual groupings can be tested empirically. We also point out that this study is a multilevel one that assesses these issues at the classroom-level. It is feasible that psychological orientations of challenge and threat manifest at the classroom-level such that some classrooms are relatively higher or lower in these orientations. Indeed, motivation research has suggested that classrooms can vary in the extent to which they are challenge- and approach-oriented *vs.* threat- and avoidance-oriented (Lau and Nie, 2008). Furthermore, these classroom-level psychological orientations may co-vary with different levels of LRI in different ways. This being the case, the present study assesses classroom-level psychological orientations alongside classroom-level LRI to identify classroom-level instructional-psychological profiles.

ASSESSING LINKS BETWEEN INSTRUCTIONAL-PSYCHOLOGICAL PROFILES AND ACADEMIC OUTCOMES

This research is conducted in the educational domain of science. Science is a challenging subject for many students (Coe et al., 2008) and there are worrying trends in students' science achievement and science participation (especially in "Western" nations). For example, in the Programme for International Student Assessment (PISA), the long-term change in the average science performance of Australia (the site of the present study) demonstrates one of the largest declines among participating nations (OECD., 2020). There are also long-term declines in science participation and enrolments among senior school students (Office of the Chief Scientist, 2014) as well as declining interest in science in high school (Tröbst et al., 2016). Inherent in these trends are three major concerns. First, students do not seem to be orienting toward science in a participatory and persistent way—Martin et al. (2012) referred to this as "switching on." Second, there are unacceptable numbers of students disengaging from science—Martin et al. (2012) referred to this phenomenon as "switching off." Third, science achievement is in decline in many nations. This being the case, it is important to: (a) initiate and foster more positive persistence ("switching on") in science, (b) arrest students' disengagement ("switching off") in science, and (c) boost students' science achievement. Therefore, we sought to explore the association between the derived network of student- and classroom-level instructional-psychological profiles and a network of student- and classroom-level outcomes in the form of science persistence, disengagement, and achievement. From a construct validation perspective (Marsh, 2002), this represents the between-network validity aspect of our investigation.

In variable-centered research, Martin et al. (2012) found that approach-oriented predictors such as self-efficacy were positively associated with "switching on" in mathematics (positive future intentions and engagement) and negatively associated with "switching off" in mathematics (disengagement). The inverse pattern of associations was found for avoidance-oriented predictors such as anxiety. Also, in variable-centered research, LRI is positively associated with achievement in mathematics (Martin and Evans, 2018). In recent person-centered research, Martin et al. (2021) found that approach (challenge)-oriented students had higher science test scores, while avoidance (threat)-oriented students had lower scores. Interestingly, and in keeping with the potential dual presence of challenge and threat (Uphill et al., 2019), Martin et al. (2021) also identified a third profile reflecting both approach (challenge) and avoidance (threat)—students in this profile scored midway between the former two profiles on the science test. Taking these recent student-level findings together, we tentatively suggest at least three student-level profiles that will be associated with our outcomes (persistence, disengagement, achievement) in a descending adaptive pattern: high approach-low avoidance-high LRI, high approach-high avoidance-high LRI, and low approach-high avoidance-low LRI. From a construct validity perspective, demonstrating associations in a predicted manner is important (Marsh et al., 2006). Regarding classroom-level findings, we believe there is not a sufficient evidence base to guide predictions and so this is a more exploratory aspect of the study.

AIMS OF THE PRESENT STUDY

We argue that to fully understand instructional cognitive load, it is important to operationalize and assess it in novel ways that can provide unique validity insights into how different students perceive and experience this load. We further suggest it is important to consider these novel assessment approaches using appropriate cutting-edge analytic models. Accordingly, we adopted a within- and between-network construct validity approach and used multilevel LPA to identify instructional-psychological profiles among students and classrooms based on students' reports of instructional cognitive load (via load reduction instruction; LRI) and their accompanying psychological challenge orientations (self-efficacy, growth goals)

and psychological threat orientations (anxiety, failure avoidance goals). In phase 1 of analyses, we sought to identify a network of instructional-psychological profiles among students (student-level within-network validity). In phase 1, we also tested the links between the derived profile network and a network of three academic outcomes (persistence, disengagement, achievement) (student-level between-network validity). In phase 2 of analysis, we extended our examination to the classroom-level where we sought to identify the network of classroom profiles based on the relative frequency of student profiles identified in phase 1 (classroom-level within-network validity). We also tested whether the derived network of different classroom profiles was associated with different levels of classroom-average persistence, disengagement, and achievement (classroom-level between-network validity). **Figure 1** demonstrates the model and parameters under investigation.

METHOD

Participants and Procedure

Participants comprised 2,071[1] Australian high school students from 188 science classrooms in eight schools. The schools were independent (non-systemic) schools, in a major capital city on the east coast of Australia. Four of these schools were single-sex boys' schools and four were single-sex girls' schools. Just over half (60%) the sample comprised girls. Participants were in Year 7 (29%), Year 8 (22%), Year 9 (24%), and Year 10 (25%). The average age was 14.02 years (SD = 1.27 years). Just under one in ten (8%) students spoke a language other than English at home. Socioeconomic status (SES) varied (range 846–1,181, M = 1,138, SD = 41, on the Australian Bureau of Statistics Index of Relative Socio-Economic Advantage and Disadvantage classification), but in aggregate was a higher SES than the Australian mean of 1,000. On average, classrooms had about eleven students (adequate for group-level effect estimation and not disproportionate to the ratio of staff to students in the independent school sector when accounting for student absences, non-participation, and students who have not received participation consent from their parents; also see Australian Bureau of Statistics, 2019). The lead researcher's university provided human ethics approval for the study. Approval was then provided by school principals for their school to participate. Then, participating students (and their parents/carers) provided consent. The online survey (that also comprised an achievement test) was administered during a regularly scheduled science class in 2018. Students completed the instrument on their own. Teachers were informed they could help students with procedural aspects of the survey, but not provide assistance in answering specific questions. The data in this investigation are shared with Martin et al. (2020a), who conducted a variable-centered study exploring the extent to which class participation, educational aspirations, and enjoyment of school mediated the relationship between LRI and achievement.

[1]There were 2,199 students in the original sample, but we removed students who did not identify their classroom (n = 90) as this information is necessary for multi-level analyses. Also removed were classes with fewer than 3 students, as we considered these class sizes too small to yield reliable estimates at the classroom level (viz. 38 students from 25 classes; see McNeish, 2014).

Materials

Indicators for the network of instructional-psychological profiles were measured by way of instructional cognitive load (load reduction instruction), psychological challenge orientation (self-efficacy, growth goals), and psychological threat orientation (anxiety, failure avoidance goals). These indicators were the within-network validity variables for this study. The network of outcome measures was assessed by way of persistence, disengagement, and achievement. These were the between-network validity variables for the study.

Instructional Cognitive Load via Load Reduction Instruction Scale—Short (LRIS-S)

As described in Martin et al. (2020a), the LRIS-S was developed to measure student perceptions of their teacher's use of instructional strategies known to reduce extraneous cognitive load (and because of this, some intrinsic cognitive load). In the LRIS-S, the five LRI factors are represented by a single item (the full LRIS has 5 items for each of the 5 factors; Martin and Evans, 2018). As detailed in Martin et al. (2020a), the factors and items (adapted to science) are: *difficulty reduction* ("When we learn new things in this science class, the teacher makes it easy at first"); *support* ("In this science class, the teacher is available for help when we need it"); *practice* ("In this science class, the teacher makes sure we practice important things we learn"); *feedback-feedforward* ("In this science class, the teacher provides frequent feedback that helps us learn"); and *independence* ("Once we know what we're doing in this science class, the teacher gives us a chance to work independently"). Responses were provided on a seven-point scale (1 = *strongly disagree* to 7 = *strongly agree*). Reliability for this scale was sound (see **Table 1**) and ICC = 0.16. Because the LRIS has an emphasis on reduction of cognitive load, Martin and Evans (2018) conducted analyses showing that the scale was significantly negatively associated with intrinsic load (load referring to task difficulty and complexity) and extraneous load (load referring to difficulty and complexity of instruction; Chandler and Sweller, 1991). They concluded that the LRIS does assess aspects of instruction impacting distinct elements of cognitive load. Martin et al. (2020a) showed that the reliability of the LRIS-S at student- and classroom-levels was high and their doubly-latent multi-level (student and classroom) factor analysis demonstrated sound fit and yielded strong factor loadings. Thus, we suggest that student-reported LRIS-S offers valid insights into the instructional practices relevant to cognitive load.

Psychological Challenge Orientation

Psychological challenge orientation was assessed via self-efficacy and growth goals. Self-efficacy (4 items; e.g., "If I try hard, I believe I can do well in this science class") was measured via the domain-specific form of the Motivation and Engagement Scale—High School (MES-HS; Martin, 2007, 2009), validated by Green et al. (2007). Growth goals (4 items; e.g., "When I do my science schoolwork I try to do it better than I've done before") were measured via the domain-specific form of the Personal Best

TABLE 1 | Descriptive statistics.

	Student-level (Level 1; L1)			Classroom-level (Level 2; L2)		
	M	*SD*	ω–Coefficient omega	*M*	*SD*	ω–Coefficient omega
Profile Network Indicators						
LRI	5.284	1.123	0.85	5.298	0.626	0.96
Self-efficacy	5.790	1.025	0.83	5.778	0.529	0.93
Growth Goals	5.202	1.152	0.90	5.200	0.655	0.99
Anxiety	4.113	1.362	0.78	4.168	0.650	0.94
Failure Avoid Goals	3.194	1.396	0.83	3.250	0.665	0.98
Outcome Network						
Persistence	5.146	1.105	0.83	5.137	0.614	0.97
Disengagement	2.387	1.332	0.87	2.401	0.765	0.98
Achievement	0.000	1.000	0.79	0.000	1.000	0.86

Achievement is standardized; LRI, load reduction instruction.

Goal Scale (Martin and Liem, 2010), for which Martin and Elliot (2016a) provided evidence of validity. Responses were provided on a seven-point scale (1 = *strongly disagree* to 7 = *strongly agree*). Reliability for the scores of both scales was sound (see **Table 1**) and ICCs = 0.08 and 0.12 for self-efficacy and growth goals, respectively.

Psychological Threat Orientation

Psychological threat orientation was assessed via anxiety (4 items; e.g., "When exams and assignments are coming up for this science class, I worry a lot") and failure avoidance goals (4 items; e.g., "Often the main reason I work in this science class is because I don't want people to think that I'm dumb"). Both were from the domain-specific form of the MES-HS (Martin, 2007, 2009), for which Green et al. (2007) provided evidence of validity. Responses were provided on a seven-point scale (1 = *strongly disagree* to 7 = *strongly agree*). Reliability for the scores of both scales was sound (see **Table 1**) and ICCs = 0.08 and 0.05 for anxiety and failure avoidance goals, respectively.

Persistence and Disengagement

In line with Martin et al. (2012), engagement was assessed via the dual dimensions of "switching on" and "switching off." Switching on was operationalized through persistence (4 items; e.g., "If I can't understand something in this science class at first, I keep going over it until I do"). Switching off was operationalized through disengagement (4 items; e.g., "Each week I'm trying less and less in this science class"). Both were from the domain-specific form of the MES-HS (Martin, 2007, 2009), validated by Green et al. (2007). Responses were provided on a seven-point scale (1 = *strongly disagree* to 7 = *strongly agree*). Reliability for scores of both scales was sound (see **Table 1**) and ICCs = 0.12 and 0.17 for persistence and disengagement, respectively.

Achievement

Achievement was measured using 12 questions in an online test (part of the online survey). Instrument piloting and development are fully described in Martin et al. (2020a). The test aligned with the science syllabus applicable to our sample; therefore, two forms were developed, one based on the Stage 4 (years 7 and 8) state science syllabus and the other based on the Stage 5 (years 9 and 10) state science syllabus. Test questions were set within the contexts of Physical World, Earth and Space, Living World, and Chemical World and addressed the following skills: questioning and predicting, planning investigations, conducting investigations, processing and analyzing data and information, and problem solving. Each question was grounded within one of the abovementioned specific science contexts and there was an ∼30/70 ratio of content-focused to skill-focused questions, with the easier questions focusing on content and the harder questions focusing on skill application. All multiple-choice test responses were recoded as dichotomous (0 = incorrect; 1 = correct). The correct answers were summed to a total achievement score (thus, a continuous scale), reflecting something of a formative construct. Achievement scores were then standardized by year level ($M = 0$; $SD = 1$). The test was reliable, as shown in **Table 1** and ICC = 0.37.

DATA ANALYSIS

Analyses were conducted using M*plus* 8.60 (Muthén and Muthén, 2017). The robust maximum likelihood (MLR) estimator was used in all models. Missing data were addressed using Full Information Maximum Likelihood (FIML; Arbuckle, 1996). Confirmatory factor analysis (CFA) was run at the student-level (and corrected for nesting within classrooms via the M*plus* "COMPLEX" command) using the standardized factor approach to identification to obtain student-level factor scores for the five profile indicators and the three outcomes. The CFA also comprised background attributes as auxiliary variables—reported in analyses in **Supplementary Materials**. Factor scores were saved and used in the LPAs. The LPA analyses comprised two phases: single-level LPA (phase 1) and multi-level LPA (phase 2).

Single-Level LPA

For the single-level LPA conducted at the student-level (Level 1; L1), we tested a range of solutions involving 1 through 9 profiles. Following Collie et al. (2020), variances and means were free to differ across profiles and indicator variables; models were estimated using at least 10,000 random start values, with 100 iterations and 1,000 final stage optimizations; and we confirmed that the best log-likelihood value was replicated for each model. Numerous indices were used to assess model fit: for the Consistent Akaike Information Criteria (CAIC), Akaike Information Criteria (AIC), Bayesian Information Criteria (BIC), and sample-size-adjusted Bayesian Information Criteria (SSA-BIC), smaller values reflect better fit. We created elbow plots of the CAIC, AIC, BIC, and SSA-BIC indices. In these plots, the profile at which point the slope noticeably flattens is another indicator of an appropriate solution (Morin et al., 2016). The *p*-value of the adjusted Lo–Mendell–Rubin Likelihood Ratio Test (*p*LMR) enabled comparison of a *k*-profile model against a *k*-1 profile model to determine if the former profile model yielded a better fit relative to the latter profile model. We also provide entropy values in **Table 2** as an indicator of classification accuracy. In addition to fit indices and where appropriate, we applied rules of parsimony, conceptual relevance, and statistical adequacy to further ascertain the optimal solution. After identifying the final network of profiles (the student-level within-network validity aspect), we then examined the network of academic outcomes (persistence, disengagement, achievement) as a function of profile membership (the student-level between-network validity aspect), controlling for background attributes. Outcomes were included using the direct approach and compared across profiles using the M*plus* "MODEL CONSTRAINT" option, which relies on the multivariate delta method for tests of statistical significance (e.g., Raykov and Marcoulides, 2004). As part of this, the outcomes were also regressed on participants' background attributes, which acted as covariate controls (McLarnon and O'Neill, 2018).

Multi-Level LPA

In phase 2, we extended the student-level (Level 1, L1) findings to determine the extent to which classroom-level (Level 2; L2) profiles could be identified; or, put another way, the extent to which we could identify classroom profiles characterized by distinct combinations of the different student profiles. Thus, phase 2 identified classroom profiles based on the relative frequency of the various L1 latent profiles. To maintain the stability of the previously identified student-level profiles (L1), we used the manual 3-step approach detailed by Litalien et al. (2019; also see Vermunt, 2010; Morin and Litalien, 2017). Multi-level LPA solutions (1–9 classroom-level profiles) were assessed. Following Collie et al. (2020), each model was estimated using at least 10,000 random start values, 100 iterations, and 1,000 final stage optimizations; replication of the best log-likelihood value was sought for each model; and the best model was selected using the same criteria as the single-level LPA (phase 1), except the *p*LMR, which is not available for multi-level LPA. After identifying the network of classroom-level profiles (the classroom-level within-network validity aspect), we then examined the network of L2 outcomes (classroom-average persistence, disengagement, and achievement) as a function of profile membership by adding classroom-average outcome variables (using the M*plus* cluster mean approach) to the final best-fitting model (the classroom-level between-network validity aspect). Outcomes were compared across profiles using the M*plus* "MODEL CONSTRAINT" option (e.g., Raykov and Marcoulides, 2004).

RESULTS

Preliminary Analyses

Table 1 shows reliability coefficients and descriptive statistics for the profile indicators and the outcome variables in the study. These data indicate acceptable reliability. The CFA used to generate factor scores yielded an excellent fit to the data, $\chi^2 = 1{,}321$, df = 378, $p < 0.001$, CFI = 0.964, TLI = 0.958, RMSEA = 0.035. Indeed, these preliminary variable-centered analyses demonstrate sound within-network validity properties (Marsh, 2002). The resulting correlation matrix is presented in **Supplementary Table 1**. These factor scores were then used in the subsequent LPAs.

Single-Level LPA

The fit statistics for the 1–9-profile solutions are shown in **Table 2** and the elbow plot is shown in **Supplementary Figure 1**. In these it is evident that the CAIC, AIC, BIC, and SSA-BIC decline with each additional profile. There appears to be slight inflection points around 4, 5, and 6 profiles. Although we do not rely on the *p*LMR, it is interesting to note it supported the 6-profile solution, but it was significant at the $p < 0.01$ level—whereas the 4–5-profile solutions were significant at $p < 0.001$. In addition, although not relying on minimum profile size as a decision criterion, we note that the 6-profile solution had a minimum profile size of <2%, whereas the 5-profile solution had a size of 8%. Taken together, we felt that additional profiles were theoretically useful and well-differentiated up to 6 profiles, but the sixth profile presented a shape that was qualitatively similar (even if it differed quantitatively) to that of the 5-profile solution. We therefore proceeded with the 5-profile solution. A graphical representation of this 5-profile solution is presented in **Figure 2**. Students corresponding to profile 1 (8% of students) reported very low LRI, very low self-efficacy, very low growth goals, neutral anxiety, and neutral failure avoidance goals. This profile was labeled Instructionally-Overburdened & Psychologically-Resigned to reflect very high instructional cognitive load and very low challenge orientation, indeed so much so they are not particularly threatened, but rather resigned. Students corresponding to profile 2 (30% of students) reported low LRI, low self-efficacy, low growth goals, high anxiety, and high failure avoidance goals. This profile was labeled Instructionally-Burdened & Psychologically-Fearful to reflect modest instructional load, low challenge orientation, and high threat orientation. Students corresponding to profile 3 (31% of students) reported above average LRI, above average self-efficacy and growth goals,

TABLE 2 | Single-level LPA fit statistics.

	1 Profile	2 Profiles	3 Profiles	4 Profiles	5 Profiles	6 Profiles	7 Profiles	8 Profiles	9 Profiles
N	2,071	2,071	2,071	2,071	2,071	2,071	2,071	2,071	2,071
Free Parameters	10	21	32	43	54	65	76	87	98
Log-likelihood	−13,965	−12,628	−12,169	−11,871	−11,628	−11,439	−11,309	−11,167	−11,078
CAIC	28,016	25,437	24,614	24,113	23,722	23,439	23,274	23,085	23,002
Akaike (AIC)	27,951	25,298	24,402	23,828	23,365	23,008	22,771	22,509	22,353
Bayesian (BIC)	28,007	25,417	24,582	24,070	23,669	23,375	23,199	22,999	22,905
S-SA BIC	27,975	25,350	24,480	23,934	23,498	23,168	22,958	22,723	22,593
Entropy	–	0.782	0.844	0.792	0.782	0.803	0.815	0.815	0.793
*p*LMR	–	<0.001	<0.001	<0.001	<0.001	0.002	0.041	0.521	0.341

N, sample size; AIC, Akaike Information Criteria; CAIC, Consistent Akaike Information Criteria; S-SA, Sample-Size Adjusted; BIC, Bayesian Information Criteria; pLMR, Lo–Mendell–Rubin Likelihood Ratio Test.

and below average anxiety and failure avoidance goals. We labeled this profile Instructionally-Supported & Psychologically-Composed to reflect low instructional load, above average challenge orientation, and below average threat orientation. Students corresponding to profile 4 (9% of students) reported very high LRI, very high self-efficacy, very high growth goals, very low anxiety, and very low failure avoidance goals. We labeled this profile Instructionally-Optimized & Psychologically-Self-Assured to reflect the very low cognitive instructional load, the very high challenge orientation, and the very low threat orientation. Students corresponding to profile 5 (22% of students) reported above average scores on each of LRI, self-efficacy, growth goals, anxiety, and failure avoidance goals. We labeled this profile Instructionally-Supported & Psychologically-Pressured to reflect low instructional load as well as dual high challenge and threat orientations.

We then assessed for differences between profiles in persistence, disengagement, and achievement (adjusted for background attribute covariates—see **Supplementary Tables 2A–E** for the predictive relationships between background attributes and the latent profiles). Mean scores are shown in **Table 3**. For persistence, findings indicated that each profile was significantly different from the other. In ascending order of persistence were: Instructionally-Overburdened & Psychologically-Resigned (lowest persistence; $M = -1.571$), then Instructionally-Burdened & Psychologically-Fearful ($M = -0.375$), then Instructionally-Supported & Psychologically-Composed ($M = 0.559$), then Instructionally-Supported & Psychologically-Pressured ($M = 0.857$), then Instructionally-Optimized & Psychologically-Self-Assured (highest persistence; $M = 1.407$).

For disengagement, findings indicated that with one exception (Instructionally-Supported & Psychologically-Composed = Instructionally-Supported & Psychologically-Pressured), each profile was significantly different from the other. In descending order of disengagement were: Instructionally-Overburdened & Psychologically-Resigned (highest disengagement; $M = 1.660$), then Instructionally-Burdened & Psychologically-Fearful ($M = 0.436$), then Instructionally-Supported & Psychologically-Composed and also Instructionally-Supported & Psychologically-Pressured ($M = -0.683$ and $M = -0.743$, respectively), then Instructionally-Optimized & Psychologically-Self-Assured (lowest disengagement; $M = -1.231$).

For achievement, findings indicated that with one exception (Instructionally-Overburdened & Psychologically-Resigned = Instructionally-Burdened & Psychologically-Fearful), each profile was significantly different from the other. In ascending order of achievement were: Instructionally-Overburdened & Psychologically-Resigned and also Instructionally-Burdened & Psychologically-Fearful (lowest achievement; $M = -0.590$ and $M = -0.409$, respectively), then Instructionally-Supported & Psychologically-Pressured ($M = 0.054$), then Instructionally-Supported & Psychologically-Composed ($M = 0.089$), then Instructionally-Optimized & Psychologically-Self-Assured (highest achievement; $M = 0.430$).

Multi-Level LPA

The fit statistics for the multi-level LPA solutions are reported in **Table 4** (the elbow plot is shown in **Supplementary Figure 2**). Here 1–9-profile solutions are presented. The 2-profile solution resulted in the consistently lowest value for the fit indices, but there was some further flattening on other indices at the third profile. Also, as described below, the 3-profile solution yielded a group that separated classrooms in qualitatively distinct ways that was beyond what was possible in a 2-profile solution that (as it turned out, and is described below) could not differentiate a Striving profile, from Thriving and Struggling profiles. Moreover, this additional profile constituted a sizeable subpopulation (22%). Morin et al. (2017) emphasize the importance of ensuring that each profile adds conceptually and practically meaningful information to a solution. Thus, while recognizing aspects of fit suggest a 2-profile solution, we concluded there was substantive and practical yield in the additional profile. Accordingly, a solution with 3 classroom-level profiles was selected as the final solution.

A graphical representation of this final 3-profile solution is presented in **Figure 3**. Examination of this 3-profile solution

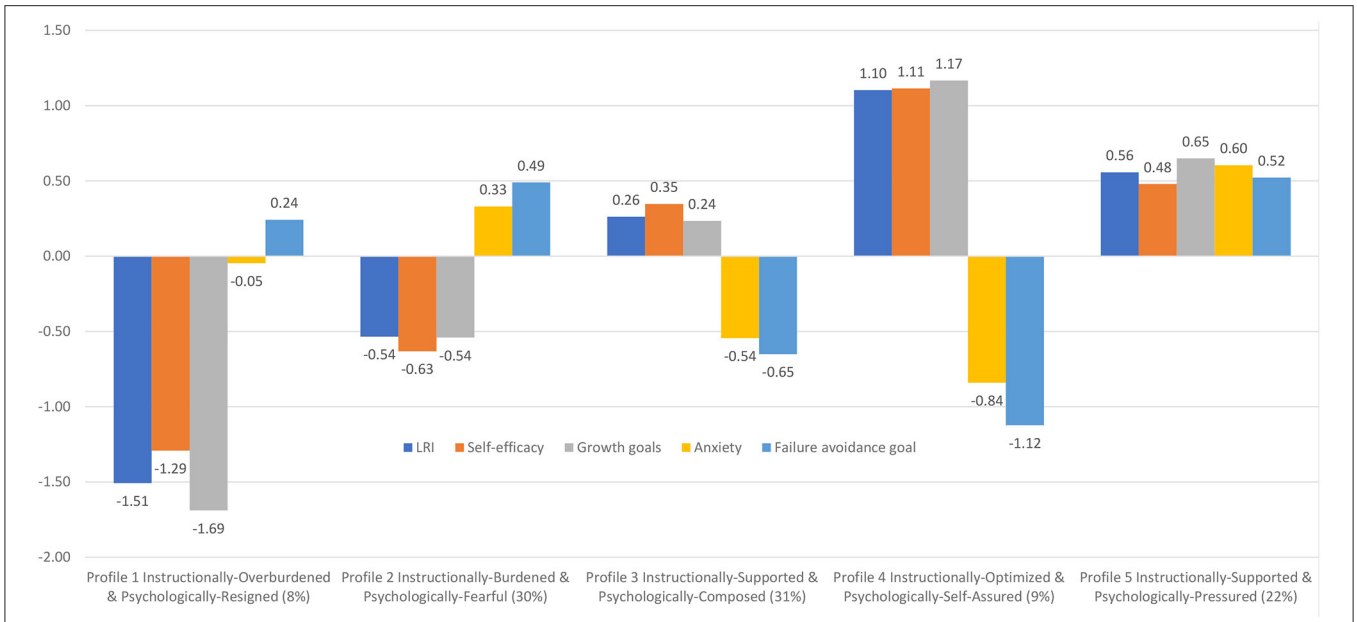

FIGURE 2 | Single-level LPA results: Instructional-psychological profile names, profile means, and % of sample.

TABLE 3 | Means (SEs and 95% CIs) on dependent variables for each instructional-psychological profile.

	Persistence				Disengagement				Achievement			
	Mean	**SE**	**95% CI**		**Mean**	**SE**	**95% CI**		**Mean**	**SE**	**95% CI**	
Instructionally-Overburdened & Psychologically-Resigned	−1.571[a]	0.097	−1.762	−1.381	1.660[a]	0.088	1.487	1.834	−0.590[a]	0.146	−0.876	−0.304
Instructionally-Burdened & Psychologically-Fearful	−0.375[b]	0.076	−0.524	−0.227	0.436[b]	0.056	0.328	0.545	−0.409[a]	0.118	−0.640	−0.179
Instructionally-Supported & Psychologically-Composed	0.559[c]	0.074	0.414	0.704	−0.683[c]	0.060	−0.800	−0.567	0.089[b]	0.115	−0.136	0.313
Instructionally-Optimized & Psychologically-Self-Assured	1.407[d]	0.072	1.222	1.549	−1.231[d]	0.040	−1.335	−1.152	0.430[c]	0.119	0.123	0.664
Instructionally-Supported & Psychologically-Pressured	0.857[e]	0.074	0.711	1.003	−0.743[c]	0.061	−0.862	−0.625	0.054[d]	0.116	−0.174	0.281

Different superscripts in a given column indicate a significant difference between means at $p < 0.05$; SE, standard error; CI, confidence interval.

suggested the presence of one Struggling classroom profile (22% of the classrooms), one Striving classroom profile (36% of the classrooms), and one Thriving classroom (42% of classrooms). The Struggling classroom had the highest proportion of students from the Instructionally-Overburdened & Psychologically-Resigned (19%) and Instructionally-Burdened & Psychologically-Fearful (49%) profiles. The Striving classroom included a high proportion of students from the Instructionally-Supported & Psychologically-Pressured (31%), Instructionally-Burdened & Psychologically-Fearful (33%), and Instructionally-Overburdened & Psychologically-Resigned (10%) profiles. The Thriving classroom had the highest proportion of students from the Instructionally-Optimized & Psychologically-Self-Assured (29%) profile, along with a high proportion of students from the Instructionally-Supported & Psychologically-Composed (14%) and Instructionally-Supported & Psychologically-Pressured (38%) profiles.

We then assessed for differences between classroom profiles in classroom-average persistence, disengagement, and achievement. Results are shown in **Table 5**. For classroom-average persistence, each classroom profile was significantly different from the other. In ascending order of classroom-average persistence were: the Struggling classroom (lowest persistence; $M = -0.610$), then the Striving classroom ($M = -0.068$), then the Thriving classroom (highest persistence; $M = 0.422$). For classroom-average disengagement, each classroom profile was significantly different from the other. In descending order of classroom-average disengagement were: the Struggling classroom (highest disengagement; $M = 0.705$), then the Striving classroom ($M = 0.048$), then the Thriving classroom

TABLE 4 | Multi-level LPA fit statistics.

	1 Profile	2 Profiles	3 Profiles	4 Profiles	5 Profiles	6 Profiles	7 Profiles	8 Profiles	9 Profiles
N	2,071	2,071	2,071	2,071	2,071	2,071	2,071	2,071	2,071
Free Parameters	4	9	14	19	24	29	34	39	44
Log-likelihood	−3,061	−3,003	−2,994	−2,989	−2,987	−2,982	−2,981	−2,981	−2,982
CAIC	6,157	6,084	6,109	6,142	6,181	6,214	6,256	6,299	6,344
Akaike (AIC)	6,131	6,025	6,016	6,017	6,023	6,022	6,030	6,041	6,053
Bayesian (BIC)	6,153	6,075	6,095	6,124	6,158	6,185	6,222	6,261	6,301
S-SA BIC	6,141	6,047	6,050	6,064	6,082	6,093	6,114	6,137	6,161
Entropy	0.669	0.664	0.687	0.655	0.617	0.633	0.607	0.662	0.649

N, sample size; AIC, Akaike Information Criteria; CAIC, Consistent Akaike Information Criteria; S-SA, Sample-Size Adjusted; BIC, Bayesian Information Criteria.

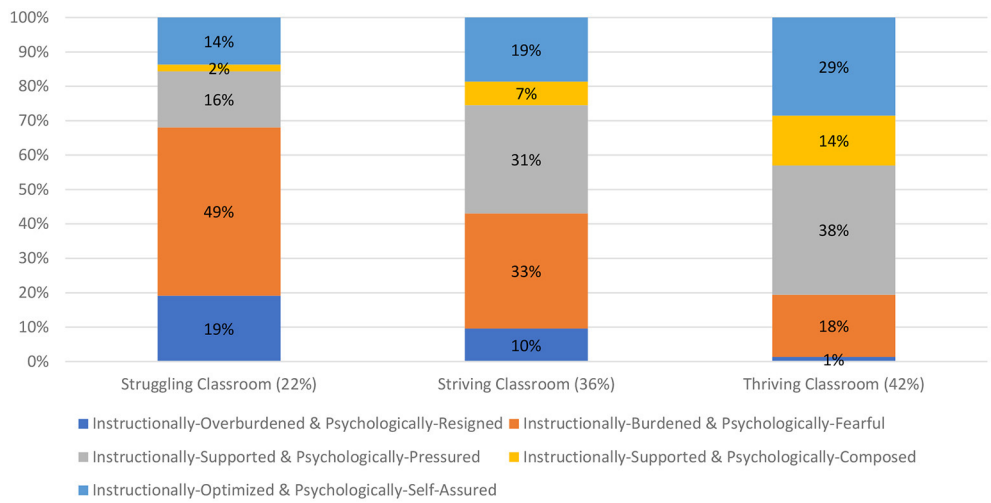

FIGURE 3 | Multi-level LPA results showing the classroom-level instructional-psychological profiles.

(lowest disengagement; $M = -0.427$). For classroom-average achievement, with one exception (Striving = Thriving), each classroom profile was significantly different from the other. In ascending order of classroom-average achievement were: the Struggling classroom (lowest achievement; $M = -0.498$), then the Striving classroom ($M = -0.032$), then the Thriving classroom (highest achievement; $M = 0.182$)—but as noted, the latter two classroom profiles were not significantly different from each other in achievement (see **Table 5**).

DISCUSSION

To best understand instructional cognitive load, we have emphasized the importance of assessing it in novel ways to reveal how different students perceive and experience this load. We have further emphasized the importance of utilizing cutting-edge analytic approaches that are appropriate to assessing these novel instrumentations. Integrating cognitive load theory and cognitive appraisal theory, we hypothesized that some students are likely to perceive cognitive load in an approach- and challenge-oriented way, and other students are likely to perceive cognitive load in an avoidant- and threat-oriented way. To the extent this is the case, we suggested that to further understand instructional cognitive load (by way of load reduction instruction; LRI) it is important to do so by also assessing students' accompanying psychological challenge and threat orientations. Adopting a novel person-centered construct validity perspective, we used latent profile analysis (LPA) to identify the network of instructional-psychological profiles based on students' reports of instructional load (LRI) and their accompanying psychological challenge orientations (self-efficacy and growth goals) and psychological threat orientations (anxiety and failure avoidance goals)—student-level within-network validity. Moreover, because students in our study were nested within (science) classrooms, we expanded our analyses to also conduct multilevel LPA to identify a network of student- and classroom-level instructional-psychological profiles—classroom-level within-network validity. We assessed student- and classroom-level between-network validity by investigating associations between the network of derived profiles and the network of student- and classroom-level persistence, disengagement, and achievement outcome variables.

TABLE 5 | Means (and SEs and 95% CIs) on dependent variables for each instructional-psychological profile.

	Persistence				Disengagement				Achievement			
	Mean	SE	95% CI		Mean	SE	95% CI		Mean	SE	95% CI	
Struggling Classrooms	−0.610[a]	0.085	−0.776	−0.443	0.705[a]	0.101	0.506	0.903	−0.498[a]	0.067	−0.630	−0.367
Thriving Classrooms	0.422[b]	0.048	0.329	0.515	−0.427[b]	0.037	−0.500	−0.354	0.182[b]	0.072	0.042	0.322
Striving Classrooms	−0.068[c]	0.074	−0.213	0.077	0.048[c]	0.086	−0.120	0.215	−0.032[b]	0.115	−0.257	0.194

Different superscripts in a given column indicate a significant difference between means at $p < 0.05$; SE, standard error; CI, confidence interval.

Summary of Findings

At the student-level, we identified five instructional-motivational profiles that represented different presentations of instructional cognitive load, challenge orientation, and threat orientation: Instructionally-Overburdened & Psychologically-Resigned students (8% of students), Instructionally-Burdened & Psychologically-Fearful students (30%), Instructionally-Supported & Psychologically-Composed students (31%), Instructionally-Optimized & Psychologically-Self-Assured students (9%), and Instructionally-Supported & Psychologically-Pressured students (22%). As we describe below, these conform to established theoretical perspectives and thus offer a student-level within-network validation perspective on the nomological network of instructional cognitive load in terms of underlying instructional-psychological orientations. We also demonstrated student-level between-network validity in that these profiles were significantly different in persistence, disengagement, and achievement (beyond the role of background attributes)—with the Instructionally-Overburdened & Psychologically-Resigned profile reflecting the most maladaptive outcomes and the Instructionally-Optimized & Psychologically-Self-Assured profile reflecting the most adaptive outcomes. In multilevel LPAs, we identified three instructional-psychological profiles among classrooms that varied in terms of instructional cognitive load, challenge orientations, and threat orientations: Striving classrooms (36% of the classrooms), Thriving classrooms (42%), and Struggling classrooms (22%). In terms of classroom-level between-network validity, we found that classroom profiles were significantly different in their levels of persistence, disengagement, and achievement—with Struggling classrooms reflecting the most maladaptive outcomes and Thriving classrooms reflecting the most adaptive outcomes, but, notably, equal to the Striving classrooms in achievement.

Findings of Particular Note

In numerous ways this study offers novel contributions to the assessment of instructional cognitive load, including: its person-centered perspective elucidating theoretically plausible student profiles based on their experience of cognitive load and their psychological orientations, the multilevel validity of the Load Reduction Instruction Scale-Short (LRIS-S) in person-centered analyses, and the validity of the links between profiles and academic outcomes. We suggest that findings hold implications for better assessing and understanding students and classrooms in terms of the cognitive load they experience through instruction. Specifically, the results show that assessing instructional load in the context of students' accompanying psychological orientations can reveal unique insights about students' learning experiences and about important differences between classrooms in terms of the instructional load that is present.

The findings supported one of the central premises of this study—namely, that similar levels of perceived instructional load can be accompanied by different levels of perceived challenge and threat. For example, at the student-level we identified two profiles that can be considered instructionally-supported but who varied in their accompanying psychological orientations. Specifically, the Instructionally-Supported & Psychologically-Composed profile experienced moderate levels of LRI, moderate challenge orientation and low threat orientation, whereas the Instructionally-Supported & Psychologically-Pressured profile experienced moderate LRI and challenge orientation but also moderate levels of threat orientation. These two profiles were also significantly different in persistence and achievement outcomes (but not disengagement), with the former profile scoring higher than the latter profile. This is notable because it shows that students with similar levels of instructional load can have different psychological experiences (i.e., differing levels of challenge and threat) that yield significant differences in academic outcomes. This underscores the yield of assessing instructional load in the context of other potentially influential accompanying factors. This requires assessment and analytic approaches that can disentangle students who perceive similar levels of instructional load but who vary on other factors (in our study, psychological challenge and threat orientations).

The Instructionally-Supported & Psychologically-Pressured profile was further illuminating in that it confirmed the existence of the contended dual challenge-threat orientation (or, approach-avoidance motive). As noted earlier, recent reviews of challenge-threat orientations have suggested the dual presence of both challenge and threat among some individuals (Uphill et al., 2019; see also Rogat and Linnenbrink-Garcia, 2019 for dual goals under approach-avoidance goal frameworks). In the case of our study, in the presence of instructional load there were some students who also reported dual challenge and threat orientations—that is, they believed they are up to the challenge of task burden but are also fearful of failure or poor performance, somewhat akin to the "overstrivers" described earlier (Covington, 2000; Martin and Marsh, 2003). Essentially, in the context of instructional load they perceive both an opportunity to succeed and a risk they may fail. Accordingly,

we identified these students as Psychologically-Pressured because even though they reflected a challenge orientation, there was an accompanying fear and avoidance (threat) inclination. Moreover, despite their threat orientation, the presence of a concomitant challenge orientation meant they experienced higher academic outcomes relative to the Instructionally-Overburdened & Psychologically-Resigned students and the Instructionally-Burdened & Psychologically-Fearful students. Nonetheless, the dual presence of challenge-threat orientations experienced by the Psychologically-Pressured profile represented a tension that we contend held them back from the more optimal academic outcomes seen in the Psychologically-Composed and Psychologically-Self-Assured profiles; this aligns with recent research that similarly demonstrates that the benefits of challenge orientation can be thwarted when there are similarly high rates of threat (Burns et al., 2020a).

Another interesting finding was that the highest instructional cognitive load (i.e., the lowest LRI scores) was not accompanied by the highest levels of threat orientation. Specifically, the Instructionally-Overburdened & Psychologically-Resigned students reflected lower levels of anxiety and failure avoidance goals than the Instructionally-Burdened & Psychologically-Fearful students and the Instructionally-Supported & Psychologically-Pressured students. It seems that in conditions where the instructional load is most poorly managed (evidenced by the lowest LRI scores), students may abandon any investment in the lesson. According to self-worth theory (Covington, 2000), when students abandon motivationally aversive conditions there can be an alleviation of anxiety and fear as their competence and academic self-worth are no longer "on the line" and under threat. Importantly, however, as they abandon their investment in cognitively burdensome instruction, their academic outcomes also decline—as evidenced by their significantly lower levels of persistence and significantly higher levels of disengagement.

Interestingly, the Instructionally-Overburdened & Psychologically-Resigned students and the Instructionally-Burdened & Psychologically-Fearful students were not significantly different in achievement. Even though the latter profile did not experience such depths of burdensome instruction, this did not yield an achievement advantage for them. Here we again point out the importance of assessing accompanying challenge and threat orientations to understand potentially counter-intuitive effects of instruction on achievement: in this study, it unearthed the fact that Instructionally-Overburdened students were not significantly different in achievement than the Instructionally-Burdened students. The former profile was Psychologically-Resigned whereas the latter profile was Psychologically-Fearful. Again drawing on self-worth theory (Covington, 2000), when students abandon investment in a task demand there can be a concomitant alleviation of anxiety and fear (discussed above) that may mean their performance can be on a par with students who are still invested in the task demand but who are highly anxious and fearful. This is yet another example of how dually assessing instructional cognitive load and psychological orientations can help us better understand instructional effects—namely, assessing concomitant challenge and threat orientations has allowed us to understand why two profiles who differ in instructional load are similar in achievement.

Another novel contribution by this study involved the multilevel analyses that enabled us to identify distinct types of classrooms differentiated in terms of how they varied in instructional load (LRI) and accompanying challenge and threat orientations. Here we unearthed three classroom profiles: Struggling, Striving, and Thriving classrooms. The Struggling classrooms were predominated by a majority of students experiencing significant instructional cognitive (over)load and psychological detachment or fear. In contrast, the Thriving classrooms had almost no students who were cognitively (over)loaded and a majority of students with adaptive challenge orientations. These two classroom profiles may be considered somewhat predictable from a binary perspective, but the third classroom profile (the Striving classroom) was more nuanced and represents both cautionary and aspirational possibilities: cautionary in the sense that if not instructionally- and psychologically-supported, these Striving classrooms risk devolving to Struggling classrooms—but aspirational in that if they are better instructionally- and psychologically-supported, they can elevate to Thriving classrooms. Where the Striving and Thriving classrooms seemed to differ most was in the number of Psychologically-Self-Assured and Psychologically-Composed students (43% of Thriving classrooms; 26% of Striving classrooms)—the implication being that educators would do well to shift students "up" from the Psychologically-Pressured profile to the Psychologically-Self-Assured and Psychologically-Composed profiles. How they can do this is now the focus of discussion.

Implications for Instructional Assessment, Evaluation, and Practice

The findings of this investigation hold implications for instructional assessment, evaluation, and practice. For instructional assessment and evaluation, the study has further demonstrated the validity of instrumentation that enables students to report on the extent to which instruction manages the cognitive load on them as they learn. The Load Reduction Instruction Scale (LRIS; and its brief form, Load Reduction Instruction Scale-Short, LRIS-S) is a student reporting tool that has been purposefully developed for in-class assessment of LRI. To date the LRI has been usefully employed in variable-centered research, and the present study has now revealed its utility in person-centered analyses. Furthermore, because the LRIS is completed in class, if enough classrooms are present in a study (as there were in our study), it can be used in multilevel analyses to gain a sense of LRI at the whole-class level. We therefore encourage the use of tools that enable in-class assessment of load-reducing instruction by students. Indeed, as Martin and Evans (2018) suggested, the LRIS may also be adapted to have teachers reflect on and attend to their own instructional practice.

Also on the matter of instructional assessment and evaluation, person-centered analyses enabled insights into how different subpopulations of students may be similar in LRI but differ in their accompanying challenge-threat orientations—and how

students may differ in LRI but be similar in challenge-threat orientations. We therefore recommend that more studies assess instructional cognitive load using person-centered approaches in order to elucidate important (and sometimes quite nuanced) subpopulations of students that would otherwise be masked in variable-centered research. This will require administering instrumentation that can assess accompanying aspects of the learner. We did so via measures of inferred challenge orientation (self-efficacy, growth goals) and threat orientation (anxiety, failure avoidance goals). However, there are other indicators of challenge and threat orientations, such as affective dimensions reflecting perceived challenge-threat (e.g., enjoyment, hope, frustration, depression, anger, boredom, etc.; Pekrun, 2006).

In terms of educational practice, because the LRIS is founded on (and assesses) an instructional framework comprising five key principles, educators can be quite specifically guided in professional learning targeting these instructional principles. Martin et al. (2020a; see also Martin, 2016; Martin and Evans, 2018, 2019) have described numerous pedagogical strategies that follow from the five principles of LRI. For example, to reduce difficulty in the initial stages of learning as appropriate to the learner's prior knowledge (principle #1), they suggest pre-testing to gain a sense of where to pitch content, pre-training, and segmenting (or, "chunking") (Pollock et al., 2002; Mayer and Moreno, 2010; Delahay and Lovett, 2019). For support and scaffolding (principle #2), suggestions include structured templates, worked examples, prompting, and advance and graphic organizers (e.g., Renkl and Atkinson, 2010; Sweller, 2012; Berg and Wehby, 2013; Renkl, 2014; Hughes et al., 2019). For sufficient practice (principle #3), deliberate practice and mental rehearsal have been recommended (e.g., Ginns, 2005; Purdie and Ellis, 2005; Nandagopal and Ericsson, 2012; Sweller, 2012). For feedback-feedforward (principle #4), corrective and improvement-oriented information has been proposed (e.g., Basso and Belardinelli, 2006; Hattie and Timperley, 2007; Shute, 2008; Hattie, 2009; Martin and Evans, 2018). For more independent and self-directed learning (principle #5), guided discovery learning has been suggested (e.g., Mayer, 2004).

There are also strategies that can foster students' challenge orientations and reduce their threat orientations. For the former, self-efficacy and growth goals were the means through which we inferred challenge orientation, and these have distinct practice implications. For self-efficacy, educators might encourage students to challenge any negative self-beliefs, especially when they are faced with difficult academic tasks (Wigfield and Tonks, 2002). Reminding students of their strengths and reiterating what they have already learned can also enhance self-efficacy (Higgins et al., 2001; Martin et al., 2019). Regarding growth goals, intervention research has demonstrated that encouraging students to set self-improvement targets (personal best goals) and teaching them how to strive to meet these targets are successful strategies (e.g., Martin and Elliot, 2016b; Ginns et al., 2018).

Anxiety and failure avoidance goals were the means through which we inferred students' threat orientation, and these also have distinct practice implications. For anxiety, there are three types of programs that tend to be offered in schools: universal programs targeting all students, selective programs targeting students at-risk of anxiety at clinical levels, and specific programs targeting students who have clinical levels of anxiety (Martin et al., 2021). Within each of these programs, cognitive–behavioral approaches tend to be successful (Neil and Christensen, 2009); here, students are specifically taught cognitive and behavioral strategies for anxiety reduction, especially for times and circumstances when anxiety is likely to strike. The use of mindfulness techniques by educators with students is another suggested strategy to reduce anxiety. In similar vein, growth mindset intervention has been found to improve individuals' stress and threat appraisals (Yeager et al., 2016). Mindfulness intervention benefits for students with negative self-beliefs have also been highlighted in several studies and reviews (Weare, 2013; Sibinga et al., 2015; McKeering and Hwang, 2019). To address students' inclination to adopt failure avoidance goals, educators are urged to reduce students' fear of failure (Covington, 2000; Martin and Marsh, 2003; Martin, 2007, 2009). Practical strategies to do this include promoting the belief that effort underpins self-improvement and does not imply a lack of ability or intelligence and making it clear that mistakes can be important ingredients for future success and do not reflect poorly on one's self-worth (Covington, 2000; Martin and Marsh, 2003).

Limitations and Future Directions

When interpreting findings there are some limitations worth noting and which have implications for future research. First, this study relied on student reports of LRI, via the LRIS. Although the validity of this methodology has previously been demonstrated (e.g., Martin and Evans, 2018) and the psychometrics in the present study were acceptable, future research might include additional indicators such as observer ratings or self-reports by teachers to triangulate with student ratings. Second, we used the short form of the LRIS, which meant we could not estimate latent profiles on the basis of the 5 LRI principles considered separately. Future research should consider this possibility and also (using the long form) look to estimate classroom profiles (L2) characterized by different levels on these 5 principles (rather than reflecting different frequencies of the L1 profiles). For example, starting from multilevel CFA models, L2 factor scores (corrected for inter-rater disagreement) can be saved, enabling more objective ratings of the classroom. Then the L1 and L2 factor scores from this model can be used to separately estimate L1 and L2 profiles. Third, there may be instructional principles that effectively manage cognitive load on learners, but which are not in the LRI framework. To the extent additional principles are identified and can be validly assessed, we recommend including them in future research. Fourth, our data were cross-sectional which means, for example, that we were unable to determine causal ordering between the profiles and the outcomes, nor whether student and classroom profile memberships change over time. Longitudinal data and modeling (e.g., latent transition analysis) will be an important avenue in future research (Collie et al., 2020). Fifth, our study included self-efficacy and growth goals to infer challenge orientation and anxiety and failure avoidance goals to infer threat orientation. There is a need for research that assesses other indicators of challenge and threat to

test the generality of our findings. For example, testing affective dimensions of perceived challenge-threat such as enjoyment, frustration, anger, boredom, etc. (Pekrun, 2006) and other challenge/approach-oriented goals such as mastery goals (Elliot, 2006) may be illuminating. There may also be potential gains in harnessing bio-psychological measures of challenge and threat in order to access real-time and more objective measures for further triangulation (Uphill et al., 2019; Martin et al., 2021). Neuro-psychological measures may additionally provide real-time indicators of experienced cognitive load. These may have the potential to deepen evaluation and understanding of LRI and its associations with challenge and threat demonstrated in this research (Berka et al., 2007; Anderson et al., 2011; Delahunty et al., 2018). Sixth, our research took place in science which is a challenging school subject (Coe et al., 2008) and one in which many students can struggle (Office of the Chief Scientist, 2014). Threat orientation may be disproportionately salient in such subjects. There is a need to explore the generalizability of our findings to other school subjects. Indeed, there is a need for research in non-science school subjects because it may be that science is more (or less) amenable to LRI. Finally, when testing for profiles in which accompanying indicators are hypothesized to be present, future research might give greater attention to real-time research methodologies. The *in-situ* dimensions of students' science engagement have been emphasized by researchers (e.g., Schneider et al., 2016) and the empirical yields of real-time engagement research has been demonstrated in other STEM subjects such as mathematics (Martin et al., 2020b).

CONCLUSION

Instructional cognitive load is perceived and experienced in different ways by different students. Some students perceive cognitive load in an approach- and challenge-oriented way, while other students perceive cognitive load in an avoidant- and threat-oriented way. To better understand instructional cognitive load, it is important to assess students' experiences of this load in the context of their accompanying psychological challenge and threat orientations. The present study did so using multilevel latent profile analysis and identified numerous instructional-psychological profiles among students and also salient instructional-psychological profiles among classrooms. These profiles were further illuminated through their associations with student- and classroom-level persistence, disengagement, and achievement. The findings of this investigation have demonstrated that assessing instructional cognitive load in the context of students' accompanying psychological orientations can reveal unique insights about students' learning experiences and about important differences between classrooms in terms of the instructional cognitive load that is present.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because they are part of an industry research partnership; consent from participants to share the dataset is not available; summative data (e.g., correlation matrix with standard deviations) are available here and elsewhere to enable analyses. Requests to access the datasets should be directed to Andrew J. Martin, andrew.martin@unsw.edu.au.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by UNSW Human Ethics Committee. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

AM led research design, data analysis, and report writing. PG and RC assisted with data analysis and report writing. EB and RK assisted with research design, interpretation of findings, and report writing. VM-S assisted with report writing. JP assisted with research design and report writing. All authors contributed to the article and approved the submitted version.

FUNDING

This study was funded by the Australian Research Council (Grant #LP170100253) and The Future Project at The King's School.

ACKNOWLEDGMENTS

The authors thank the participating schools for assisting with data collection and Carolyn Imre and Brad Papworth for advice on study design and analysis.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.656994/full#supplementary-material

REFERENCES

Anderson, E. W., Potter, K. C., Matzen, L. E., Shepherd, J. F., Preston, G. A., and Silva, C. T. (2011). A user study of visualization effectiveness using EEG and cognitive load. *Comp. Graph. Forum* 30, 791–800. doi: 10.1111/j.1467-8659.2011.01928.x

Arbuckle, J. L. (1996). "Full information estimation in the presence of incomplete data," in *Advanced Structural Equation Modeling: Issues and Techniques,* eds G. A. Marcoulides and R. E. Schumacker (Lawrence Erlbaum Associates),243–278.

Australian Bureau of Statistics (2019). *Schools Australia*. ABS.

Bandura, A. (1997). *Self-Efficacy: The Exercise of Control*. Freeman.

Basso, D., and Belardinelli, M. O. (2006). The role of the feedforward paradigm in cognitive psychology. *Cogn. Process.* 7, 73–88. doi: 10.1007/s10339-006-0034-1

Bauer, D. J., and Curran, P. J. (2004). The integration of continuous and discrete latent variable models: potential problems and promising opportunities. *Psychol. Methods* 9, 3–29. doi: 10.1037/1082-989X.9.1.3

Berg, J. L., and Wehby, J. (2013). Preteaching strategies to improve student learning in content area classes. *Interv. School Clinic* 49, 14–20. doi: 10.1177/1053451213480029

Berka, C., Levendowski, D. J., Lumicao, M. N., Yau, A., Davis, G., Zivkovic, V. T., et al. (2007). EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks. *Aviat. Space Environ. Med.* 78, B231–B244.

Britton, J. C., Lissek, S., Grillon, C., Norcross, M. A., and Pine, D. S. (2011). Development of anxiety: the role of threat appraisal and fear learning. *Depress. Anxiety* 28, 5–17. doi: 10.1002/da.20733

Burns, E. C., Martin, A. J., and Collie, R. J. (2018). Adaptability, personal best (PB) goal setting, and gains in students' academic outcomes: a longitudinal examination from a social cognitive perspective. *Contemp. Educ. Psychol.* 53, 57–72. doi: 10.1016/j.cedpsych.2018.02.001

Burns, E. C., Martin, A. J., and Collie, R. J. (2019). Understanding the role of personal best (PB) goal setting in students' declining engagement: a latent growth model. *J. Educ. Psychol.* 111, 557–572. doi: 10.1037/edu0000291

Burns, E. C., Martin, A. J., Kennett, R. K., Pearson, J., and Munro-Smith, V. (2020a). Optimizing science self-efficacy: a multilevel examination of the moderating effects of anxiety on the relationship between self-efficacy and achievement in science. *Contemp. Educ. Psychol.* doi: 10.1016/j.cedpsych.2020.101937

Burns, E. C., Martin, A. J., Mansour, M., Anderson, M., Gibson, R., and Liem, G. A. D. (2020b). Motivational processes that support arts participation: An examination of goal orientations and aspirations. *Psychol. Aesthet. Creat. Arts* 14, 384–400. doi: 10.1037/aca0000242

Campbell, D. T., and Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol. Bull.* 56, 81–105. doi: 10.1037/h0046016

Chandler, P., and Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cogn. Instruct.* 8, 293–332. doi: 10.1207/s1532690xci0804_2

Chen, O., Kalyuga, S., and Sweller, J. (2017). The expertise reversal effect is a variant of the more general element interactivity effect. *Educ. Psychol. Rev.* 29, 393–405. doi: 10.1007/s10648-016-9359-1

Coe, R., Searle, J., Barmby, P., Jones, K., and Higgins, S. (2008). *Relative Difficulty of Examinations in Different Subjects. Report for SCORE (Science Community Supporting Education)*. Centre for Evaluation and Modelling; Durham University.

Collie, R. J., Malmberg, L.-E., and Martin, A. J., Sammons, P., and Morin, A. (2020). A multilevel person-centered examination of teachers' workplace demands and resources: links with work-related well-being. *Front. Psychol.* 11:626. doi: 10.3389/fpsyg.2020.00626

Collie, R. J., Shapka, J. D., Perry, N. E., and Martin, A. J. (2015). Teachers' beliefs about social-emotional learning: identifying teacher profiles and their relations with job stress and satisfaction. *Learn. Instruct.* 39, 148–157. doi: 10.1016/j.learninstruc.2015.06.002

Covington, M. V. (2000). Achievement: an integrative review. *Annu. Rev. Psychol.* 51, 171–200. doi: 10.1146/annurev.psych.51.1.171

Delahay, A. B., and Lovett, M. C. (2019). "Distinguishing two types of prior knowledge that support novice learners," in *Proceedings of the 41st Annual Meeting of the Cognitive Science Society, CogSci,* eds A. Goel, C. M. Seifert, and C. Freska Goel (Montreal, QC).

Delahunty, T., Seery, N., and Lynch, R. (2018). Exploring the use of electroencephalography to gather objective evidence of cognitive processing during problem solving. *J. Sci. Educ. Technol.* 27, 114–130. doi: 10.1007/s10956-017-9712-2

Elliot, A., Murayama, K., Kobeisy, A., and Lichtenfeld, S. (2015). Potential-based achievement goals. *Br. J. Educ. Psychol.* 85, 192–206. doi: 10.1111/bjep.12051

Elliot, A. J. (2006). The hierarchical model of approach-avoidance motivation. *Motiv. Emot.* 30, 111–116. doi: 10.1007/s11031-006-9028-7

Elliot, A. J., Murayama, K., and Pekrun, R. (2011). A 3 × 2 achievement goal model. *J. Educ. Psychol.* 103, 632–648. doi: 10.1037/a0023952

Evans, P., and Martin, A. J. (Submitted). Load reduction instruction: Multilevel effects on motivation, engagement, and achievement in mathematics.

Ghaderyan, P., Abbasi, A., and Ebrahimi, A. (2018). Time-varying singular value decomposition analysis of electrodermal activity: a novel method of cognitive load estimation. *Measurement* 126, 102–109. doi: 10.1016/j.measurement.2018.05.015

Ginns, P. (2005). Meta-analysis of the modality effect. *Learn. Instruct.* 15, 313–331. doi: 10.1016/j.learninstruc.2005.07.001

Ginns, P., Martin, A. J., Durksen, T. L., Burns, E. C., and Pope, A. (2018). Personal Best (PB) goal-setting enhances arithmetical problem-solving. *Aust. Educ. Res.* 45, 533–551. doi: 10.1007/s13384-018-0268-9

Goldstein, H. (2003). *Multilevel Statistical Models, 3rd Edn*. Hodder Arnold.

Green, J., Martin, A. J., and Marsh, H. W. (2007). Motivation and engagement in English, mathematics and science high school subjects: towards an understanding of multidimensional domain specificity. *Learn. Ind. Diff.* 17, 269–279. doi: 10.1016/j.lindif.2006.12.003

Hattie, J. (2009). *Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement*. Routledge.

Hattie, J., and Timperley, H. (2007). The power of feedback. *Rev. Educ. Res.* 77, 81–112. doi: 10.3102/003465430298487

Higgins, E. T., Friedman, R. S., Harlow, R. E., Idson, L. C., Ayduk, O. N., and Taylor, A. (2001). Achievement orientations from subjective histories of success: promotion pride versus prevention pride. *Eur. J. Soc. Psychol.* 31, 3–23. doi: 10.1002/ejsp.27

Howard, S. J., Burianov,á, H., Ehrich, J., Kervin, L., Calleia, A., Barkus, E., et al. (2015). Behavioral and fMRI evidence of the differing cognitive load of domain-specific assessments. *Neuroscience* 297, 38–46. doi: 10.1016/j.neuroscience.2015.03.047

Hughes, M. D., Regan, K. S., and Evmenova, A. (2019). A computer-based graphic organizer with embedded self-regulated learning strategies to support student writing. *Intervent. School Clinic.* 55, 13–22. doi: 10.1177/1053451219833026

Ikehara, C. S., and Crosby, M. E. (2005). "Assessing cognitive load with physiological sensors," in *Proceedings of the 38th Annual Hawaii International Conference on System Sciences* (IEEE), 295a−295a.

Kalyuga, S., Rikers, R., and Paas, F. (2012). Educational implications of expertise reversal effects in learning and performance of complex cognitive and sensorimotor skills. *Educ. Psychol. Rev.* 24, 313–337. doi: 10.1007/s10648-012-9195-x

Klepsch, M., and Seufert, T. (2020). Understanding instructional design effects by differentiated measurement of intrinsic, extraneous, and germane cognitive load. *Instruct. Sci.* 48, 45–77. doi: 10.1007/s11251-020-09502-9

Krell, M. (2017). Evaluating an instrument to measure mental load and mental effort considering different sources of validity evidence. *Cogent Educ.* 4:1280256. doi: 10.1080/2331186X.2017.1280256

Lau, S., and Nie, Y. (2008). Interplay between personal goals and classroom goal structures in predicting student outcomes: a multilevel analysis of person-context interactions. *J. Educ. Psychol.* 100, 15–29. doi: 10.1037/0022-0663.100.1.15

Lazarus, R. S., and Folkman, S. (1984). *Stress, Appraisal, and Coping*. Springer.

Leppink, J., Paas, F., Van der Vleuten, C. P., Van Gog, T., and Van Merriënboer, J. J. (2013). Development of an instrument for measuring different types of cognitive load. *Behav. Res. Methods* 45, 1058–1072. doi: 10.3758/s13428-013-0334-1

Litalien, D., Gillet, N., Gagné, M., Ratelle, C. F., and Morin, A. J. S. (2019). Self-determined motivation profiles among undergraduate students: A robust test of profile similarity as a function of gender and age. *Learn. Ind. Diff.* 70, 39–52. doi: 10.1016/j.lindif.2019.01.005

Marsh, H. W. (2002). A multidimensional physical self-concept: a construct validity approach to theory, measurement, and research. *Psychol. J. Hellen. Psychol. Soc.* 9, 459–493.

Marsh, H. W., Lüdtke, O., Nagengast, B., Trautwein, U., Morin, A. J., Abduljabbar, A. S., et al. (2012). Classroom climate and contextual effects: conceptual and methodological issues in the evaluation of group-level effects. *Educ. Psychol.* 47, 106–124. doi: 10.1080/00461520.2012.670488

Marsh, H. W., Lüdtke, O., Robitzsch, A., Trautwein, U., Asparouhov, T., Muthén, B., et al. (2009). Doubly-latent models of school contextual effects: integrating multilevel and structural equation approaches to control measurement and sampling error. *Multivariate Behav. Res.* 44, 764–802. doi: 10.1080/00273170903333665

Marsh, H. W., Martin, A. J., and Hau, K. T. (2006). "A multiple method perspective on self-concept research in educational psychology: a construct validity approach," in *Handbook of Multimethod Measurement in Psychology*, eds M. Eid and E. Diener (American Psychological Association Press), 441–456. doi: 10.1037/11383-030

Martin, A. J. (2007). Examining a multidimensional model of student motivation and engagement using a construct validation approach. *Br. J. Educ. Psychol.* 77, 413–440. doi: 10.1348/000709906X118036

Martin, A. J. (2009). Motivation and engagement across the academic lifespan: a developmental construct validity study of elementary school, high school, and university/college students. *Educ. Psychol. Meas.* 69, 794–824. doi: 10.1177/0013164409332214

Martin, A. J. (2016). Using Load Reduction Instruction (LRI) to boost motivation and engagement. *Br. Psychol. Soc*.

Martin, A. J., Anderson, J., Bobis, J., Way, J., and Vellar, R. (2012). Switching on and switching off in mathematics: an ecological study of future intent and disengagement amongst middle school students. *J. Educ. Psychol.* 104, 1–18. doi: 10.1037/a0025988

Martin, A. J., and Elliot, A. J. (2016a). The role of personal best (PB) and dichotomous achievement goals in students' academic motivation and engagement: a longitudinal investigation. *Educ. Psychol.* 36, 1285–1302. doi: 10.1080/01443410.2015.1093606

Martin, A. J., and Elliot, A. J. (2016b). The role of personal best (PB) goal setting in students' academic achievement gains. *Learn. Ind. Diff.* 45, 222–227. doi: 10.1016/j.lindif.2015.12.014

Martin, A. J., and Evans, P. (2018). Load reduction instruction: exploring a framework that assesses explicit instruction through to independent learning. *Teach. Teach. Educ.* 73, 203–214. doi: 10.1016/j.tate.2018.03.018

Martin, A. J., and Evans, P. (2019). "Load reduction instruction: Sequencing explicit instruction and guided discovery to enhance students' motivation, engagement, learning, and achievement," in *Advances in Cognitive Load Theory: Rethinking Teaching*, eds S. Tindall-Ford, S. Agostinho, and J. Sweller (Routledge). 15–29.

Martin, A. J., Ginns, P., Burns, E., Kennett, R., and Pearson, J. (2020a). Load reduction instruction in science and students' science engagement and science achievement. *J. Educ. Psychol*. doi: 10.1037/edu0000552

Martin, A. J., Kennett, R., Pearson, J., Mansour, M., Papworth, B., and Malmberg, L.-E. (2021). Challenge and threat appraisals in high school science: investigating the roles of psychological and physiological factors. *Educ. Psychol*. doi: 10.1080/01443410.2021.1887456

Martin, A. J., and Liem, G. A. (2010). Academic personal bests (PBs), engagement, and achievement: a cross-lagged panel analysis. *Learn. Ind. Diff.* 20, 265–270. doi: 10.1016/j.lindif.2010.01.001

Martin, A. J., Malmberg, L.-E., Kennett, R., Mansour, M., Papworth, B., and Pearson, J. (2019). What happens when students reflect on their self-efficacy during a test? Exploring test experience and test outcome in science. *Learn. Ind. Diff.* 73, 59–66. doi: 10.1016/j.lindif.2019.05.005

Martin, A. J., Mansour, M., and Malmberg, L.-E. (2020b). What factors influence students' real-time motivation and engagement? An experience sampling study of high school students using mobile technology. *Educ. Psychol.* 40, 1113–1135. doi: 10.1080/01443410.2018.1545997

Martin, A. J., and Marsh, H. W. (2003). Fear of failure: friend or foe? *Aust. Psychol.* 38, 31–38. doi: 10.1080/00050060310001706997

Mayer, R. E. (2004). Should there be a three-strikes rule against pure discovery learning? The case for guided methods of instruction. *Am. Psychol.* 59, 14–19. doi: 10.1037/0003-066X.59.1.14

Mayer, R. E., and Moreno, R. (2010). "Techniques that reduce extraneous cognitive load and manage intrinsic cognitive load during multimedia learning," in *Cognitive Load Theory,* eds J. L. Plass, R. Moreno, and R. Brunken (Cambridge University Press), 131–152. doi: 10.1017/CBO9780511844744.009

McKeering, P., and Hwang, Y. S. (2019). A systematic review of mindfulness-based school interventions with early adolescents. *Mindfulness* 10, 593–610. doi: 10.1007/s12671-018-0998-9

McLarnon, M. J. W., and O'Neill, T. A. (2018). Extensions of auxiliary variable approaches for the investigation of mediation, moderation, and conditional effects in mixture models. *Organ. Res. Methods* 21, 955–982. doi: 10.1177/1094428118770731

McNeish, D. M. (2014). Modeling sparsely clustered data: design-based, model-based, and single-level methods. *Psychol. Methods* 19, 552–563. doi: 10.1037/met0000024

Moors, A., Ellsworth, P. C., Scherer, K. R., and Frijda, N. H. (2013). Appraisal theories of emotion: state of the art and future development. *Emot. Rev.* 5, 119–124. doi: 10.1177/1754073912468165

Morin, A. J. S., Boudrias, J. S., Marsh, H. W., McInerney, D. M., Dagenais-Desmarais, V., and Litalien, D. (2017). Complementary variable- and person-centered approaches to the dimensionality of psychometric constructs: Approaches to psychological wellbeing at work. *J. Bus. Psychol.* 32, 395–419. doi: 10.1007/s10869-016-9448-7

Morin, A. J. S., and Litalien, D. (2017). *Longitudinal Tests of Profile Similarity and Latent Transition Analyses*. Montreal, QC: Substantive Methodological Synergy Research Laboratory.

Morin, A. J. S., Meyer, J. P., Creusier, J., and Biétry, F. (2016). Multiple-group analysis of similarity in latent profile solutions. *Organ. Res. Methods* 19, 231–254. doi: 10.1177/1094428115621148

Muthén, L. K., and Muthén, B. O. (2017). *Mplus [computer software]*. Muthén & Muthén.

Nandagopal, K., and Ericsson, K. A. (2012). "Enhancing students' performance in traditional education: implications from the expert performance approach and deliberate practice," in *APA Educational Psychology Handbook,* eds K. R. Harris, S. Graham, and T. Urdan (American Psychological Association), 257–293. doi: 10.1037/13273-010

Neil, A. L., and Christensen, H. (2009). Efficacy and effectiveness of school-based prevention and early intervention programs for anxiety. *Clin. Psychol. Rev.* 29, 208–215. doi: 10.1016/j.cpr.2009.01.002

OECD. (2020). *Education GPS*. OECD

Office of the Chief Scientist (2014). *Benchmarking Australian Science*. Canberra: Australian Government.

Paas, F., Renkl, A., and Sweller, J. (2003). Cognitive load theory and instructional design: recent developments. *Educ. Psychol.* 38, 1–4. doi: 10.1207/S15326985EP3801_1

Paas, F. G., Van Merriënboer, J. J., and Adam, J. J. (1994). Measurement of cognitive load in instructional research. *Percept. Motor Skills* 79, 419–430. doi: 10.2466/pms.1994.79.1.419

Pekrun, R. (2006). The control-value theory of achievement emotions: assumptions, corollaries, and implications for educational research and practice. *Educ. Psychol. Rev.* 18, 315–341. doi: 10.1007/s10648-006-9029-9

Pollock, E., Chandler, P., and Sweller, J. (2002). Assimilating complex information. *Learn. Instruct.* 12, 61–86. doi: 10.1016/S0959-4752(01)00016-0

Purdie, N., and Ellis, L. (2005). *A Review of the Empirical Evidence Identifying Effective Interventions and Teaching Practices for Students With Learning Difficulties in Years 4, 5, and 6*. Australian Council for Educational Research. Available online at: https://research.acer.edu.au/tll_misc/7

Putwain, D. W., Remedios, R., and Symes, W. (2015). Experiencing fear appeals as a challenge or a threat influences attainment value and academic self-efficacy. *Learn. Instruct.* 40, 21–28. doi: 10.1016/j.learninstruc.2015.07.007

Putwain, D. W., and Symes, W. (2014). Subjective value and academic self-efficacy: the appraisal of fear appeals used prior to a high-stakes test as threatening or challenging. *Soc. Psychol. Educ.* 17, 229–248. doi: 10.1007/s11218-014-9249-7

Putwain, D. W., and Symes, W. (2016). The appraisal of value-promoting messages made prior to a high-stakes mathematics examination: the interaction of message-focus and student characteristics. *Soc. Psychol. Educ.* 19, 325–343. doi: 10.1007/s11218-016-9337-y

Putwain, D. W., Symes, W., and Wilkinson, H. M. (2017). Fear appeals, engagement, and examination performance: the role of challenge and threat appraisals. *Br. J. Educ. Psychol.* 87, 16–31. doi: 10.1111/bjep.12132

Raudenbush, S. W., and Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods, 2nd Edn*. Sage.

Raykov, T., and Marcoulides, G. A. (2004). Using the delta method for approximate interval estimation of parameter functions in SEM. *Struct. Equat. Model.* 11, 621–637. doi: 10.1207/s15328007sem1104_7

Renkl, A. (2014). Toward an instructionally oriented theory of example-based learning. *Cogn. Sci.* 38, 1–37. doi: 10.1111/cogs.12086

Renkl, A., and Atkinson, R. K. (2010). "Learning from worked-out examples and problem solving," in *Cognitive Load Theory,* eds J. L. Plass, R.

Moreno, and R. Brunken (Cambridge University Press), 91–108. doi: 10.1017/CBO9780511844744.007

Rogat, T. K., and Linnenbrink-Garcia, L. (2019). Demonstrating competence within one's group or in relation to other groups: a person-oriented approach to studying achievement goals in small groups. *Contemp. Educ. Psychol.* 59:101781. doi: 10.1016/j.cedpsych.2019.101781

Roseman, I. J., and Smith, C. A. (2001). "Appraisal theory: overview, assumptions, varieties, controversies," in *Series in Affective Science. Appraisal Processes in Emotion: Theory, Methods, Research,* eds K. R. Scherer, A. Schorr, and T. Johnstone (Oxford University Press), 3–19.

Rosenshine, B. V. (2009). "The empirical support for direct instruction," in *Constructivist Instruction: Success or Failure?,* eds S. Tobias and T. M. Duffy (Routledge) 201–220.

Schneider, B., Krajcik, J., Lavonen, J., Salmela-Aro, K., Broda, M., Spicer, J., et al. (2016). Investigating optimal learning moments in US and Finnish science classes. *J. Res. Sci. Teach.* 53, 400–421. doi: 10.1002/tea.21306

Schunk, D. H., and DiBenedetto, M. K. (2014). "Academic self-efficacy," in *Handbook of Positive Psychology in Schools,* eds M. J. Furlong, R. Gilman, and E. S. Huebner (Elsevier), 115–521.

Shute, V. J. (2008). Focus on formative feedback. *Rev. Educ. Res.* 78, 153–189. doi: 10.3102/0034654307313795

Sibinga, E. M., Webb, L., Ghazarian, S. R., and Ellen, J. M. (2015). School-based mindfulness instruction: An RCT. *Pediatrics* 137, 1–8. doi: 10.1542/peds.2015-2532

Sweller, J. (2004). Instructional design consequences of an analogy between evolution by natural selection and human cognitive architecture. *Instruct. Sci.* 32, 9–31. doi: 10.1023/B:TRUC.0000021808.72598.4d

Sweller, J. (2012). "Human cognitive architecture: why some instructional procedures work and others do not," in *APA Educational Psychology Handbook*, K. R. Harris, S. Graham, and T. Urdan (American Psychological Association), 295–325. doi: 10.1037/13273-011

Sweller, J., Ayres, P., and Kalyuga, S. (2011). *Cognitive Load Theory*. Springer. doi: 10.1007/978-1-4419-8126-4

Tröbst, S., Kleickmann, T., Lange-Schubert, K., Rothkopf, A., and Möller, K. (2016). Instruction and students' declining interest in science: an analysis of German fourth-and sixth-grade classrooms. *Am. Educ. Res. J.* 53, 162–193. doi: 10.3102/0002831215618662

Uphill, M. A., Rossato, C., Swain, J., and O'Driscoll, J. M. (2019). Challenge and threat: a critical review of the literature and an alternative conceptualization. *Front. Psychol.* 10:1255. doi: 10.3389/fpsyg.2019.01255

Van Yperen, N. W., Blaga, M., and Postmes, T. (2015). A meta-analysis of the impact of situationally induced achievement goals on task performance. *Hum. Perf.* 28, 165–182. doi: 10.1080/08959285.2015.1006772

Vermunt, J. K. (2010). Latent class modeling with covariates: two improved three-step approaches. *Polit. Anal.* 18, 450–469. doi: 10.1093/pan/mpq025

Wang, Q., Yang, S., Liu, M., Cao, Z., and Ma, Q. (2014). An eye-tracking study of website complexity from cognitive load perspective. *Decis. Support Syst.* 62, 1–10. doi: 10.1016/j.dss.2014.02.007

Weare, K. (2013). Developing mindfulness with children and young people: a review of the evidence and policy context. *J. Child. Serv.* 8, 141–153. doi: 10.1108/JCS-12-2012-0014

Wigfield, A., and Tonks, S. (2002). "Adolescents' expectancies for success and achievement task values during middle and high school years," in *Academic Motivation of Adolescents,* eds F. Pajares and T. Urdan (Information Age Publishing), 58–82.

Yeager, D. S., Lee, H. Y., and Jamieson, J. P. (2016). How to improve adolescent stress responses: insights from integrating implicit theories of personality and biopsychosocial models. *Psychol. Sci.* 27, 1078–1091. doi: 10.1177/0956797616649604

Yu, K., and Martin, A. J. (2014). Personal best (PB) and 'classic' achievement goals in the Chinese context: their role in predicting academic motivation, engagement, and buoyancy. *Educ. Psychol.* 34, 635–658. doi: 10.1080/01443410.2014.895297

Conflict of Interest: It is appropriate to note that one of the measures (the MES) in the study is a published instrument attracting a small fee (approx. US$110 per 1,000 respondents) part of which is put toward its ongoing development and administration and part of which is also donated to UNICEF. However, for this study, there was no fee involved for its use.

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Education

ORIGINAL RESEARCH
published: 14 July 2021
doi: 10.3389/feduc.2021.705551

Comparing Two Subjective Rating Scales Assessing Cognitive Load During Technology-Enhanced STEM Laboratory Courses

Michael Thees[1]*, *Sebastian Kapp*[1], *Kristin Altmeyer*[2], *Sarah Malone*[2], *Roland Brünken*[2] and *Jochen Kuhn*[1]

[1]Department of Physics, Physics Education Research Group, Technische Universität Kaiserslautern, Kaiserslautern, Germany, [2]Department of Education, Saarland University, Saarbrücken, Germany

Cognitive load theory is considered universally applicable to all kinds of learning scenarios. However, instead of a universal method for measuring cognitive load that suits different learning contexts or target groups, there is a great variety of assessment approaches. Particularly common are subjective rating scales, which even allow for measuring the three assumed types of cognitive load in a differentiated way. Although these scales have been proven to be effective for various learning tasks, they might not be an optimal fit for the learning demands of specific complex environments such as technology-enhanced STEM laboratory courses. The aim of this research was therefore to examine and compare the existing rating scales in terms of validity for this learning context and to identify options for adaptation, if necessary. For the present study, the two most common subjective rating scales that are known to differentiate between load types (the cognitive load scale by Leppink et al. and the naïve rating scale by Klepsch et al.) were slightly adapted to the context of learning through structured hands-on experimentation where elements such as measurement data, experimental setups, and experimental tasks affect knowledge acquisition. N = 95 engineering students performed six experiments examining basic electric circuits where they had to explore fundamental relationships between physical quantities based on the observed data. Immediately after the experimentation, the students answered both adapted scales. Various indicators of validity, which considered the scales' internal structure and their relation to variables such as group allocation as participants were randomly assigned to two conditions with a contrasting spatial arrangement of the measurement data, were analyzed. For the given dataset, the intended three-factorial structure could not be confirmed, and most of the a priori-defined subscales showed insufficient internal consistency. A multitrait–multimethod analysis suggests convergent and discriminant evidence between the scales which could not be confirmed sufficiently. The two contrasted experimental conditions were expected to result in different ratings for the extraneous load, which was solely detected by one adapted scale. As a further step, two new scales were assembled based on the overall item pool and the given dataset. They revealed a three-factorial structure in accordance

OPEN ACCESS

Edited by:
Moritz Krell,
Freie Universität Berlin, Germany

Reviewed by:
Kim Ouwehand,
Erasmus University Rotterdam, Netherlands
Jeroen Van Merrienboer,
Maastricht University, Netherlands

***Correspondence:**
Michael Thees
theesm@physik.uni-kl.de

Specialty section:
This article was submitted to Assessment, Testing and Applied Measurement, a section of the journal Frontiers in Education

Received: *05 May 2021*
Accepted: *14 June 2021*
Published: *14 July 2021*

Citation:
Thees M, Kapp S, Altmeyer K, Malone S, Brünken R and Kuhn J (2021) Comparing Two Subjective Rating Scales Assessing Cognitive Load During Technology-Enhanced STEM Laboratory Courses. Front. Educ. 6:705551. doi: 10.3389/feduc.2021.705551

with the three types of load and seemed to be promising new tools, although their subscales for extraneous load still suffer from low reliability scores.

Keywords: cognitive load, differential measurement, rating scale, validity, split-attention effect, STEM laboratories, multitrait–multimethod analysis

INTRODUCTION

Experimentation in laboratory-like environments is an integral aspect of higher science education (Trumper, 2003; Hofstein and Lunetta, 2004; Lunetta et al., 2005). Guided by a predefined task, learners manipulate experimental setups and observe scientific phenomena in order to explore or verify functional relationships between specific quantities in interaction with their theoretical background (American Association of Physics Teachers, 2014; Lazonder and Harmsen, 2016). Although this inquiry-based format allows for unique hands-on learning experiences, various empirical studies revealed contrary results concerning the learning gain of laboratory courses (Volkwyn et al., 2008; Zacharia and Olympiou, 2011; de Jong et al., 2013; Wilcox and Lewandowski, 2017; Husnaini and Chen, 2019; Kapici et al., 2019). In response, technology-based approaches are applied to support students during experimentation and thereby ensure essential learning and raise the effectiveness of experimentation as a learning scenario (de Jong et al., 2013; Zacharia and de Jong, 2014; de Jong, 2019; Becker et al., 2020).

The most common way to evaluate the effectiveness of new approaches is to apply conceptual knowledge tests to measure learning gains based on content-related knowledge (Etkina et al., 2006; Vosniadou, 2008; de Jong, 2019). However, this procedure does not account for learning as a complex cognitive process. Since the focus of conceptual knowledge tests is merely on learning outcomes, it remains unclear whether and how the learning effects could be further increased and learning processes made more efficient. This gap is closed by considering cognitive load theory (CLT; Sweller et al., 1998, 2019; Sweller, 2020), which provides a useful framework to describe learning in terms of information processing and which respects human cognitive architecture as well as learners' prior knowledge and the demands of the instruction. Hence, to evaluate the effects of a learning scenario, investigations should not only solely consider the effectiveness in terms of higher scores in knowledge tests but also the efficiency in terms of an optimal level of cognitive demands. This integration of cognitive processes as a key element of learning scenarios requires sensitive and valid measurement instruments to determine the cognitive load.

CLT outlines the working memory and the long-term memory as those entities that are central for processing information and building up knowledge structures (Sweller et al., 1998, Sweller et al., 2019) called schemata (Sweller et al., 1998). Already stored knowledge can be retrieved from long-term memory to support information processing in working memory. While the long-term memory is considered permanent and unlimited in terms of capacity, working memory is limited by the number of information elements that can be processed simultaneously (Baddeley, 1992; Sweller et al., 1998, Sweller et al., 2019; Cowan, 2001). Consequently, learners cannot process information with any desired complexity, which means that to ensure successful learning, this limited capacity should be respected. Any processing of information requires mental processes that consume working memory capacity, which is called cognitive load. CLT distinguishes three types of cognitive load (Sweller, 2010; Sweller et al., 1998, Sweller et al., 2019): intrinsic cognitive load (ICL), extraneous cognitive load (ECL), and germane cognitive load (GCL). ICL is related to the complexity of the learning content and depends on the learner's prior knowledge as already built-up schemata reduce the number of elements that must be processed simultaneously in working memory. ECL refers to processes that are not essential and therefore hamper learning such as searching for relevant information within the environment or maintaining pieces of information in mind over a longer time (Mayer and Moreno, 2003). GCL represents the amount of cognitive resources devoted to processing information into knowledge structures. The amount of ECL imposed by a task affects the remaining resources that can be devoted to germane processing. Current theoretical considerations suggest that GCL cannot be essentially distinguished from ICL as both are closely related to processes of schema acquisition (Kalyuga, 2011; Jiang and Kalyuga, 2020). As a consequence, a reinterpretation of CLT as a two-factor model (ICL/ECL) is discussed. GCL is integrated into this model as a function of working memory resources needed to deal with the ICL of a task instead of representing an independent source of working memory load (Sweller, 2010; Sweller et al., 2019).

One of the main goals of CLT is to derive design guidelines for learning materials and environments that ensure that learning processes can proceed efficiently and undisturbed by irrelevant processing steps (Sweller et al., 2019). This can be achieved by removing unnecessary and distracting information as well as by a reasonable presentation format to avoid split-attention that consumes cognitive capacities and impairs essential learning (Mayer and Moreno, 1998; Ayres and Sweller, 2014). Therefore, elements of information that need to be associated with each other in learning should be presented without delay and in spatial proximity as described by the multimedia design principles of temporal and spatial contiguity (Mayer and Fiorella, 2014). These principles are empirically proven to reduce ECL and support learning in multimedia learning scenarios (Schroeder and Cenkci, 2018).

Scientific experimentation in STEM laboratory courses is assumed to be a highly complex learning scenario since learners are confronted with numerous sources of information such as experimental setups and measurement data which are presented in various representational forms. Although most of the given elements are typical features of the laboratory situation,

not all of them are essential for the learning process. As CLT is considered universal and applicable to various learning scenarios, its framework can also be applied to laboratory courses (Thees et al., 2020).

Since cognitive load has rarely been seen as a main variable to investigate the impact of hands-on laboratory courses, there existed no valid measurement instruments that 1) addressed the aforementioned characteristics of scientific hands-on experimentation including context-specific load-inducing sources and 2) provided results that allowed for a differentiated interpretation of the three load types. Former investigations by Kester et al. (Kester et al., 2005; Kester et al., 2010) used the one-item scale by Paas (1992) in the context of virtual science experiments, i.e., screen-based electricity simulations, to rate mental effort as a measure of cognitive load. There, the authors revealed higher transfer performance for learning with integrated rather than split-source formats. However, no differences concerning mental effort were found, which could be due to the limitations of the one-item cognitive load measurement (Kester et al., 2010). We intended to address this gap for real hands-on experiments by considering existing instruments that are known to differentiate load types and adapting them to fit the context of lab courses.

Even though current theoretical approaches integrate GCL in a dual intrinsic-extraneous load typology of cognitive load, Klepsch et al. (2017) argued that creating supportive learning scenarios requires a comprehensive understanding of task-related aspects of cognitive load (ICL/ECL) as well as of a learner's deliberately devoted germane resources (GCL) and their interactions. On these grounds, a differentiated measurement of cognitive load capturing its three-partite nature is still considered expedient.

The search for adequate instruments to measure the three types of cognitive load has a long history in cognitive load research. The most common approaches use subjective rating scales where participants rate their perceived cognitive load by evaluating their agreement with predefined statements (Brünken et al., 2003; Krell, 2017; Jiang and Kalyuga, 2020). There exist essentially two different rating scales that are proven to differentially measure the three types of load. These are the cognitive load scale (CLS; 10-item questionnaire) developed by Leppink et al. (2013) and the (second version of the) naïve rating scale (NRS; 8-item questionnaire) by Klepsch et al. (2017). Both scales were applied in various learning contexts (Leppink et al., 2014; Altmeyer et al., 2020; Andersen and Makransky, 2021a; Andersen and Makransky, 2021b; Becker et al., 2020; Kapp et al., 2020; Klepsch and Seufert, 2020, Klepsch and Seufert, 2021; Skulmowski and Rey, 2020; Thees et al., 2020), while the reliability of the subscales and the valid measurement of the three load types were confirmed multiple times (Klepsch et al., 2017; Becker et al., 2020; Klepsch and Seufert, 2020; Thees et al., 2020; Andersen and Makransky, 2021a; Andersen and Makransky, 2021b). However, their application in different contexts usually requires moderate adaptations.

With the objective of identifying an appropriate scale to measure the three types of cognitive load in the complex context of STEM laboratory courses, we adapted two existing cognitive load scales. We based our work on the original scales as presented in Leppink et al. (2013) and Klepsch et al. (2017) as well as former adaptations of the CLS in the target context by Thees et al. (2020). In this process, both scales were adapted regarding terminology and partly extended to take various characteristics of the laboratory environment into account. Although these adaptations are highly plausible, they require empirical, evidence-based validation of the resulting scales in the intended learning context. Accordingly, the main research question of the present study was whether the adapted scales can be considered as valid measurement instruments of cognitive load for the context of STEM laboratory courses.

Validity is defined as the appropriateness of interpreting test scores in an intended manner (Kline, 2000; AERA et al., 2011; Kane, 2013). The presented analyses followed the concepts given by the *Standards for Educational and Psychological Testing* (AERA et al., 2011) where the overall evidence for validity is based on considering multiple sources of evidence such as content, internal structure, relation to other variables, and response processes. As mentioned before, the main emphases of the application and interpretation of the scales are the suitability for the special context and the differentiated measurement of the three types of cognitive load. Based on this, the following sources of evidence were considered and evaluated during the presented analyses.

A prerequisite for interpreting test scores in the target context of STEM laboratory courses is that the items adequately represent the addressed constructs (ICL, ECL, and GCL) in terms of their formulation. In this sense, adequate items must match the sources of cognitive load that are part of STEM experiments as a learning environment. This *evidence based on content* (AERA et al., 2011) was considered during the item development, i.e., the adaptation of the original items toward the target context. In order to successfully distinguish between the three types of cognitive load, each adapted scale is expected to show a three-partite internal structure that matches the structure inherited by the original scales. This *evidence based on internal structure* (AERA et al., 2011) was considered during the analysis of the presented dataset. The simultaneous application of two adapted scales that are intended to measure the same constructs allowed for evaluating convergent and discriminant evidence to determine whether the same constructs were addressed by the respective subscale and whether different types of load could be clearly distinguished. The evaluation of properly addressing the intended constructs was further addressed by inducing group-specific differences by an external factor. By varying the presentation format of crucial information that was relevant to the learning process, ECL was varied, and the analyses evaluated whether the adapted scales could detect these induced differences. In addition, the scales should not indicate any differences in ICL since the complexity of the content and the experimental tasks as well as the representational forms were equal for both groups. Furthermore, a negative correlation between prior knowledge and ICL was expected, which is intended to verify the reduction of perceived content-related complexity due to the already built-up knowledge structures. These aspects related to an outer criterion and were considered *evidence based on relations to other variables*

(AERA et al., 2011). As both scales are applied as rating scales and the individual process of rating each item is not considered part of the analyses, *evidence based on response processes* (AERA et al., 2011) was not considered in the present analyses.

In the present study, both adapted scales were applied after learners had participated in a technology-enhanced laboratory course unit examining hands-on experiments in the context of electricity. The experimental tasks and the overall procedure followed the study design of Altmeyer et al. (2020). Participants had to explore basic physical quantities by setting up several electric circuits and observing automatically provided measurement data while manipulating fundamental parameters. To induce differences in ECL by an external factor, two experimental learning conditions were included to contrast the spatial arrangement of the learning-relevant measurement data as a between-subject factor. One group received a split-source format where the data were anchored as virtual displays to their corresponding component using augmented reality and therefore spread across the learning environment. The other group received an integrated format where the data were grouped together on a single display. Former studies in the context of hands-on electricity laboratory courses have emphasized that measurement values, which have to be compared and related to each other in order to learn successfully, should be presented in spatial proximity (Altmeyer et al., 2020; Kapp et al., 2020; Thees et al., 2020) to avoid the well-known split-attention effect (Schroeder and Cenkci, 2018). Hence, the split-source format was expected to trigger unnecessary search processes, and the corresponding group was expected to rate higher ECL than the group with the integrated format. Both groups received the same experimental tasks and equal representational forms of the data to avoid differences in the complexity of the learning material. In terms of the evaluation of validity sources, this leads to the following hypotheses.

Hypothesis based on the internal structure is as follows:

(H1) *Since both adapted scales are intended to differentiate the three types of cognitive load, confirmatory factor analyses are expected to prove their three-partite internal structure.*

Hypotheses based on relation to other variables are as follows:

(H2) *Since both adapted scales include subscales that are intended to measure the same latent variable, high correlations between corresponding subscales (convergent evidence) and low correlations between different subscales (discriminant evidence) are expected.*
(H3) *The integrated presentation of measurement data reduces perceived ECL compared to the split-source format.*
(H4) *Since the complexity of the learning material was not varied and participants were randomly assigned to the conditions, equal ratings for ICL are expected.*
(H5) *Since ICL depends on learners' prior knowledge, negative correlations between prior knowledge scores and ICL ratings are expected.*

Furthermore, insufficient evidence for the internal structure might cast doubt on the appropriateness of the respective adaptations and challenge validity evidence based on content or other variables. In reaction, the construction of a new scale based on the overall item pool is considered a useful procedure to contribute to scale development for the target context, leading to the following research question:

(RQ) *Is it possible to merge both scales into a new scale that fulfills the intended three-partite structure as well as detects the induced differences in ECL?*

MATERIALS AND METHODS

Item Development

While the NRS was already available in German (Klepsch et al., 2017), the CLS had to be translated to implement it in German university courses. We translated the scale with an emphasis on maintaining the meaning of the original items while applying comprehensible and grammatically correct formulations. We have already implemented the translated scale in previous studies (Altmeyer et al., 2020; Thees et al., 2020), where it has proven useful in principle, and we have further refined it for the present study. As both scales were not originally intended to be used in the context of STEM laboratory courses, all the items had to be adapted. The most important aspect was to emphasize the experiment itself consisting of the experimental tasks and procedures as well as all the components of the experimental setup and the learning environment, such as data displays and instruments. The adaptation intended to point out that the scales are referring to the cognitive load induced by the experimental tasks and not any accompanying activities such as pre- or posttests or preparation phases which are mandatory for graded laboratory courses. Hence, any formulations referring to general terms such as "lecture," "lesson," or "activity" were replaced by "experiment" or "experimental task." The results can be found in **Tables 1**, **2**.

Concerning the NRS (**Table 1**), the items of the ICL and GCL subscales were adapted by replacing the term "activity" as mentioned before. For the ECL subscale, the term "information" was specified as "measurement data." These data are seen as the crucial information of the scientific context and the basis for any learning process as the information about the mutual dependencies between the physical quantities of the behavior of experimental components is solely represented by the data. The 7-point Likert scale level was adopted from the original work by Klepsch et al. (2017), including the labeling of the scale range as "absolutely wrong" (left endpoint; German: "Stimme überhaupt nicht zu") and "absolutely right" (right endpoint; German: "Stimme voll zu").

Concerning the CLS (**Table 2**), the references within the items were also adjusted to the "experiment." Furthermore, for the ICL and GCL subscales, the contents of the learning scenario (formerly statistics and corresponding formulas) were replaced by "measurement procedure," "representations," and "physical

TABLE 1 | Original and adapted NRS, based on the work of Klepsch et al. (2017).

Type of load	Original scale		Adapted scale		#
	Item—German	Item—English	Item—German	Item—English	
ICL	Bei der Aufgabe musste man viele Dinge gleichzeitig im Kopf bearbeiten	For this task, many things needed to be kept in mind simultaneously	Beim Experimentieren musste man viele Dinge gleichzeitig im Kopf bearbeiten	During experimentation, many things needed to be kept in mind simultaneously	NRS-1
	Diese Aufgabe war sehr komplex	This task was very complex	Das Experimentieren war sehr komplex	Experimentation was very complex	NRS-2
ECL	Bei dieser Aufgabe ist es mühsam, die wichtigsten Informationen zu erkennen	During this task, it was exhausting to find the important information	Beim Experimentieren war es mühsam, die wichtigsten Informationen zu erkennen	During experimentation, it was exhausting to find the important information	NRS-3
	Die Darstellung bei dieser Aufgabe ist ungünstig, um wirklich etwas zu lernen	The design of this task was very inconvenient for learning	Die Darstellung der Messwerte beim Experimentieren war ungünstig um wirklich etwas zu lernen	The presentation of measurement data was very inconvenient for learning	NRS-4
	Bei dieser Aufgabe ist es schwer, die zentralen Inhalte miteinander in Verbindung zu bringen	During this task, it was difficult to recognize and link the crucial information	Beim Experimentieren war es schwierig, die richtigen Messwerte und Bauteile miteinander in Verbindung zu bringen	During experimentation, it was difficult to link appropriate data and components	NRS-5
GCL	Ich habe mich angestrengt, mir nicht nur einzelne Dinge zu merken, sondern auch den Gesamtzusammenhang zu verstehen	I made an effort, not only to understand several details but also to understand the overall context	Beim Experimentieren habe ich mich angestrengt, mir nicht nur einzelne Dinge zu merken, sondern auch den Gesamtzusammenhang zu verstehen	During experimentation, I made an effort, not only to understand several details but also to understand the overall context	NRS-6
	Es ging mir beim Bearbeiten der Lerneinheit darum, alles richtig zu verstehen	My point while dealing with the task was to understand everything correct	Es ging mir beim Experimentieren darum, alles richtig zu verstehen	My point while experimenting was to understand everything correct	NRS-7
	Die Lerneinheit enthielt Elemente, die mich unterstützten, den Lernstoff besser zu verstehen	The learning task consisted of elements supporting my comprehension of the task	Die Aufgaben, die ich während dem Experimentieren bearbeiten musste, haben mich dabei unterstützt, den Lernstoff besser zu verstehen	The experimental task supported my comprehension of the content	NRS-8

laws." This resulted in one additional item for each subscale (CLS-3 and CLS-12). Another item was added to the ICL subscale referring to the complexity of the experimental setup (CLS-4). For ECL, the term "instructions" was directed to the "experimental task" and the "work booklet." There, another item was added concerning the operation of the experimental setup (CLS-7). Hence, the original 10-item scale was expanded to a 14-item scale in order to capture various facets of the context. Furthermore, the scale range was adjusted to a six-point Likert scale. Within this step, the term "very" was excluded from each item. The labeling of the scale range was adopted from the original work by Leppink et al. (2013), ranging from "not at all" (left endpoint; German: "Trifft gar nicht zu") to "completely the case" (right endpoint; German: "Trifft voll und ganz zu").

Participants

The sample originally consisted of $N = 117$ engineering students from a medium-sized German university (approximately 14,000 students in total) who attended the same introductory physics lecture. Six of them had to be excluded due to language problems, and another 16 students had to be excluded due to missing values in the overall dataset. The remaining $N = 95$ students constitute the sample for all further analyses. Participants were randomly assigned to group 1, receiving an integrated presentation format ($N = 48$; 15% female, 81% male; age: $M = 19.8$, $SD = 1.3$; semester: $M = 1.9$, $SD = 1.3$), and group 2 ($N = 47$; 15% female, 74% male; age: $M = 20.1$, $SD = 1.5$; semester: $M = 2.3$, $SD = 1.7$), receiving a split-source presentation format. The investigation was conducted during the winter semester 2019. Participation was reimbursed with a bonus percentage of 5% for the final examination score.

Materials

During the intervention, participants performed structured physics experiments for which they had to construct several electrical circuits and analyze measurement data to derive fundamental laws for voltage and current (well known as Kirchhoff's laws), which are based on a former study by Altmeyer et al. (2020). This inquiry process was guided by structured task descriptions in which six different circuits were examined. Learners had to build up these circuits with typical educational equipment (i.e., cables, a voltage source, and resistors) based on a given circuit diagram and answered a set of single-choice items concerning the relation of voltage or amperage at all components based on the observed data. To observe a variety of data in order to derive physical laws, learners were encouraged to manipulate fundamental parameters of the experiment, i.e., the source voltage (**Figure 1**). The data were

TABLE 2 | Original and adapted CLS, based on the work of Leppink et al. (2013).

Type of load	Original scale	Translated	Adapted scale		#
	Item—English	Item—German	Item—English	Item—German	
ICL	The topic/topics covered in the activity was/were very complex	Die während der Aktivität behandelten Themen waren sehr komplex	The experiment covered topics that I perceived as complex	Die beim Experimentieren thematisierten Inhalte empfinde ich als komplex	CLS-1
	The activity covered formulas that I perceived as very complex	Die Aktivität behandelte Formeln, welche ich als sehr komplex empfand	I perceived the measurement procedure as complex	Das Aufnehmen der Messwerte habe ich als komplex empfunden	CLS-2
			The experiment covered representations that I perceived as complex	Die beim Experimentieren verwendeten Darstellungen habe ich als komplex empfunden	CLS-3
			I perceived the experimental setup as complex	Die experimentellen Aufbauten habe ich inhaltlich als komplex empfunden	CLS-4
	The activity covered concepts and definitions that I perceived as very complex	Die Aktivität behandelte Konzepte und Definitionen, welche ich als sehr komplex empfand	The experiment covered physical laws that I perceived as complex	Die beim Experimentieren betrachteten physikalischen Zusammenhänge habe ich als komplex empfunden	CLS-5
ECL	The instructions and/or explanations during the activity were very unclear	Die Arbeitsaufträge und/oder Erklärungen zur Aktivität waren sehr unklar	The instructions during the experiment were unclear	Die Arbeitsaufträge zum Experimentieren waren unklar	CLS-6
			The operation of the experimental setup was unclear	Das Bedienen des Experiments war unklar	CLS-7
	The instructions and/or explanations were, in terms of learning, very ineffective	Die Arbeitsaufträge und/oder Erklärungen waren sehr ungeeignet für den Lernfortschritt	The instruction during the experiment was, in terms of learning, ineffective	Die Arbeitsaufträge zum Experimentieren waren für meinen persönlichen Lernfortschritt ungeeignet.	CLS-8
	The instructions and/or explanations were full of unclear language	Die Arbeitsaufträge und/oder Erklärungen enthielten viele sprachliche Unklarheiten	The work booklet was full of unclear language	Die Experimentieranleitung enthielt viele sprachliche Unklarheiten	CLS-9
GCL	The activity really enhanced my understanding of the topic(s) covered	Die Aktivität hat mein Verständnis zu den betrachteten Themen wirklich gefördert	The experiment enhanced my understanding of the topic covered	Das Experimentieren heute hat mein Verständnis zu dem betrachteten Themengebiet gefördert	CLS-10
	The activity really enhanced my knowledge and understanding of statistics	Die Aktivität hat mein Wissen und Verständnis zu Statistik wirklich gefördert	The experiment enhanced my understanding of the measurement procedures	Das Experimentieren heute hat mein Verständnis zur Aufnahme von Messwerten gefördert	CLS-11
	The activity really enhanced my understanding of the formulas covered	Die Aktivität hat mein Verständnis zu den betrachteten Formeln wirklich gefördert	The experiment enhanced my understanding of the physical laws covered	Das Experimentieren heute hat mein Wissen zu den betrachteten physikalischen Zusammenhängen gefördert	CLS-12
			The experiment enhanced my understanding of the representations covered	Das Experimentieren heute hat mein Verständnis zu den verwendeten Darstellungen gefördert	CLS-13
	The activity really enhanced my understanding of concepts and definitions	Die Aktivität hat mein Verständnis zu Konzepten und Definitionen wirklich gefördert	The experiment enhanced my general understanding of physical concepts and definitions	Das Experimentieren heute hat mein allgemeines Verständnis zu physikalischen Konzepten und Definitionen gefördert	CLS-14

provided automatically via a technology-enhanced measuring system and were visualized in real time. Hence, every interaction with the experiment that led to a change in its physical properties could be immediately observed as a change in the displayed data. The experimental tasks were, in terms of the complexity of the examined circuits and the required prior knowledge, comparable to such experiments that are part of the corresponding introductory physics laboratory courses which are mandatory for university STEM programs. Hence, the learning content and the complexity of the laboratory work instructions matched the curriculum of university engineering students.

The learning environment consisted of the following: a work booklet that detailed the experimental tasks and circuit diagrams and the experimental components such as wires, a range of resistors, a voltage source, and a device that virtually displayed the automatically gathered measurement data (**Figures 2**, **3**). For group 1, the measurement data were presented in a clearly arranged matrix on a tablet display (**Figure 2**). For group 2,

1.3: Serial circuit with three different resistors

Turn off the power supply and change your serial circuit by replacing two of the resistors from the previous experiment with two new resistors with different resistances ($R_2 = 100\ \Omega$, $R_3 = 150\ \Omega$)!

(Your supervisor will check the circuit bevor you are allowed to turn on the power supply)

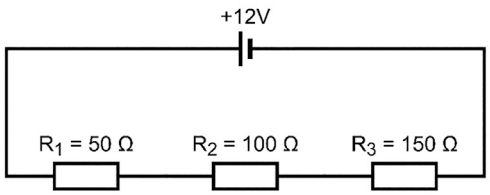

Abb. 5: Circuit diagram for experiment 1.3

Turn on the power supply and observe the behavior of voltage and current at all resistors and the power supply! Vary the power supply's voltage and examine whether this leads to changes in your observations!

[...]

FIGURE 1 | Example of the experimental task description (translated for this publication, corresponds to the circuits given in **Figures 2**, **3**).

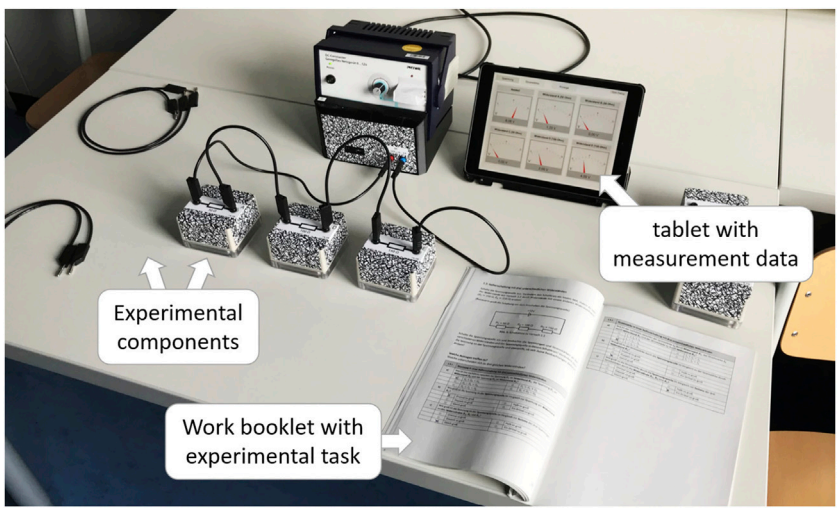

FIGURE 2 | The learning environment as experienced by group 1 (presentation of the measurement data via separate display on a tablet).

smartglasses (Microsoft HoloLens, first-generation developer edition) were used as a see-through head-mounted augmented reality device, and the measurement values were presented as virtual 3D components next to the corresponding real parts of the electric circuits within the visual field of the smartglasses using visual marker recognition (**Figure 3**). Both groups received equal representational forms, i.e., numerical values and a virtual needle deflection. Accordingly, the only difference between the two groups was the spatial arrangement of the virtual real-time measurement displays. Further information on the technical implementation of the learning environment was described by Altmeyer et al. (2020) and Kapp et al. (2020).

Both adapted subjective rating scales were applied as shown in **Tables 1**, **2** in order to measure cognitive load in a differentiated way.

Prior knowledge was determined via conceptual knowledge consisting of 10 single-choice items, which were also used in a similar form by Altmeyer et al. (2020). These items were selected from a conceptual knowledge test originally developed by Urban-Woldron and Hopf (2012) and Burde (2018) based on their compatibility with the physical concepts (i.e., voltage and current

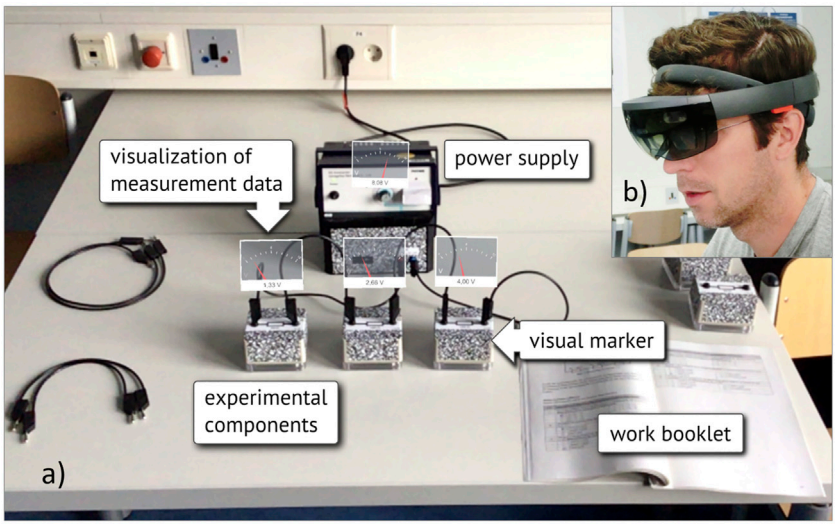

FIGURE 3 | (a) Representation of the AR view as seen through the smartglasses by participants. **(b)** Researcher wearing smartglasses.

Consider the electrical circuit below.

How large is the current through both resistors?

Answer:

- There is a current through both resistors. The current through R_1 is larger than the current through R_2.
- There is a current through both resistors. The current through R_2 is larger than the current through R_1.
- The current through both resistors is the same.
- There is a current through R_1 but not through R_2.
- There is a current through R_2 but not through R_1.

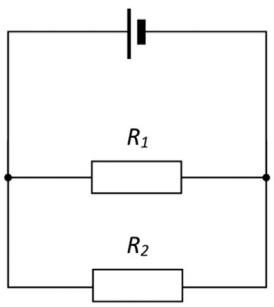

FIGURE 4 | Example of the conceptual knowledge items as presented to the participants (Urban-Woldron and Hopf, 2012, translated for this publication).

in simple circuits, Kirchhoff's laws) and the complexity of the circuits (i.e., parallel and serial circuits with few components) addressed during the experimentation phase. Five of the items were directly related to circuits that were part of the experimental tasks and were therefore considered "instruction-related" in subsuming analyses. The items were already available in German, but to match the formal representation of the instructions from the experiment, we adapted the symbols of the circuit diagrams (symbols for resistors, voltage source, etc.). An example item can be found in **Figure 4**.

Furthermore, knowledge tests concerning concrete measurement data and a usability questionnaire were applied, but these were excluded from the presented analyses (Thees et al., in preparation). Eventually, students were asked for demographic data on a voluntary basis.

Procedure

After receiving general information about the study and data protection as well as providing written consent for participation, the students completed the prior knowledge test (pretest). All the items were presented consecutively on a computer screen, and completion took approximately 10 min.

Afterward, participants were introduced to the actual learning environment, i.e., the work booklet, the experimental

components, and the operation of the displaying device (tablet or smartglasses). They were randomly assigned to one of the two intervention groups. Students using the smartglasses were able to wear their own glasses or contact lenses at the same time without any limitation.

The introduction was followed by the experimentation phase, in which students conducted the six experimental tasks as presented in the work booklet. After setting up each circuit, a supervisor checked and corrected the wiring in order to ensure safe experimentation. Students did not prepare for this experiment, and no further guidance or support was provided. The experimentation phase lasted approximately 30 min.

Subsequently, participants consecutively completed the subjective cognitive load rating scales as paper–pencil tests, starting with the adapted CLS. Each student received the same order of items, but the items were presented in a randomized order so that they were not grouped by their intended three-partite structure. Answering both questionnaires took less than 10 min.

Eventually, students answered questions concerning demographic data on a voluntary basis in a paper–pencil format.

Data Analysis

For each subscale, the mean values were calculated as scores, which were scaled to [0; 1] afterward.

To provide evidence based on the internal structure, the reliability of each subscale for both scales was calculated as internal consistency (Cronbach's alpha; α_c) with the conventional threshold of $\alpha_c = 0.70$ for acceptable reliability (Kline, 2000). In addition, confirmatory factor analyses (CFA) were conducted for both scales, evaluating their intended three-factorial structure representing the three types of cognitive load (addressing H1). There, correlations between the factors (i.e., the subscales) were allowed.

To provide evidence based on relations to other variables, both scales were compared following the procedure of a traditional multitrait–multimethod analysis (Campbell and Fiske, 1959) in order to search for convergent and discriminant evidence as each method (scale) addresses each trait (type of load). There, the correlations between the subscale scores for the two applied methods as well as the reliability scores in terms of internal consistency were considered and compared via a correlation table called MTMM matrix (addressing H2). Although there are no clear guidelines concerning thresholds, strong evidence is indicated if the correlations between the same traits measured by different methods are higher than the correlations between different traits measured by different methods. The traditional evaluation of the correlation table was complemented with a subsuming confirmatory MTMM, which was calculated as a correlated trait–correlated method model via a CFA, which allowed for correlations between all components (Eid, 2000). Furthermore, it was checked whether the scales could detect differences in the subscales between the two intervention types (grouping variable) during the study. Therefore, group-specific ECL scores were compared using a two-sided independent sample *t*-test (addressing H3). An equivalent *t*-test was conducted to compare group-specific ICL scores (addressing H4). In addition, the correlations between the ICL subscales and the score in the pretest were included. There, a negative correlation was expected as higher prior knowledge is assumed to reduce the complexity of the content due to already existing knowledge schemata (addressing H5).

TABLE 3 | Correlation table for MTMM analysis (MTMM matrix; only correlations with $p < 0.05$ are displayed).

		Method A: NRS			Method B: CLS		
	Trait	**ICL**	**ECL**	**GCL**	**ICL**	**ECL**	**GCL**
NRS	ICL	(0.55)					
	ECL	0.30[2]	(0.53)				
	GCL	0.26[2]	−0.24[2]	(0.62)			
CLS	ICL	0.53[1]	0.37[3]	*n.s*[3]	(0.85)		
	ECL	0.20[3]	0.55[1]	−0.35[3]	0.36[2]	(0.43)	
	GCL	*n.s*[3]	*n.s*[3]	0.48[1]	0.22[2]	−0.22[2]	(0.89)

n.s. = not significant ($p > 0.05$).
( ) : reliability (Cronbach's alpha).
[1]Monotrait–heteromethod coefficients.
[2]Heterotrait–monomethod coefficients.
[3]Heterotrait–heteromethod coefficients.

Going one step further, we intended to combine both scales in order to merge them into a new scale with better model fit concerning the tripartite structure (addressing RQ). This was based on an exploratory factor analysis (EFA), which was conducted using all items of both scales together. In this instance, the Kaiser–Meyer–Olkin measure revealed a good sampling adequacy with an overall $KMO = 0.79$. The individual KMO_j values were in the range of [0.65; 0.89]. Furthermore, Bartlett's test of sphericity, $\chi^2(231) = 1{,}006.4$, $p < 0.001$, revealed adequate item correlations. The scree plot and a parallel analysis were taken into account to determine the optimal number of factors, which was found to be three. Since the factors to be extracted were allowed to correlate with each other, an oblique factor rotation ("oblimin") was applied. As the intention was to find a short and concise scale, we limited the number of items included for each subscale to three. Two new models were developed based on the factor loadings and the relation to the group variable in the presented study. Both scales were evaluated by conducting a confirmatory factor analysis with their intended three-factorial structure.

In general, the significance level for type I errors was considered as $\alpha = 0.05$. For each confirmatory analysis, the following indices were applied with their corresponding cutoff values indicating acceptable model fit: the comparative fit index (CFI) and the Tucker–Lewis index (TLI), each ≥0.95, as well as the root mean square error of approximation (RMSEA) and the standardized root mean square residual (SRMR), each ≤ 0.08.

All the confirmatory analyses were conducted using the lavaan package (version 0.6-6) in the R programming language (version 3.6.0). For the EFA, the psych package (version 1.8.12) was used.

RESULTS

Validity Evidence Based on Internal Structure

The reliability analyses revealed insufficient values for the NRS, $\alpha_c(\text{ICL}) = 0.52$, $\alpha_c(\text{ECL}) = 0.53$, and $\alpha_c(\text{GCL}) = 0.62$, and mixed

TABLE 4 | Group-specific results for both adapted scales.

Scale	Subscale	Group 1 (*M(SD)*)	Group 2 (*M(SD)*)	*t*-test		
				t	*df*	*p*
NRS	ICL	0.25 (0.15)	0.25 (0.15)	0.00	93.0	1
	ECL	0.14 (0.14)	0.21 (0.17)	2.17	89.3	0.03
	GCL	0.72 (0.19)	0.71 (0.16)	-0.43	91.7	0.67
CLS	ICL	0.20 (0.14)	0.21 (0.14)	0.12	92.9	0.91
	ECL	0.13 (0.09)	0.14 (0.11)	0.44	87.6	0.66
	GCL	0.68 (0.19)	0.67 (0.21)	-0.11	91.0	0.92

results for the CLS, α_c(ICL) = 0.86, α_c(ECL) = 0.43, and α_c(GCL) = 0.90. Concerning the NRS, all the subscales did not reach the common threshold of $\alpha_c = 0.70$. In contrast, the subscales of the CLS for ICL and GCL showed satisfying results, but not for ECL.

The subsuming CFA also revealed no clear results. Concerning the NRS, the model fit indices did not reach the conventional thresholds, CFI = 0.83, TLI = 0.72, RMSEA = 0.11, and SRMR = 0.09. Concerning the CLS, RMSEA = 0.07 indicated an acceptable model fit, while the other indices narrowly missed the range for acceptable values, CFI = 0.94, TLI = 0.93, and SRMR = 0.09. In sum, there was no consistent indication of an acceptable model fit for both scales concerning the assumed structure with three inherent factors, which contradicts Hypothesis 1.

Validity Evidence Based on Relations to Other Variables

In order to compare the behavior of both adapted scales in terms of an MTMM approach, a correlation table based on Pearson's correlation was calculated (MTMM matrix; **Table 3**). Here, the correlations between the two methods concerning each trait (monotrait–heteromethod coefficients) became significant ($p < 0.05$) with a range of $r = 0.48$ to $r = 0.55$ (Cohen, 1988), indicating convergent evidence between the two scales. These correlations were higher than those significant correlations between different traits measured by different methods (heterotrait–heteromethod coefficients), emphasizing discriminant evidence. The same results were found concerning the correlations between different traits measured by the same method (heterotrait–monomethod coefficients), which were also lower than the monotrait–heteromethod coefficients. Furthermore, the patterns (ranks and sign of correlations) of the monomethod–heterotrait blocks were comparable for both methods. In contrast, the reliability values (Cronbach's α_c) showed high variance. In sum, based on the correlation table (**Table 3**), these findings emphasized convergent and discriminant evidence.

The subsuming confirmatory MTMM analysis revealed acceptable values for RMSEA = 0.06 and SRMR = 0.08. In contrast, CFI = 0.93 and TLI = 0.91 were slightly below the range for acceptable model fit.

Table 4 shows the group-dependent scores for each subscale. The results from the independent-sample *t*-test revealed for the adapted NRS a significant difference in favor of group 1 (lower ECL) in accordance with Hypothesis 3, while the CLS showed no group-specific differences. However, both NRS and CLS indicate no differences between groups concerning ICL in accordance with Hypothesis 4. Details of the test statistics can be found in **Table 4**.

Furthermore, there were no significant correlations between the pretest results and the ICL-related subscales, both for the full pretest scores, $r = -0.12$ and $p = 0.25$ for the NRS, $r = -0.02$ and $p = 0.82$ for the CLS, and the intervention-related items, $r = -0.07$ and $p = 0.51$ for the NRS, $r = 0.01$ and $p = 0.89$ for the CLS. These results contradict Hypothesis 5.

Evaluation of Combined Scales

All the items of both scales were taken into account to merge them into a new scale. First, an exploratory factor analysis was conducted to evaluate which items group together. Both the scree plot and a parallel analysis indicate a three-factorial structure. The items with the highest loading indicate conformity with the types of load known from theory, although some items with lower loadings are not grouped in accordance with their intended position. **Table 5** displays the extracted factor loadings.

For the first new model (referred to as model 1), the three items with the highest (positive) loadings were included because they represent their respective factor in a reliable manner. Hence, the ICL consisted of the items CLS-2, CLS-3, and CLS-4, the ECL subscale consisted of CLS-9, NRS-4, and CLS-6, and the GCL subscale consisted of CLS-10, CLS-12, and CLS-13. The subsuming CFA revealed adequate to good model fit, CFI = 0.98, TLI = 0.98, RMSEA = 0.05, and SRMR = 0.06.

In this way, model 1 corresponds directly to the structure revealed by the EFA for the given dataset. In terms of validity, it therefore meets the evidence source of the internal structure. The second model (referred to as model 2) aimed to integrate another source of evidence (evidence based on relation to other variables) by including those items in the ECL subscale that had proven to be sensitive toward the induced differences between the groups. Hence, for model 2, the same items as in model 1 were used to merge the ICL and GCL subscales because of their high loadings. For the ECL subscale, we used the full subscale of the NRS (NRS-3, NRS-4, and NRS-5) in order to incorporate the ability to detect a significant difference in terms of ECL. A subsuming confirmatory factor analysis also revealed adequate to good model fit, CFI = 1.0, TLI = 1.0, RMSEA = 0.00, and SRMR = 0.06.

Since both new models shared the same items for ICL and GCL, they reached the same (sufficient) level of reliability for

TABLE 5 | Results of the EFA for all items of both NRS and CLS.

Item	Factor 1 interpreted as ICL	Factor 2 interpreted as GCL	Factor 3 interpreted as ECL
CLS-3	**0.75**	-0.04	0.03
CLS-2	**0.74**	-0.11	0.05
CLS-4	**0.73**	0.03	-0.01
NRS-2	**0.72**	-0.02	-0.16
CLS-5	**0.66**	0.13	0.14
NRS-1	**0.53**	-0.03	-0.16
CLS-1	**0.48**	0.36	0.25
NRS-3	**0.40**	-0.04	0.22
CLS-7	**0.31**	-0.06	0.27
CLS-10	-0.05	**0.95**	0.02
CLS-12	0.02	**0.82**	-0.11
CLS-13	-0.09	**0.79**	0.13
CLS-14	0.05	**0.72**	-0.01
CLS-11	0.10	**0.62**	-0.16
NRS-8	0.11	**0.48**	-0.21
CLS-9	0.10	0.11	**0.60**
NRS-6	0.27	0.22	**-0.49***
NRS-7	-0.01	0.23	**-0.48***
NRS-4	0.09	-0.12	**0.47**
CLS-6	0.31	-0.05	**0.45**
NRS-5	0.23	0.01	**0.35**

Highest item loadings are given in bold.
**Negative loadings were not considered for combined scales.*

these subscales, α_c(ICL) = 0.79 and α_c(GCL) = 0.90. They slightly differed concerning the reliability of their ECL subscales, α_c(ECL, model 1) = 0.54 and α_c(ECL, model 2) = 0.57 which are still below the desired cutoff value α_c = 0.70. Furthermore, the sensitivity toward group-specific differences in ECL seemed to be inherited as model 1 showed no significant difference, $t(90.3) = -0.64$ and $p = 0.52$, while model 2 adopted the significant differences from the full adapted NRS, $t(89.3) = 2.17$ and $p = 0.033$.

DISCUSSION

Validity Based on Content

Both scales had to be adapted, and the CLS had to be expanded to fit the desired context. Since experimenting in STEM laboratory courses has been commonly based on generating and interpreting the measurement data, the measurement procedure and the corresponding quantities as well as their functional relationships and scientific laws are the main source of the information that has to be processed in order to generate new knowledge structures. Especially concerning the adapted and expanded CLS, the item development included all those relevant sources of content-related complexity in the subscales dedicated to measure ICL as well as GCL, whereas the items of the NRS merely consisted of general expressions. Hence, the adapted CLS appears to be slightly advantageous as a higher number of typical aspects from the learning scenario were directly addressed within the items.

Following the concept of ECL as presented by CLT, processes that do not contribute to essential learning originate from irrelevant and distracting elements. These include language issues and presentation formats that demand unnecessary search processes and representational holding. While the CLS originally included text comprehension as a source of ECL, the adapted version was not expanded toward the presentation formats (e.g., by addressing distracting search processes in the items), though this was a specific part of the presented study. In this case, the adapted CLS could be limited in its ability to cover all relevant load-inducing aspects that learners face throughout the experimental procedure. In contrast, the NRS already addressed presentational aspects, which were retained for the adapted version.

In sum, all subscales covered relevant aspects of the learning environment, but each with a specific main emphasis toward instructional design aspects. Based on the item formulation, the adapted CLS seems to address more precisely ICL and GCL, while the adapted NRS seems to address ECL in a more sensitive way for the context of laboratory learning scenarios. Furthermore, this emphasizes a general need for developing and validating specific instruments that directly address the characteristics of learning scenarios and include all crucial load-inducing elements. A more general item formulation might be too abstract, which could result in participants not being able to relate the items to the given situation without being further introduced to the intention and the meaning of the respective scale (e.g., Klepsch et al., 2017).

Validity Evidence Based on Internal Structure

Concerning their internal consistency for the given dataset, the subscales of the adapted NRS and adapted CLS cannot be seen as sufficiently reliable. Moreover, these low indices are far below those of the original work by Klepsch et al. (2017) and therefore

challenge the benefits and appropriateness derived from the content analysis (*Validity Based on Content*). It is probable that the a priori specification of load-inducing content will not fit the subjective impressions of the learners during the experimentation phase. In contrast, the subscales for ICL and GCL of the adapted CLS show a good internal consistency. Except for ECL, the values are in the range of the original work by Leppink et al. (2013) or former adaptions of the scale (Thees et al., 2020; Andersen and Makransky, 2021a, Andersen and Makransky, 2021b). Here again, the insufficient reliability for ECL casts doubt on whether the items of this specific subscale are appropriate to measure the intended type of cognitive load. Especially in comparison with the findings of Thees et al. (2020), who used a very similar formulation of the ECL items in another scientific context (thermodynamics instead of electricity), these results challenge a broad applicability of a simple adaption of the original CLS and raise the question of how to integrate context-specific sources of load while the overall pedagogical approach remains comparable (e.g., inquiry-based learning).

The results of the CFA also undermine the intended internal structure of each scale as the model fit indices do not provide sufficient formal evidence for the three assumed factors. Hence, the confirmatory analysis strengthens criticisms of the appropriateness of the three-factorial structure as intended during the item development. This might be a consequence of a rather small sample size because the conventional rule of thumb that the number of participants should be more than 10 times the number of items is only reached for the adapted NRS, but not for the CLS. Another limiting factor might be the reduction of the scale range from a 10-point to a six-point scale for the adapted CLS.

In sum, these findings reveal that the intended internal structure of the instruments is not fully represented in the data, which constrains the interpretation of the single subscales. We must therefore reject the first hypothesis and question the appropriateness of the adapted scales to differentiate between three different types of load in the context of technology-enhanced laboratory courses. Although the CFAs mostly narrowly missed the acceptable range for the fit indices, which can be interpreted as a case of a too small sample size, the low indices for internal consistency as the reliability measure for four out of six subscales remain the main issue for the internal structure.

Validity Based on Relations to Other Variables

Assuming the three-factorial structure of the scales as validated in various former studies, a traditional MTMM matrix based on a correlation table was analyzed. Although the reliability of all the subscales adapted from the NRS and for EL adapted from the CLS was not sufficient, significant correlations and repeating patterns indicate convergent and discriminant validity between the two scales. This means that the corresponding subscales in both approaches have meaningful coincidence and that each subscale can be distinguished from the others according to their interpretation as different types of cognitive load. These findings preliminarily emphasize the scales' appropriateness as load-measuring instruments. However, the strength of evidence is limited due to missing cutoff values for the traditional interpretation of correlation patterns. Furthermore, the results could not be sufficiently reproduced by a confirmatory MTMM approach as not all indices indicate an acceptable model fit. Hence, although there are promising findings based on the traditional comparison of correlation patterns, we cannot provide sufficient formal evidence for convergent and discriminant validity, which means that the second hypothesis is not clearly supported by the data of the present study. Thus, the MTMM analysis does not support the internal structure of both scales as being directed to the same three different latent variables.

Concerning the contrasted presentation formats, a sensitive scale was expected to reflect group-specific differences in ECL in favor of group 1. For the given dataset, only the adapted NRS revealed a significant difference between the two intervention groups. As expected, group 1 reported lower scores for ECL. Hence, the findings support the third hypothesis for the adapted NRS and emphasize it as the more sensitive scale toward the contrasted presentation formats and the accompanying load sources, i.e., the spatial split of related information elements. The missing sensitivity of the adapted CLS toward differences in ECL might be the consequence of a biased focus on language issues and an insufficient adaptation toward other load-inducing sources for this specific subscale. However, these findings are in accordance with a study conducted by Skulmowski and Rey (2020), who also revealed that the NRS is more likely to detect differences in ECL than the CLS. In their research, the authors also argued that the original items of the CLS might focus too much on the verbal aspects of the learning scenario, while the NRS addresses information processing in a more generalized way.

Both adapted scales did not show significant differences concerning ICL scores, which is in accordance with the intention to provide both groups with equal content, experimental setups, and representational forms of the measurement data. Hence, the fourth hypothesis is supported for both scales. However, there is no significant correlation between the scores of both ICL subscales and the specific or full prior knowledge scores, which contradicts the theory-based expectation that learners with lower prior knowledge will perceive a higher ICL. Eventually, a missing correlation might indicate that learners' prior knowledge was sufficient as a conceptual prerequisite to successfully conduct the experimental tasks. However, this leads to a rejection of the fifth hypothesis because this result does not support the compliance of the ICL subscale with the theoretical concept of ICL in terms of the CLT.

In sum, the direct comparison between the two adapted scales via the MTMM matrix emphasizes but does not prove convergent and discriminant evidence due to insufficient support by the confirmatory model fit. The relation to the grouping variable for the given study emphasizes the adapted NRS as more sensitive toward differences in ECL, which is in accordance with previous findings. As expected, both scales reveal equal ICL ratings for both groups. However, the relation between ICL and prior knowledge could not be verified. Eventually, the relation to

other variables revealed mixed to rather unfavorable results as most of the underlying hypotheses had to be rejected.

Combined Scales

As the internal structure of each adapted scale remains challenged after considering evidence based on the reliability scores as well as after the CFAs and the MTMM approaches, we decided to construct a combined instrument based on the given item pool of both scales. As the first step, the EFA revealed a three-factorial structure for the combined dataset. In addition, the items with the highest factor loadings indicate accordance with the expected underlying latent variable so that the three factors can be interpreted as related to the three types of cognitive load (**Table 5**). Given the self-imposed restriction of using only three items per factor to obtain a concise scale, two models were derived that considered those items with the highest positive factor loadings and findings from validity evidence based on the relation to the grouping variable.

Both models showed acceptable to good model fit in subsuming CFAs concerning their three-factorial internal structure, emphasizing their capability to differentiate between the three types of load. While the subscales for ICL and GCL are equal in both models and consist of items from the adapted CLS, the models differ concerning the ECL subscale. While model 1 follows the ranking of factor loadings from the EFA, resulting in a mix of items from both adapted NRS and CLS, model 2 inherits the full ECL subscale from the adapted NRS. This step is not based on the findings from the EFA, but respects the fact that this particular subscale was able to detect group-specific differences in ECL which are likely to exist in studies that contrast presentation formats to address well-known multimedia effects such as split-attention (Schroeder and Cenkci, 2018). Hence, model 2 constitutes a further development as it integrates validity evidence based on the relation to other variables.

However, both models still suffer from low internal consistency concerning ECL, which reduces the reliability of the acceptable model fits. This issue might result from the fact that the items dedicated to measuring ECL cover different load-inducing elements such as data presentation or verbal components. Hence, they cannot be expected to equally contribute to the score, and so, reaching a high internal consistency remains difficult. Andersen and Makransky (2021b) even considered ECL as a multidimensional variable, which presents a plausible reason for our low internal consistency findings. Eventually, we follow the results of the presuming EFA by considering ECL as an unidimensional factor which addresses multiple learning-irrelevant elements.

In sum, model 2 is considered the best scale based on the given item pool and the given dataset. Concerning the content of the new subscales for ICL and ECL, the items refer to concrete aspects of the experimental tasks, i.e., those components that are a priori determined the basis of the learning process. Hence, the combination of both NRS and CLS showed that the most valuable items for the given dataset were taken from the CLS, but the NRS provided a meaningful supplement. Furthermore, the restriction to three items per subscale emphasizes the need to focus on those elements of the learning environment that are mandatory to deal with during the learning process.

Future Work

To address a wider range of technology-based learning scenarios, our adapted versions could be enhanced by integrating items from other adaptations. For example, Andersen and Makransky (2021a) included the term "information display format" as a source of load in their ECL subscale, which was based on the original CLS. This term would directly address the contrasted presentation formats in our study without any bias toward a certain technology. On the contrary, such general formulations require a clarification as to what they are referring to, such as by a short introduction prior to the subjective rating, where the term is specified for each intervention group.

As most of the samples used in comparable studies consist of university students, studies validating the application of the considered scales in school contexts are missing. At the school level, learners are expected to have a different amount of prior knowledge and metacognitive skills. Hence, the measurement of cognitive load based on subjective experiences could be much more challenging (Brünken et al., 2003; Klepsch et al., 2017). Therefore, the scales have to be adapted concerning the item formulation as well as the scale levels and the endpoint labeling. In addition, items on passive load (mental load) and active load (mental effort), developed by Klepsch and Seufert (2021), could be added. The authors could show that the item on passive load related to the ICL factor of their scale and the item on active load related to the GCL factor. Klepsch and Seufert recommended the use of these additional items with children and tasks that require learners' self-regulation (e.g., laboratory work). Such adaptation might demand further investigations toward validity evidence. Future work might consider expert ratings for the item content to strengthen the explanatory power of the content-related evidence (Brünken et al., 2003; Klepsch et al., 2017). Furthermore, it will be essential to validate the new and further developed scales on a large sample as well as to consider that further (back-) translations of the presented (German) scales might affect validity aspects.

In the present study, only some of all possible sources providing evidence for the validation of the cognitive load scales were examined. Future studies should not only experimentally manipulate the ECL but also systematically manipulate all three types of cognitive load and verify whether the developed scales can also reflect variations in the ICL and GCL. Manipulation of ICL could be achieved by contrasting laboratory tasks with different levels of complexity or by contrasting groups with different levels of prior knowledge (evidence based on the relation to other variables). However, previous research suggesting that a subject's ability to reliably differentiate between ICL and ECL depends on a sufficient level of prior knowledge (Zu et al., 2021) should also be considered. GCL could be manipulated by providing or not providing self-regulation prompts during student experimentation.

Another option for analyzing validity evidence based on the relation to other variables could be a direct comparison between subjective ratings and objectives measures such as eye-tracking

data. Recent developments of mobile eye-tracking devices allow for collecting data in dynamic situations such as laboratory courses and might even be applied to augmented reality-based learning scenarios (Kapp et al., 2021) so that various approaches of technology-enhanced learning scenarios can be accompanied by both the subjective rating scales and the objective gaze-based measures. Nevertheless, the interpretation should consider prior research indicating that there might be no linear relationship between objective and subjective measures but that they rather cover different facets of cognitive load (Minkley et al., 2021).

Conclusion

In this article, we present supporting and critical points regarding the validity of two popular subjective cognitive load-rating scales in the context of technology-enhanced science experiments. Although the content of the adapted items seemed to be promising in terms of addressing various facets of the learning environment, the low internal consistency and the insufficient evidence for the intended three-factorial structure negate the appropriateness of the adapted scales. However, based on the correlations between the subscales, there are various indications that the addressed latent variables (i.e., ICL, ECL, and GCL) are comparable in both scales and can be distinguished from each other. Again, these assumptions cannot be formally confirmed based on the given dataset. In sum, three of five deduced hypotheses toward different sources of evidence in terms of validity had to be rejected due to insufficient formal evidence. Hence, there are no sufficient results that favor either the adapted NRS or the adapted CLS, although they seem to be convincing regarding their content.

The interpretation of this conflict is twofold. First, for the learning context under investigation, we question the current state of the adapted scales as they are not appropriate to measure different types of cognitive load. This would explain the insufficient reliability and the insufficient model fits concerning the assumed internal structure. In contrast, one could assume that the items of both scales are capable of representing the real load-inducing elements, but each scale addresses some but not all facets of the learning environment. Hence, solely by combining both item pools, it was possible to reach an adequate scale (model 2). At this point, the advantages of both adapted scales were combined to form a promising new scale for the context of complex science learning scenarios (although this scale is not without its flaws). The internal consistency of the ECL subscales is not acceptable but can be made plausible via the inherent multiple aspects covered by the items.

The presented study is an example of applying known and empirically validated scales to an essential and realistic learning scenario from STEM education. Since inquiry-based learning scenarios contain multiple information sources, researchers must develop new instruments to be able to correctly measure cognitive load. Moreover, the issues raised in the analyses show that it is necessary to seek for validity based on different sources such as content, internal structure, and relation to other variables. In this sense, we want to encourage the community to contribute to the question of how to create valid and suitable questionnaires to determine cognitive load in specific complex learning scenarios.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

Ethical review and approval were not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

MT: conceptualization, methodology, formal analysis, investigation, writing, and supervision; SK: software, methodology, formal analysis, and investigation; KA: methodology and investigation; SM: conceptualization, methodology, and writing; RB: conceptualization, resources, and funding acquisition; JK: conceptualization, resources, writing, project administration, and funding acquisition.

FUNDING

The dataset this paper draws upon was collected as part of the research projects GeAR (grant no. 01JD1811B) and gLabAssist (grant no. 16DHL1022), both funded by the German Federal Ministry of Education and Research (BMBF). The funding source had no involvement in preparing and conducting the study or in preparing the manuscript.

REFERENCES

AERA; APA; NCME (2011). *Report and Recommendations for the Reauthorization of the institute of Education Sciences*. Washington D.C: American Educational Research Association.

Altmeyer, K., Kapp, S., Thees, M., Malone, S., Kuhn, J., and Brünken, R. (2020). The Use of Augmented Reality to foster Conceptual Knowledge Acquisition in STEM Laboratory Courses-Theoretical Background and Empirical Results. *Br. J. Educ. Technol.* 51, 611–628. doi:10.1111/bjet.12900

American Association of Physics Teachers (2014). *AAPT Recommendations for the Undergraduate Physics Laboratory Curriculum,* https://www.aapt.org/Resources/upload/LabGuidlinesDocument_EBendorsed_nov10.pdf.

Andersen, M. S., and Makransky, G. (2021a). The Validation and Further Development of a Multidimensional Cognitive Load Scale for Virtual Environments. *J. Comput. Assist. Learn.* 37, 183–196. doi:10.1111/jcal.12478

Andersen, M. S., and Makransky, G. (2021b). The Validation and Further Development of the Multidimensional Cognitive Load Scale for Physical and Online Lectures (MCLS-POL). *Front. Psychol.* 12, 642084. doi:10.3389/fpsyg.2021.642084

Ayres, P., and Sweller, J. (2014). "The Split-Attention Principle in Multimedia Learning," in *The Cambridge Handbook of Multimedia Learning*. Editor R. E. Mayer. Second edition (New York: Cambridge University Press), 206–226.

Baddeley, A. (1992). Working Memory. *Science* 255, 556–559. doi:10.1126/science.1736359

Becker, S., Klein, P., Gößling, A., and Kuhn, J. (2020). Using mobile Devices to Enhance Inquiry-Based Learning Processes. *Learn. Instruction* 69, 101350. doi:10.1016/j.learninstruc.2020.101350

Brünken, R., Plass, J. L., and Leutner, D. (2003). Direct Measurement of Cognitive Load in Multimedia Learning. *Educ. Psychol.* 38, 53–61. doi:10.1207/S15326985EP3801_7

Burde, J.-P. (2018). *Konzeption und Evaluation eines Unterrichtskonzepts zu einfachen Stromkreisen auf Basis des Elektronengasmodells*. Berlin: Logos. doi:10.30819/4726

Campbell, D. T., and Fiske, D. W. (1959). Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix. *Psychol. Bull.* 56, 81–105. doi:10.1037/h0046016

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* 2. ed. Hillsdale, NJ: Erlbaum.

Cowan, N. (2001). The Magical Number 4 in Short-Term Memory: A Reconsideration of Mental Storage Capacity. *Behav. Brain Sci.* 24, 87–114. doi:10.1017/s0140525x01003922

de Jong, T., Linn, M. C., and Zacharia, Z. C. (2013). Physical and Virtual Laboratories in Science and Engineering Education. *Science* 340, 305–308. doi:10.1126/science.1230579

de Jong, T. (2019). Moving towards Engaged Learning in STEM Domains; There Is No Simple Answer, but Clearly a Road Ahead. *J. Comput. Assist. Learn.* 35, 153–167. doi:10.1111/jcal.12337

Eid, M. (2000). A Multitrait-Multimethod Model with Minimal Assumptions. *Psychometrika* 65, 241–261. doi:10.1007/bf02294377

Etkina, E., van Heuvelen, A., White-Brahmia, S., Brookes, D. T., Gentile, M., Murthy, S., et al. (2006). Scientific Abilities and Their Assessment. *Phys. Rev. ST Phys. Educ. Res.* 2, 113. doi:10.1103/PhysRevSTPER.2.020103

Hofstein, A., and Lunetta, V. N. (2004). The Laboratory in Science Education: Foundations for the Twenty-First century. *Sci. Ed.* 88, 28–54. doi:10.1002/sce.10106

Husnaini, S. J., and Chen, S. (2019). Effects of Guided Inquiry Virtual and Physical Laboratories on Conceptual Understanding, Inquiry Performance, Scientific Inquiry Self-Efficacy, and Enjoyment. *Phys. Rev. Phys. Educ. Res.* 15, 31. doi:10.1103/PhysRevPhysEducRes.15.010119

Jiang, D., and Kalyuga, S. (2020). Confirmatory Factor Analysis of Cognitive Load Ratings Supports a Two-Factor Model. *TQMP* 16, 216–225. doi:10.20982/tqmp.16.3.p216

Kalyuga, S. (2011). Cognitive Load Theory: How Many Types of Load Does it Really Need? *Educ. Psychol. Rev.* 23, 1–19. doi:10.1007/s10648-010-9150-7

Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores. *J. Educ. Meas.* 50, 1–73. doi:10.1111/jedm.12000

Kapici, H. O., Akcay, H., and de Jong, T. (2019). Using Hands-On and Virtual Laboratories Alone or Together—Which Works Better for Acquiring Knowledge and Skills? *J. Sci. Educ. Technol.* 28, 231–250. doi:10.1007/s10956-018-9762-0

Kapp, S., Barz, M., Mukhametov, S., Sonntag, D., and Kuhn, J. (2021). ARETT: Augmented Reality Eye Tracking Toolkit for Head Mounted Displays. *Sensors* 21, 2234. doi:10.3390/s21062234

Kapp, S., Thees, M., Beil, F., Weatherby, T., Burde, J.-P., Wilhelm, T., et al. (2020). "The Effects of Augmented Reality: A Comparative Study in an Undergraduate Physics Laboratory Course," in Proceedings of the 12th International Conference 1on Computer Supported Education, May 2–4, 2020 (SciTePress - Science and Technology Publications), Vol. 2, 197–206. doi:10.5220/0009793001970206

Kester, L., Kirschner, P. A., and van Merriënboer, J. J. G. (2005). The Management of Cognitive Load during Complex Cognitive Skill Acquisition by Means of Computer-Simulated Problem Solving. *Br. J. Educ. Psychol.* 75, 71–85. doi:10.1348/000709904X19254

Kester, L., Paas, F., and van Merriënboer, J. J. G. (2010). "Instructional Control of Cognitive Load in the Design of Complex Learning Environments," in *Cognitive Load Theory*. Editors J. L. Plass, R. Moreno, and R. Brunken (Cambridge: Cambridge University Press), 109–130.

Klepsch, M., Schmitz, F., and Seufert, T. (2017). Development and Validation of Two Instruments Measuring Intrinsic, Extraneous, and Germane Cognitive Load. *Front. Psychol.* 8, 1997. doi:10.3389/fpsyg.2017.01997

Klepsch, M., and Seufert, T. (2021). Making an Effort versus Experiencing Load. *Front. Educ.* 6 (56). doi:10.3389/feduc.2021.645284

Klepsch, M., and Seufert, T. (2020). Understanding Instructional Design Effects by Differentiated Measurement of Intrinsic, Extraneous, and Germane Cognitive Load. *Instr. Sci.* 48, 45–77. doi:10.1007/s11251-020-09502-9

Kline, P. (2000). *The Handbook of Psychological Testing*. 2. ed. London: Routledge.

Krell, M. (2017). Evaluating an Instrument to Measure Mental Load and Mental Effort Considering Different Sources of Validity Evidence. *Cogent Edu.* 4, 1280256. doi:10.1080/2331186X.2017.1280256

Lazonder, A. W., and Harmsen, R. (2016). Meta-Analysis of Inquiry-Based Learning. *Rev. Educ. Res.* 86, 681–718. doi:10.3102/0034654315627366

Leppink, J., Paas, F., van der Vleuten, C. P. M., van Gog, T., and Van Merriënboer, J. J. G. (2013). Development of an Instrument for Measuring Different Types of Cognitive Load. *Behav. Res.* 45, 1058–1072. doi:10.3758/s13428-013-0334-1

Leppink, J., Paas, F., van Gog, T., van der Vleuten, C. P. M., and van Merriënboer, J. J. G. (2014). Effects of Pairs of Problems and Examples on Task Performance and Different Types of Cognitive Load. *Learn. Instruction* 30, 32–42. doi:10.1016/j.learninstruc.2013.12.001

Lunetta, V. N., Hofstein, A., and Clough, M. P. (2005). "Learning and Teaching in the School Science Laboratory: An Analysis of Research, Theory, and Practice," in *Handbook of Research on Science Education*. Editors S. K. Abell and N. G. Lederman (New York, NY: Lawrence Erlbaum; Routledge), 393–441.

Mayer, R. E., and Moreno, R. (1998). A Split-Attention Effect in Multimedia Learning: Evidence for Dual Processing Systems in Working Memory. *J. Educ. Psychol.* 90, 312–320. doi:10.1037/0022-0663.90.2.312

Mayer, R. E., and Moreno, R. (2003). Nine Ways to Reduce Cognitive Load in Multimedia Learning. *Educ. Psychol.* 38, 43–52. doi:10.1207/s15326985ep3801_6

Mayer, R., and Fiorella, L. (2014). "Principles for Reducing Extraneous Processing in Multimedia Learning: Coherence, Signaling, Redundancy, Spatial Contiguity, and Temporal Contiguity Principles," in *The Cambridge Handbook of Multimedia Learning*. Editor R. E. Mayer. Second edition (New York: Cambridge University Press), 279–315.

Minkley, N., Xu, K. M., and Krell, M. (2021). Analyzing Relationships between Causal and Assessment Factors of Cognitive Load: Associations between Objective and Subjective Measures of Cognitive Load, Stress, Interest, and Self-Concept. *Front. Educ.* 6 (56). doi:10.3389/feduc.2021.632907

Paas, F. G. W. C. (1992). Training Strategies for Attaining Transfer of Problem-Solving Skill in Statistics: A Cognitive-Load Approach. *J. Educ. Psychol.* 84, 429–434. doi:10.1037/0022-0663.84.4.429

Schroeder, N. L., and Cenkci, A. T. (2018). Spatial Contiguity and Spatial Split-Attention Effects in Multimedia Learning Environments: a Meta-Analysis. *Educ. Psychol. Rev.* 30, 679–701. doi:10.1007/s10648-018-9435-9

Skulmowski, A., and Rey, G. D. (2020). Subjective Cognitive Load Surveys lead to Divergent Results for Interactive Learning media. *Hum. Behav. Emerg. Tech.* 2, 149–157. doi:10.1002/hbe2.184

Sweller, J. (2020). Cognitive Load Theory and Educational Technology. *Education Tech. Res. Dev.* 68, 1–16. doi:10.1007/s11423-019-09701-3

Sweller, J. (2010). Element Interactivity and Intrinsic, Extraneous, and Germane Cognitive Load. *Educ. Psychol. Rev.* 22, 123–138. doi:10.1007/s10648-010-9128-5

Sweller, J., van Merriënboer, J. J. G., and Paas, F. (2019). Cognitive Architecture and Instructional Design: 20 Years Later. *Educ. Psychol. Rev.* 31, 261–292. doi:10.1007/s10648-019-09465-5

Sweller, J., van Merriënboer, J. J. G., and Paas, F. G. W. C. (1998). Cognitive Architecture and Instructional Design. *Educ. Psychol. Rev.* 10, 251–296. doi:10.1023/a:1022193728205

Thees, M., Kapp, S., Strzys, M. P., Beil, F., Lukowicz, P., and Kuhn, J. (2020). Effects of Augmented Reality on Learning and Cognitive Load in university Physics Laboratory Courses. *Comput. Hum. Behav.* 108, 106316. doi:10.1016/j.chb.2020.106316

Trumper, R. (2003). The Physics Laboratory - A Historical Overview and Future Perspectives. *Sci. Edu.* 12, 645–670. doi:10.1023/a:1025692409001

Urban-Woldron, H., and Hopf, M. (2012). Entwicklung eines Testinstruments zum Verständnis in der Elektrizitätslehre [Development of a diagnostic instrument for testing studen understanding of basic electricity concepts]. *Z. für Didaktik der Naturwissenschaften* 18, 201–227.

Volkwyn, T. S., Allie, S., Buffler, A., and Lubben, F. (2008). Impact of a Conventional Introductory Laboratory Course on the Understanding of

Measurement. *Phys. Rev. ST Phys. Educ. Res.* 4, 4. doi:10.1103/PhysRevSTPER.4.010108

Vosniadou, S. (2008). *International Handbook of Research on Conceptual Change*. New York: Routledge.

Wilcox, B. R., and Lewandowski, H. J. (2017). Developing Skills versus Reinforcing Concepts in Physics Labs: Insight from a Survey of Students' Beliefs about Experimental Physics. *Phys. Rev. Phys. Educ. Res.* 13, 65. doi:10.1103/PhysRevPhysEducRes.13.010108

Zacharia, Z. C., and de Jong, T. (2014). The Effects on Students' Conceptual Understanding of Electric Circuits of Introducing Virtual Manipulatives within a Physical Manipulatives-Oriented Curriculum. *Cogn. Instruction* 32, 101–158. doi:10.1080/07370008.2014.887083

Zacharia, Z. C., and Olympiou, G. (2011). Physical versus Virtual Manipulative Experimentation in Physics Learning. *Learn. Instruction* 21, 317–331. doi:10.1016/j.learninstruc.2010.03.001

Zu, T., Munsell, J., and Rebello, N. S. (2021). Subjective Measure of Cognitive Load Depends on Participants' Content Knowledge Level. *Front. Educ.* 6 (56). doi:10.3389/feduc.2021.647097

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 31 August 2021
doi: 10.3389/fpsyg.2021.703857

Validation of Cognitive Load During Inquiry-Based Learning With Multimedia Scaffolds Using Subjective Measurement and Eye Movements

Marit Kastaun[1], Monique Meier[1], Stefan Küchemann[2] and Jochen Kuhn[2]*

[1] Department of Biology Education, Institute for Biology, Universität Kassel, Kassel, Germany, [2] Physics Education Research Group, Department of Physics, Technische Universität Kaiserslautern, Kaiserslautern, Germany

OPEN ACCESS

Edited by:
Moritz Krell,
Freie Universität Berlin, Germany

Reviewed by:
Andreas Korbach,
Saarland University, Germany
Paul Ayres,
University of New South Wales, Australia
Kun Yu,
University of Technology Sydney, Australia

***Correspondence:**
Marit Kastaun
m.kastaun@uni-kassel.de

Specialty section:
This article was submitted to Educational Psychology, a section of the journal Frontiers in Psychology

Received: *30 April 2021*
Accepted: *31 July 2021*
Published: *31 August 2021*

Citation:
Kastaun M, Meier M, Küchemann S and Kuhn J (2021) Validation of Cognitive Load During Inquiry-Based Learning With Multimedia Scaffolds Using Subjective Measurement and Eye Movements. Front. Psychol. 12:703857. doi: 10.3389/fpsyg.2021.703857

Subject-method barriers and cognitive load (CL) of students have a particular importance in the complex learning process of scientific inquiry. In this work, we investigate the valid measurement of CL as well as different scaffolds to reduce it during experimentation. Specifically, we examine the validity of a subjective measurement instrument to assess CL [in extraneous cognitive load (ECL), intrinsic cognitive load, and germane cognitive load (GCL)] during the use of multimedia scaffolds in the *planning* phase of the scientific inquiry process based on a theoretical framework of the CL theory. The validity is analyzed by investigating possible relationships between causal (e.g., cognitive abilities) and assessment (e.g., eye-tracking metrics) factors in relation to the obtained test scores of the adapted subjective measurement instrument. The study aims to elucidate possible relationships of causal factors that have not yet been adequately investigated in relation to CL. Furthermore, a possible, still inconclusive convergence between subjective test scores on CL and objectively measured indicators will be tested using different eye-tracking metrics. In two studies ($n = 250$), 9th and 11th grade students experimentally investigated a biological phenomenon. At the beginning of the *planning* phase, students selected one of four multimedia scaffolds using a tablet (Study I: $n = 181$) or a computer with a stationary eye-tracking device (Study II: $n = 69$). The subjective cognitive load was measured *via* self-reports using a standardized questionnaire. Additionally, we recorded students' gaze data during learning with the scaffolds as objective measurements. Besides the causal factors of cognitive-visual and verbal abilities, reading skills and spatial abilities were quantified using established test instruments and the learners indicated their representation preference by selecting the scaffolds. The results show that CL decreases substantially with higher grade level. Regarding the causal factors, we observed that cognitive-visual and verbal abilities have a significant influence on the ECL and GCL in contrast to reading skills. Additionally, there is a correlation between the representation preference and different types of CL. Concerning the objective measurement data, we found that the absolute

fixation number is predictive for the ECL. The results are discussed in the context of the overall methodological research goal and the theoretical framework of CL.

Keywords: cognitive load, cognitive abilities, representation preference, scaffolding, eye tracking, scientific inquiry

INTRODUCTION

Cognitive load (CL) is a theoretical, psychological construct that describes the individual loads of learners during the processing, construction, and memorizing of (new) information (Sweller et al., 2019). Within the Cognitive Load Theory (CLT), one of many basic assumptions is that only a limited number of information elements can be processed simultaneously due to the limited working memory capacity (Paas et al., 2010). Therefore, the processing of information that is relevant to learning should be optimized and loads irrelevant to learning are to be minimized (among others van Merriënboer et al., 2006; Paas and van Merriënboer, 2020). Another assumption claims the processing of information in two different channels – the verbal and the visual channel (dual channel assumption; Skuballa et al., 2019; Sweller et al., 2019; Scheiter et al., 2020). Furthermore, it is assumed that CL is composed of three distinct categories – intrinsic (ICL), extraneous (ECL), and germane cognitive load (GCL; among others Sweller, 1994; Zu et al., 2020). ICL is the cognitive load associated with high or low element interactivity during the processing of information/learning material (Sweller et al., 2011, 2019). Accordingly, the degree of interactivity between essential information elements required for learning determines the ICL. If the element interactivity is low, only a few elements are processed by the learner in working memory at the same time. If the element interactivity increases, the load on working memory also rises (Sweller et al., 2011). In contrast, ECL includes the cognitive load generated by suboptimal instructional design. Consequently, the learner needs to spend more cognitive resources on irrelevant or poorly constructed information during task performance or information processing than on the actual task solution (Sweller et al., 2019). The ECL and ICL together determine the total CL in the working memory by the learner. They are additively related because the resources for processing come from the same working memory pool (Sweller et al., 2011). The third category of cognitive load is GCL. This is expended by the learner to construct schemas and form mental models (Paas et al., 2003; Sweller et al., 2019), and it is therefore often described as the load that generates learning. In recent years, the study of CL has become increasingly important. To test and consequently optimize instructional designs or as a control variable for, e.g., self-regulated learning processes, such as inquiry-based learning (see "Environment (E) & Task (T) as causal factor – scaffolding & inquiry learning"; Kaiser et al., 2018), the valid measurement of CL represents a central goal of educational and psychological research (among others Kirschner et al., 2006; Minkley et al., 2018; Sweller et al., 2019). Using subjective self-assessments (Cierniak et al., 2009b; Leppink et al., 2013; Klepsch et al., 2017; Krell, 2017) or objective measures as indicators of CL, such as heart rate (Paas and van Merriënboer, 1994; Minkley et al., 2021), pupil dilation (Chen and Epps, 2014; Huh et al., 2019), blink rate (Chen and Epps, 2014), or gaze behavior (Korbach et al., 2018; Zu et al., 2020), different approaches to measuring CL have been investigated. Possible relationships and convergences between subjective and objective measurement tools for identifying CL are also coming more into the focus of research (among others Minkley et al., 2021). However, the use of subjective measurement instruments by self-assessment is still an established way to assess CL based on a learning task or instructional learning approach (Cierniak et al., 2009b; Leppink et al., 2013; Klepsch et al., 2017; Krell, 2017; Zu et al., 2020; Thees et al., 2021). Although the subjective measurement approach is controversial in research regarding valid measurement of CL (de Jong, 2010; Kirschner et al., 2011; Sweller et al., 2011), studies, such as the one by Klepsch et al. (2017), demonstrate a proven measurement accuracy. Other studies also reveal that measuring the three types of CL can pose different difficulties (DeLeeuw and Mayer, 2008; Ayres, 2018). First, it is unclear how, for example, the different CLs (ICL, GCL, and ECL) can be affected by the wording of items, since variations of the items lead to different results in performance and CL (Sweller et al., 2011; Klepsch and Seufert, 2021). Second, besides the wording, it is uncertain whether learners are able to assess their competences in relation to the differentiated items (ICL; GCL; and ECL; Sweller et al., 2011). Nevertheless, approaches to measure the three types of CL *via* self-assessments, e.g., *via* the design of controlled conditions (among others Sweller et al., 2011; Klepsch and Seufert, 2021), suggest that a categorical differentiation and measurement of the individual CLs are possible. However, Thees et al. (2021) also showed that relying on established subjective instruments to measure CL may lead to various difficulties. Composing a subjective measurement instrument from established items could lead to a more valid measurement of the three types of cognitive load (ECL, GCL, and ICL). However, these established self-assessments have mostly been used with university students, whereas research on subjective instruments measuring the different types of CL for middle-school students in instructional teaching-learning settings has been scarce (e.g., van de Weijer-Bergsma and van der Ven, 2021). Therefore, the focus of this research is to design and test a self-assessment questionnaire to measure the three types of CL for 9th and 11th grade students. To examine a possible valid measurement in combination with a convergence between subjective and objective measurement instruments of CL, various methods from validity research can be used. Consistent with the Standards for Educational and Psychological Testing (AERA et al., 2014), more recent validity approaches open possible opportunities for proving evidence and convergence

via argumentative, theory-based test score interpretations (Messick, 1995; Kane, 2006, 2013). According to Kane (2013, p. 13), validity is "an integrated evaluation judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment". Based on the validation framework of Cronbach and Meehl (1955), the Argument-Based Approach to Validation follows the examination of test score interpretation. This consists of making different validity assumptions from an interpretation-usage argument and testing them using different methods as well as relating them back to the existing theories (Kane, 2013). Hence, validity investigations do not represent a mere testing of formalized theories. They rather show the establishment and testing of coherent and plausible chains of reasoning that examine and explain the results of a measurement instrument in a theory- and evidence-based manner. Within these chains of reasoning, other validity criteria should be integrated and the test score should be examined from different perspectives (Messick, 1995; Kane, 2013). Regarding the validity testing of a subjective instrument to measure the three types of CL in this study (ICL, ECL, and GCL), the theory-based analysis of possible influences on CL and interactions within the learning environment needs to be investigated (causal factors: environment, task, and learner). The identification of possible relationships between these causal factors and the occurring ICL, ECL, and GCL (assessment factors) when using multimedia scaffolds is therefore required first. Then, a possible convergence between the results of the subjective instrument and objective indicators of CL will be examined in more detail. Therefore, the theoretical assumptions between the causal and the assessment factors will be presented first in order to deduct assumptions with regard to the validation of the measurement instrument that will be tested within the performed studies.

CONSTRUCT OF COGNITIVE LOAD WITH CAUSAL AND ASSESSMENT FACTORS

The distinction between causal and assessment factors in relation to CL was described by Paas and van Merriënboer (1994) based on study results in the context of problem solving. Choi et al. (2014) extended this theory-based construct of CL (**Figure 1**). In addition to the influencing factors of the task (task: e.g., complexity and format) and learner-specific characteristics (learner: e.g., cognitive abilities or prior knowledge), the model was extended by the factor of learning environment. Like the task as well as the learner characteristics, the environment can influence the load that occurs, e.g., physical conditions. The causal factors are in close interaction with each other which can significantly influence the expression of individual CL (Choi et al., 2014; Paas and van Merriënboer, 2020). The interaction between the causal factors can be seen, for example, between the task and the characteristics of the learner. Thus, different studies show the close interaction among prior knowledge and task complexity or task success (expertise-reversal effect: Cierniak et al., 2009a; Kalyuga, 2013).

Mental load, mental effort, and task performance are the assessment factors by which individual CL can be judged in relation to, for example, mastering a task or instruction. Both mental load and mental effort generally describe the CL that must be exerted by the learner to complete a given task. Mental load is the term used to summarize the subject-independent CL that relates solely to the characteristics of the task (Paas and van Merriënboer, 1994; Choi et al., 2014). Therefore, mental load is equivalent to the construct of ICL (e.g., Minkley et al., 2021). In contrast, mental effort describes the loads "which refers to the amount of capacity or resources that is actually allocated by the learner to accommodate the task demands" (Choi et al., 2014, p. 228). Mental effort therefore refers to the cognitive resources that are used during the specific problem solving/task and can accordingly be equated with GCL and ECL (e.g., Paas and van Merriënboer, 1994; Choi et al., 2014; Krell, 2017). Regardless of the classification of the individual types of CL, these are related to the causal factors that can influence both task performance and the CL of the learner. The level of prior knowledge (L), for example, has an influence on the expression of the ICL (see "Prior Knowledge"), whereas the ECL is considered to be mainly dependent on the instructional design (T). Both types of loads are perceived by the learner as rather passive (Klepsch and Seufert, 2021). The GCL, on the other hand, represents the perceived by the learner as rather active load (Seufert, 2018). The assessment factors can be measured by subjective (self-assessment) and objective (for example, heart rate) instruments as an indicator for CL (see "Introduction" and **Figure 1**). Subjective instruments, such as the ones of Zu et al. (2020); Klepsch et al. (2017); Krell (2017); or Leppink et al. (2013), have the benefit that they can be easily used *via* paper-pencil questionnaires within the intervention. However, contrary to objective measures, such as Heart Rate or Gaze Behavior, they depend on the self-assessment competences of learners/participants who assess and report their own load based on the item formulations (see "Introduction"). The objective instruments provide another way to measure CL. Although they are technical and costly, objective measurement methods are increasingly used. Although a valid measurement of CL cannot yet be granted, various studies show close correlations between subjective and objective measurement of CL categories (Korbach et al., 2018; Solhjoo et al., 2019; Minkley et al., 2021). Eye-movement measurements represent one research approach to investigate CL during learning with visual (and auditory) material or problem solving. Thus, different metrics, such as fixations duration (Zu et al., 2020) or transitions (Korbach et al., 2018), are used to derive learners' CL. Previous work on the relationship of students' gaze data and their subjective CL ratings demonstrated that certain eye-tracking metrics discriminated between the three types of CL (Zu et al., 2020). Zu et al. (2020) found that there is a significant relation of the mean fixation duration of students with low prior knowledge in a physics context and the ECL. Whereas the ICL significantly relates to the transitions between an animation and a text, the GCL is linked

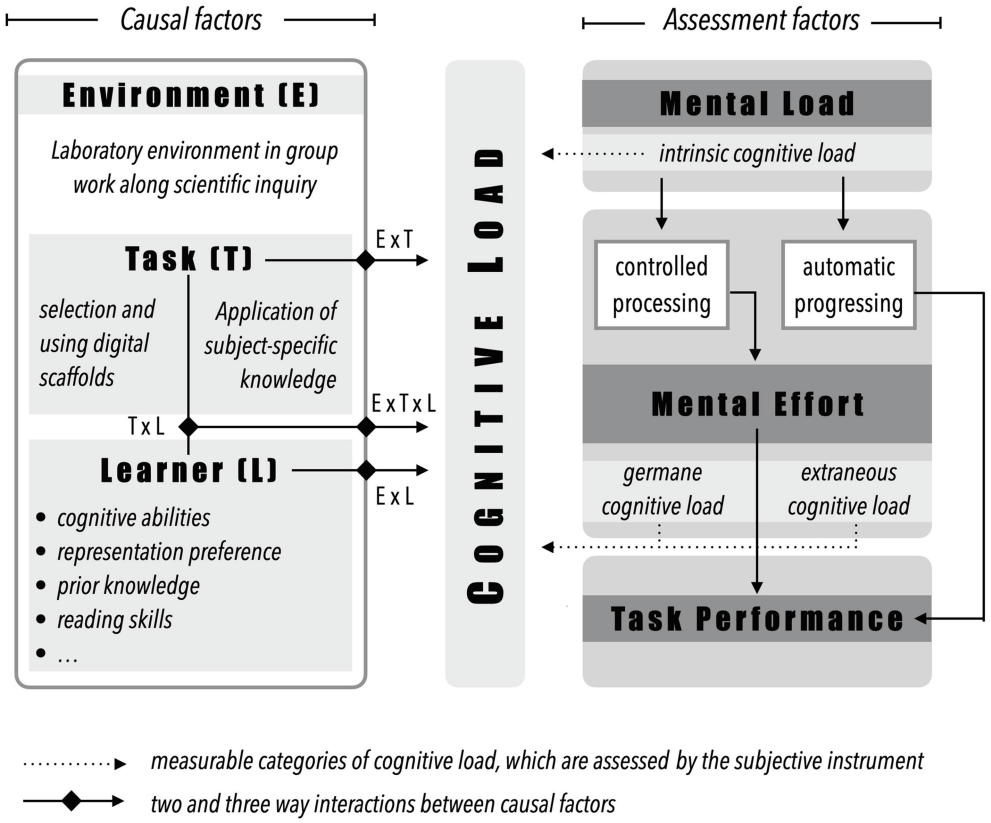

FIGURE 1 | Construct of cognitive load with for the study-specific causal and assessment factors (adapted from Choi et al., 2014).

to the dwell time on an animation. These results in the context of electric motors were observed for students with low prior knowledge and they were independent on the working memory capacity.

Environment (E) & Task (T) as Causal Factor – Scaffolding & Inquiry Learning

As shown in **Figure 1**, the learning environment and task are causal factors affecting CL. As Choi et al. (2014) noted, these two factors usually cannot be clearly separated in (terms of) practical implementation which also becomes clear by looking at the learning environment and the task in our study. In science education, inquiry-based learning is an approach to promote scientific reasoning as an integral part of science education (e.g., OECD, 2007). Within inquiry-based learning, learners are asked to actively apply different facets of knowledge to investigate a scientific phenomenon along the different phases of the scientific knowledge process (Kind and Osborne, 2017; de Jong, 2019). Through the generation of a research question and testable hypotheses, the planning of a scientific investigation to the execution and interpretation of the experimentally obtained results, learners actively use their content-related knowledge as well as their methodological skills (inquiry skills/ scientific reasoning skills) to investigate the phenomenon/ problem. Studies show that this problem-solving process (open inquiry learning: Bell et al., 2005), which is usually self-regulated and cooperative, has higher learning effects and can contribute better to long-term learning of scientific skills than direct forms of instruction (Alfieri et al., 2011; Furtak et al., 2012). However, other studies also suggest that learning along the scientific inquiry process can cause student-specific problems, as the complexity and openness of the learning process exceed the individual capacity of working memory (Leutner et al., 2005; Kirschner et al., 2006). As a result, supporting methods, such as protocol sheets provided with tasks, or the use of scaffolds, such as prompts and feedback, can minimize these occurring loads (Hmelo-Silver et al., 2007) and support the learner in his or her inquiry-based learning process (guided inquiry learning: Bell et al., 2005). Studies show, among other things, that short, direct prompts do not hinder the constructivist learning method of inquiry-based learning. However, depending on the design and the use, they can minimize individual CLs (e.g., Kirschner et al., 2006). Based on this, different approaches to scaffolding have been explored empirically (e.g., Arnold et al., 2014; Kaiser and Mayer, 2019; Meier and Kastaun, 2021). Using digital media, scaffolds can enable individual learning in a variety of forms.

In addition to prior knowledge and motivational factors, learners bring in further characteristics that offer starting points for differentiation. Especially when dealing with multimedia-based representation combinations (e.g., video or animation),

cognitive characteristics, such as learning preferences for visual or verbal learning material (representation preference; Mayer and Massa, 2003), spatial ability (Höffler and Leutner, 2011), or cognitive style (Höffler and Schwartz, 2011; Koć-Januchta et al., 2019), can have an impact on the learning process. Regarding the construction of multimedia scaffolds, these may not only be composed of descriptive and depictive representations (Ainsworth, 2006), but also of verbal audio tracks or symbolic texts in combination with static and dynamic images (among others Nitz et al., 2014). For a targeted, effective as well as preferably individualized use of technology-enhanced scaffolds, multimedia design principles based on CLT and the theory of multimedia learning (Mayer, 2014) should be resorted to minimize ECL and create free capacities in the working memory. Therefore, the use of multimedia scaffolds in the complex problem-solving process of inquiry-based learning (**Figure 1**: E + T) along with learner characteristics as causal factors of CL should be elicited.

Learners' (L) Characteristics as Causal Factor

Learner-specific characteristics are elementary in the context of CL (**Figure 1**) and thus also in the long-term memorization of new knowledge elements (among others Choi et al., 2014). Besides motivational and affective factors, interest (Baars et al., 2017) or fun (van de Weijer-Bergsma and van der Ven, 2021), the influence of cognitive characteristics, such as prior knowledge (Kalyuga et al., 2003; Cierniak et al., 2009a; Kalyuga, 2013), in relation to CL has been widely investigated. Interactions with the theory of multimedia learning as well as distinction of further cognitive characteristics aiming at cognitive abilities in relation to representation-based information processes, such as the development of reading skills, spatial ability, or cognitive-verbal and visual abilities, were examined so far only little in relation to CL or with unclear results (among others Ho et al., 2014; Chen et al., 2018; Lehmann and Seufert, 2020). However, based on theoretical frameworks of multimedia learning (Moreno, 2010; Mayer, 2014), human cognitive architecture (Paas and van Merriënboer, 2020), and empirical studies of individual cognitive abilities, assumptions can be made that identify a relationship between specific cognitive, representation-based, abilities, the learning material, and the resulting CL. In the following, specific (cognitive) characteristics are identified and elicited in terms of a possible causal factor in connection with CL and their possible of measurement within the present study.

Prior Knowledge

Studies investigating prior knowledge have shown that it significantly influences learning performance (Kalyuga, 2013; Chen et al., 2018; Richter et al., 2018; Seufert, 2019). Based on the cognitive architecture of knowledge processing and memorization, increased prior knowledge contributes to chunking (new) information and minimizing element interaction in the working memory (Pollock et al., 2002; Kalyuga, 2011). However, this is only correct up to a certain level of prior knowledge and dependent on the design of learning materials. For example, very detailed worked-out examples can minimize the learning outcome for students with a high level of prior knowledge, as their own individual knowledge construction is inhibited by the content specifications, whereas learners with a low level of prior knowledge can particularly benefit from these scaffolds (expertise-reversal effect: Kalyuga, 2013; Chen et al., 2016). In the context of inquiry-based learning, studies show that novices benefit less from individual knowledge construction than experts because they are missing prior experience or concrete information about the individual knowledge facets (e.g., Shrager and Siegler, 1998; Siegler, 2005; Kendeou and van den Broek, 2007). Reasons for this could be associated with high levels of CL and a consequent reduction of GCL, because the resources of the working memory for processing the learning content are exceeded (element interactivity, see "Introduction"). If the complexity of the learning task increases, explicit instruction is beneficial for long-term retention (Chen et al., 2016, 2018).

Cognitive Abilities

Along with the investigation of instructional forms and its pros and cons for specific learners, different correlations regarding CL can also be identified (e.g., Mayer and Massa, 2003). In addition to prior knowledge, spatial ability is increasingly a focus of investigation (e.g., Quaiser-Pohl et al., 2001). Basically, a positive correlation can be found between the expression of spatial ability and learning performance (Höffler, 2010). Regarding dynamic and static representations, such as animations or static images, spatial ability can act like a compensator according to the ability-as-compensator hypothesis (e.g., Höffler, 2010; Münzer, 2012; Sanchez and Wiley, 2014), but only if this ability is also actively needed for processing the learning task (Höffler and Leutner, 2011). Hence, it can be assumed that the higher spatial ability the better the learning performance when dealing with static images, whereas when learning with animated images, lower spatial ability may contribute to deeper understanding (e.g., Huk, 2006). Further determinants of how individual learning abilities are influenced can be identified using Jäger's (1984) intelligence model. The recognition of patterns, the agreement of similarities, or other linguistic as well as visual contexts allow conclusions to be drawn about logical comprehension as well as verbal and visual thinking skills. A few studies show that here, the expression is also related to learning performance (Kuhn et al., 1988; Kaiser and Mayer, 2019). Along with verbal reasoning skills, reading skills can have an impact on individual learning performance, especially for text-based instructional materials (Plass and Homer, 2002; Plass et al., 2003; Jäger et al., 2017). In connection with the preference for specific, material-related instructional formats (see "Learning preference/Representation Preference"), reading skills can also have an influence on the processing of the task or the provided material (Peterson et al., 2015; Lehmann and Seufert, 2020). On the one hand, reading skills are also closely related to cognitive (verbal) skills. On the other hand, the expression of reading skills can have an influence on information processing especially when learning with text-based (monomodal) materials (Schneider et al., 2017).

Learning Preference/Representation Preference

Preferences with respect to a specific instructional material, modality, or representation fundamentally describe a person's preferences to do or use something in a certain way than in another way (Kürschner et al., 2005; Lehmann and Seufert, 2020). The optional, student-specific selection of a learning material according to an individual's preference, may result in positive effects in terms of, for example, motivation to complete the task (in reference to the self-regulation: Rheinberg et al., 2000; Baars et al., 2017). Mayer and Massa (2003) describe learning preference in (multimedia) learning situations as, e.g., visual or verbal learners showing a preference for a certain multimedia representation/instruction (representation preference). So, in specific learning situations, learners often prefer one multimedia representation for learning instruction (e.g., image and audio vs. image-text) over others (Mayer and Massa, 2003; Choi and Sadar, 2011). In the study of Lehmann and Seufert (2020), it was investigated whether the preference of a modality has an influence on learning performance and CL. Although this study did not examine multimedia learning materials, but only used monomodal text as visual or auditory stimuli, the results suggest a possible relationship between learning preference and CL. Their findings provide insights into the relationships between the (learning-) preference, the use of the respective instructional material (only modality), and the expression of CL. Thus, learners who associated their preferred modality with visual texts benefited and reported a lower ECL than those who did not use a modality matching their indicated preference (Lehmann and Seufert, 2020).

Aims of the Study, Research Question, Assumptions, & Hypotheses

The aim of this work is to validate, along the Argument-Based Approach to Validation by Kane (2013) and Standards for Educational and Psychological Testing (AERA et al., 2014), a subjective instrument for students of the 9th and 11th grade for measuring the three different types of CL (ECL, ICL, and GCL):

(a) Based on the theoretical construct according to Choi et al. (2014), relationships between causal and assessment factors can be identified, which need to be tested in order to validate a subjective measurement instrument for the three different types of CL (**Figure 1**). In particular, the assessment of cognitive load resulting from instructional design (ECL; multimedia scaffolds) and its interactions/relationships with causal factors (representation preference; cognitive abilities) is of great interest for the present study.

(b) A possible convergence between the subjective measurement of CL with an objective instrument (eye-tracking metrics) will be checked. Both measures the load during the use of multimedia scaffolds and the related task for the *planning* phase of the inquiry process. Furthermore, a deeper investigation of possible correlations and influences of causal factors (cognitive-verbal and visual abilities, reading skills, and representation preference) through different analysis steps will be proofed.

According to this, the following overall, methodical research question will be investigated: *Do subjective ratings of cognitive load (ECL, ICL, and GCL) correlate with other measures or indicators of CL?* To investigate the research question, we made different validation assumptions that are differentiated *via* hypotheses and empirically tested in two studies. However, only assumptions and hypotheses are listed where we expect significant correlations, influences, and effects in respect to the CL derived from the theory.

Assumption 1) The subjective measurement has an internal consistency and separates the relevant subscales – ECL, GCL, and ICL.

H1.1: By adapting the items from previously tested instruments of CL (Cierniak et al., 2009b; Leppink et al., 2013; Klepsch et al., 2017), the three subscales ECL, GCL, and ICL can be identified both verbally and by statistical characteristics (internal structure and internal consistency).

H1.2: The CL subscales show positive and negative correlations to each other. The ECL scale correlates negatively with the GCL; the ICL scale correlates positively with the GCL (Leppink et al., 2013; Klepsch et al., 2017).

Assumption 2) The subjective instrument leads to different cognitive load levels for the respective students in grades 9 and 11.

H2.1: Higher prior knowledge, which students should have in grade 11, should result in lower CL than that of grade 9 students (Kalyuga, 2013; Chen et al., 2018; Richter et al., 2018; Seufert, 2019).

Assumption 3) Students' specific scores in the causal factors – cognitive-verbal and visual abilities, reading skills, spatial ability, and representation preference affect the individual CL (ECL, GCL, and ICL).

H3.1: There is a correlation between cognitive-verbal ability scales and CL, especially for text-based (monomodal) scaffolds (Kuhn et al., 1988; Mayer and Massa, 2003; Lehmann and Seufert, 2020).

H3.2: Cognitive-visual ability correlates negatively with ECL and positively with GCL (Jäger, 1984; Jäger et al., 2017).

H3.3: For the selected text-based (monomodal) scaffolds, reading-skill scores correlate negatively with ECL scores and positively with GCL scores (Plass et al., 2003; Lehmann and Seufert, 2020).

H3.4: Spatial ability may have an impact on CL when using static scaffolds (static image). It is possible that a high spatial ability minimizes the ECL and contributes to the increase of the GCL (ability-as-compensator hypothesis; Höffler, 2010; Münzer, 2012).

H3.5: The individual representation preference will condition the choice of a multimedia digital scaffold. The choice and the resulting possible fit between representation preference and use of the scaffold will be reflected in a low average ECL (Kürschner et al., 2005; Lehmann and Seufert, 2020).

Assumption 4) As an objective instrument, visual attention in terms of specific eye-tracking metrics (total fixation count; mean fixation

duration; and total fixation count) shows a high convergence with the levels of subjectively measured cognitive load (ECL, ICL, and GCL).

H4.1: There is a significant correlation between the total fixation count and the level of CL. The total fixation count predicts the ECL (Zu et al., 2020).

H4.2: The mean fixation duration and total fixation duration are significantly related to the level of CL. The mean fixation duration and total fixation duration predict the ECL (Zu et al., 2020).

MATERIALS AND METHODS

General Design of the Study I and II

To test the subjective measurement of CL, two studies were conducted within FLOX, a teaching-learning laboratory of the University of Kassel. In both studies, the intervention, measurement times, and tasks were identical. Thus, students experimented in small groups (three students) along the scientific inquiry process on a selected topic of metabolic physiology (see "Environment (E) & Task (T) as causal factor – scaffolding & inquiry learning"; "Procedure"). The experiment as well as the experimental components, the scaffolds, the instructional material, and group size did not differ in both sub-studies. However, the difference was in the playback device of the scaffolds and in the time duration of the experiment (see "Procedure"; "Scaffolds & Task"). Whereas in study I, the scaffolds were provided to the participants *via* a tablet, in study II, the students watched their scaffold on a computer. In contrast to study I, which was performed with an entire class (environment), study II could only be conducted with three students at a time (in a similar total duration of the laboratory day in study I), and therefore, the experiment designed by the students was not performed in study II. The total duration of the laboratory day was 3–4 h.

Participants

Participants in both studies ($n = 250$, 53.2% female) were 9th ($n = 142$) and 11th ($n = 108$) grade students from schools in Kassel, Germany (see **Table 1** for details). Divided into the two sub-studies, study I involved 181 students (9th grade: $n = 111$; 11th grade: $n = 70$), whereas the second study employed a total of 69 students (9th grade: $n = 31$; 11th grade: $n = 38$). Both student groups and their legal guardians were given written and verbal information about the study. Based on this, both parties had to give their written consent to participate in the study. This was voluntary and all student data during the intervention were recorded anonymously.

Procedure

One week before each experimental-laboratory day, a pre-instruction including a pre-assessment was conducted within regular school hours. At this first measurement point, the students' demographic data, their cognitive abilities (Cognitive Ability Test: Heller and Perleth, 2000), and reading skills (Reading Speed and Reading Comprehension: Schneider et al., 2017) were assessed in a written form (**Table 2**). On the laboratory day, small groups of three students were first randomly assigned. Following this, the general procedure of the laboratory day was explained with the participants. In addition, the individual scaffolds (static image-text = image-text; static image-audio = image-audio, moving image-text = animation; and moving image-audio = video) were introduced by content-remote examples since the subsequent selection of the preferred scaffold was made independently by the students. After examining the phenomena "A mushroom in the pizza dough" in a video jointly in the classroom/laboratory (study 1) or in groups (study 2), the independent, experimental work along the inquiry-based learning process within the small groups followed. After observing the phenomenon, the students were first invited to independently pose a research question (for example: *What influence does temperature have on the activity of yeast?*). Afterward, the groups formulated two well-founded hypotheses in line with the research question, which they would like to test in their experiment. The hypothesis formulation was followed by the *planning* phase of the experiment, which is also the intervention phase. At the beginning of the intervention, the students were initially asked to continue working individually. Along their protocol, in which the students documented all results of the experimentation phases, they were first instructed to individually select one of the four provided scaffolds. In study I, they viewed them on a tablet. In study 2, they used a computer with a stationary eye-tracking system (Tobii-Pro X3-120; 120 HZ; <0.4°; 22-inch screen). Before each eye-tracking recording, a short technical briefing and an individual 9-point calibration were performed. To optimize the eye-tracking recording and to minimize any unnecessary distractions of the participants, some preparations were made, such as the light controlling of the eye-tracking areas and the separation of working areas (Kastaun et al., 2020; Kastaun and Meier, 2021). Directly after using the multimedia scaffolds, a subjective measurement of individual CL was integrated into their protocol. After completing the CL test, students had the task to operationalize the variables for their experiment before going into the group work again. Then, they developed a complete plan in the group to test one of their hypotheses. Afterward, the groups conducted their experiment together (study I) or used their planning anticipated videos that showed the experimental implementation and matching value tables (study II). After that, the students interpreted the results in reference to the research question and hypotheses.

Scaffolds & Task

In both studies, the students were required to choose one option from the same list of multimedia scaffolds (**Figure 2**). These differed in the combination and modality of their respective individual representations. Image-text and animation represent the monomodal scaffolds, whereas image-audio and video can be summarized as multimodal scaffolds with spoken text. The design of the graphics and the content of all scaffolds are identical and comparable according to cognitive-psychological principles (among others Mayer, 2014). The scaffolds have about the same time duration. The length is between 2.5 and 3 min. The content consisted of written and spoken text as well as

TABLE 1 | Demographic data from the studies.

		Total sample (Study I+II)			*Study I*			*Study II*		
		M	***SD***	***n***	***M***	***SD***	***n***	***M***	***SD***	***n***
Age, years		15.82	1.10	250	15.77	1.13	181	15.95	1.05	69
Sex, % female		53.2		250	48.1		181	66.7		69
Grade	**9th**			142			111			31
	11th			108			70			38

TABLE 2 | Descriptive data of used measurement of assessment and causal factors for the total sample.

	M	***SD***	***Cronbachs α***	***Items per Scale***	***References***
ASSESSMENT FACTORS					
Extraneous cognitive load (ECL)	2.17	1.01	0.891	5	Klepsch et al., 2017;
Germane cognitive load (GCL)	4.04	1.06	0.814	3	Cierniak et al., 2009b;
Intrinsic cognitive load (ICL)	3.50	1.34	0.696	2	Leppink et al., 2013
CAUSAL FACTORS					
Cognitive abilities					
Cognitive-visual ability (Figure classification, N01)	13.93	8.88	0.895	24	Heller and Perleth, 2000
Cognitive-visual ability (Figure Analogy, N02)	12.92	4.22	0.830	24	
Spatial ability (Paper Folding, N03)	7.37	3.31	0.790	14	
Cognitive-verbal ability (Word Analogy, V03)	7.5	2.6	0.700	19	
Reading skills					
Reading comprehension	46.75	26.26	0.82	47	
Reading speed	41.62	28.36	0.80		Schneider et al., 2017
Reading accuracy	64.72	25.32	0.82	47	

graphical representations to summarize the methodological skills/knowledge for the *planning* phase. Especially, the concept of variable operationalization and the creation of measurement series and associated measurement concept are focused without reference to any specific domain content (Meier and Kastaun, 2021). After the use of a scaffold and the cognitive load measurement, the students were asked to identify the dependent and independent variables (see "Environment (E) & Task (T) as causal factor – scaffolding & inquiry learning"). To do so, the students needed to use their prior knowledge as well as the information provided in the scaffold and transfer it to the specific subject context of the experiment. The use of the scaffold as well as the identification of the dependent and independent variables constitutes the task of the environment (**Figure 1**).

Instruments & Methods

Different instruments and evaluation procedures were used to test the validation hypotheses. **Table 2** lists the individual test instruments and associated scales.

Cognitive Load

The subjective instrument for measuring individual CL was adapted and composed of different items from validated test instruments (Cierniak et al., 2009b; Leppink et al., 2013; Klepsch et al., 2017; see **Supplementary Table A**). With the aim to measure the different load categories, ECL, GCL, and ICL, during learning with multimedia scaffolds, the paper-based questionnaire initially consisted of 15 different items (ECL = 7 items; GCL = 4 items; and ICL = 4 items). These items were composed and adapted in such a way that they can be independently reflected upon and assessed by 9th and 11th grade students and have a primary relationship to the scaffolds. In addition, the subjective instrument is primarily intended to assess the CL based on the multimedia scaffolds. Therefore, the items were linguistically adapted and composed in this manner. Using a 6-point Likert scale from *strongly disagree (1)* to *strongly agree (6)*, the students' self-reports were recorded in relation to the respective items. After the survey, we performed a Principal Component Analysis of the obtained datasets (using Varimax rotation) and the CL scales were formed by merging discrimination indices. The Cronbach's α-values were determined to verify the internal consistency.

Reading Speed and Reading Comprehension Test

The Reading Speed and Reading Comprehension Test (Schneider et al., 2017) is an established test instrument in school diagnostics and is used for the differentiated determination of reading speed, reading accuracy, and reading comprehension

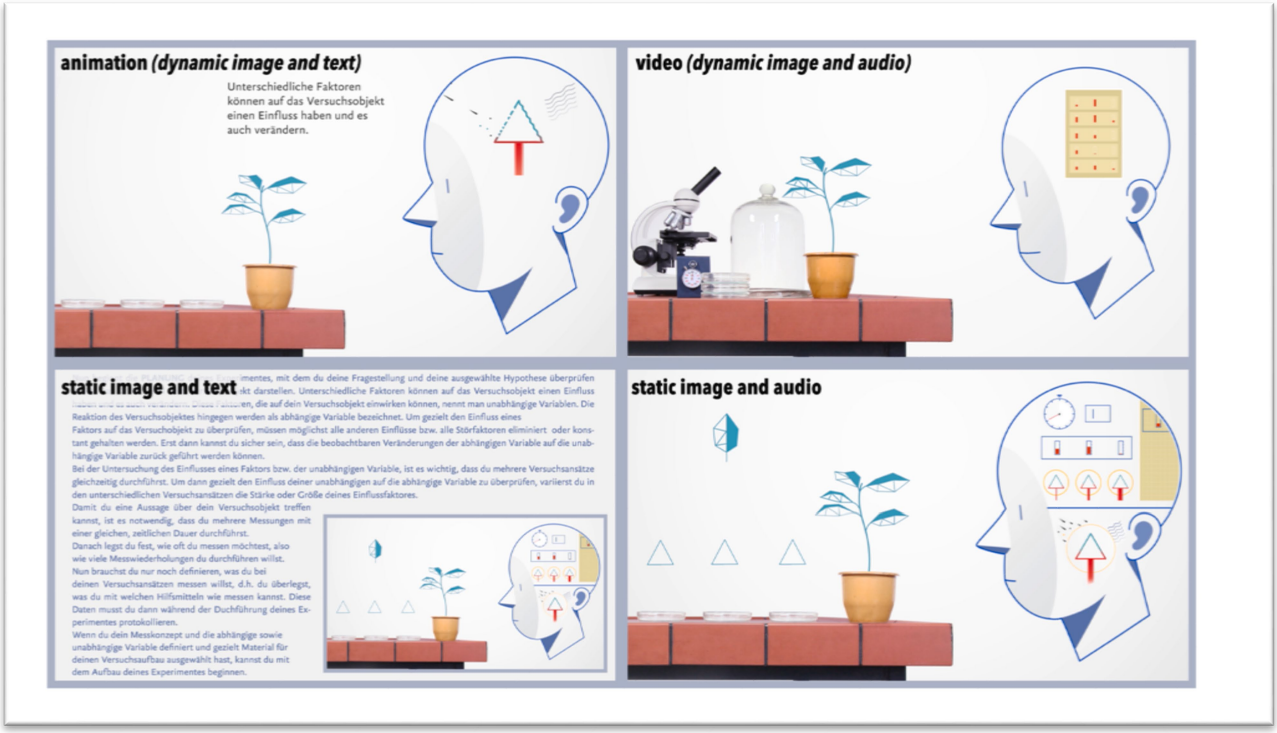

FIGURE 2 | Image sections of the scaffolds in the different multimedia representation combinations.

from grades 5 to 12. It is a speed-power test in which a given text must be read as quickly and accurately as possible within a fixed time (6 min). In the text, items are integrated at individual points, which must be used to complete the sentence by selecting the correct word. By analyzing the correctly crossed words, the reading accuracy across the text as well as the reading comprehension can be evaluated by a point system. To record reading speed, the number of all words read within the prescribed time was counted and evaluated using an evaluation scale.

Cognitive Ability Test

The Cognitive Ability Test (Heller and Perleth, 2000) is also an established test instrument in school diagnostics. It is used for the differential measurement of cognitive ability dimensions that are particularly relevant for learning at school. Scales of cognitive-verbal (V03) and visual abilities (N01, N02) as well as spatial ability (N03) were used from this. The individual instruments primarily aim to measure differential content-based, processing capacity. The Cognitive Ability Test is also a speed-power test. Within a prescribed time (8 min), participants had to solve the individual test items of the respective scales. The visual, series tasks (N01), visual, and verbal relation tasks (N02 & V03) as well as paper-folding tasks measuring the spatial ability (N03: paper-folding test) were analyzed by an evaluation scale (correct/wrong), and the expressions of the individual constructs were identified by the formation of total scores.

Representation Preference

Representation preference was generated by an independent and free choice of the scaffold. The evaluation is generated *via* a frequency analyses of the individual usage behavior.

Gaze Behavior

Within study II ($n = 69$), students' gaze data were recorded using a stationary eye-tracking system. Using the software Tobii-Pro Studio, all students with a poor calibration were eliminated from the sample. In this work, we analyzed the learning stimuli as a whole, i.e., there was a single AOI that covered all information provided in the learning material. Afterward, the total fixation count, mean fixation duration, and total fixation duration were extracted using the software. Since the scaffolds have different characteristics in information retrieval and processing, the data were first z-standardized.

Data Analysis

In the following analysis, we excluded datasets that showed missing values or too large measurement inaccuracies (outliers and dropout: $n = 23$) or calibration errors (eye-tracking datasets, $n = 4$; Gaze Sample under 65%) in the main intervention. Measurement data from the Reading Speed and Reading Comprehension Test and the Cognitive Ability Test may have different samples to the overall or sub-study

as these datasets are partially incomplete. Nevertheless, these subjects were included in the main analysis regarding representation preference and CL since the complete datasets were available. The IBM SPSS Statistic 27 software was used for the statistical analyses.

To investigate the validity assumptions and research hypotheses, correlation and regression analyses as well as methods of inferential statistics (*t*-test and ANOVA) were applied for both studies. First, the internal consistency was examined in more detail by factor analytic procedures and correlation analyses of the individual scales of the CL test (validity assumption 1). To test the second validity assumption, different *t*-tests were conducted to analyze the assumed differences in the types of cognitive load (ECL; GCL; and ICL) between the two grade levels (9th and 11th grade). In addition, group differences (ANOVA) between the selected scaffolds (representation preference) and the expressions of the ECLs, GCLs, and ICLs were analyzed for each grade in order to exclude possible variance due to the different multimedia representations. Along the third validity assumption, correlations with the individual scales of the Reading Speed and Reading Comprehension Test as well as the Cognitive Ability Test in relation to the individual expressions of CL were conducted. Multiple regression analyses were used to further examine these correlations to identify predictors of CL during the use of the various scaffolds. No violations of the predictors were identified for any of the regressions. The Durbin-Watson statistics for all regressions ranged from 1 to 3 and no autocorrelations between variables were present (VIF < 10). Furthermore, possible correlations (multiple regressions with categorical variables) and differences (*t*-test) between representation preference and the expression of representation preference were investigated. The quantitative data from the second study were used to test the fourth validity assumption. For this purpose, the *z*-standardized eye-tracking metrics (total fixation count; mean fixation duration; and total fixation count) were correlated with the individual expressions of CL.

TABLE 3 | Parallel-test reliability *via* Pearson-correlations ($n = 250$).

		GCL	ICL
ECL	*r*	−0.38	0.23
	p	**0.00**	**0.00**
GCL	*r*		0.03
	p		0.61

p < 0.05.

RESULTS

Validity Assumption I

We performed a Principal Component Analysis (Varimax rotation) to extract the most important independent factors (KMO = 0.751; the Bartlett Test was highly significant; $p < 0.001$). Kaiser's criteria and the scree-plot yielded empirical justification for retaining three components (value ≥1) which accounted for 65.76% of the total variances. From these dimension-related items, the individual scales of CL were composed and reduced from an initial set of 15 items to 10 (ECL = 5 items, GCL =3 items, and ICL = 2 items; see **Supplementary Table A**). Items were excluded that either had too low factor loadings (<0.3) or loaded too high on another factor that was not integrated in the construct. These resulting 10 items and three scales were checked for internal consistency using Cronbach's α (**Table 2**) and correlation analyses of the individual subscales (**Table 3**). The examination of the internal consistency by the Cronbach's α-values of the individual subscales and the total scale for the CL shows satisfactory characteristic values (**Table 2**). As expected, statistically significant correlations between the ECL and GCL ($p < 0.001$) and ECL and ICL ($p < 0.001$) were found. However, the scales of the GCL and ICL did not show a statistically significant relationship ($p = 0.61$; **Table 3**).

Validity Assumption II

To test the second validity assumption stating that CL decreases in progressive, school-based education due to an increase in methodological skills, group comparisons ($n = 250$) were made regarding the CL between 9th and 11th grade students.

TABLE 4 | Differences between 9th and 11th grade in cognitive load (*t*-test, $n = 250$).

	Grade 9th ($n = 142$)		Grade 11th ($n = 108$)		
	M	***SD***	***M***	***SD***	
ECL	2.41	1.06	1.87	0.87	(95% − CI[0.29, 0.78])/$t(248) = 4.26$, $p < 0.001$, $d = 0.552$, $r = 0.26$
GCL	3.90	1.07	4.22	1.03	(95% − CI[−0.57, −0.04])/$t(248) = -2.27$, $p = 0.024$, $d = -0.295$, $r = 0.14$
ICL	3.94	1.27	2.92	1.22	(95% − CI[0.70, 1.33])/$t(248) = 6.36$, $p < 0.001$, $d = 0.822$, $r = 0.38$

TABLE 5 | Pearson correlation between causal and assessment factors for the total sample.

		Cognitive-verbal ability (V03)	Cognitive-visual abilities (N01/02)		Spatial ability (N03)	Reading comprehension	Reading speed	Reading accuracy
ECL	*r*	−0.167	0.066	−0.085	−0.055	−0.112	−0.111	0.023
	p	**0.013**	0.323	0.202	0.409	0.143	0.148	0.766
GCL	*r*	0.143	−0.111	−0.001	0.067	0.163	0.190	0.044
	p	**0.034**	0.095	0.985	0.320	**0.034**	**0.013**	0.571
ICL	*r*	−0.062	0.118	−0.125	−0.082	−0.107	−0.125	−0.025
	p	0.359	0.079	0.063	0.226	0.164	0.102	0.744

p < 0.05.

TABLE 6 | Multiple regression with ECL and cognitive abilities ($n = 222$).

Coefficient	*B*	*SE(B)*	*BETA*	*P*	*VIF*
(Constant)	2.460	0.300			
V03	−0.068	0.028	−0.172	**0.016**	1.155
N03	−0.007	0.025	−0.024	0.765	1.446
N02	−0.020	0.020	−0.083	0.310	1.515
N01	0.040	0.020	0.152	**0.046**	1.315

$R^2 = 0.047$, Durbin-Watson-Statistic = 1.870. $p < 0.05$.

TABLE 7 | Multiple regression with GCL and cognitive abilities ($n = 222$).

Coefficient	*B*	*SE(B)*	*BETA*	*P*	*VIF*
(Constant)	4.081	0.314			
V03	0.069	0.029	0.168	**0.019**	1.155
N03	0.033	0.026	0.101	0.203	1.446
N02	−0.005	0.021	−0.021	0.798	1.515
N01	−0.052	0.021	−0.189	**0.013**	1.315

$R^2 = 0.052$, Durbin-Watson-Statistic = 1.677. $p < 0.05$.

TABLE 8 | Multiple regression with ICL und cognitive abilities ($n = 222$).

Coefficient	*B*	*SE(B)*	*BETA*	*P*	*VIF*
(Constant)	3.471	0.397			
V03	−0.024	0.037	−0.046	0.516	1.144
N03	−0.033	0.032	−0.080	0.314	1.435
N02	−0.052	0.026	−0.162	**0.048**	1.509
N01	0.079	0.026	0.228	**0.003**	1.310

$R^2 = 0.055$, Durbin-Watson-Statistic = 2.133. $p < 0.05$.

The groups differed statistically significantly in all scales of CL (**Table 4**). A large effect was shown according to Cohen (1988) regarding the ICLs ($d = 0.822$). The 11th grade students rated their intrinsic CL about one value lower than the 9th grade students. Medium effects were identified in the ECLs ($d = 0.552$). Regarding GCL, the groups differed statistically significantly but only with a small effect size ($d = 0.295$). Thus, the GCL conducive to learning increases as expected in the higher grade level by a value of 0.3. To support the results and to strengthen the link to prior knowledge as an influencing variable, single factor ANOVAs with Bonferroni-correction were conducted on the expression of the individual types of CL and the representation preference (choice of scaffold). The results show that only the ECL ($F(3,3) = 3.359$, $p = 0.021$, $\eta_p^2 = 0.068$) in the 9th grade differed significantly between the learners who chose an image-text combination ($n = 37$; $M = 2.82$; $SD = 1.08$) and those who, in contrast, chose a video as a scaffold ($n = 56$; $M = 2.13$; $SD = 1.04$). Furthermore, there were significant differences within the 11th grade with respect to the GCL ($F(3,3) = 3.187$, $p = 0.027$, $\eta_p^2 = 0.085$) between the groups of animations ($n = 29$; $M = 4.56$; $SD = 0.86$) and picture text ($n = 38$; $M = 3.86$; $SD = 0.97$). For the ICL and between the other groups regarding the ECL and GCL, no significant differences could be identified (see **Supplementary Material**).

Validity Assumption III

To test the third validity assumption, correlation and regression analyses were conducted and evaluated for the total sample ($n = 250$; **Tables 5–11**) and the subsamples (study I: $n = 181$; study II: $n = 69$; see **Supplementary Material**). Negative correlations between cognitive-verbal abilities (V03) and the expressions of the ECL ($r = 0.17$; **Table 5**) as well as positive ones with the GCL ($r = 0.14$, **Table 5**) were found. Moreover, showing a small correlation coefficient, the expressions of reading speed ($r = 0.19$) and reading comprehension ($r = 0.16$) are positively related to the scale of the GCL's. Regarding the cognitive-visual and spatial scales (N01, N02, and N03) as well as the reading accuracy scale, no other statistically significant correlations could be identified (**Table 5**). Correlation analyses within the sample of the first study ($n = 181$, see **Supplementary Material**) further revealed statistically significant correlations between the ICL and the expression of reading comprehension ($r = 0.21$). Moreover, significant correlations between reading speed ($r = 0.21$) and cognitive-visual ability (N01; $r = 0.16$) and the expression of the GCL were identified. However, for the second study ($n = 69$, see **Supplementary Material**), only one significant correlation could be found. Cognitive-visual ability (N02) correlated significantly negatively with the ICL scale ($r = 0.29$). Apart from that, multiple linear regressions were used to examine the cognitive abilities (N01, N02, N03, and V03) and reading skill scales as possible predictor variables for CL in the use of multimedia scaffolds. Regarding the reading-skill scales for reading comprehension, reading accuracy and reading speed have no influence on the individual CL (see **Supplementary Material**). In contrast, the cognitive-verbal and visual abilities (V03, N01, and N02) scales showed a significant

TABLE 9 | Selection frequencies of scaffolds.

Scaffolds	*Selection frequencies*					
	Total sample (n = 249)		*study I (n = 180)*		*study II (n = 69)*	
	nj	*hj*	*nj*	*hj*	*nj*	*hj*
image-text	75	30.1	53	29.4	22	31.9
image-audio	20	8.0	15	8.3	5	7.2
animation	62	24.9	46	25.6	16	23.2
video	92	36.9	66	36.7	26	37.7

nj = absolute frequency = sum of used scaffolds in the respective format; hj = relative frequency (in percent) = relative share of the selected scaffolds in relation to the total number of scaffolds used.

TABLE 10 | Multiple regression between ECL and representation preference ($n = 250$).

Coefficient	*B*	*SE(B)*	*BETA*	*P*	*VIF*
(Constant)	1.951	0.105			
image-text	0.416	0.156	0.188	**0.008**	1.268
image-audio	0.449	0.248	0.120	0.072	1.120
animation	0.416	0.165	0.165	0.138	1.257

$R^2 = 0.033$, Durbin-Watson-Statistic = 1.939. $p < 0.05$.

TABLE 11 | Multiple regression between GCL and representation preference ($n = 250$).

Coefficient	*B*	*SE(B)*	*BETA*	*P*	*VIF*
(Constant)	4.170	0.110			
image-text	–0.346	0.165	–0.149	**0.037**	1.268
image-audio	–0.270	0.261	–0.069	0.302	1.120
animation	–0.009	0.174	–0.004	0.959	1.257

$R^2 = 0.023$, Durbin-Watson-Statistic = 1.873. $p < 0.05$.

influence on the ECL (**Table 6**), GCL (**Table 7**), and ICL (**Table 8**) within the total sample. Multiple linear regressions with categorical variables were performed to investigate the influence of representation preference on the different types of CL. The results show a significant difference regarding the expression of the ECL and GCL between the learners who used scaffolds with static image and text and those who learned with the video (**Tables 10**, **11**). The ECL of students who learned with static images and text was significantly higher than for the video users ($p = 0.008$), whereas GCL decreased significantly ($p = 0.037$). As expected from the validity assumption III, the representation preference that is equivalent to the voluntary choice of the scaffold had no further significant influence on the expression of CL (see **Supplementary Material** and **Tables 10**, **11**). For an in-depth analysis of possible correlations or differences between representation preference and the expression of reading skills, cognitive abilities and CL, group differences between those who used multimodal and those who used monomodal scaffolds in the *planning* phase were examined. Differences were found with respect to the expression of reading speed

TABLE 12 | Pearson correlation with eye-tracking metrics and cognitive load scales ($n = 69$).

		Total fixation count	**Mean fixation duration**	**Total fixation duration**
ECL	*r*	0.247	–0.233	0.065
	p	**0.041**	0.055	0.596
GCL	*r*	–0.079	0.169	0.089
	p	0.517	0.165	0.467
ICL	*r*	–0.020	0.140	0.056
	p	0.871	0.252	0.646

$p < 0.05$.

TABLE 13 | Linear regression between ECL and total fixation count ($n = 69$)

Coefficient	*B*	*SE(B)*	*BETA*	*P*	*VIF*
(Constant)	–0.118	0.103			
Total fixation count	0.216	0.04	0.247	**0.041**	1.000

$R^2 = 0.061$, Durbin-Watson-Statistic = 2.220. $p < 0.05$.

(95%-CI[0.31, 17.59]/$t(167) = 2.05$, $p = 0.042$, $d = 0.318$) and ECL (95%-CI[0.01, 0.51]/$t(247) = 2.01$, $p = 0.046$, $d = 0.256$, $r = 0.132$). Accordingly, users of a monomodal scaffolds exhibited significantly higher reading speed as well as ECL than those who selected and used a multimedia one. Furthermore, no statistically significant differences between the multimedia and monomodal users could be identified (see **Supplementary Material**). In addition, group differences between good and poor readers were analyzed regarding the GCL, ECL, and ICL scores for the text-based (monomodal) scaffolds. Again, no significant differences between the individual groups and the expression of cognitive load were found (see **Supplementary Material**).

Validity Assumption IV

To test the fourth validity assumption, correlations (**Table 12**) were first calculated between the z-standardized eye-tracking metrics and the expressions of CL. Within the sample ($n = 69$), a significant correlation between the total fixation count and the expression of ECL was found ($r = 0.247$). Although no other significant correlations were evident, the correlation coefficient between the mean fixation duration and the expression of ECL, which is just above the significance level, also indicated a possible relationship. An examination of the total fixation count as a predictor for the individual CL revealed no significant results for GCL and ICL. In contrast, the total fixation count showed a significant effect on ECL (**Table 13**).

DISCUSSION

The aim of the study was to test a subjective instrument for measuring differential CL for 9th and 11th grade students across different validity assumptions. Hereby, possible relationships between causal factors (learner characteristics: e.g., cognitive, verbal and visual abilities, reading skills, and representation preference) and

assessment factors of CL were investigated (**Figure 1**). Furthermore, an objective measurement method (eye-tracking: total fixation count, fixation duration, fixation, and mean fixation duration) was used. With the constant inclusion of the overarching methodological research question (see "Introduction"), different analyses were carried out for the respective assumptions (see "Aims of the Study, Research Question, Assumptions & Hypotheses"). In the following, the individual validity assumptions will be discussed and combined in a final concluding discussion in relation to the research question.

Discussion of Validation Assumptions I–IV

The first assumption targets the internal consistency as well as significant correlations between the expected scales of the ECL, ICL, and GCL for validity testing. The constructed subjective measurement instrument is closely based on existing test items (Cierniak et al., 2009b; Leppink et al., 2013; Klepsch et al., 2017). Even with language and content adaptation, the evidence-based structure to the CL is confirmed in the dimensions of the ECLs, ICLs, and GCLs (H1.1). The examination of internal consistency shows satisfactory to very excellent Cronbachs-alpha values (see "Validity Assumption I") and, except for the values from the ICL, coincides with the results from other studies (Klepsch et al., 2017). In this context, possible reasons can be identified in the linguistic construction as these items of the ICL (*for example: "I found the identification of the dependent and independent variables to be difficult.,"* see **Supplementary Table A**) did not directly refer to the use of scaffold but to the definition of the individual variables, and thus, to the complexity of variable operationalization (Leppink et al., 2013; Klepsch et al., 2017; Klepsch and Seufert, 2021). In contrast, the GCL and ECL items refer almost exclusively to the scaffolds and less to the task within the *planning* phase of the experiment (see **Supplementary Table A**). This might be one reason why the GCL and ECL scales are not significantly correlated with each other, in contrast to the ECL and ICL scales (Leppink et al., 2013; see "Introduction"). Furthermore, the results confirm the additive relationship between the ICL and ECL (see "Introduction"; Sweller et al., 2011).

An examination of possible differences between the two grades 9th and 11th regarding the expression of CL comprises the second validity assumption. The group comparisons show results in line with expectations that the expression of CL and its categories ECL, GCL, and ICL is less pronounced in 11th grade students than in 9th grade. This can be attributed to differently developed prior knowledge regarding the content, and it is likely to constitute a consolidation of the more developed self-regulation of the learners from 11th grade (among others Richter et al., 2018; Eitel et al., 2020; H2.1). Considering the group differences in the respective grade level regarding the used scaffold and the expression of the ECL, ICL, and GCL, the results do not only indicate a difference in subject-related knowledge. In contrast to the older 11th grade students, who did not differ significantly in ECL, significant differences appear among the younger students between those who used an image-text combination and those who chose a video (see "Validity Assumption II"). This suggests that either the instructions from the video-based scaffold were easier to relate to each other due to their auditory and visual layout (Paivio, 2006; Mayer, 2014) or that the learners tended to overestimate their own abilities in dealing with the image-text representation combination. The last is of course also strongly related to the subject-specific complexity as well as the formulation of the ECL items, since, for example, "*unclear language*" (see **Supplementary Table A**) can be understood ambiguously. Furthermore, the significant difference in grade 11 in relation to the GCL among the learners who chose an image-text combination or an animation can be interpreted in such a way that the students encountered a hurdle during the linking and extraction of the content from the scaffolds (see "Validity Assumption II"). The dynamic presentation of text and images in the animations may have minimized the mental interconnection of the information. Another explanation could be that the image-text combination supports repetitive reading and self-regulated comprehension due to the static presentation (Kastaun and Meier, 2021).

Regarding the third validity assumption that the individual expression of the cognitive-visual and verbal abilities, reading skills, spatial ability, and representation preference (causal factors) is related to the CL, a differential overall assessment emerges. In comparison with the results of the spatial ability, there are correlations to the other causal factors from which different explanatory approaches can be derived in relation to the overarching question regarding the validity test of the subjective test instrument (H3.4). Cognitive-verbal and visual abilities (N01, N02, and V03) as well as reading skills are included as variables in some studies on CL but are not further investigated in a differentiated analysis of CL (e.g., Cierniak et al., 2009b). Furthermore, different test instruments exist and can be used to measure this, which makes it difficult to transfer the results in respect to the measured construct (e.g., verbal reasoning ability; e.g., van de Weijer-Bergsma and van der Ven, 2021). Thus, it is currently not clear how cognitive abilities relate to the individual scales of CL. In the present study, relationships and influences of cognitive-verbal and visual abilities as well as influences on CL and its differentiated categories could be identified (**Tables 5–11**; H3.1; H3.2; and H3.3). Possible explanations can be found, among others, in the construction of the items and the item structure (see "Cognitive Ability Test," relationship and series tasks) of the individual scales of the cognitive ability test used in this study (Heller and Perleth, 2000). These scales (N01, N02, and V03) measure content-based skills by determining logical thinking using content-based, here, visual and verbal examples *via* series/classification (N01) and ratio/analogy tasks (N02 & V03). According to Jäger et al.'s (2006) intelligence model, in which the scales measure verbal and visual abilities under a constant reference to the individual thinking ability, content-based cognitive abilities contribute to a person's general intelligence. In terms of CL, this means that the correlations are not only due to students' verbal or visual level, but also due to the ability to easily recognize links between different information. Furthermore, the results can also be explained by the construction of the scaffolds. The primary information carriers of the multimedia scaffolds are the verbal or textual representations,

i.e., the visual representations also contribute to content development but without the verbal or text-based elements they do not explain the main content.

Considering the multimedia learning theory (Moreno, 2010; Mayer, 2014), which states that verbal and text-based information can be reorganized in working memory *via* images and sounds, high verbal cognitive reasoning ability can help minimize ECL and increase GCL. This is also evident from a closer look at the subsamples (see "Validity assumption III" and **Supplementary Material**). The explanatory approach, which implies that those with a higher expression of cognitive-verbal abilities benefit more than those who show an increased visual ability due to the design characteristics (among others Choi and Sadar, 2011), is supported by the significant negative correlations present in study I and II (see "Validity assumption III") as well as by the findings of the multiple regression analyses. These show a strong to moderate effect of cognitive abilities on the individual load categories. In respect to the scaffolds used, cognitive-verbal abilities predicted low ECL and higher GCL, whereas cognitive-visual abilities scales did not predict any of the CL categories (see "Validity assumption III").

Concerning the study of reading skills as a causal factor to CL, the results show that positive relationships exist between reading comprehension as well as reading speed and GCL. On the one hand, this can be attributed to the fact that cognitive-verbal reasoning skills are closely and positively related to the individual reading skills (Schneider et al., 2017). On the other hand, it shows that the reading process is an active construction activity (Artelt et al., 2007; Scheerer-Neumann, 2018). Thus, the significant correlations may result from the higher load based on the selected scaffold (monomodal) but also from the text-based self-assessment of the CL. Accordingly, the text-based test itself may have elicited a correlation between the GCL and reading comprehension or reading speed. Contrary to the identified correlations, the results of the multiple regression analyses do not show any significant influences of the reading skills on the three types of CL (see **Supplementary Material**). Further analyses between the use of the monomodal/multimodal scaffolds and the expression of cognitive load in relation to the reading skills also showed differentiated results. It is possible that the reading skills do not have a significant influence on the decision of the selected scaffold but still show advantages in the processing of the information (based on the modality) about the load.

The relationship between spatial ability and CL (H3.4), which has been supported and assumed by many studies (e.g., Höffler, 2010; Anvari et al., 2013), is not evident in the present study. As already stated by Höffler and Leutner (2011), among others, it also becomes clear with the present results that the spatial ability only has a significant influence on task performance if the task itself requires it. Consequently, it can be concluded that spatial awareness is not required for processing the information of the scaffolds used here.

Finally, the assumed correlations between the causal factor of representation preference and CL can be partially confirmed (H3.5). Independent of the selected scaffold, the ECL shows a very low expression (see "Procedure"). Multiple regression analyses confirm that the representation preference, i.e., the choice of scaffold, influences ECL and GCL. Thus, those students who show a preference for the image-text representation combination are more likely to experience significantly higher ECL than those who use the video. This is also evident in the GCL, which negatively predicts preference for image-text compared to video users. This can be attributed to the level of complexity of the text-based or auditory text (Leahy and Sweller, 2011; Lehmann and Seufert, 2020) and/or to the multimedia presentation mode of the verbal text in terms of simultaneous, dynamic visualizations (Moreno, 2010; Mayer, 2014) or due to the students' failure to assess what kind of multimedia representation they could best handle in the situation (instruction).

The examination of a possible convergence between the subjective and objective instruments for CL (validity assumption IV) showed results partially confirming to expectations. The positive correlation (**Table 12**) and the significant influence (**Table 13**) of the absolute fixation count support the assumed convergence (H4.1; Zu et al., 2020). Zu et al. (2020) found a moderate negative correlation between the ECL and the mean fixation duration and a weak correlation between the total visit duration on an animation and the GCL. In our work, we found a weak negative correlation between the ECL and the mean fixation duration that was just above the significance level ($p = 0.55$) and a significant positive correlation between total fixation count and the ECL. This implies that previously reported relationships between eye-tracking measures and cognitive load do not translate to our setting. The examination of other eye-tracking metrics, such as transitions between the AOIs, was not possible based on the material used here since possible selection and integration processes cannot be investigated unambiguously *via* visual attention when including auditory scaffolds.

Limitations and Implications for Future Work

Along the individual validity assumptions based on the theoretical interpretation-use argumentation (Kane, 2013), arguments for a possible valid measurement *via* the test scores of the measurement instruments are used (causal & assessment factors; objective instruments), and their relationships to each other are identified. The central goal of validating a subjective measurement instrument for 9th and 11th grade students was partially achieved by confirming individual validity assumptions under study-specific conditions. The results support a valid measurement but also indicate that a validity test of a subjective measurement instrument for CL is associated with different difficulties. On the one hand, CL can be influenced by diverse causal factors in different ways (**Figure 1**; Choi et al., 2014), which is particularly related to the construction of the task and/or environment as well as its interaction with CL. On the other hand, a difficulty arises from the fact that there are no comprehensive, tested relationships between the individual subjective and objective measurement instruments for CL (assessment factors) in an evidence-based manner yet. Particularly regarding the cognitive abilities for visual and verbal reasoning, more detailed investigation is needed to determine whether these factors influence CL in a similar way as prior knowledge (expertise-reversal effect: Cierniak et al.,

2009a; Kalyuga, 2013). Another limitation of our study is the missing inclusion of learning performance. Since the CLT assumes that CL also influences learning performance (see **Figure 1**), it would have been appropriate to include this in relation to the manifestations of cognitive load in relation to the causal factors. Due to our aims and the corroborating, temporally extensive design of the study, we were not able to consider learning performance. The focus was on the design of the scaffolds and the resulting cognitive load. In the future, this should be captured by valid instruments for the methodological skills/knowledge for the *planning* (content of scaffolds) to extend the analyzed relationships between causal and assessment factors. However, the results of this study demonstrate that performance-based (cognitive abilities and reading skills) and objective (eye-tracking metrics) instruments are linked to the subjectively obtained test scores on individual CL. Therefore, further analyses are needed in the future to examine the relationships between the individual causal factors in relation to different tasks and the resulting CL. However, not only the findings between the causal and assessment factors show further approaches of investigation, but also the subjective measurement instrument from its linguistic composition by itself. Thus, the results indicate that especially the linguistic construction of the items has a great influence on the actual measurability of the individual constructs. On the one hand, there is the difficulty of phrasing the items (especially GCL and ICL) in such a way that they reflect the construct to be measured (Wording, see "Introduction"). Above all, the measurement of the GCL must be considered with reservations. In addition to the term "knowledge," the word "understanding" is used in the items. It is questionable whether students' understanding and comprehension can actually be equated with GCL (Ayres, 2018), and how the corresponding items can be worded differently to express quantify GCL. On the other hand, there is the challenge of adapting the linguistic level to the learners in such a way that they can understand and assess it independently (Sweller et al., 2011; Klepsch and Seufert, 2021). In conjunction with the validation approach according to Kane (2013), which considers the generalizability of the test instrument as an elementary part of validity, the existing subjective test instruments should be examined in the future regarding a possible standardization with tested adaptation possibilities that fit the investigated setting (self-regulated learning or problem-based learning).

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

MM developed the basic idea for the present study and supervised the project. MK and MM contributed to the conception and design of the studies. MK developed the used material and questionnaires and led data collection for studies, analyzed all the data and interpreted it with the help of MM, SK, and JK, and was mainly responsible for writing the paper. SK and JK supported in the execution and analysis of the eye-tracking data. This article is part of the Ph.D. thesis of MK supervised by MM. All authors agreed to the final submitted version of the manuscript and agreed to be responsible for all aspects of the work, ensuring that questions regarding the accuracy or integrity of any part of the work were adequately investigated and resolved.

FUNDING

The project on which this article is based and the associated supervision by MM was funded by the Deutsche Telekom Stiftung within the framework of the program Fellowship Fachdidaktik MINT.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.703857/full#supplementary-material

REFERENCES

AERA, APA, and NCME (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Ainsworth, S. (2006). DeFT: a conceptual framework for considering learning with multiple representations. *Learn. Instruct.* 16, 183–198. doi: 10.1016/j.learninstruc.2006.03.001

Alfieri, L., Brooks, P. J., Aldrich, N. J., and Tenenbaum, H. R. (2011). Does discovery-based instruction enhance learning? *J. Educ. Psychol.* 103, 1–18. doi: 10.1037/a0021017

Anvari, F., Tran, H., and Kavakli, M. (2013). Using cognitive load measurement and spatial ability test to identify talented students in three-dimensional computer graphics programming. *Int. J. Inf. Educ. Technol.* 3:94. doi: 10.7763/IJIET.2013.V3.241

Arnold, J. C., Kremer, K., and Mayer, J. (2014). Understanding students' experiments – what kind of support do they need in inquiry tasks? *Int. J. Sci. Educ.* 36, 2719–2749. doi: 10.1080/09500693.2014.930209

Artelt, C., McElvany, N., Christmann, U., Richter, T., Groeben, N., Köster, J., et al. (2007). *Förderung von Lesekompetenz – Expertise (Bildungsforschung, 17)*. Berlin: Bundesministerium für Bildung und Forschung.

Ayres, P. (2018). "Subjective measures of cognitive load: what can they reliably measure?," in *Cognitive Load Measurement and Application: A Theoretical Framework for Meaningful Research and Practice*. ed. R. Z. Zheng (New York: Routledge/Taylor & Francis Group), 9–28.

Baars, M., Wijnia, W., and Paas, F. (2017). The association between motivation, affect, and self-regulated learning when solving problems. *Front. Psychol.* 8:1346. doi: 10.3389/fpsyg.2017.01346

Bell, R., Smetana, L. K., and Binns, I. (2005). Simplifying inquiry instruction. *Sci. Teach.* 72, 30–33.

Chen, S., and Epps, J. (2014). Using task-induced pupil diameter and blink rate to infer cognitive load. *Hum. Comput. Interact.* 29, 390–413. doi: 10.1080/07370024.2014.892428

Chen, O., Kalyuga, S., and Sweller, J. (2016). Relations between the worked example and generation effects on immediate and delayed tests. *Learn. Instruct.* 45, 20–30. doi: 10.1016/j.learninstruc.2016.06.007

Chen, H., Ning, X., Wang, L., and Yang, J. (2018). Acquiring new factual information: effect of prior knowledge. *Front. Psychol.* 9:1734. doi: 10.3389/fpsyg.2018.01734

Choi, J., and Sadar, S. (2011). An empirical investigation of the relationships Among cognitive abilities, cognitive style, and learning preferences in students enrolled in specialized degree courses at a Canadian college. *Can. J. Scholarsh. Teach. Learn.* 2:5. doi: 10.5206/cjsotl-rcacea.2011.1.5

Choi, H. H., van Merriënboer, J. J. G., and Paas, F. (2014). Effects of the physical environment on cognitive load and learning: towards a new model of cognitive load. *Educ. Psychol. Rev.* 26, 225–244. doi: 10.1007/s10648-014-9262-6

Cierniak, G., Gerjets, P., and Scheiter, K. (2009a). Expertise reversal in multimedia learning: subjective load ratings and viewing behavior as cognitive process indicators. *Proc. Annu. Meet. Cogn. Sci. Soc.* 31, 1906–1911.

Cierniak, G., Scheiter, K., and Gerjets, P. (2009b). Explaining the split-attention effect: is the reduction of extraneous cognitive load accompanied by an increase in germane cognitive load? *Comput. Hum. Behav.* 25, 315–324. doi: 10.1016/j.chb.2008.12.020

Cohen, J. (1988). The effect size. *Statistical power analysis for the behavioral sciences*, 77–83.

Cronbach, L. J., and Meehl, P. E. (1955). Construct validity in psychological tests. *Psychol. Bull.* 52, 281–302. doi: 10.1037/h0040957

de Jong, T. (2010). Cognitive load theory, educational research, and instructional design: some food for thought. *Instr. Sci.* 38, 105–134. doi: 10.1007/s11251-009-9110-0

de Jong, T. (2019). Moving towards engaged learning in STEM domains; there is no simple answer, but clearly a road ahead. *J. Comput. Assist. Learn.* 35, 153–167. doi: 10.1111/jcal.12337

DeLeeuw, K. E., and Mayer, R. E. (2008). A comparison of three measures of cognitive load: evidence for separable measures of intrinsic, extraneous, and germane load. *J. Educ. Psychol.* 100, 223–234. doi: 10.1037/0022-0663.100.1.223

Eitel, A., Endres, T., and Renkl, A. (2020). Self-management as a bridge between cognitive load and self-regulated learning: the illustrative case of seductive details. *Educ. Psychol. Rev.* 32, 1073–1087. doi: 10.1007/s10648-020-09559-5

Furtak, E. M., Seidel, T., Iverson, H., and Briggs, D. C. (2012). Experimental and quasi-experimental studies of inquiry-based science teaching: a meta-analysis. *Rev. Educ. Res.* 82, 300–329. doi: 10.3102/0034654312457206

Heller, K. A., and Perleth, C. (2000). *KFT 4–12+ R. Kognitiver Fähigkeitstest für 4. bis 12. Klassen, Revision*. Göttingen: Beltz Test.

Hmelo-Silver, C. E., Duncan, R. G., and Chinn, C. A. (2007). Scaffolding and achievement in problem-based and inquiry learning: a response to Kirschner, Sweller, and Clark (2006). *Educ. Psychol.* 42, 99–107. doi: 10.1080/00461520701263368

Ho, H. N. J., Tsai, M. -J., Wang, C.-Y., and Tsai, C. C. (2014). Prior knowledge and online inquiry-based science reading: evidence from eye tracking. *Int. J. Sci. Math. Educ.* 12, 525–554. doi: 10.1007/s10763-013-9489-6

Höffler, T. N. (2010). Spatial ability: its influence on learning with visualizations - a meta-analytic review. *Educ. Psychol. Rev.* 22, 245–269. doi: 10.1007/s10648-010-9126-7

Höffler, T. N., and Leutner, D. (2011). The role of spatial ability in learning from instructional animations – evidence for an ability-as-compensator hypothesis. *Comput. Hum. Behav.* 27, 209–216. doi: 10.1016/j.chb.2010.07.042

Höffler, T. N., and Schwartz, R. N. (2011). Effects of pacing and cognitive style across dynamic and non-dynamic representations. *Comput. Educ.* 57, 1716–1726. doi: 10.1016/j.compedu.2011.03.012

Huh, D., Kim, J. H., and Jo, I. H. (2019). A novel method to monitoring changes in cognitive load in videobased learning. *J. Comput. Assist. Learn.* 35, 721–730. doi: 10.1111/jcal.12378

Huk, T. (2006). Who benefits from learning with 3D models? The case of spatial ability. *J. Comput. Assist. Learn.* 22, 392–404. doi: 10.1111/j.1365-2729.2006.00180.x

Jäger, A. O. (1984). Intelligenzstrukturforschung: konkurrierende Modelle, neue Entwicklungen, Perspektiven. *Psychol. Rundsch.* 35, 21–35.

Jäger, A. O., Holling, H., Preckel, F., Schulze, R., Vock, M., Süß, H.-M., et al. (2006). *BIS-HB: Berliner Intelligenzstrukturtest für Jugendliche: Begabungs- und Hochbegabungs-diagnostik – Manual*. Göttingen: Hogrefe.

Jäger, D., Itsios, C., Franz, T., and Müller, R. (2017). Cognitive Load und Aufgabenmerkmale–Verwendung von Zusatzfragen bei authentischen Problemen. *Didaktik der Physik - Beiträge zur DPG-Frühjahrstagung PhyDid. B, DD*, 13.1, 125–130.

Kaiser, I., and Mayer, J. (2019). The long-term benefit of video modeling examples for guided inquiry. *Front. Educ.* 4:104. doi: 10.3389/feduc.2019.00104

Kaiser, I., Mayer, M., and Malai, D. (2018). Self-generation in context of inquiry-based learning. *Front. Psychol.* 9:2440. doi: 10.3389/fpsyg.2018.02440

Kalyuga, S. (2011). Cognitive load theory: how many types of load does it really need? *Educ. Psychol. Rev.* 23, 1–19. doi: 10.1007/s10648-010-9150-7

Kalyuga, S. (2013). Effects of learner prior knowledge and working memory limitations on multimedia learning. *Procedia. Soc. Behav. Sci.* 83, 25–29. doi: 10.1016/j.sbspro.2013.06.005

Kalyuga, S., Ayres, P., Chandler, P., and Sweller, J. (2003). The expertise reversal effect. *Educ. Psychol.* 38, 23–31. doi: 10.1207/S15326985EP3801_4

Kane, M. (2006). "Content-related validity evidence in test development," in *Handbook of Test Development*. eds. S. M. Downing and T. M. Haladyna (Mahwah, New Jersey: Lawrence Erlbaum Associates Publishers), 131–153.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *J. Educ. Meas.* 50, 1–73. doi: 10.1111/jedm.12000

Kastaun, M., Hunold, C., and Meier, M. (2020). Eye-Tracking – Visuelle Wahrnehmung sichtbar machen. *Biologie in unserer Zeit* 3, 172–173. doi: 10.1002/biuz.202070311

Kastaun, M., and Meier, M. (2021). "Eine qualitative Analyse von Blickdaten bei statischen und dynamischen Repräsentationen im naturwissenschaftlichen Erkenntnisprozess," in *Eye Tracking als Methode in der Mathematik- und Naturwissenschaftsdidaktik: Forschung und Praxis*. eds. P. Klein, M. Schindler, N. Graulich and J. Kuhn (Heidelberg: Springer Spectrum).

Kendeou, P., and van den Broek, P. (2007). The effects of prior knowledge and text structure on comprehension processes during reading of scientific texts. *Mem. Cogn.* 35, 1567–1577. doi: 10.3758/BF03193491

Kind, P., and Osborne, J. (2017). Styles of scientific reasoning—a cultural rationale for science education? *Sci. Educ.* 101, 8–31. doi: 10.1002/sce.21251

Kirschner, P. A., Ayres, P., and Chandler, P. (2011). Contemporary cognitive load theory research: The good, the bad and the ugly. *Comput. Hum. Behav.* 27, 99–105. doi: 10.1016/j.chb.2010.06.025

Kirschner, P. A., Sweller, J., and Clark, R. E. (2006). Why minimal guidance during instruction does not work: an analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educ. Psychol.* 41, 75–86. doi: 10.1207/s15326985ep4102_1

Klepsch, M., Schmitz, F., and Seufert, T. (2017). Development and validation of two instruments measuring intrinsic, extraneous, and germane cognitive load. *Front. Psychol.* 8:1997. doi: 10.3389/fpsyg.2017.01997

Klepsch, M., and Seufert, T. (2021). Making an effort versus experiencing load. *Front. Educ.* 6:645284. doi: 10.3389/feduc.2021.645284

Koć-Januchta, M. M., Höffler, T. N., Eckhardt, M., and Leutner, D. (2019). Does modality play a role? Visual-verbal cognitive style and multimedia learning. *J. Comput. Assist. Learn.* 35, 747–757. doi: 10.1111/jcal.12381

Korbach, A., Brünken, R., and Park, B. (2018). Differentiating different types of cognitive load: a comparison of different measures. *Educ. Psychol. Rev.* 30, 503–529. doi: 10.1007/s10648-017-9404-8

Krell, M. (2017). Evaluating an instrument to measure mental load and mental effort considering different sources of validity evidence. *Cogent Educ.* 4:1280256. doi: 10.1080/2331186X.2017.1280256

Kuhn, D., Amsel, E., and O'Loughlin, M. (1988). *The Development of Scientific Thinking Skills*. San Diego, CA: Academic Press.

Kürschner, C., Schnotz, W., Eid, M., and Hauck, G. (2005). Individuelle Modalitätspräferenzen beim Textverstehen: Präferenzen für auditive oder visuelle Sprachverarbeitung in unterschiedlichen Bevölkerungsgruppen. *Z. Entwicklungspsychol. Padagog. Psychol.* 37, 2–16. doi: 10.1026/0049-8637.37.1.2

Leahy, W., and Sweller, J. (2011). Cognitive load theory, modality of presentation and the transient information effect. *Appl. Cogn. Psychol.* 25, 943–951. doi: 10.1002/acp.1787

Lehmann, J., and Seufert, T. (2020). The interaction between text modality and the learner's modality preference influences comprehension and cognitive load. *Front. Psychol.* 10:2820. doi: 10.3389/fpsyg.2019.02820

Leppink, J., Paas, F., van der Vleuten, C. P. M., van Gog, T., and Van Merriënboer, J. J. G. (2013). Development of an instrument for measuring different types of cognitive load. *Behav. Res. Ther.* 45, 1058–1072. doi: 10.3758/s13428-013-0334-1

Leutner, D., Funke, J., Klieme, E., and Wirth, J. (2005). "Problemlösefähigkeit als fächerübergreifende Kompetenz," in *Problemlösekompetenz von Schülerinnen und Schülern*. eds. E. Klieme, D. Leutner and J. Wirth (Wiesbaden: VS Verlag für Sozialwissenschaften), 11–19.

Mayer, R. E. (2014). "Introduction to multimedia learning," in *The Cambridge Handbook of Multimedia Learning*. ed. R. E. Mayer (Cambridge MA: Cambridge University), 1–24.

Mayer, R. E., and Massa, L. J. (2003). Three facets of visual and verbal learners: cognitive ability, cognitive style, and learning preference. *J. Educ. Psychol.* 95, 833–846. doi: 10.1037/0022-0663.95.4.833

Meier, M., and Kastaun, M. (2021). "Lernunterstützungen als Werkzeug individualisierter Förde-rung im naturwissenschaftlichen Erkenntnisprozess," in *Vielfältige Wege biologiedidaktischer Forschung*. eds. M. Meier, C. Wulff and K. Ziepprecht (Münster: Waxmann), 95–116.

Messick, S. (1995). Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *Am. Psychol.* 50, 741–749. doi: 10.1037/0003-066X.50.9.741

Minkley, N., Kärner, T., Jojart, A., Nobbe, L., and Krell, M. (2018). Students' mental load, stress, and performance when working with symbolic or symbolic–textual molecular representations. *J. Res. Sci. Teach.* 55, 1162–1187. doi: 10.1002/tea.21446

Minkley, N., Xu, K. M., and Krell, M. (2021). Analyzing relationships between causal and assessment factors of cognitive load: associations between objective and subjective measures of cognitive load, stress, interest, and self-concept. *Front. Educ.* 6:632907. doi: 10.3389/feduc.2021.632907

Moreno, R. (2010). Cognitive load theory: more food for thought. *Instr. Sci.* 38, 135–141. doi: 10.1007/s11251-009-9122-9

Münzer, S. (2012). Facilitating spatial perspective taking through animation: evidence from an aptitude–treatment-interaction. *Learn. Individ. Differ.* 22, 505–510. doi: 10.1016/j.lindif.2012.03.002

Nitz, S., Ainsworth, S. E., Nerdel, C., and Prechtl, H. (2014). Do student perceptions of teaching predict the development of representational competence and biological knowledge? *Learn. Instruct.* 31, 13–22. doi: 10.1016/j.learninstruc.2013.12.003

OECD (2007). *PISA 2006: Science Competencies for Tomorrow's World. Vol. 1*. Paris: OECD Publishing.

Paas, F., Tuovinen, J. E., Tabbers, H., and van Gerven, P. W. M. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educ. Psychol.* 38, 63–71. doi: 10.1207/S15326985EP3801_8

Paas, F., van Gog, T., and Sweller, J. (2010). Cognitive load theory: new conceptualizations, specifications, and integrated research perspectives. *Educ. Psychol. Rev.* 22, 115–121. doi: 10.1007/s10648-010-9133-8

Paas, F., and Van Merriënboer, J. J. G. (1994). Instructional control of cognitive load in the training of complex cognitive tasks. *Educ. Psychol. Rev.* 6, 351–371. doi: 10.1007/BF02213420

Paas, F., and van Merriënboer, J. J. G. (2020). Cognitive-load theory: methods to manage working memory load in the learning of complex tasks. *Curr. Dir. Psychol. Sci.* 29, 394–398. doi: 10.1177/0963721420922183

Paivio, A. (2006). *Mind and its Evolution; A Dual Coding Theoretical Interpretation*. Mahwah: Lawrence Erlbaum Associates, Inc.

Peterson, K., DeCato, L., and Kolb, D. A. (2015). Moving and learning: expanding style and increasing flexibility. *J. Exp. Educ.* 38, 228–244. doi: 10.1177/1053825914540836

Plass, J. L., Chun, D. M., Mayer, R. E., and Leutner, D. (2003). Cognitive load in reading a foreign language text with multimedia aids and the influence of verbal and spatial abilities. *Comput. Hum. Behav.* 19, 221–243. doi: 10.1016/S0747-5632(02)00015-8

Plass, J. L., and Homer, B. D. (2002). "Cognitive load in multimedia learning: the role of learner preferences and abilities," in *Proceedings of the International Conference on Computers in Education*. ed. B. Werner (USA: IEEE Computer Society), 564–568.

Pollock, E., Chandler, P., and Sweller, J. (2002). Assimilating complex information. *Learn. Instruct.* 12, 61–86. doi: 10.1016/S0959-4752(01)00016-0

Quaiser-Pohl, C., Lehmann, W., and Schirra, J. (2001). Sind Studentinnen der Computervisualistik besonders gut in der Raumvorstellung? Psychologische Aspekte bei der Wahl eines Studienfachs. *FIfF Kommunikation* 18, 42–46.

Rheinberg, F., Vollmeyer, R., and Burns, B. D. (2000). "Motivation and self-regulated learning," in *Motivational Psychology of Human Development: Developing Motivation and Motivating Development*. ed. J. Heckhausen (Amsterdam: Elsevier Science), 81–108.

Richter, J., Scheiter, K., and Eitel, A. (2018). Signaling text–picture relations in multimedia learning: the influence of prior knowledge. *J. Educ. Psychol.* 110, 544–560. doi: 10.1037/edu0000220

Sanchez, C. A., and Wiley, J. (2014). The role of dynamic spatial ability in geoscience text comprehension. *Learn. Instruct.* 31, 33–45. doi: 10.1016/j.learninstruc.2013.12.007

Scheerer-Neumann, G. (2018). *Lese-Rechtschreib-Schwäche und Legasthenie. Grundlagen, Diagnostik und Förderung*. Stuttgart: Verlag W. Kohlhammer.

Scheiter, K., Ackerman, R., and Hoogerheide, V. (2020). Looking at mental effort appraisals through a metacognitive lens: are they biased? *Educ. Psychol. Rev.* 32, 1003–1027. doi: 10.1007/s10648-020-09555-9

Schneider, W., Schlagmüller, M., and Ennemoser, M. (2017). *LGVT 5–12+. Lesegeschwindigkeits- und Verständnistest für die Klassen 5–12*. Göttingen: hogrefe.

Seufert, T. (2018). The interplay between self-regulation in learning and cognitive load. *Educ. Res. Rev.* 24, 116–129. doi: 10.1016/j.edurev.2018.03.004

Seufert, T. (2019). Training for coherence formation when learning from text and picture and the interplay with learners' prior knowledge. *Front. Psychol.* 10:193. doi: 10.3389/fpsyg.2019.00193

Shrager, J., and Siegler, R. S. (1998). SCADS: A model of children's strategy choices and strategy discoveries. *Psychol. Sci.* 9, 405–410. doi: 10.1111/1467-9280.00076

Siegler, R. S. (2005). Children's learning. *Am. Psychol.* 60, 769–778. doi: 10.1037/0003-066X.60.8.769

Skuballa, I. T., Xu, K. M., and Jarodzka, H. (2019). The impact of co-actors on cognitive load: when the mere presence of others makes learning more difficult. *Comput. Hum. Behav.* 101, 30–41. doi: 10.1016/j.chb.2019.06.016

Solhjoo, S., Haigney, M. C., McBee, E., van Merriënboer, J. J. G., Schuwirth, L., Artino, A. R., et al. (2019). Heart rate and heart rate variability correlate with clinical reasoning performance and self-reported measures of cognitive load. *Sci. Rep.* 9, 1–9. doi: 10.1038/s41598-019-50280-3

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learn. Instruct.* 4, 295–312. doi: 10.1016/0959-4752(94)90003-5

Sweller, J., Ayres, P. L., and Kalyuga, S. (2011). Intrinsic and extraneous cognitive load. *In Cognitive Load Theory*. New York, NY: Springer. 57–69.

Sweller, J., van Merriënboer, J. J. G., and Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educ. Psychol. Rev.* 31, 261–292. doi: 10.1007/s10648-019-09465-5

Thees, M., Kapp, S., Altmeyer, K., Malone, S., Brünken, R., and Kuhn, J. (2021). Comparing two subjective rating scales assessing cognitive load During technology-enhanced STEM laboratory courses. *Front. Educ.* 6:705551. doi: 10.3389/feduc.2021.705551

van de Weijer-Bergsma, E., and van der Ven, S. H. G. (2021). Why and for whom does personalizing math problems enhance performance? Testing the mediation of enjoyment and cognitive load at different ability levels. *Learn. Individ. Differ.* 87:101982. doi: 10.1016/j.lindif.2021.101982

van Merriënboer, J. J. G., Kester, L., and Paas, F. (2006). Teaching complex rather than simple tasks: balancing intrinsic and germane load to enhance transfer of learning. *Appl. Cogn. Psychol.* 20, 343–352. doi: 10.1002/acp.1250

Zu, T., Hutson, J., Loschky, L. C., and Rebello, N. S. (2020). Using eye movements to measure intrinsic, extraneous, and germane load in a multimedia learning environment. *J. Educ. Psychol.* 112, 1338–1352. doi: 10.1037/edu0000441

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Psychology

REVIEW
published: 10 September 2021
doi: 10.3389/fpsyg.2021.702538

The Validity of Physiological Measures to Identify Differences in Intrinsic Cognitive Load

Paul Ayres[1], Joy Yeonjoo Lee[2], Fred Paas[3,4] and Jeroen J. G. van Merriënboer[2]*

[1] School of Education, University of New South Wales, Sydney, NSW, Australia, [2] School of Health Professions Education, Maastricht University, Maastricht, Netherlands, [3] Department of Psychology, Education and Child Studies, Erasmus University, Rotterdam, Netherlands, [4] School of Education/Early Start, University of Wollongong, Wollongong, NSW, Australia

A sample of 33 experiments was extracted from the Web-of-Science database over a 5-year period (2016–2020) that used physiological measures to measure intrinsic cognitive load. Only studies that required participants to solve tasks of varying complexities using a within-subjects design were included. The sample identified a number of different physiological measures obtained by recording signals from four main body categories (heart and lungs, eyes, skin, and brain), as well as subjective measures. The overall validity of the measures was assessed by examining construct validity and sensitivity. It was found that the vast majority of physiological measures had some level of validity, but varied considerably in sensitivity to detect subtle changes in intrinsic cognitive load. Validity was also influenced by the type of task. Eye-measures were found to be the most sensitive followed by the heart and lungs, skin, and brain. However, subjective measures had the highest levels of validity. It is concluded that a combination of physiological and subjective measures is most effective in detecting changes in intrinsic cognitive load.

Keywords: intrinsic cognitive load, physiological measures, validity, working-memory load, workload, cognitive load theory

OPEN ACCESS

Edited by:
Jon-Chao Hong, National Taiwan Normal University, Taiwan

Reviewed by:
Giovanna Bubbico, University of Studies G. d'Annunzio Chieti and Pescara, Italy
Bruce Mehler, Massachusetts Institute of Technology, United States

****Correspondence:***
Fred Paas
paas@essb.eur.nl

Specialty section:
This article was submitted to Educational Psychology, a section of the journal Frontiers in Psychology

Received: *29 April 2021*
Accepted: *13 August 2021*
Published: *10 September 2021*

Citation:
Ayres P, Lee JY, Paas F and van Merriënboer JJG (2021) The Validity of Physiological Measures to Identify Differences in Intrinsic Cognitive Load. Front. Psychol. 12:702538. doi: 10.3389/fpsyg.2021.702538

INTRODUCTION

The main aim of this study was to examine the validity of using physiological techniques to measure cognitive load by examining construct validity (see Gravetter and Forzano, 2018) and sensitivity (see Longo and Orru, 2018). More specifically to investigate the ability of physiological measures to detect differences in intrinsic cognitive load caused by tasks of varying complexity. To meet this aim we examined the findings from a number of studies drawn from a 5-year sample that measured cognitive load using physiological techniques. In particular, we were interested in examining a sample of contemporary studies that had access to the most up-to-date technology.

Researchers across many fields have been interested in the amount of mental resources invested in attempting a task. One such field is human factors, where studies have focused on everyday tasks such as driving a motor vehicle. In particular, of much interest has been the mental resources required to not only drive the car but deal with other requirements or distractors. Measurement of these mental resources has received much attention, for example, Wickens (2002) developed a model based on multiple resource theory that predicted dual-task interference.

A number of descriptors have been used to represent these types of mental investments such as working memory load, mental workload, and cognitive load, according to the context used and/or the theoretical influences of the researchers. Although such labels are often used interchangeably and have similar meanings, we will use the term cognitive load because of its close association with cognitive load theory (CLT) and the identification of different types of cognitive load (see Sweller et al., 2011), which is an important consideration in the current paper.

Cognitive load theory emerged in the 1980s as a theory that made predictions about learning and problem solving based on the amount of mental resources and effort (cognitive load) invested in the tasks. An integral part of CLT (see Sweller et al., 2011; Sweller et al., 2019) has therefore been to find instruments that measure cognitive load in order to provide direct evidence for the assumptions made and to advance the theory further. Up until now a number of self-rating subjective measures (see Paas, 1992; van Gog and Paas, 2008; Ayres, 2018) have been developed as the preferred methods. However, despite their widescale use, many researchers have argued for more objective alternatives such as working memory dual-tasks (see Park and Brünken, 2015) or physiological measures (see Antonenko and Niederhauser, 2010).

In more recent times, helped by advances and the availability of new technologies, physiological measures have become more popular in CLT research (see Ayres, 2020). Many physiological measures have already been used based on theoretical arguments and experimental data to confirm individual validity (see Kane, 2013). However, our study was a broader investigation to gain a more global overview of physiological measures in order to gain an up-to-date picture of their validity.

Cognitive load theory has identified three types of cognitive load: intrinsic, extraneous, and germane (see Sweller et al., 2019). According to CLT intrinsic cognitive load is generated by the element interactivity (elements that need to be processed simultaneously) of the task (Sweller et al., 2011); whereas in instructional settings, cognitive load can also be generated by learners dealing with the instructional designs (extraneous cognitive load) and the actual learning processes generated (germane cognitive load). Complex tasks typically have many interacting elements creating high levels of intrinsic cognitive load. The intrinsic cognitive load generated by any task is not fixed *per se* as it is dependent upon the expertise of the problem-solver or learner. Learners with high levels of expertise possess knowledge structures (schemas) that enable them to chunk together many elements (see Chi et al., 1982; Sweller et al., 2011) thus reducing intrinsic cognitive load. It should be noted that CLT emphasizes the importance of interacting elements for creating complexity in learning settings; however, increased cognitive load does not always depend on interactivity for other non-learning types of tasks. For example, trying to recall 12 random numbers after a brief observation requires more mental effort than recalling five numbers. Nevertheless, interactivity between elements can be a major cause of intrinsic cognitive load.

Subjective measures have been reasonably successful by requiring learners to rate different aspects of the learning process using multi-item scales (see Leppink et al., 2013), although it is debatable whether learners are able to identify different forms of cognitive load when many cognitive processes are interacting. In contrast, as far as we know no physiological measures have been developed to distinguish the types of cognitive load due to a number of confounding factors such as interactions between task complexity and learning or stress. To make a consistent analysis by controlling for possible confounding factors, the present study focused on studies that generated changes in intrinsic cognitive load caused by variations in problem complexity rather than by additional learning factors. By focusing on studies that require participants to solve tasks without instruction, and no requirement for learning, cognitive load can be narrowed down to one source (task complexity). Under these conditions we assume total cognitive load to be equivalent to intrinsic cognitive load as no instruction or learning, or other interacting factors are directly involved (see Ayres, 2006). Hence for our sample of studies, we expected the various physiological techniques used to measure only intrinsic load as no other types of cognitive load were present.

Mental arithmetic tasks provide a good example of tasks that vary in complexity and generate different levels of intrinsic cognitive load. They require the simultaneous storage and processing of information, which generates working memory load (cognitive load). Because completing a task serves as a different function to learning about the task, the source of cognitive load is intrinsic and dependent upon element interactivity only, as no instructional steps are included. Consider the following two problems: (a) calculate 3*4 + 7; and (b) 12*8 + 14 + 11*2. The first problem is fairly straightforward with only two simple calculations involved; whereas the latter has four more difficult calculations with greater element interactivity. The second problem requires more working memory resources to overcome the intrinsic cognitive load generated.

From a validity perspective, Borsboom et al. (2004) argue that "A test is valid for measuring an attribute if: (a) the attribute exists; and (b) variations in the attribute causally produce variation in the measurement outcomes" (p. 1061). Clearly, cognitive load exists for mental arithmetic tasks (and all tasks that require working memory load), so any instrument that purports to measure cognitive load should find differences in cognitive load between the two problems described above. Based on cognitive load theory we hypothesize that within-participants comparisons of tasks with different levels of complexity reveal differences in intrinsic load that should become visible in physiological measures. The more studies support this hypothesis for a specific physiological measure, the higher the construct validity of this measure (Gravetter and Forzano, 2018).

Borsboom et al. (2004) also suggest that detecting changes in the attribute is integral to validity. Using Messick's argument that validity (see Messick, 1989) does not simply have black or white outcomes, Borsboom et al. (2004) suggest that there can be different levels of validity dependent upon how sensitive the measure is to detecting changes. Longo and Orru (2018) refer to an ability to detect changes as *sensitivity*. Hence, we consider sensitivity as a second important feature in establishing the validity of a physiological measure. Note that the concept of validity can encompass the concept of sensitivity, since validity

in general is a broader concept. Thus, we treat both construct validity and sensitivity as aspects of validity.

As previously mentioned many studies have used physiological techniques to measure cognitive load. Four main categories have been identified that correspond with major body organs: the heart (i.e., cardiovascular) and lungs (i.e., respiratory), eyes, skin (i.e., electrodermal), and brain. The following sections summarize some of the key ways that data has been extracted from these four categories to measure cognitive load.

CARDIOVASCULAR AND RESPIRATORY MEASURES OF COGNITIVE LOAD

Heart Rate

Cardiovascular measures or heart rate (HR) parameters have a long history as psychophysiological indices of cognitive load (Paas et al., 2003). However, HR data can be problematic because it is affected by many psychological and physiological factors, such as emotions and physical activities (Jorna, 1992). The challenge has been to disentangle these factors to isolate the factors that can be indicative of cognitive load or use well-controlled experimental designs. An example of a well-controlled design can be found in the study by Mehler et al. (2012), who showed that HR is a highly sensitive physiological measure for detecting systematic variations in cognitive load.

Heart Rate Variability

Blitz et al. (1970) showed that heart rate variability (HRV) can be used to differentiate between different levels of cognitive load. HRV can be assessed through measurement of the electrical activity of the heart, which can be visualized in an electrocardiogram (Mulder, 1992), or measurement of blood volume changes in the microvascular bed tissue using a light-based technology called photoplethysmography (Challoner, 1979).

According to the Task Force of the European Society of Cardiology and the North American Society of Pacing Electrophysiology (1996), HRV can be described as the oscillation in the interval between consecutive heartbeats as well as the oscillations between consecutive instantaneous HRs (i.e., variability in time between the successive R-tops of the cardiogram). In a study on the usefulness of HRV as an index of operator effort, Aasman et al. (1987; see also Mulder, 1992) further specified the HRV measure based on the knowledge that the time between successive heartbeats is determined by three different feedback mechanisms, connected with respiration, blood pressure (BP), and body temperature regulation. Using a special mathematical technique (i.e., spectral analysis) to investigate periodical components of the HRV, Aasman et al. (1987) were able to show that cognitive load is specifically related to the short-term regulation of arterial BP. The relationship between cognitive load and HRV is indirect (Solhjoo et al., 2019), because an increase in cognitive load will lead to an increase in BP, which will lead to a decrease in HRV. A similar indirect relationship has also been identified for respiratory activity with increasing cognitive load increasing the respiratory frequency (e.g., Grassmann et al., 2016b), which will lead to a decrease in HRV (e.g., Song and Lehrer, 2003). The spectral analysis technique can be used to separate the effects of respiratory rate (high frequency band, 0.15–0.40 Hz) and thermoregulation (low-frequency band, 0.02–0.06 Hz) from the mid-frequency band (0.07–0.14 Hz), which is determined by the arterial BP regulation and related to cognitive load.

The HRV measure is generally accepted as a measure of cognitive load (e.g., Finsen et al., 2001; De Rivecourt et al., 2008; Thayer et al., 2012). However, Paas et al. (1994) have argued that it has mainly been used successfully with short-duration basic task (e.g., binary decision tasks) under well-controlled conditions. Paas et al. (1994) showed that with longer-lasting learning tasks typically used in educational research, the validity and sensitivity of the spectral analysis technique of the HRV was low. The technique was only sensitive to relatively large differences in cognitive load, i.e., differences between mentally inactive and mentally active periods. According to Aasman et al. (1987) the high intrinsic variability in the HR signal is one of the sources of its low reliability and relative insensitivity to small differences in processing load between tasks. The studies that Paas et al. (1994) analyzed to determine the sensitivity of the HRV technique used relatively long duration learning tasks. In contrast to basic tasks, which mainly consist of mentally active periods, such tasks naturally contain both mentally active and inactive periods, and consequently create a rather noisy signal. In addition, although spectral analysis of HRV allows cognitive load to be measured at a higher rate than subjective measures, it cannot be considered a real-time measure, because it needs time to process. From this real-time measurement perspective, HR can also be argued to have an advantage over HRV, because HR changes can be detected in a much shorter period than changes in HRV.

Respiratory Measures

In contrast to cardiovascular measures, the use of respiratory measures has received much less research attention (for a review, see Grassmann et al., 2016a). Similar to the cardiovascular measures, respiratory measures are also influenced by and reflective for metabolic, psychological, and behavioral processes (Wientjes et al., 1998). Grassmann et al. (2016a) reviewed studies that used respiratory indices of cognitive load as a function of task difficulty, task duration, and concurrent performance feedback.

Research into the relationship between respiratory measures and cognitive load has used measures based on time (e.g., number of breaths per minute, i.e., respiration rate), volume (amount of air inhaled during one respiratory cycle, i.e., tidal volume), gas exchange (e.g., proportion of released CO_2 to inhaled O_2, i.e., respiratory exchange ratio), and variability parameters of these measures. In their review, Grassmann et al. (2016a) found that those respiratory parameters can be measured with breathing monitors, respiratory inductive plethysmography, strain gauges, impedance-based methods, capnography, and metabolic analyzers. Results of this study revealed that cognitive load was positively related to respiration rate (e.g., Backs and Seljos, 1994), and frequency of sighing (e.g., Vlemincx et al., 2011). In addition, negative relationships with

cognitive load were found for both inspiratory and expiratory time (e.g., Pattyn et al., 2010) and partial pressure of end-tidal carbon dioxide ($petCO_2$; Grassmann et al., 2016a). For the respiratory measures it is important to note that they can be disrupted by verbal activities, such as effects of talking on the breathing pattern, which can be a confound when using breathing pattern as an index of cognitive load (e.g., Tininenko et al., 2012).

EYE ACTIVITY MEASURES OF COGNITIVE LOAD

For decades, eye-tracking indices have been used to measure cognitive load in various fields (Rosch and Vogel-Walcutt, 2013; Glaholt, 2014; Wilbanks and McMullan, 2018). Thanks to its portability and unobtrusiveness, eye tracking supports more natural task environments (Eckstein et al., 2017). Moreover, its indices correspond to not only autonomic responses (e.g., pupil dilation, blinks) but also consciously modulated processes (e.g., eye fixation), facilitating the investigation of diverse indicators of cognitive load. The following sections describe some key measures of cognitive load drawn from eye data and important conditions.

Pupil Dilation

Pupil dilation reflects noradrenergic activity of the autonomous nervous system that regulates arousal and mental activity (Eckstein et al., 2017). A large number of studies have shown that pupil dilation is positively correlated with cognitive demands imposed by the tasks (Hess and Polt, 1964; Kahneman and Beatty, 1966; Hyönä et al., 1995; Van Orden et al., 2001). Most modern image-based eye-trackers can readily collect pupil data. However, multiple factors (e.g., light reflex, gaze position, pupil dilation latency) can affect this data, which challenges proper measurement of the cognitive effects on pupil size. Thus, researchers must take extra precautions to establish well-controlled experimental setups, for instance, maintaining a constant luminance of stimuli, using baseline data, or employing computational correction methods (Hayes and Petrov, 2016; Chen et al., 2017).

Blink Rate

The frequency of spontaneous blinks, or blink rate, is modulated by dopaminergic activity in the central nervous system that involves goal-oriented behavior and reward processing (Eckstein et al., 2017). Studies have shown that blink rate significantly increases as a function of time on task, fatigue, and workload (Stern et al., 1994; Tsai et al., 2007). However, this measure for assessing cognitive load is task-specific. When the task requires intensive visual processing (e.g., reading, air traffic control), blink rate is rather inhibited, resulting in a decreased rate (Van Orden et al., 2001; Recarte et al., 2008). Blinks can be easily detected by eye trackers, while several artifacts (e.g., reflections in glasses, participant motion) must be regulated (Holmqvist and Andersson, 2017, Chapter 15). Moreover, blink data is not continuous and its distribution is often non-Gaussian, requiring auxiliary calculation methods (Siegle et al., 2008).

Fixation

Eye fixation is a more consciously modulated behavior compared to pupil dilation and blinks. Three types of fixation measures have been frequently used for assessing cognitive load, reflecting different aspects of visual information processing: fixation rate (the number of fixations divided by a given time), fixation duration (time span when the eye is relatively still), and transition rate (the number of gaze shifts per second from one area of interest to another). Note that the first two, fixation rate and duration, are inversely correlated given the same trial duration, which makes the interpretation of the results highly task-specific. For instance, if the task requires frequent searching of different locations (e.g., scene perception, surveillance), increased fixation rate is associated with high cognitive load, accompanying short fixation duration (Van Orden et al., 2001; Chen et al., 2018). If the task includes deep and effortful processing of particular visual targets, long fixation duration would indicate high cognitive load (Callan, 1998; Henderson, 2011; Reingold and Glaholt, 2014).

When the task involves integration of information between different areas of interest (AOIs), transition rate can be a suitable measure (Schmidt-Weigand et al., 2010). Studies have shown that high cognitive load increases transition rate in static task environments, while it decreases the rate in dynamic task environments (van Meeuwen et al., 2014; Lee et al., 2019). For fixation data analysis, data quality and AOI definition are critical. Valid fixations should be detected after assuring data quality in terms of accuracy, precision, and tracking loss (Orquin and Holmqvist, 2018). Researchers should then carefully define AOIs relevant to sources of cognitive load, and select the measures pertinent to given task characteristics through task analysis and piloting.

Other Eye Measures

More eye-tracking and ocular indices have been explored in various research contexts. Variability in horizontal gaze position reduced as cognitive load increased in driving simulation tasks (He et al., 2019). In a simulated surgical task, intraocular pressure (i.e., fluid pressure inside the eye) was positively correlated with cognitive load (Vera et al., 2019). Ocular astigmatism aberration (i.e., deviation of optic elements of the eye), mediated by the intraocular pressure, was also shown to increase as a function of cognitive load (Jiménez et al., 2018). Since different measures can demonstrate different aspects of cognitive load, combining multiple eye measures may provide a higher construct validity and sensitivity as a cognitive load measure (Van Orden et al., 2000; Ryu and Myung, 2005; Mehler et al., 2009).

ELECTRODERMAL MEASURES OF COGNITIVE LOAD

Electrodermal measures have a long history as psychophysiological measures of emotional or cognitive stress and arousal (for a review see Posada-Quintero and Chon, 2020). The basis of the measurement of electrodermal activity (EDA) is the change in electrical activity in the eccrine sweat glands on the plantar and palmar sides of the hand, which are particularly responsive to psychological stimuli imposing stress.

Increased stress leads to increased sweating, which lowers the resistance and augments the electrical conductance of the skin (Dawson et al., 2000).

Within the EDA signal, two components can be distinguished. Firstly, a tonic skin conductance component (i.e., skin conductance level) that changes slowly over time. This component is considered a measure of psychophysiological activation. Secondly, a phasic skin conductance component (i.e., skin conductance response) that changes abruptly. These fast changes are reflected in the peaks in the electrodermal signal and are also called the galvanic skin response (Braithwaite et al., 2013). This component is impacted by stress and arousal (e.g., Hoogerheide et al., 2019).

Based on the knowledge that cognitive load is one of the cognitive states that causes people to experience stress it is assumed that changes in cognitive load affect the galvanic skin response through changes in skin conductance, with increases in cognitive load leading to increases in the galvanic skin response. Several studies have confirmed the positive relation between cognitive load and the galvanic skin response (e.g., Setz et al., 2009; Nourbakhsh et al., 2012; Larmuseau et al., 2019). Mehler et al. (2012), who investigated the sensitivity of skin conductance level to cognitive load, studied three age groups (20–29, 40–49, 60–69) working on a working memory task at three levels of cognitive load. They found a significant pattern of incremental increase of skin conductance level as a function of increasing cognitive load, thereby confirming the sensitivity of the measure of skin conductance level.

Vanneste et al. (2020) recently argued that the usability of EDA measures (i.e., skin conductance response rate and skin conductance response duration) as measurement instrument for cognitive load is limited, because it could only explain a limited proportion of the variance in cognitive load (approx. 22%). Charles and Nixon (2019) have suggested that EDA may be sensitive to sudden, but not gradual changes in cognitive load. However, this is only seems to apply to skin conductance response measures, because skin conductance level measures have been shown to increase with a gradual changes in cognitive load (e.g., Mehler et al., 2012).

Skin conductance response activity is commonly measured by the frequency of the peaks (i.e., skin conductance response rate), the duration of the peaks (skin conductance response duration), and the magnitude of the peaks (i.e., skin conductance magnitude) in the signal. Recently, Posada-Quintero et al. (2016) have introduced the spectral analysis technique to process skin conductance response activity data. This analysis method is commonly used to investigate periodical components of the HRV signal (e.g., Aasman et al., 1987). The newly developed index, incorporating the components between 0.08 and 0.24 Hz, was found to be highly sensitive to cognitive stress.

BRAIN ACTIVITY MEASURES OF COGNITIVE LOAD

In the past, measures of activity of the brain have mainly been used to assess cognitive load in controlled laboratory settings because they required advanced, unportable equipment for electroencephalography (EEG) or functional magnetic resonance imaging (fMRI). But nowadays, the use of brain measures is becoming more popular because new apparatus, such as wireless EEG caps and portable fNIRS devices (functional near infrared spectroscopy) are mobile, easy to use, and less obtrusive.

Electroencephalography

Electroencephalography records electrical activity of the brain. Multiple electrodes are placed on the scalp (typically using the 10/20 system; Jasper, 1958) and measure voltage fluctuations resulting from ionic current within the neurons of the brain. Assessments typically focus on the type of neural oscillations ('brain waves') that can be observed in EEG signals in the frequency domain. Spectral analysis gives insight into information contained in the frequency domain, distinguishing waveforms such as gamma (>35 Hz), beta (12–35 Hz), alpha (8–12 Hz), theta (4–8 Hz), and delta (0.5–4 Hz). Electrodes are placed on different locations of the scalp so that they read from different lobes or regions of the brain: Pre-frontal, frontal, temporal, parietal, occipital, and central. Cognitive load has mainly been found to be correlated with an increase in the parietal alpha band power and frontal-midline theta band power (Antonenko et al., 2010).

Functional Magnetic Resonance Imaging

Functional magnetic resonance imaging measures brain activity by detecting changes associated with cerebral blood flow, which is directly coupled to neuronal activation. Typically, it uses the blood oxygen level dependent (BOLD) contrast, which images the changes in blood flow related to energy use of brain cells. Statistical procedures are necessary to extract the underlying signal because it is frequently corrupted by noise from various sources. The level of brain activation in the whole brain or its specific regions can be graphically represented by color-coding its strength (e.g., showing fMRI BOLD signal increases in red and decreases in blue). Cognitive load has been found to be correlated with increased activation of neural regions associated with working memory, such as the fronto-parietal attention network (e.g., Tan et al., 2016; Mäki-Marttunen et al., 2019).

Functional Near Infrared Spectroscopy

Functional near infrared spectroscopy is an optical brain monitoring technique which uses near-infrared spectroscopy to measure brain activity by estimating cortical hemodynamic activity which occurs in response to neural activity. The signal can be compared with the BOLD signal measured by fMRI and is capable of measuring changes both in oxy- and deoxyhemoglobin concentration from regions near the cortical surface; local increases of oxyhemoglobin as well as decreases in deoxyhemoglobin are indicators of cortical activity. A typical system contains pairs of optical source and detector probes that are placed on the scalp with a lightweight headband typically using the same locations as EEG electrodes (i.e., the 10/20 system). As for fMRI, cognitive load is correlated with increased activation of the fronto-parietal network (Hosseini et al., 2017). In addition, using a combination of EEG and fNIRS signals has been shown to improve the sensitivity of cognitive load measurements (Aghajani et al., 2017).

RESEARCH QUESTIONS

Our aim was to gain an overview of the validity of physiological measures of intrinsic cognitive load from a collection of contemporary studies by examining construct validity and sensitivity. In particular, we were interested in identifying the different types of measures in use in such studies within the four main categories identified above: namely the heart and lungs, eyes, skin, and brain. The more studies that find the expected effects for a specific physiological measure, the stronger the support for its construct validity. Also by comparing the various measures with each other we aimed to identify any variations in levels of sensitivity. Furthermore, as we assumed that such variations could be influenced by specific tasks, we also investigated how validity was influenced by the types of tasks used. Our main research questions were:

RQ1: Do the physiological measures have construct validity in detecting changes in intrinsic cognitive load across tasks?

RQ2: How sensitive are the physiological measures to detecting changes in intrinsic cognitive load?

RQ3: Does the type of task impact on overall validity?

MATERIALS AND METHODS

In the present study we included only experiments that had within-subject designs (*Criterion 1*) comparing tasks with different levels of complexity (*Criterion 2*). Criterion 1 ensures that potential moderating factors such as prior knowledge that can impact on intrinsic cognitive load were limited. Criterion 2 ensured that intrinsic cognitive load would vary across tasks and thus could be detected by a valid physiological measure. Consistent with the main focus of this study we only included experiments that actually measured cognitive load using physiological measures (*Criterion 3*). In order to have a sufficiently large sample based on contemporary studies we examined data from the last 5 years. Although there are many databases, we chose the Web-of-Science as it is one of the most prestigious and authentic, and provided a sufficient enough sample to fulfill our aims.

Selecting the Sample of Studies

Step 1

To find an up-to-date sample of studies that featured physiological measures of cognitive load, a search was conducted in the *Web of Science* using the keywords "Physiological measures cognitive load" for the previous 5 years up until 30 November 2020. This search included articles, book chapters, and books. Some slight variations of the keywords were used such as working memory load, which produced little if any differences. In total 208 studies were initially identified.

Step 2

The abstract of each source was read to filter out any study that clearly did not meet our essential criteria of physiological measures of cognitive load (Criterion 3), within-subject designs (Criterion 1) with problems of varying complexity (Criterion 2). This analysis narrowed down the sample to 98 studies.

Step 3

From Step 2 it was possible to eliminate many studies that clearly did not meet Criterion 1–3, but many abstracts did not have sufficient information to make a definitive decision. Hence, a more thorough reading and analysis of each study was conducted to ensure each condition was satisfied. In particular to satisfy Criterion 2 it was necessary to include only studies that found significant differences on performance scores across the tasks. Using the example given above, recalling 12 random numbers after a brief observation is more mentally demanding than recalling five numbers. It is expected that more errors would be made recalling the 12-number task than the 5-number task, and this difference in errors would be caused by variations in cognitive load generated by complexity rather than prior knowledge about the domain. Hence, studies that did not report such significant differences between tasks were excluded.

A small number of studies were included that did not report score differences because this information was provided in previous studies or based on expert opinion (N = 3), or time to completion (N = 1). Studies that manipulated factors such as anxiety, stress, and other confounding factors were also eliminated (N = 4) as these factors are known to impact on working memory. For example, pupil dilation can indicate emotional arousal and therefore indicate extraneous cognitive load caused through distraction if the emotion is not part of the task (Lee et al., 2020).

These processes led to a final sample of 28 studies with 33 experiments (note: these studies are starred in the reference section). The mean participant size per experiment was 29.2 (SD = 14.8) with 53% males; 26 of the experiments consisted of adults with a mean age between 20 and 30, five included adults with no recorded mean statistics, and two studies focused on older adults (mean ages of 58 and 70).

Data Analysis

For each study a record was made of: (a) what cognitive load measures were used; (b) the type of tasks used; (c) the number of within-subject tasks; and (d) how many significant differences were found on performance tasks and cognitive load measures, and if these differences matched each other. It was notable that nearly all studies included a number of different measures of cognitive load.

Physiological Measures of Cognitive Load

The physiological measures as expected could be grouped into four categories consistent with the major organs of the body: the heart and lungs, eyes, skin, and brain. It was notable that 62.5% of all studies used measures from one category, 28.1% from two categories, 6.3% from three categories, and 3.1% from all four categories, indicating a battery of tests were utilized. In each of the four categories different types of measures were used, often in the same study. For example, a study might record both HR and HRV. Results for each category are described next.

Heart and Respiration Measures

Information (see **Table 1**) was collected from ECGs (N = 5), HR monitors (N = 3), a fNIRS (N = 1), a BP monitor (N = 1), a plethysmograph (N = 1), breathing monitors (N = 2), and a multi-purpose Shimmer GSR + device (N = 1). The main heart measures used were HR and HRV. There was also vascular response index measure used by calculating the ratio of specific amplitudes of the signals. Respiration rates and BP were also used. Eleven studies included heart and respiration measures, eight of which recorded different types of measurement within this category.

Eye Measures

Information (see **Table 2**) was collected from eye-tracking devices (N = 11), EEG (N = 1), EOG (N = 1), a web-camera (N = 1), and from tonometry (N = 1). The two most frequent measures in this category were pupil diameter and blink rates. The former is based on increased size of pupil averaged over time and the latter the number of blinks per time period. There were also measures of gaze fixations, saccades (gaze transitions), astigmatism, and ocular pressure. Fifteen studies included eye measures, five of which recorded different types of measurement connected to this category.

Skin Measures

Eight studies (see **Table 3**) collected GSR data measuring EDA from sensors attached to the foot (N = 1), hand (N = 4), fingers (N = 2), or wrist (N = 1). This data was difficult to divide into sub-categories because studies often did not provide sufficient information or used a variety of different signal features. For example, the majority of studies did not indicate which of the EDA signal components (phasic and tonic; see Vanneste et al., 2020) were measured. Although, it could be assumed that the phasic data was most likely used because of the shorter time intervals involved. Furthermore, several studies reported mean and accumulation GSR statistics based on different features such as amplitudes, total power, gradients, peak numbers, spectral density, and wave rises. Even though they had some common labels such as GSR-mean, it was not necessarily means of the same data. Some studies also transformed their data using Fourier analysis techniques. Hence, it was not possible to form consistent subgroups and therefore all GSR data was grouped together under the heading Skin-GSR.

In two studies skin temperature was recorded. Although we note that changes in temperature due to stress or arousal can be caused by vasoconstrictions (see Vanneste et al., 2020) and therefore could be considered under the Heart and Respiration category, we reported this measure here in the Skin category because of the direct reference to, and measurement of, this part of the body. We did not include this measure in the skin category summary because it did fit easily with the other measures, which all were based on GSR signals. Its exclusion provides a more meaningful grouping for further analysis.

Brain Measures

Thirteen studies (see **Table 4**) reported measures based on the five sub-bands (alpha, beta, gamma, delta, and theta) of brain electrical activity obtained from EEGs (N = 10), fMRIs (N = 2), and an fNIRS (N = 1) data and was recorded according to the five sub-bands (see **Table 4**). Typically, band power or amplitude was recorded following power spectrum density analysis. Eleven studies recorded more than one sub-band of data. In some studies information was collected on various lobes of the brain such as the frontal, occipital, and parietal lobes. In these cases, data was classified according to the sub-bands (e.g., alpha). In some limited studies, signal data were combined or transformed (e.g., alpha and theta data were combined) and this was recorded under the category 'Other.'

Subjective Measures of Cognitive Load

Even though our main aim was to investigate physiological measures of cognitive load, many studies in the sample also included subjective measures. In particular, the NASA-TLX scale (see Hart and Staveland, 1988) was often used as a comparative tool as it was considered the gold standard in measuring workload in human–computer interaction studies. This scale requires task participants to subjectively rate: mental demand, physical demand, temporal demand, performance, effort, and frustration. All studies (N = 12) that used this instrument aggregated the six items to get an overall mental workload rating, which we documented. Many studies also recorded and analyzed the six sub-scales separately. In these cases, we also recorded the data for mental demand and effort, as they more closely resembled the single-item rating scales used in cognitive load theory. Further, some studies (N = 8) also used single-item measures of effort, difficulty, and demand, that were not part of the NASA-TLX survey, and more consistent with the scale devised by Paas (1992). This data was also recorded and included in **Table 5** under the heading Single item. Although one exception to this was an aggregated measure of intrinsic cognitive load using 3 items based on the survey developed by Leppink et al. (2013).

Types of Tasks

For each type of cognitive load measure, a record was made of the type of tasks used in the study. These could be categorized into arithmetic, working memory, simulations, object-shape manipulations and word tasks. Individual analysis of task types was recorded in **Tables 1–5**. Arithmetic tasks were constructed from mental arithmetic problems; simulations used specialized equipment to mimic (simulate) real-life tasks that involved motorcar driving, engineering skills, military exercises and surgery; memory tasks included n-back and digit-span tasks; object/shapes included visual object tracking and shape construction tasks; and word tasks were based on both reading and writing tasks.

Assessing Validity

The following steps were conducted to assess the validity of the identified measures. The first step was to confirm whether the various physiological measures were capable of detecting *any* significant differences in cognitive load across the different tasks in each study. Based on cognitive load theory, it is predicted that higher complexity tasks yield higher intrinsic load that should thus be reflected in the physiological measure; the more studies provide evidence for that prediction, the higher the construct validity of the measure. Therefore, for each use of a

physiological measure, a record was made of whether a significant difference was found across the tasks. This information was recorded in Column-4 (labeled *At least 1 significant difference*) for **Tables 1–5**. For example, Column-4 in **Table 1** for HR measures collected during simulation tasks, indicates that significant differences between tasks in HR were found in four of the four experiments (100%).

The next step was to document to what extent the cognitive load measures matched the performance test scores, providing information on levels of construct validity and sensitivity. In studies that feature tasks of different complexities, it is assumed that performance scores will vary in accordance with cognitive load. As more demand is made on working memory, less correct answers would be expected (see Ayres and Sweller, 1990). For example, if a study used 3 tasks (T1, T2, and T3) there are three possible pair-wise comparisons (T1–T2, T1–T3, T2–T3). If the three tasks had significantly different levels of complexity then it would be expected that test scores would give three significant differences in pair-wise comparisons. It would then be expected that the measures of cognitive load would also detect three significant differences, consistent with the nature of the physiological measure (e.g., as task complexity increases and scores decrease, HR increases). In the case of three significant test differences and three corresponding cognitive load differences in an experiment, a match was recorded. If the cognitive load measure only found two significant differences this was considered a non-match. This information was recorded in Column-5 (labeled *Matches with task performance*) for **Tables 1–5**. For the simulation-HR example in **Table 1** (Column-5), cognitive load measures matched test scores in three of the four studies (75% matches), indicating that for the other study cognitive load measures failed to find as many differences as the test scores.

It should be noted that in two studies in the sample no pair-wise comparisons were made, only an overall ANOVA was

TABLE 1 | Heart and respiration measures.

Tasks	Type of measure	No. of experiments	At least 1 significant difference	Matches with task performance	Pair-wise deviations
Arithmetic					
	HR	1	1 (100%)	1 (100%)	0
	HRV	1	1 (100%)	1 (100%)	0
	Vasc. Response	2	2 (100%)	2 (100%)	0
	BP	1	0 (0%)	0 (0%)	1.00
	Total	5	4 (80%)	4 (80%)	0.20
Simulations					
	HR	4	4 (100%)	3 (75%)	–0.25
	HRV	3	0 (0%)	0 (0%)	–1.67
	Respiration	3	2 (67%)	2 (67%)	–0.33
	Total	10	6 (60%)	5 (40%)	–0.70
Memory					
	HR	1	1 (100%)	1 (100%)	0
	HRV	3	3 (100%)	3 (100%)	0
	Respiration	1	1 (100%)	1 (100%)	0
	Total	5	5 (100%)	5 (100%)	0
Word tasks					
	HR	1	0 (0%)	0 (0%)	–3.00
	HRV	1	0 (0%)	0 (0%)	–3.00
	Total	2	0 (0%)	0 (0%)	–3.00
All					
	HR	7	6 (86%)	5 (71%)	–0.57
	HRV	8	4 (50%)	4 (50%)	–1.00
	Vasc. response	2	2 (100%)	2 (100%)	0
	Respiration	4	3 (75%)	3 (75%)	–0.25
	BP	1	0 (0%)	0 (0%)	–1.00
	Totals	22	15 (68%)	14 (64%)	–0.64

TABLE 2 | Eyes.

Tasks	Type of measure	No. of experiments	At least 1 significant difference	Matches with task performance	Pair-wise deviations
Arithmetic Simulations	–	–	–	–	–
	Pupil diam.	6	3 (50%)	3 (50%)	–0.80
	Blink rate	4	4 (100%)	4 (100%)	0
	Fixations	3	3 (100%)	3 (100%)	+0.33
	Ocular press.	1	1 (100%)	1 (100%)	0
	Total	14	11 (79%)	11 (79%)	–0.29
Memory	–	–	–	–	–
	Pupil diam.	1	1 (100%)	1 (100%)	0
	Astigmatism	1	1 (100%)	1 (100%)	0
	Saccades	1	1 (100%)	1 (100%)	0
	Total	3	3 (100%)	3 (100%)	0
Objects/ shapes	Pupil diam.	3	3 (100%)	2 (67%)	–0.67
	Blink rate	1	1 (100%)	1 (100%)	0
	Total	4	4 (100%)	3 (75%)	–0.50
Word tasks	–	–	–	–	–
	Pupil diam.	1	1 (100%)	1 (100%)	0
	Blink rate	1	1 (100%)	0 (0%)	1
	Total	2	2 (100%)	1 (50%)	–0.50
All	–	–	–	–	–
	Pupil diam.	11	8 (73%)	7 (64%)	–0.60
	Blink rate	6	6 (100%)	5 (83%)	–0.17
	Other	6	6 (100%)	6 (100%)	0
	Totals	23	20 (87%)	18 (78%)	–0.30

TABLE 3 | Skin measures.

Tasks	Type of measure	No. of experiments	At least 1 significant difference	Matches with task performance	Pair-wise deviations
Arithmetic	GSR	6	6 (100%)	3 (50%)	–1
Simulation	GSR	1	1 (100%)	1 (100%)	+1
Memory	GSR	2	1 (50%)	1 (50%)	–0.50
Objects/ shapes	GSR	2	1 (50%)	1 (50%)	–0.50
Written	GSR	2	2 (100%)	2 (100%)	0
	Totals	13	11 (85%)	8 (62%)	–0.54

conducted. In these cases, only one overall comparison could be made. On some rare occasions, the cognitive load measure found more significant differences than the test performance scores suggesting a more sensitive measure, so these were also considered a match because they found at least the same number of differences.

To measure the variations between the test scores and the cognitive load scores, the differences between them were recorded. For example, if the test scores indicated three significant pairwise comparisons, but the cognitive load measure only found one, the variation was recorded as –2. For each sub-category of measure, this information was averaged and recorded in Column-6 (labeled *Pair-wise deviations*) for each table. For example, –1.67 is recorded for HRV for simulations in **Table 1**. Over the three studies (each individual experiment counted separately if studies contained multiple experiments) total variations short of what was expected summed to five giving an average of –1.67. On the rare occasions that the cognitive load measure found more significant differences than the test scores, it was possible to have a positive (+) average (see **Table 2**: fixations for simulations).

The data reported in columns 4–6 enabled differences between the sensitivity of the various measures to be compared. For example, if for three different cognitive tasks measure-A records no cognitive load differences, then scores in columns 4–6 would be recorded as (0, 0, –3). Whereas if a more sensitive measure-B records two cognitive load differences then it would be recorded as (1, 0, –1). B is clearly more sensitive than A which is reflected in this scoring rubric. Note a perfect match of three significant differences in cognitive load would be recorded as (1, 1, 0).

Finally, to make some comparisons between the different measures, overall summaries are reported in **Table 6**.

RESULTS

Analysis of Individual Measures

Heart and Respiration Measures

Heart and respiration measures were recorded in 11 experiments (see Rendon-Velez et al., 2016; Wong and Epps, 2016; Reinerman-Jones et al., 2017; Wu et al., 2017; Lyu et al., 2018; Alrefaie et al., 2019; He et al., 2019; Jaiswal et al., 2019; Ahmad et al., 2020; Digiesi et al., 2020; Gupta et al., 2020; Zakeri et al., 2020). For this category (see **Table 1**), the most common forms of measures were HR and HRV. Respiration rates and blood pressure (BP) were also measured along with some novel indices, such as the Vascular response index, calculated from the ratio of different amplitudes taken from a photoplethysmogram waveform (see Lyu et al., 2018). For 8 of the 11 studies, more than one sub-category of measures were collected. With such a small

TABLE 4 | Brain measures.

Tasks	Type of measure	No. of experiments	At least 1 significant difference	Matches with task performance	Pair-wise deviations
Arithmetic	Alpha	1	1 (100%)	1 (100%)	0
	Beta	1	1 (100%)	0 (0%)	–2.00
	Theta	1	1 (100%)	0 (0%)	–1.0
	Total	3	3 (100%)	1 (33%)	1.00
Simulations	Alpha	5	4 (80%)	2 (40%)	–0.80
	Beta	5	3 (60%)	0 (0%)	–1.20
	Gamma	3	2 (67%)	1 (33%)	–1.33
	Delta	3	2 (67%)	1 (33%)	–1.33
	Theta	4	2 (50%)	1 (25%)	–1.25
	Other	2	0 (0%)	0 (0%)	–2.00
	Total	22	13 (59%)	5 (23%)	–1.36
Memory	Alpha	2	2 (100%)	2 (100%)	0
	Gamma	1	1 (100%)	1 (100%)	0
	Theta	1	0 (0%)	0 (0%)	–3.00
	Other	1	1 (100%)	1(100%)	0
	Total	5	4 (80%)	4 (80%)	–0.60
Objects/ shapes	Alpha	1	0 (0%)	0 (0%)	–3.00
	Beta	1	1 (100%)	0 (0%)	–2.00
	Other	2	2 (100%)	1 (50%)	–1.00
	Total	4	3 (75%)	1 (25%)	–1.50
All	Alpha	9	7 (78%)	5 (56%)	–0.78
	Beta	7	5 (71%)	0 (0%)	–1.43
	Gamma	4	3 (75%)	2 (50%)	–1.00
	Delta	3	2 (67%)	1 (33%)	–1.33
	Theta	6	3 (50%)	1 (17%)	–1.50
	Other	5	3 (60%)	2 (40%)	–1.00
	Totals	34	23 (68%)	11 (32%)	–1.15

TABLE 5 | Subjective measures.

Tasks	Type of measure	No. of experiments	At least 1 significant difference	Matches with task performance	Pair-wise deviations
Arithmetic	Single item	3	3 (100%)	3 (100%)	0
Simulations	NASA-overall	8	7 (88%)	5 (63%)	–0.38
	NASA-effort	6	5 (83%)	3 (50%)	–0.50
	NASA-demand	5	5 (100%)	4 (100%)	0
	Total	19	17 (89%)	12 (63%)	–0.32
Memory	NASA-overall	2	2 (100%)	2 (100%)	0
	NASA-effort	1	1 (100%)	1 (100%)	0
	NASA-demand	2	2 (100%)	2 (100%)	0
	Single item	1	1 (100%)	1 (100%)	0
	Total	6	6 (100%)	6 (100%)	0
Objects/ shapes	Single item	4	4 (100%)	4 (100%)	0
All	NASA-overall	10	9 (90%)	7 (70%)	–0.30
	NASA-effort	7	6 (86%)	4 (57%)	–0.43
	NASA-demand	7	7 (100%)	6 (86%)	0
	Single item	8	8 (100%)	8 (100%)	0
	Totals	32	30 (94%)	25 (78%)	–0.19

sample size it was not possible to conduct meaningful statistical tests between individual measures; however, some observations could be made. Firstly, HR, the vascular response index and respiration rates produced high levels of consistency with test results (indicating more sensitivity), even though the two latter cases involved a small number of cases. Secondly HR measures were more sensitive than HRV measures, the latter being totally ineffective on simulation tasks. From the perspective of task type, memory and arithmetic tasks produced high levels of consistency, whereas the more complex simulations did not. The two studies with word tasks produced no significant cognitive load measures for HR or HRV. Combined across the different tasks, the heart measures were capable of finding at least one significant difference in cognitive load in 68% of studies, with exact matches of 64%.

TABLE 6 | Summaries.

Tasks	Type of measure	No. of collections	At least 1 significant difference	Matches with task performance	Pair-wise deviations
Arithmetic	Subjective	3	3 (100%)	3 (100%)	0
	Skin-GSR	6	6 (100%)	3 (50%)	–1.00
	Brain	3	3 (100%)	1 (33%)	–1.00
	Heart	5	4 (80%)	4 (80%)	0.20
	Eyes	–	–	–	–
	All	17	16 (94%)	11 (65%)	–0.59
Simulations	Skin-GSR	1	1 (100%)	1 (100%)	+1.00
	Subjective	19	17 (89%)	12 (63%)	–0.32
	Eyes	14	11 (79%)	11 (79%)	–0.29
	Heart	10	6 (60%)	5 (40%)	–0.70
	Brain	22	13 (59%)	5 (23%)	–1.36
	All	66	48 (73%)	34 (52%)	–0.70
Memory	Subjective	6	6 (100%)	6 (100%)	0
	Heart	5	5 (100%)	5 (100%)	0
	Eyes	3	3 (100%)	3 (100%)	0
	Brain	5	4 (80%)	4 (80%)	–0.60
	Skin-GSR	2	1 (50%)	1 (50%)	–0.50
	All	20	18 (90%)	18 (90%)	–0.20
Objects/ shapes	Subjective	4	4 (100%)	4 (100%)	0
	Eyes	4	3 (75%)	3 (75%)	–0.50
	Brain	4	3 (75%)	1 (25%)	–1.50
	Skin-GSR	4	1 (25%)	1 (25%)	–1.25
	Heart	–	–	–	–
	All	16	11 (69%)	9 (56%)	–0.81
Word	Skin-GSR	2	2 (100%)	2 (100%)	0
	Eyes	2	2 (100%)	1 (50%)	–0.50
	Heart	2	0 (0%)	0 (0%)	–3.00
	Subjective	–	–	–	–
	Brain	–	–	–	–
	All	6	4 (67%)	3 (50%)	–1.17
All	Subjective	32	30 (94%)	25 (78%)	–0.19
	Eyes	23	20 (87%)	18 (78%)	–0.30
	Skin-GSR	13	11 (85%)	8 (62%)	–0.54
	Heart	22	15 (68%)	14 (64%)	–0.64
	Brain	34	23 (68%)	11 (32%)	–1.05

Eye Measures

Eye measures were recorded in 15 experiments (see Mazur et al., 2016; Rendon-Velez et al., 2016; Wong and Epps, 2016; Hosseini et al., 2017; Yan et al., 2017; Jiménez et al., 2018; Alrefaie et al., 2019; He et al., 2019; Hossain et al., 2019; Vera et al., 2019; Ahmad et al., 2020; Maki-Marttunen et al., 2020; van Acker et al., 2020; Vanneste et al., 2020; Zakeri et al., 2020). The two most popular measures were pupil diameter and blink rates. Measuring the number of saccades, astigmatism, gaze positions, and ocular pressure were also used in a small number of experiments. For five of the 15 studies, more than one sub-category of eye measures were collected. Overall, blink rates had a very high level of being able to detect changes in cognitive load indicating greater sensitivity; whereas pupil diameter was less successful, especially on simulation tasks compared with other measures. Combined across the different tasks, the eye measures were capable of finding at least one significant difference in cognitive load in 87% of studies, with exact matches of 78%.

Skin Measures

Skin measures were reported in 10 experiments (see Nourbakhsh et al., 2017; Ghaderyan et al., 2018; Lyu et al., 2018; He et al., 2019; Hossain et al., 2019; Larmuseau et al., 2019; Gupta et al., 2020; Vanneste et al., 2020). Six of the 10 studies used multiple skin measures. The two most frequent forms of GSR-accumulation measure and GSR-other were able to detect differences in cognitive load at the 100% level. In contrast, GSR-average found cognitive load differences in only 50% of the time. In terms of exact matches, GSR-other had a success rate of 83% suggesting a high level of sensitivity, compared with 50% (GSR-average) and 33% (GSR-accumulation). The two cases where skin temperature was recorded found no cognitive load differences. Overall, the more infrequent measures of signal data (e.g., wave rises) produced the best results.

Brain Measures

Brain measures were reported in 13 experiments (see Mazur et al., 2016; Tan et al., 2016; Wang et al., 2016; Hosseini et al., 2017; Reinerman-Jones et al., 2017; Wu et al., 2017; Katahira et al., 2018; He et al., 2019; Abd Rahman et al., 2020; Gupta et al., 2020; Maki-Marttunen et al., 2020; Vanneste et al., 2020). Eleven of the 13 studies recorded multiple types of signals (e.g., alpha, beta, and gamma) that were often taken from different parts of the brain. As can be seen from **Table 4**, the overall brain signal measures were only able to detect differences in cognitive load at the 68% level and could only match performance differences at 32%. Measures of alpha and gamma waves were the most promising. Beta waves were able to detect differences at the 71% level but had zero matches with test scores. Delta and theta waves generally had low levels of sensitivity.

Subjective Measures

Subjective measures were recorded in 18 studies (see Rendon-Velez et al., 2016; Nourbakhsh et al., 2017; Reinerman-Jones et al., 2017; Yan et al., 2017; Wu et al., 2017; Ghaderyan et al., 2018; Jiménez et al., 2018; He et al., 2019; Jaiswal et al., 2019; Larmuseau et al., 2019; Vera et al., 2019; Abd Rahman et al., 2020;

Digiesi et al., 2020; Gupta et al., 2020; van Acker et al., 2020; Zakeri et al., 2020). As can be seen from **Table 5**, subjective measures were very successful in identifying differences in cognitive load. In terms of identifying any changes in cognitive load, accuracy scores ranged from 86–100%. In terms of exact matches with test scores, single-item measures scored at 100% accuracy. The NASA-demand score was high at 86%, followed by the overall NASA at 70% and the NASA-effort at 57%.

Comparison of the Different Categories

By combining the data for each category of measure (see **Table 6**) sample sizes became sufficiently large to make meaningful statistical comparisons. Because the data did not fit normal distributions non-parametric tests were completed. For the three data sets summarized in columns 4–6, Kruskal–Wallis tests were completed for the five categories.

At Least 1-Significant Difference in Cognitive Load Detected

Comparing the five measures for this data there was an overall significant difference between the five categories (Kruskal–Wallis $\chi^2 = 9.67$, $df = 4$, $p = 0.046$). *Post hoc* tests using the Benjamini and Hochberg (1995) correction method indicated no significant pairwise differences. However, it is worth noting that subjective measures had an accuracy rate of 94% compared with Heart and Brain measures that were below 70%.

Matches With Task Performance

For this data there was a significant between-group difference (Kruskal–Wallis $\chi^2 = 18.52$, $df = 4$, $p < 0.001$). *Post hoc* tests indicated that both the subjective measures ($M = 0.78$, $SD = 0.42$, $p = 0.002$) and the eye measures ($M = 0.78$, $SD = 0.42$, $p = 0.004$) had significantly more matches with task performance than the brain measures ($M = 0.32$, $SD = 0.47$).

Matching Deviations

For this data there was a significant between-group difference (Kruskal–Wallis $\chi^2 = 19.44$, $df = 4$, $p < 0.001$). *Post hoc* tests indicated that both the subjective measures ($M = 0.19$, $SD = 0.82$, $p < 0.001$) and eye measures ($M = 0.30$, $SD = 0.76$, $p = 0.006$) had significantly less deviations from the task performances than the brain measures ($M = 1.15$, $SD = 1.05$).

In summary, the analyses provided in this section suggest that overall the subjective and eye measures were the most sensitive indicators of cognitive load differences across tasks. Clearly the least effective were brain measures. Measures associated with the skin and heart were located between the other indicators. Nevertheless, all subcategories of measures were able to detect some differences in cognitive loads, and also some matches with test scores. Only the heart measures on word problems failed to detect any significant differences or matches.

Task Comparisons

The data in **Table 6** were examined in terms of task differences. The memory and arithmetic tasks recorded the best results. For both tasks, the combined measures were able to detect at least one cognitive load difference at the 90% level. For matches with performance, memory tasks achieved a 90% match, whereas arithmetic tasks had lower match rates at the 65% level. Simulations, object/shape manipulations and word tasks were overall at a lower level of accuracy, although overall matches were at least 50%.

Individual Measure Comparisons

The previous analysis in this section was on aggregated data. A closer look at individual measures was achieved by examining those specific measures (no other/combined categories were included) that had been used at least six times over the sample. By averaging the % scores for the data in columns 4 (at least one significant cognitive load difference) and 5 (matches with test performance) of **Tables 1–5** the following ranking was found: single-item subjective measures (100%), NASA-demand (93%), Blink-rates (91.5%), NASA-overall (80%), HR (78.5%), GSR (73.5%), NASA-effort (71.5%), Pupil-diameter (68.5%), Alpha waves (67%), and HRV (50%). Each of these measures, across all five general categories scored at least at the 50% accuracy level. There were four scores above 90% indicating very high levels of accuracy at detecting cognitive load differences, but three of those four scores were subjective rather than physiological measures.

Battery of Tests

As reported above nearly all studies ($N = 30$) included a battery of tests from within and/or between the categories (including subjective measures) to measure cognitive load. Examination of this data revealed that in all but three cases, at least one of the tests had an exact match with the expected number of cognitive load differences.

DISCUSSION

There is considerable evidence that supports the argument that changes in cognitive load can be detected by physiological measures that utilizes data signals collected from the heart and lungs, eyes, skin, and brain. Between and within these four categories, signals react differently according to the measure. For example, as cognitive load increases, pupil dilation is expected to increase, but HRV to decrease. From the perspective of a validity study, results should be consistent with the predictions made for each individual measure. The more studies find evidence for this prediction, the higher the construct validity of the measure. The data analyzed here indicate that physiological measures taken from all four parts of the body are capable of detecting changes in cognitive load generated by differences in task complexity. Of all the measures cataloged in **Tables 1–4** only BP and skin temperature failed to find a significant difference. However, these measures were only reported in three studies and therefore these non-effects cannot be generalized. Overall measures associated with the heart and lungs, eyes, skin, and brain measures were all capable of finding significant cognitive load differences across tasks, providing a level of construct validity (see Borsboom et al., 2004). Therefore, in answer to RQ1 (*Do the physiological measures have construct validity in detecting changes in intrinsic cognitive load across tasks?),* we conclude that nearly all the measures

identified in this sample have some level of construct validity because of their capacity to detect changes frequently.

The information reported in **Tables 1–4**, **6** indicated that some physiological measures were more sensitive than others to finding cognitive load differences across tasks. The most sensitive physiological measures were blink rates, HR, pupil dilations, and alpha waves. Therefore, in answer to RQ2 (*How sensitive are the physiological measures in detecting changes in intrinsic cognitive load?),* we conclude that the different measures were sensitive enough to identify differences in cognitive load but the levels of sensitivity varied significantly between measures.

By examining the different types of tasks used in these studies we found that memory and mental arithmetic tasks led to higher degrees of sensitivity for the cognitive load measures than simulations and object/shape manipulations. The former are short tasks relying completely on working memory capacity; whereas the latter are more authentic and specialized tasks that may depend more on long-term memory (e.g., knowledge of medical functions and procedures). Although no studies were included that manipulated stress and anxiety, these affective influences may have been generated by more authentic high-stakes tasks (e.g., surgery simulations) leading to confounding factors. Nevertheless, our evidence found that eye measures were more sensitive than other physiological tests for the more specialized tasks, perhaps due to the importance of information received visually. Clearly the type of task does impact on the different levels of sensitivity across the physiological measures and therefore overall validity (RQ3 – *Does the type of task impact on overall validity?*).

Although not the main aim of our study the sample provided data on subjective measures of cognitive load, which also showed validity. We found that compared with most of the physiological measures included in our sample, the subjective measures had a higher level of sensitivity. As reported above, single-item measures of cognitive load such as asking to rate the amount of effort, difficulty or demand experienced, were found to have the highest sensitivity. Only blink-rate data was at the same level. The favorable finding for subjective measures is interesting as in more recent time, CLT researchers have started to use more physiological tests for measuring cognitive load (Ayres, 2020). Many commentators have expressed concern over the high use of subjective measures suggesting more objective methods are required (see Schnotz and Kürschner, 2007; Kirschner et al., 2011).

Our data suggests that it may be premature to abandon subjective methods, but the question arises should we be more circumspect in using physiological measures? They are objective, but in our sample evidence emerged that many are not as sensitive as subjective measures. In only one study in the sample did a physiological measure (alpha signals) identify more variations in cognitive load than the subjective measure (overall NASA-TLX). It is also notable that nearly 30 years ago Paas and van Merriënboer (1994; see also Paas et al., 1994) found a self-rating measure of mental effort to be superior to HR data, which led to wide-scale adoption of subjective measures of effort and difficulty (see van Gog and Paas, 2008; Ayres, 2018). However, there are cases in the literature where subjective measures have also been found to lack sensitivity. For example, Lee et al. (2020) found pupil dilation can be more sensitive than self-rating measures. This was the case when the task included the management of emotions, and the confounding factors were well-controlled. Hence, the picture is not definitive and more examination of the influences on physiological measures is required, as reported next.

Clearly some of the physiological measures in this sample lacked sensitivity especially with specific types of tasks. However, this can be because these physiological measures did not optimally match with the type of task. The choice of the measure types is important for the question of sensitivity: some measures are more sensitive than others for specific tasks, even within the same category of physiological measures. For instance, in driving tasks, horizontal gaze dispersion showed a larger effect size than other eye-tracking measures (Wang et al., 2014). Another explanation described in the introduction is that there are a number of other conditions that can have a negative impact on physiological measures. For example, HR and eye measures can be influenced by participant motion (e.g., driving a car, see Lohani et al., 2019) EDA may be more sensitive to sudden rather than gradual changes, and some areas of the brain (e.g., the fronto-parietal attention network) are more conducive to measuring cognitive load than other areas. It is possible that some studies may have been influenced by such factors, and therefore some caution should be shown in interpreting the results unconditionally. In a review into using physiological measures of more wide-ranging causes of mental workload, Charles and Nixon (2019) concluded that there was no single preferred measure that could be used across all tasks and domains, but more evidence was emerging in how best to utilize each type. Our study to some extent supports this finding; however, we suggest further that until more research is completed, the best outcomes may be found in a well-chosen combination of tests.

In conclusion, we believe there is a solid case for including both physiological and subjective methods to measure cognitive load. Like other studies (see Johannessen et al., 2020) using a battery of different tests was found to be effective as in most cases at least one of the tests identified all the changes in cognitive load. Consequently, we suggest two points to take into account in using physiological measure to measure cognitive load: (1) select a measure based on the understanding of the given task; and (2) triangulate by combining different physiological measures as well as subjective measures. As our study showed, physiological measures can be valid in some setups (shown by results for RQ1), but also the sensitivity can vary across different measures depending on the task types (RQ2 and RQ3). Thus, careful selection of the right measures for the given task is essential. A thorough task analysis (e.g., cognitive task analysis) could be helpful in achieving this aim.

Triangulation is an effective research method to gain a comprehensive perspective and validation of data (Patton, 1999). Studies have shown that combining multiple physiological measures may present a higher sensitivity in measuring cognitive load than using a single measure (Van Orden et al., 2000; Aghajani et al., 2017). Combining physiological measures with subjective measures may show either positive convergence

(Aldekhyl et al., 2018) or diverging sensitivity as aforementioned. However, such inconsistency does not necessarily represent an incredibility of data, but rather provides a deeper insight into the results (Patton, 1999).

We also found in this study that the nomenclature used by some authors was not always consistent with other authors. In a small number of studies, insufficient detail was provided in the method to make a definitive judgment on what exactly was used for the basis of the data calculations. Most notably this inconsistency was sometimes found in the GSR measures. For example, an 'average' GSR value would be reported without a clear indication of what part of the signal the average referred to. Although this did not impact on our overall findings, future studies should ensure that all necessary information is reported, and conform to a standardization of labels such as previously suggested by Fowles et al. (1981).

The sole focus of this study was on tasks that generated intrinsic cognitive load only. No attempt was made to include learning studies that would generate other types of cognitive load. Detecting changes in intrinsic load is important to CLT but so is extraneous cognitive load as CLT is predominantly interested in the impact of all types of cognitive load on learning. Hence, further research is required of this nature to explore validity during learning experiments that also featured between-subject experiments. As well as task complexity changes in cognitive load caused by other factors such as distractors (e.g., noise), and affective factors (e.g., emotion) should also be researched.

In addition, the studies in this sample were predominantly completed by young adults whose mean ages ranged from 20 to 30. However, two studies (see Tan et al., 2016; Abd Rahman et al., 2020) focused on elderly adults. Although these studies did not report any data that was inconsistent with the other studies in the sample, it is known that aging adults experience cognitive decline and are more susceptible to cognitive load variations (see van Gerven et al., 2000; Klencklen et al., 2017). Similarly, the sample did not include any young children. Subsequently, we cannot generalize our results to other age groups and therefore more research is needed in both older and younger populations to explore potential differences.

Our main focus was to examine the validity of using physiological measures of cognitive load. We conducted the study from a measurement and theoretical perspective especially that of cognitive load theory. It is worth mentioning, however, that there are real-world applications of measuring total cognitive load; often referred to as mental workload. For example, how to manage cognitive load during medical training and real-life practice (see Fraser et al., 2015; Johannessen et al., 2020; Szulewski et al., 2020) is critical. Similarly, in driving a car it can be advantageous from a safety perspective to be able to monitor the driver's cognitive states and implement interventions accordingly (see Lohani et al., 2019; Meteier et al., 2021). When driving, real-time continuous data is required which can be collected through physiological measures, but using subjective measures is impossible. From this important perspective, physiological measures have a clear edge.

By analyzing a snapshot of studies taken from a recent 5-year period we have clearly not included some of the key studies from the period before this time. Although we cite a number of significant studies in our introduction, it was never our intention to go beyond our specified period. We wanted to investigate exactly what a contemporary collection of studies would reveal about measuring cognitive load. Our sample, like previous studies confirmed that different types of physiological measures are capable of measuring cognitive load.

There were a number of novel aspects to the study. Firstly, we included only studies that investigated the capacity of physiological measures to identify changes in intrinsic cognitive load by manipulating task complexity. Most studies in measuring cognitive load using physiological methods have not made this distinction. Secondly, by taking a broad sample we were able to compare a wide variety of physiological measures from four main categories, and to compare them with each other from both a construct validity and sensitivity perspective. An added bonus was that we were also able to benchmark them against a number of subjective rating scales. Thirdly, by examining the influence of task types, and the highlighting of technological precautions and other influences outlined in the literature review, we were able to identify some major factors that impact cognitive load measures.

In conclusion, we found that nearly all the physiological measures identified in this sample had some level of validity. However, there were wide variations in sensitivity to detect changes in intrinsic cognitive load, which was impacted by task specificity. In contrast, subjective measures generally had high levels of validity. We recommend that a battery of tests (physiological and/or subjective) are required to obtain the best indicators of changes in intrinsic cognitive load.

AUTHOR CONTRIBUTIONS

PA conducted the literature search and collected, compiled, and analyzed the data. PA, JL, FP, and JM designed the study and wrote the manuscript. All authors contributed to the article and approved the submitted version.

REFERENCES

Aasman, J., Mulder, G., and Mulder, L. J. (1987). Operator effort and the measurement of heart-rate variability. *Hum. Fact.* 29, 161–170. doi: 10.1177/001872088702900204

Abd Rahman, N. I., Dawal, S. Z. M., and Yusoff, N. (2020). Driving mental workload and performance of ageing drivers. *Transp. Res. Part F* 69, 265–285.

Aghajani, H., Garbey, M., and Omurtag, A. (2017). Measuring mental workload with EEG+fNIRS. *Front. Hum. Neurosci.* 11:359. doi: 10.3389/fnhum.2017.00359

Ahmad, M. I., Keller, I., Robb, D. A., and Lohan, K. S. (2020). A framework to estimate cognitive load using physiological data. *Person. Ubiq. Comput.* doi: 10.1007/s00779-020-01455-7

Aldekhyl, S., Cavalcanti, R. B., and Naismith, L. M. (2018). Cognitive load predicts point-of-care ultrasound simulator performance. *Perspect. Med. Educat.* 7, 23–32. doi: 10.1007/s40037-017-0392-7

Alrefaie, M. T., Summerskill, S., and Jackon, T. W. (2019). In a heartbeat: Using driver's physiological changes to determine the quality of a takeover in highly automated vehicles. *Accid. Anal. Prev.* 131, 180–190. doi: 10.1016/j.aap.2019.06.011

Antonenko, P. D., and Niederhauser, D. S. (2010). The influence of leads on cognitive load and learning in a hypertext environment. *Comput. Hum. Behav.* 26, 140–150. doi: 10.1016/j.chb.2009.10.014

Antonenko, P., Paas, F., Grabner, F., and van Gog, T. (2010). Using electroencephalography to measure cognitive load. *Educat. Psychol. Rev.* 22, 425–438. doi: 10.1007/s10648-010-9130-y

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

Ayres, P. (2018). "Subjective measures of cognitive load: What can they reliably measure?," in *Cognitive Load Measurement and Application: A Theoretical Framework for Meaningful Research and Practice*, ed. R. Zheng (New York, NY: Routledge).

Ayres, P. (2020). Something old something new for cognitive load theory. *Comput. Hum. Behav.* 113:106503. doi: 10.1016/j.chb.2020.106503

Ayres, P., and Sweller, J. (1990). Locus of difficulty in multistage mathematics problems. *Am. J. Psychol.* 103, 167–193. doi: 10.2307/1423141

Backs, R. W., and Seljos, K. A. (1994). Metabolic and cardiorespiratory measures of mental effort: the effects of level of difficulty in a working memory task. *Int. J. Psychophysiol.* 16, 57–68. doi: 10.1016/0167-8760(94)90042-6

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc.* 57, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Blitz, P. S., Hoogstraten, J., and Mulder, G. (1970). Mental load, heart rate and heart rate variability. *Psychol. Forsch.* 33, 277–288. doi: 10.1007/bf00424555

Borsboom, D., Mellenbergh, G. J., and van Heerden, J. (2004). The concept of validity. *Psychol. Rev.* 111, 1061–1071.

Braithwaite, J., Watson, D., Jones, R., and Row, M. (2013). A guide for analysing electrodermal activity (EDA) & skin conductance responses (SCRs) for psychological experiments. *Psychophysiology* 49, 1017–1034. doi: 10.1111/j.1469-8986.2012.01384.x

Callan, D. J. (1998). "Eye movement relationships to excessive performance error in aviation," in *Proceedings of the Human Factors and Ergonomics Society annual meeting*. (New York, NY: Routledge).

Challoner, A. V. J. (1979). "Photoelectric plethysmography for estimating cutaneous blood flow," in *Non Invasive Physiological Measurements*, Vol. l, ed. P. Rolfe (London: Academic), 125–151.

Charles, R. L., and Nixon, J. (2019). Measuring mental workload using physiological measures: A systematic review. *Appl. Ergon.* 74, 221–232. doi: 10.1016/j.apergo.2018.08.028

Chen, H., Dey, A., Billinghurst, M., and Lindeman, R. W. (2017). "Exploring pupil dilation in emotional virtual reality environments," in *Proceedings of the International Conference on Artificial Reality and Telexistence Eurographics Symposium on Virtual Environments* (New York, NY: Routledge).

Chen, Y., Yan, S., and Tran, C. C. (2018). Comprehensive evaluation method for user interface design in nuclear power plant based on mental workload. *Nucl. Engin. Technol.* 30, 1–10.

Chi, M., Glaser, R., and Rees, E. (1982). "Expertise in problem solving," in *Advances in the Psychology of Human Intelligence*, ed. R. Sternberg (Hillsdale, NJ: Erlbaum), 7–75.

Dawson, M. E., Schell, A. M., and Filion, D. L. (2000). "The electrodermal system," in *Handbook of Psychophysiology*, 2nd Edn, eds J. T. Cacioppo, L. G. Tassinary, and G. C. Berntson (Cambridge, MA: Cambridge University Press), 200–223.

De Rivecourt, M., Kuperus, M. N., Post, W. J., and Mulder, L. J. M. (2008). Cardiovascular and eye activity measures as indices for momentary changes in mental effort during simulated flight. *Ergonomics* 51, 1295–1319. doi: 10.1080/00140130802120267

Digiesi, S., Manghisi, V. M., Facchini, F., Klose, E. M., Foglia, M. M., and Mummolo, C. (2020). Heart rate variability based assessment of cognitive workload in smart operators. *Manag. Product. Engin. Rev.* 11, 56–64.

Eckstein, M. K., Guerra-Carrillo, B., Miller Singley, A. T., and Bunge, S. A. (2017). Beyond eye gaze: What else can eyetracking reveal about cognition and cognitive development? *Devel. Cogn. Neurosci.* 25, 69–91. doi: 10.1016/j.dcn.2016.11.001

Finsen, L., Søgaard, K., Jensen, C., Borg, V., and Christensen, H. (2001). Muscle activity and cardiovascular response during computer-mouse work with and without memory demands. *Ergonomics* 44, 1312–1329. doi: 10.1080/00140130110099065

Fowles, D. C., Christie, M. J., Edelberg, R., Grings, W. W., Lykken, D. T., and Venables, P. H. (1981). Publication recommendations for electrodermal measurements. *Psychophysiology* 18, 232–239. doi: 10.1111/j.1469-8986.1981.tb03024.x

Fraser, K. L., Ayres, P., and Sweller, J. (2015). Cognitive load theory for the design of medical simulations. *Simul. Healthc.* 10, 295–307. doi: 10.1097/SIH.0000000000000097

Ghaderyan, P., Abbasi, A., and Ebrahimi, A. (2018). Time-varying singular value decomposition analysis of electrodermal activity: A novel method of cognitive load estimation. *Measurement* 126, 102–109. doi: 10.1016/j.measurement.2018.05.015

Glaholt, M. G. (2014). *Eye Tracking in the Cockpit: A Review of the Relationships Between Eye Movements and the Aviators Cognitive State*. Canada: Defense Research & Development Toronto.

Grassmann, M., Vlemincx, E., Von Leupoldt, A., Mittelstädt, J. M., and Van den Bergh, O. (2016a). Respiratory changes in response to cognitive load: a systematic review. *Neural Plast.* 2016:8146809.

Grassmann, M., Vlemincx, E., von Leupoldt, A., and Van den Bergh, O. (2016b). The role of respiratory measures to assess mental load in pilot selection. *Ergonomics* 59, 745–753. doi: 10.1080/00140139.2015.1090019

Gravetter, F. J., and Forzano, L.-A. B. (2018). *Research Methods for the Behavioural Sciences*. Boston, MA: Cengage.

Gupta, K., Hajika, R., Pai, Y. S., Duenser, A., Lochner, M., and Billinghurst, M. (2020). "Measuring human trust in a virtual assistant using physiological sensing in virtual reality," in *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)* (Netherland: IEEE), 756–765.

Hart, S. G., and Staveland, L. E. (1988). "Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research," in *Human Mental Workload. Advances in Psychology*, Vol. 52, eds P. A. Hancock and N. Meshkati (Amsterdam: North-Holland), 139–183. doi: 10.1016/s0166-4115(08)62386-9

Hayes, T. R., and Petrov, A. A. (2016). Mapping and correcting the influence of gaze position on pupil size measurements. *Behav. Res. Methods* 48, 510–527. doi: 10.3758/s13428-015-0588-x

He, D., Donmez, B., Liu, C. C., and Plataniotis, K. N. (2019). High cognitive load assessment in drivers through wireless electroencephalography and the validation of a modified N-Back task. *IEEE Transact. Hum. Mach. Syst.* 49, 362–371. doi: 10.1109/thms.2019.2917194

Henderson, J. M. (2011). *Eye Movements and Scene Perception. In the Oxford Handbook of Eye Movements*. Oxford: Oxford University Press, 593–606.

Hess, E. H., and Polt, J. M. (1964). Pupil size in relation to mental activity during simple problem-solving. *Science* 143, 1190–1192. doi: 10.1126/science.143.3611.1190

Holmqvist, K., and Andersson, R. (2017). *Eye-Tracking: A Comprehensive Guide to Methods, Paradigms and Measures*. New Jersey, NJ: Lund Eye-Tracking Research Institute.

Hoogerheide, V., Renkl, A., Fiorella, L., Paas, F., and Van Gog, T. (2019). Enhancing example-based learning: Teaching on video increases arousal and improves retention and transfer test performance. *J. Educat. Psychol.* 111, 45–56. doi: 10.1037/edu0000272

Hossain, D., Salimullah, S. M., Chowdhury, A. N., Hasan, S. M. N., Kabir, E., Mahmudi, R., et al. (2019). "Measurement of cognitive load for writing tasks using Galvanic Skin Response," in *6th International Conference on Networking, Systems and Security (NSYSS 2019) Dhaka* (Bangladesh: Association for Computing Machinery).

Hosseini, S. M. H., Bruno, J. L., Baker, J. M., Gundran, A., Harbott, L. K., Gerdes, C., et al. (2017). Neural, physiological, and behavioral correlates of visuomotor cognitive load. *Scient. Rep.* 7:8866.

Hyönä, J., Tommola, J., and Alaja, A.-M. (1995). Pupil dilation as a measure of processing load in simultaneous interpretation and other language tasks. *Q. J. Exp. Psychol.* 48, 598–612. doi: 10.1080/14640749508401407

Jaiswal, D., Chowdhury, A., Banerjee, T., and Chatterjee, D. (2019). "Effect of mental workload on breathing pattern and heart rate for a working memory task: A pilot study," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (Netherland: IEEE), 2202–2206.

Jasper, H. H. (1958). The ten twenty electrode system of the international federation. *Electroencephalogr. Clin. Neurophysiol.* 10, 371–375.

Jiménez, R., Cárdenas, D., González-Anera, R., Jiménez, J. R., and Vera, J. (2018). Measuring mental workload: ocular astigmatism aberration as a novel objective index. *Ergonomics* 61, 506–516. doi: 10.1080/00140139.2017.1395913

Johannessen, E., Szulewski, A., Radulovic, N., White, M., Braund, H., Howes, D., et al. (2020). Psychophysiologic measures of cognitive load in physician team leaders during trauma resuscitation. *Comput. Hum. Behav.* 111:106393. doi: 10.1016/j.chb.2020.106393

Jorna, P. G. (1992). Spectral analysis of heart rate and psychological state: a review of its validity as a workload index. *Biol. Psychol.* 34, 237–257. doi: 10.1016/0301-0511(92)90017-o

Kahneman, D., and Beatty, J. (1966). Pupil diameter and load on memory. *Science* 154, 1583–1585. doi: 10.1126/science.154.3756.1583

Kane, M. T. (2013). Validating teh interpretations and uses of test scores. *J. Educat. Measur.* 50, 1–73. doi: 10.1111/jedm.12000

Katahira, K., Yamazaki, Y., Yamaoka, C., Ozaki, H., Nakagawa, S., and Nagata, N. (2018). EEG correlates of the flow state: a combination of increased frontal theta and moderate frontocentral alpha rhythm in the mental arithmetic task. *Front. Psychol.* 9:300. doi: 10.3389/fpsyg.2018.00300

Kirschner, P., Ayres, P., and Chandler, P. (2011). Contemporary cognitive load theory: the good, bad and the ugly. *Comput. Hum. Behav.* 27, 99–105. doi: 10.1016/j.chb.2010.06.025

Klencklen, G., Lavenex, P. B., Brandner, C., and Lavenex, P. (2017). Working memory decline in normal aging: Memory load and representational demands affect performance. *Learn. Motivat.* 60, 10–22. doi: 10.1016/j.lmot.2017.09.002

Larmuseau, C., Vanneste, P., Desmet, P., and Depaepe, F. (2019). "Multichannel data for understanding cognitive affordances during complex problem solving," in *Proceedings of the 9th International Conference on Learning Analytics & Knowledge March 2019* (Bangladesh: Association for Computing Machinery). 61–70.

Lee, J. Y., Donkers, J., Jarodzka, H., and van Merriënboer, J. J. (2019). How prior knowledge affects problem-solving performance in a medical simulation game: Using game-logs and eye-tracking. *Comput. Hum. Behav.* 99, 268–277. doi: 10.1016/j.chb.2019.05.035

Lee, J. Y., Donkers, J., Jarodzka, H., Sellenraad, G., and Van Merriënboer, J. J. G. (2020). Different effects of pausing on cognitive load in a medical simulation game. *Comput. Hum. Behav.* 110:106385. doi: 10.1016/j.chb.2020.106385

Leppink, J., Paas, F., van der Vleuten, C. P. M., van Gog, T., and van Merriënboer, J. J. G. (2013). Development of an instrument for measuring different types of cognitive load. *Behav. Res. Methods* 45, 1058–1072. doi: 10.3758/s13428-013-0334-1

Lohani, M., Payne, B. R., and Strayer, D. L. (2019). A review of psychophysiological measuresto assess cognitive states in real-world driving. *Front. Hum. Neurosci.* 13:57. doi: 10.3389/fnhum.2019.00057

Longo, L., and Orru, G. (2018). "An evaluation of the reliability, validity, and sensitivity of three human mental workload measures under different instructional conditions in third-level education," in *Computer Supported Education. CSEDU 2018. Communications in Computer and Information Science*, Vol. 1022, eds B. McLaren, R. Reilly, S. Zvacek, and J. Uhomoibhi (Cham: Springer).

Lyu, Y. Q., Zhang, X., Luo, X. M., Hu, Z. Y., Zhang, J. Y., and Shi, Y. C. (2018). Non-invasive measurement of cognitive load and stress based on the reflected stress-induced vascular response index. *ACM Transact. Appl. Percept.* 15:17.

Mäki-Marttunen, V., Hagen, T., and Espeseth, T. (2019). Task context load induces reactive cognitive control: an fMRI study on cortical and brain stem activity. *Cogn. Affec. Behav. Neurosci.* 19, 945–965. doi: 10.3758/s13415-019-00691-6

Maki-Marttunen, V., Hagen, T., Laeng, B., and Espeseth, T. (2020). Distinct neural mechanisms meet challenges in dynamic visual attention due to either load or object spacing. *J. Cogn. Neurosci.* 32, 65–84. doi: 10.1162/jocn_a_01469

Mazur, L. M., Mosaly, P. R., Moore, C., Comitz, E., Yu, F., Falchook, A. D., et al. (2016). Toward a better understanding of task demands, workload, and performance during physician-computer interactions. *J. Am. Med. Informat. Assoc.* 23, 1113–1120. doi: 10.1093/jamia/ocw016

Mehler, B., Reimer, B., and Coughlin, J. F. (2012). Sensitivity of physiological measures for detecting systematic variations in cognitive demand from a working memory task: an on-road study across three age groups. *Hum. Fact.* 54, 396–412. doi: 10.1177/0018720812442086

Mehler, B., Reimer, B., Coughlin, J. F., and Dusek, J. A. (2009). Impact of incremental increases in cognitive workload on physiological arousal and performance in young adult drivers. *J. Transport. Res. Board* 2138, 6–12. doi: 10.3141/2138-02

Messick, S. (1989). "Validity," in *Educational Measurement*, ed. R. L. Linn (Washington, DC: American Council on Education), 13–103.

Meteier, Q., Capallera, M., Ruffieux, S., Angelini, L., Abou Khaled, O., Mugellini, E., et al. (2021). Classification of drivers' workload using physiological signals in conditional automation. *Front. Psychol.* 12:596038. doi: 10.3389/fpsyg.2021.596038

Mulder, L. J. (1992). Measurement and analysis methods of heart rate and respiration for use in applied environments. *Biol. Psychol.* 34, 205–236. doi: 10.1016/0301-0511(92)90016-N

Nourbakhsh, N., Chen, F., Wang, Y., and Calvo, R. A. (2017). Detecting users' cognitive load by galvanic skin response with affective interference. *ACM Transact. Interact. Intell. Syst.* 2017:12.

Nourbakhsh, N., Wang, Y., Chen, F., and Calvo, R. A. (2012). "Using galvanic skin response for cognitive load measurement in arithmetic and reading tasks," in *Proceedings of the 24th Conference on Australian Computer-Human Interaction OzCHI* (New York, NY: Association for Computing Machinery).

Orquin, J. L., and Holmqvist, K. (2018). Threats to the validity of eye-movement research in psychology. *Behav. Res. Methods* 50, 1645–1656. doi: 10.3758/s13428-017-0998-z

Paas, F. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *J. Educat. Psychol.* 84, 429–434. doi: 10.1037/0022-0663.84.4.429

Paas, F., and van Merriënboer, J. J. G. (1994). Variability of worked examples and transfer of geometrical problem-solving skills: A cognitive-load approach. *J. Educat. Psychol.* 86, 122–133. doi: 10.1037/0022-0663.86.1.122

Paas, F., Tuovinen, J. E., Tabbers, H., and Van Gerven, P. W. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educat. Psychol.* 38, 63–71. doi: 10.1207/s15326985ep3801_8

Paas, F., van Merriënboer, J. J. G., and Adam, J. J. (1994). Measurement of cognitive load in instructional research. *Percept. Motor Skills* 79, 419–430. doi: 10.2466/pms.1994.79.1.419

Park, B., and Brünken, R. (2015). The rhythm method: A new method for measuring cognitive load—An experimental dual-task study. *Appl. Cogn. Psychol.* 29, 232–243. doi: 10.1002/acp.3100

Patton, M. Q. (1999). Enhancing the quality and credibility of qualitative analysis. *Health Ser. Res.* 34:1189.

Pattyn, N., Migeotte, P. F., Neyt, X., van den Nest, A., and Cluydts, R. (2010). Comparing real-life and laboratory-induced stress reactivity on cardio-respiratory parameters: differentiation of a tonic and a phasic component. *Physiol. Behav.* 101, 218–223. doi: 10.1016/j.physbeh.2010.04.037

Posada-Quintero, H. F., and Chon, K. H. (2020). Innovations in electrodermal activity data collection and signal processing: A systematic review. *Sensors* 20:479. doi: 10.3390/s20020479

Posada-Quintero, H. F., Florian, J. P., Orjuela-Canþoìn, A. D., Aljama-Corrales, T., Charleston-Villalobos, S., and Chon, K. H. (2016). Power spectral density analysis of electrodermal activity for sympathetic function sssessment. *Annal. Biomed. Engin.* 44, 3124–3135. doi: 10.1007/s10439-016-1606-6

Recarte, M. Á, Pérez, E., Conchillo, Á, and Nunes, L. M. (2008). Mental workload and visual impairment: Differences between pupil, blink, and subjective rating. *Span. J. Psychol.* 11, 374–385. doi: 10.1017/s1138741600004406

Reinerman-Jones, L., Barber, D. J., Szalma, J. L., and Hancock, P. A. (2017). Human interaction with robotic systems: performance and workload evaluations. *Ergonomics* 60, 1351–1368. doi: 10.1080/00140139.2016.1254282

Reingold, E. M., and Glaholt, M. G. (2014). Cognitive control of fixation duration in visual search: The role of extrafoveal processing. *Vis. Cogn.* 22, 610–634. doi: 10.1080/13506285.2014.881443

Rendon-Velez, E., van Leeuwen, P. M., Happee, R., Horvath, I., van der Vegte, W. F., and de Winter, J. C. F. (2016). The effects of time pressure on driver performance and physiological activity: A driving simulator study. *Transp. Res. Part F* 2016, 150–169. doi: 10.1016/j.trf.2016.06.013

Rosch, J. L., and Vogel-Walcutt, J. J. (2013). A review of eye-tracking applications as tools for training. *Cogn. Technol. Work* 15, 313–327. doi: 10.1007/s10111-012-0234-7

Ryu, K., and Myung, R. (2005). Evaluation of mental workload with a combined measure based on physiological indices during a dual task of tracking and mental arithmetic. *Int. J. Industr. Ergon.* 35, 991–1009. doi: 10.1016/j.ergon.2005.04.005

Schmidt-Weigand, F., Kohnert, A., and Glowalla, U. (2010). A closer look at split visual attention in system-and self-paced instruction in multimedia learning. *Learn. Instruct.* 20, 100–110. doi: 10.1016/j.learninstruc.2009.02.011

Schnotz, W., and Kürschner, C. (2007). A reconsideration of cognitive load theory. *Educat. Psychol. Rev.* 19, 469–508. doi: 10.1007/s10648-007-9053-4

Setz, C., Arnrich, B., Schumm, J., La Marca, R., Tröster, G., and Ehlert, U. (2009). Discriminating stress from cognitive load using a wearable EDA device. *IEEE Transact. Inform. Technol. Biomed.* 14, 410–417. doi: 10.1109/titb.2009.2036164

Siegle, G. J., Ichikawa, N., and Steinhauer, S. (2008). Blink before and after you think: Blinks occur prior to and following cognitive load indexed by pupillary responses. *Psychophysiology* 45, 679–687. doi: 10.1111/j.1469-8986.2008.00681.x

Solhjoo, S., Haigney, M. C., McBee, E., van Merrienboer, J. J. G., Schuwirth, L., Artino, A. R., et al. (2019). Heart rate and heart rate variability correlate with clinical reasoning performance and self-reported measures of cognitive load. *Scientific Rep.* 9, 1–9.

Song, H. S., and Lehrer, P. M. (2003). The effects of specific respiratory rates on heart rate and heart rate variability. *Appl. Psychophysiol. Biofeedb.* 28, 13–23.

Stern, J. A., Boyer, D., and Schroeder, D. (1994). Blink rate: a possible measure of fatigue. *Hum. Fact.* 36, 285–297. doi: 10.1177/001872089403600209

Sweller, J., Ayres, P., and Kalyuga, S. (2011). *Cognitive Load Theory*. New York, NY: Springer.

Sweller, J., van Merriënboer, J. J. G., and Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educat. Psychol. Rev.* 31, 261–292. doi: 10.1007/s10648-019-09465-5

Szulewski, A., Howes, D., van Merrienboer, J. J. G., and Sweller, J. (2020). From theory to practice: The application of cognitive load theory to the practice of medicine. *Acad. Med.* 96, 24–30. doi: 10.1097/acm.0000000000003524

Tan, C. H., Low, K. A., Schneider-Garces, N., Zimmerman, B., Fletcher, M. A., Maclin, E. L., et al. (2016). Optical measures of changes in cerebral vascular tone during voluntary breath holding and a Sternberg memory task. *Biol. Psychol.* 118, 184–194. doi: 10.1016/j.biopsycho.2016.05.008

Task Force of the European Society of Cardiology and the North American Society of Pacing Electrophysiology. (1996). Heart rate variability: standards of measurement, physiological interpretation and clinical use. *Circulation* 93, 1043–1065. doi: 10.1161/01.cir.93.5.1043

Thayer, J. F., Ahs, F., Fredrikson, M., Sollers, J. J. III, and Wager, T. D. (2012). A meta-analysis of heart rate variability and neuroimaging studies: Implications for heart rate variability as a marker of stress and health. *Neurosci. Biobehav. Rev.* 36, 747–756. doi: 10.1016/j.neubiorev.2011.11.009

Tininenko, J. R., Measelle, J. R., Ablow, J. C., and High, R. (2012). Respiratory control when measuring respiratory sinus arrhythmia during a talking task. *Biol. Psychol.* 89, 562–569. doi: 10.1016/j.biopsycho.2011.12.022

Tsai, Y.-F., Viirre, E., Strychacz, C., Chase, B., and Jung, T.-P. (2007). Task performance and eye activity: predicting behavior relating to cognitive workload. *Aviat. Space Environ. Med.* 78, B176–B185.

van Acker, B. B., Bombeke, K., Durnez, W., Parmentier, D. D., Mateus, J. C., Biondi, A., et al. (2020). Mobile pupillometry in manual assembly: A pilot study exploring the wearability and external validity of a renowned mental workload lab measure. *Int. J. Industr. Ergon.* 75:102891. doi: 10.1016/j.ergon.2019.102891

van Gerven, P. W. M., Paas, F., Van Merriënboer, J. J. G., and Schmidt, H. G. (2000). Cognitive load theory and the acquisition of complex cognitive skills in the elderly: Towards an integrative framework. *Educat. Gerontol.* 26, 503–521. doi: 10.1080/03601270050133874

van Gog, T., and Paas, F. (2008). Instructional efficiency: Revisiting the original construct in educational research. *Educat. Psychol.* 43, 16–26. doi: 10.1080/00461520701756248

van Meeuwen, L. W., Jarodzka, H., Brand-Gruwel, S., Kirschner, P. A., de Bock, J. J., and van Merriënboer, J. J. (2014). Identification of effective visual problem solving strategies in a complex visual domain. *Learn. Instruct.* 32, 10–21. doi: 10.1016/j.learninstruc.2014.01.004

Van Orden, K. F., Jung, T.-P., and Makeig, S. (2000). Combined eye activity measures accurately estimate changes in sustained visual task performance. *Biol. Psychol.* 52, 221–240. doi: 10.1016/s0301-0511(99)00043-5

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

Vanneste, P., Raes, A., Morton, J., Bombeke, K., Van Acker, B. B., Larmuseau, C., et al. (2020). Towards measuring cognitive load through multimodal physiological data. *Cogn. Technol. Work* 23, 567–585. doi: 10.1007/s10111-020-00641-0

Vera, J., Diaz-Piedra, C., Jiménez, R., Sanchez-Carrion, J. M., and Di Stasi, L. L. (2019). Intraocular pressure increases after complex simulated surgical procedures in residents: an experimental study. *Surg. Endosc.* 33, 216–224. doi: 10.1007/s00464-018-6297-7

Vlemincx, E., Taelman, J., De Peuter, S., Van Diest, I., and Van Den Bergh, O. (2011). Sigh rate and respiratory variability during mental load and sustained attention. *Psychophysiology* 48, 117–120. doi: 10.1111/j.1469-8986.2010.01043.x

Wang, S., Gwizdka, J., and Chaovalitwongse, W. A. (2016). Using Wireless EEG Signals to Assess Memory Workload in the n-Back Task. *IEEE Transact. Hum. Mach. Syst.* 46, 424–435. doi: 10.1109/thms.2015.2476818

Wang, Y., Reimer, B., Dobres, J., and Mehler, B. (2014). The sensitivity of different methodologies for characterizing drivers' gaze concentration under increased cognitive demand. *Transp. Res. Part F* 26, 227–237. doi: 10.1016/j.trf.2014.08.003

Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoret. Issues Ergon. Sci.* 3, 159–177. doi: 10.1080/14639220210123806

Wientjes, C. J., Grossman, P., and Gaillard, A. W. (1998). Influence of drive and timing mechanisms on breathing pattern and ventilation during mental task performance. *Biol. Psychol.* 49, 53–70. doi: 10.1016/s0301-0511(98)00026-x

Wilbanks, B. A., and McMullan, S. P. (2018). A review of measuring the cognitive workload of electronic health records. *CIN* 36, 579–588. doi: 10.1097/cin.0000000000000469

Wong, H. K., and Epps, J. (2016). Pupillary transient responses to within-task cognitive load variation. *Comput. Methods Progr. Biomed.* 137, 47–63. doi: 10.1016/j.cmpb.2016.08.017

Wu, Y. B., Miwa, T., and Uchida, M. (2017). Using physiological signals to measure operator's mental workload in shipping - an engine room simulator study. *J. Mar. Engin. Technol.* 16, 61–69. doi: 10.1080/20464177.2016.1275496

Yan, S., Tran, C. C., Chen, Y., Tan, K., and Habiyaremye, J.-L. (2017). Effect of user interface layout on the operators' mental workload in emergency operating procedures in nuclear power plants.Automation in Construction. *Nucl. Engin. Design* 82, 179–192. doi: 10.1016/j.nucengdes.2017.07.012

Zakeri, Z., Mansfield, N., Sunderland, C., and Omurtag, A. (2020). Physiological correlates of cogntive load in laparoscopic surgery. *Scient. Rep.* 10:12927. doi: 10.1038/s41598-020-69553-3

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Education

ORIGINAL RESEARCH
published: 20 September 2021
doi: 10.3389/feduc.2021.702616

Measuring Cognitive Load: Are There More Valid Alternatives to Likert Rating Scales?

Kim Ouwehand[1], Avalon van der Kroef[1], Jacqueline Wong[2] and Fred Paas[1,2]*

[1]Department of Psychology, Education, and Child Studies, Erasmus University Rotterdam, Rotterdam, Netherlands, [2]Delft Institute of Applied Mathematics, Delft University of Technology, Delft, Netherlands, [3]School of Education/Early Start, University of Wollongong, Wollongong, NSW, Australia

Cognitive load researchers have used varying subjective techniques based on rating scales to quantify experienced cognitive load. Although it is generally assumed that subjects can introspect on their cognitive processes and have no difficulty in assigning numerical values to the imposed cognitive load, little is known about how visual characteristics of the rating scales influence the validity of the cognitive load measure. In this study we look at validity of four subjective rating scales (within groups) differing in visual appearance by participants rating perceived difficulty and invested mental effort in response to working on simple and complex weekday problems. We used two numerical scales (the nine-point Likert scale most often used in Cognitive load theory research and a Visual Analogue Scale ranging between 0–100%) and two pictorial scales (a scale consisting of emoticons ranging from a relaxed blue-colored face to a stressed red-colored face and an "embodied" scale picturing nine depicted weights from 1–9 kg). Results suggest that numerical scales better reflect cognitive processes underlying complex problem solving while pictorial scales Underlying simple problem solving. This study adds to the discussion on the challenges to quantify cognitive load through various measurement methods and whether subtleties in measurements could influence research findings.

Keywords: cognitive load, measurement methodology, subjective rating scales, visualization, problem solving

OPEN ACCESS

Edited by:
Lu Wang,
University of Georgia, United States

Reviewed by:
Petar Radanliev,
University of Oxford, United Kingdom
Savio W. H. Wong,
The Chinese University of Hong Kong, China

****Correspondence:***
Kim Ouwehand
ouwehand@essb.eur.nl

Specialty section:
This article was submitted to Educational Psychology, a section of the journal Frontiers in Education

Received: *29 April 2021*
Accepted: *02 September 2021*
Published: *20 September 2021*

Citation:
Ouwehand K, Kroef Avd, Wong J and Paas F (2021) Measuring Cognitive Load: Are There More Valid Alternatives to Likert Rating Scales?. Front. Educ. 6:702616. doi: 10.3389/feduc.2021.702616

INTRODUCTION

Cognitive load theory (CLT) centralizes the characteristics of human cognitive architecture, and especially the limitations of working memory in time and capacity (Baddeley, 1992, 2000), as a prerequisite for the optimization of learning. Cognitive-load researchers focus on instructional methods that can be used to manage working memory load (i.e., cognitive load). Cognitive load has been conceptualized as a multidimensional construct consisting of three types of cognitive load (e.g. Sweller, 2010), namely 1) intrinsic load that is imposed by the learning task itself, 2) extraneous load that is imposed by the design of the instruction, and 3) germane load that is related to the amount of cognitive resources that learners have available for learning. All three types of load have been proposed to be influenced by element interactivity (Sweller 2010); how many separate parts of information need to be integrated for learning to occur. During learning, initially separate information elements are categorized, organized and chunked into schemata in long-term memory, which after construction can be treated as one information element in working

memory (Sweller, et al., 1998, 2019; van Merriënboer & Sweller, 2005). This process is called schematization and is a core mechanism underlying successful learning in CLT. When trying to successfully learn materials with high element interactivity, it is proposed that more mental effort needs to be invested by the learner than for materials with low element interactivity. Therefore, having a valid indication of cognitive load experienced/spent during a specific task or activity could provide crucial information on the development of a learning process and quality of an instruction.

The most widely used measures of cognitive load are subjective measures based on ratings of perceived mental effort and task difficulty (Paas, et al., 2003; Sweller, et al., 2019). There are two main assumptions underlying subjective measures of cognitive load. Firstly, it is assumed that all learners have similar clear understanding of what is meant by "invested mental effort" and "difficulty of a task". Secondly, all learners are assumed to possess the metacognitive ability to monitor how much mental effort they have invested. Based on these assumptions, this common understanding or knowledge of the terms "invested mental effort" and "task difficulty" as well as the accuracy of individuals' monitoring skills are not tested or controlled for when using the rating scales. Therefore, the reliability and validity of such subjective measures are debatable (e.g., Ayres, 2018). One of the issues that can arise with the cognitive-load rating scale concerns the way the scale is represented. We suggest that whether the scale represents symbols or numbers might affect the mental effort and task difficulty ratings, for examplr by imposing additional (extraneous) cognitive load. To investigate the effect of the symbolic/numerical representation of cognitive load in the rating scales on ratings of mental effort and task difficulty, we identified three alternatives to the original 9-point Likert rating classic 9-point scale in. One of these alternatives is also a symbolic representation, the second is a more affective one, representing the emotional aspect of effort and task difficulty, and the third is a more embodied one representing effort and task difficulty as weight. The central focus in the study is on construct validity; do the ratings (i.e., scores) on the measurement scale reflect the construct we intend to measure and are there differences between the scales?

One of the first subjective measures of cognitive load was developed by Paas (1992). In this study, learners were asked to indicate on a 9-point Likert scale "how much mental effort they have invested in a task", ranging from 1 (very, very low mental effort) to 9 (very, very high mental effort). Subjective measures are advantageous for cognitive load research since they do not require a complicated experimental set-up and can be easily implemented and used multiple times in most research designs (Sweller and Paas, 2017). However, subjective measures of cognitive load have faced criticism, mainly for being implemented in research in an inconsistent way (Sweller et al., 2011). One of the inconsistencies concerns the verbal labels used to assess cognitive load. For example, instead of *mental effort,* learners were asked to rate the *difficulty of the task* by indicating on a 7- or 9-point scale "how difficult or easy the learning task was for them", ranging from 1 (very, very easy) to 7 or 9 (very, very difficult). Studies have shown that subjective task difficulty ratings, like mental effort ratings, varied according to the level of element interactivity of a task (e.g., Ayres, 2006; Ouwehand et al., 2014). However, research suggested that the two verbal labels (i.e., *mental effort* and *task difficulty*) measure different aspects of the cognitive load De Leeuw and Mayer (2008). More specifically, (De Leeuw and Mayer, 2008), found that task difficulty ratings were related to intrinsic load and perceived mental effort to germane load, indicating the way verbal labels are being phrased in the rating scales can influence the measurement of cognitive load. Another inconsistency is the timing and frequency of measurement. Research showed that perceived mental effort and task difficulty were significantly higher when measured at the end of the learning phase (i.e., delayed) than when taking the average of the ratings obtained after each learning task (i.e., immediate) (van Gog et al., 2012; Schmeck et al., 2015).

Despite the variations in the way cognitive load has been subjectively measured in research (i.e., verbal labels and timing of measurement), both mental effort and task difficulty have been found to reliably reflect differences in the complexity of the instructional design in numerous studies (e.g., Hadie and Yusoff, 2016; Ouwehand et al., 2014; for an overview see; Paas et al., 2003). While previous research has focused on the type of measurement (e.g. physiological measurement and self-reports) and differences in the timing and verbal labels, little is known about whether the way in which the Likert scales are formatted influences the measurement of cognitive load. Sung and Wu (2018) argued that there are several issues inherent to the design of Likert scales, particularly the ambiguous numbers of the response categories and the response style underlying the ordinal measurement of data. Therefore, the aim of the current study was to explore the validity of alternative representations of Likert rating scales to measure subjective cognitive load.

The measurement validity was examined by the relationship between the subjective measures (i.e., mental effort and perceived difficulty) and the performance measures (i.e., accuracy and time on task) for simple and complex problems. Three alternative representation formats (i.e., Visual Analogue Scale, affective, and embodied) were investigated and compared with the original 9-point Likert scale for measuring cognitive load (Paas, 1992).

The first type of visual representation employed in this study was a Visual Analogue Scale (VAS). The VAS presents the numbers on a line continuum and participants can move a bar (or pin a point) between 0 and 100% to determine their level of cognitive load. Therefore, a VAS transforms ordinal-level measurement data from the discrete response categories in a Likert scale to continuous and interval-level measurement data (Sung and Wu, 2018). Research indicated that the VAS has a high test-retest reliability and a small measurement error (e.g., Alghadir et al., 2018). In addition, the VAS is a well-known measurement scale in the domain of judgments of learning (JoL) in which learners have to predict their future performance by indicating on a VAS how likely they think they will remember a just learned item on a future test (e.g., Rhodes, 2016). Recent research in the field of educational psychology called for an integration between cognitive load and self-regulated learning theories to better understand the dynamic relations between

cognitive resources available for managing one's own learning process and for the learning process itself. One of the challenges to the integration of theories of cognitive load and self-regulated learning is in measurement (Sweller and Paas, 2017). Therefore, findings from the current study can potentially pave the way for future research to determine whether VAS can be used as a common scale to measure concepts in cognitive load (i.e., mental effort and task difficulty) and self-regulated learning (i.e., JoL).

Besides the well-known and widely used original cognitive load scale and the VAS, we were interested in examining visual characteristics that display internal processes (i.e., mental and affective states). Numerical representations, which are characteristic of the original scale and the VAS, are a rather abstract reflection of internal processes. Given that grounding mental representations can support understanding (Barsalou, 2008, 2016), it is possible that a scale that reflects internal processes used when working on a task will improve the validity of the scale. Therefore, the first question of interest is whether a better reflection of internal processes could increase the validity of the rating scale. To this end, we designed two pictorial scales as a reflection of internal processes: an affective scale with icons to represent a range of emotions (i.e., emoticons) and an embodied scale with pictures of weights to represent a range of physical load.

Although scales with affective stimuli are frequently used in the media and medical practice (e.g., satisfaction reviews on products or services or pain rating scales), literature on the affect in learning and subjective rating in CLT research seems scarce. Interestingly, in one of the earliest lines of research on learning, using *operant conditioning* (e.g. Skinner, 1963), affect plays a central role in the learning process. According to the operant conditioning theory we (humans, but also other animals) learn from pleasant or unpleasant consequences of our actions. Put simply, actions with pleasant outcomes tend to be repeated and actions with unpleasant outcomes avoided. Since then, a lot of support for the operant learning theory has been gathered (for reviews see, Gordan and Amutan, 2014; Staddon and Cerutti, 2003). Mechanisms for the role of affect in learning has been extended by neuropsychological evidence of a reward circuit in the brain in which more primitive brain areas dealing with emotions highly interact with more recent brain areas more involved in higher-order cognitive processes such as executive processing (for a review, see O'Doherty et al., 2017). In a recent review by Shenhav et al. (2017), cognitive load (which these authors refer to as mental effort) is approached from an affective perspective by looking at a costs/benefits ratio of invested cognitive load. Because humans are limited in their resource capacity, they need to be efficient in their allocation of cognitive resources. In their review, these authors argue that investing (high) mental effort is a negative experience in terms of affect. As a consequence, a task requiring high mental effort would be experienced with more negative affect than a task requiring low mental effort. Following this perspective, affect can be a direct reflection of cognitive load. In line with this view, Sitzmann et al. (2010) showed that people are better at self-assessing affective processes (i.e., motivation and satisfaction) than purely cognitive processes. This suggests that human learners are more capable of defining emotional processes than purely cognitive processes. Based on these reasonings, we propose that a subjective rating scale depicting affect from negative (i.e. aversive) to positive, might represent the experience of mental effort of learners better than a more abstract numerical scale, and therefore, might be a more valid manner to measure invested mental effort and perceived task difficulty.

A second aspect we would like to explore is inspired by the embodied cognition theory. This theory states that cognition is grounded in perception and action (for a recent theoretical overview, see Barsalou, 2016). In other words, this theory claims that our cognitive processes and functions are shaped by the way we interact with our surroundings. Since the nineties a lot of evidence has been gathered, suggesting that for a substantial part, our cognition is tightly bound to how we perceive and interact in the world (Barsalou, 2008). However, there is an ongoing debate on whether embodied cognition only applies to the lower level cognitive abilities (i.e. procedural, motor learning) or also to more higher level cognitive abilities (i.e. conceptual learning) (for a critical review, see Caramazza et al., 2014). For the present research, the embodied view is still interesting, because the term mental or cognitive ***load***, metaphorically indicates that a certain weight is related to the task (i.e., how "heavy" or "burdensome" a task is). Indeed the analogy of physical weight for mental effort is also used by Shenhav et al. (2017), to explain how effort mediates between capacity and performance. Interestingly, abstract metaphors such as the 'heaviness' of a task in our study are also empirically found to be connected to embodied cognition. For example, Zanolie et al. (2012) found an attentional bias for abstract concepts such as "power" on a vertical axis. In their experiment, participants were presented with a power-related word (either related to high or low power) after which they had to identify objects either presented on the top or bottom of a computer screen. It was found that target identification was faster for items that were presented on a semantically congruent location compared to an incongruent location. This result was explained by the process of mental simulation: humans tend to imagine perceptual and motoric features evoked by a stimulus in such a way that a single stimulus can elicit a rich image and or action plan for a situation in which the stimulus is normally encountered. Drawing further on these findings that metaphors can also facilitate cognitive processing, we added a scale depicting nine weights ranging from light (small 1 kg) to heavy (large 9 kg). In this way, we expressed the idea of mental "load" in a more concrete manner. For instance, a heavy problem or task (load) would correspond to a heavier weight. This might fit the experiences of mental effort and task difficulty (i.e., load on cognition) better than a more abstract numeric scale.

To investigate the construct validity of the different scales, we adopted the dominant approach used in cognitive load research; we used the relation between performance/learning and cognitive load ratings (for a meta-analysis see, Naismith and Cavalcanti, 2015). More specifically, we inspected correlation analyses between cognitive load ratings measured by the mental effort and task difficulty ratings on each of the four different scales (the original nine-point Likert scale, the VAS, an affective scale illustrated with emoticons, and an embodied scale illustrated with weights) with performance measured by accuracy of problem-solving and time on task for simple and complex problems. Although the current research is exploratory in nature, we would like to put forward some hypotheses.

First, based on cognitive load studies that showed that subjective measures are a valid way of measuring cognitive load, we hypothesized that perceived mental effort and difficulty would be rated higher for the complex problems than for the simple problem across all four measurement scales. Secondly, we explored whether there are differences between the different scales in differentiating effort and difficulty between simple and complex problem solving. Building on literature stating that affect is tightly related to the learning process (Shenhav et al., 2017), we suggest that emoticons, representing affect might reflect the perceived mental effort and difficulty better than a numeric scale and therefore correlate higher and more significantly to performance. Also, the relation is expected to be stronger for the complex problems, since these might be more arousing and frustrating. In support of the embodied cognition theory (e.g., Barsalou, 2008), we expect that pictures of increasing weights might represent a more concrete picture of "load" and therefore might represent perceived mental effort and difficulty better than a numeric scale. Following the argument that embodied cognition might be more related to more lower level cognitive abilities, it is expected that ratings on this scale correlate higher and (more) significantly with the performance on the simple problems. Finally, to gather insight into participants' experience when using the different scales to rate their cognitive load, participants were asked to vote for their most and least favorite scale and give some reasoning on their (dis)likes.

MATERIALS AND METHODS

The present study was conducted in accordance with the guidelines of the ethical committee of the host University. Below we describe our sample, design, materials, procedure and analysis plan.

Participants

Participants were 46 healthy young adults (Psychology students, 39 women; M_{age} = 22.4 years, SD = 2.45) who participated in this study as part of a course requirement. All participants gave written consent before participation. A power test using G*power3.1 software (Faul et al., 2009) showed that for the 2 (complexity) x 4 (scale type) within-subject design we use when aiming for an effect size of f = 0.25, power = 0.95 and α = 0.05, a sample with minimal 23 participants is required.

Design

In a 2 (Complexity: simple and complex questions) x 4 (Scale Type: original nine point, visual analogue, affect, and embodied) within-subjects design, participants were presented with simple and complex problems that they had to solve. Four types of rating scales were used to measure perceived difficulty and mental effort.

MATERIALS

Problem-solving task. Sixteen weekday problems were used (Sweller, 1993; see also; van Gog et al., 2012) of which eight were low in element interactivity (simple problems) and eight high (complex problems). **Table 1** shows an example of a simple and a complex problem in the problem-solving task. The simple problems consisted of two elements that students needed to consider when solving the problem while the complex problems consisted of five elements (i.e., elements are underlined in **Table 1**). The complex problems would be harder to solve than the simple problems because of the higher element interactivity.

Rating scales. Cognitive load was measured by two self-report items, one item was to assess perceived difficulty *"Indicate on this scale how difficult you found the problem*" and the other was to assess invested mental effort *"Indicate on this scale how much mental effort it cost you to solve the problem*? Four types of rating scales were investigated in this study. The first type of rating scale was the original 9-point Likert rating scale of Paas (1992) ranging from 1 (very, very easy) to 9 (very, very difficult). The second type of rating scale was a Visual Analogue Scale (VAS) in which participants could move a bar on a line representing a 0–100% continuum. The third type of rating scale, also termed as the affective scale, was a self-designed 9-point scale made up of emoticons ranging from a blue emoticon depicting a relaxed expression to a red emoticon depicting a stressed/aroused expression. The fourth type of rating scale, also termed as the embodied scale, was a self-designed 9-point scale presenting pictures of increasing weights ranging from 1–9 kg **Figure 1** illustrates the four types of rating scales.

Procedure

The experiment was constructed using an online survey platform, Qualtrics (https://www.qualtrics.com/). A link to access the experiment was shared with the participants. **Figure 2** illustrates the procedure of the experiment. At the start of the experiment, participants were instructed to not use paper and pencil or other external tools for the problem-solving task. They were also informed that there was a time limit to solve the problems. After this instruction, they were given a practice problem to gain familiarity with the task before proceeding to the first block of questions.

Altogether, participants had to solve four blocks of questions. In each block, participants had to solve four problems comprising two simple and two complex problems. After each problem, participants had to rate their rate mental effort and perceived difficulty (i.e., 16 times in total). The rating scale presented to the participants varied in each block (original 9-point Likert scale, VAS, emoticon, weights). The four blocks were counterbalanced so that the order in which the different type of rating scales that were presented to the participants were not the same. At the end of the experiment, participants were asked to indicate the rating scale that they liked best and the one that they liked the least out of the four types of rating scales.

Scoring of the dependent variables

Performance. Performance was measured by accuracy scores and time on task. For each correctly solved problem, one point was assigned. The mean accuracy was determined for each difficulty level (simple and complex) within each scale type (original 9-point

TABLE 1 | Example of a simple and complex problem used in the problem-solving task.

Problem type	Question
Simple	If today is Friday, which day of the week is it in 2 days?[a]
Complex	What day is 2 days after the day after tomorrow if the day before yesterday was 7 days after Wednesday?[a]

[a]*Note Number of elements needed to solve the problems are underlined.*

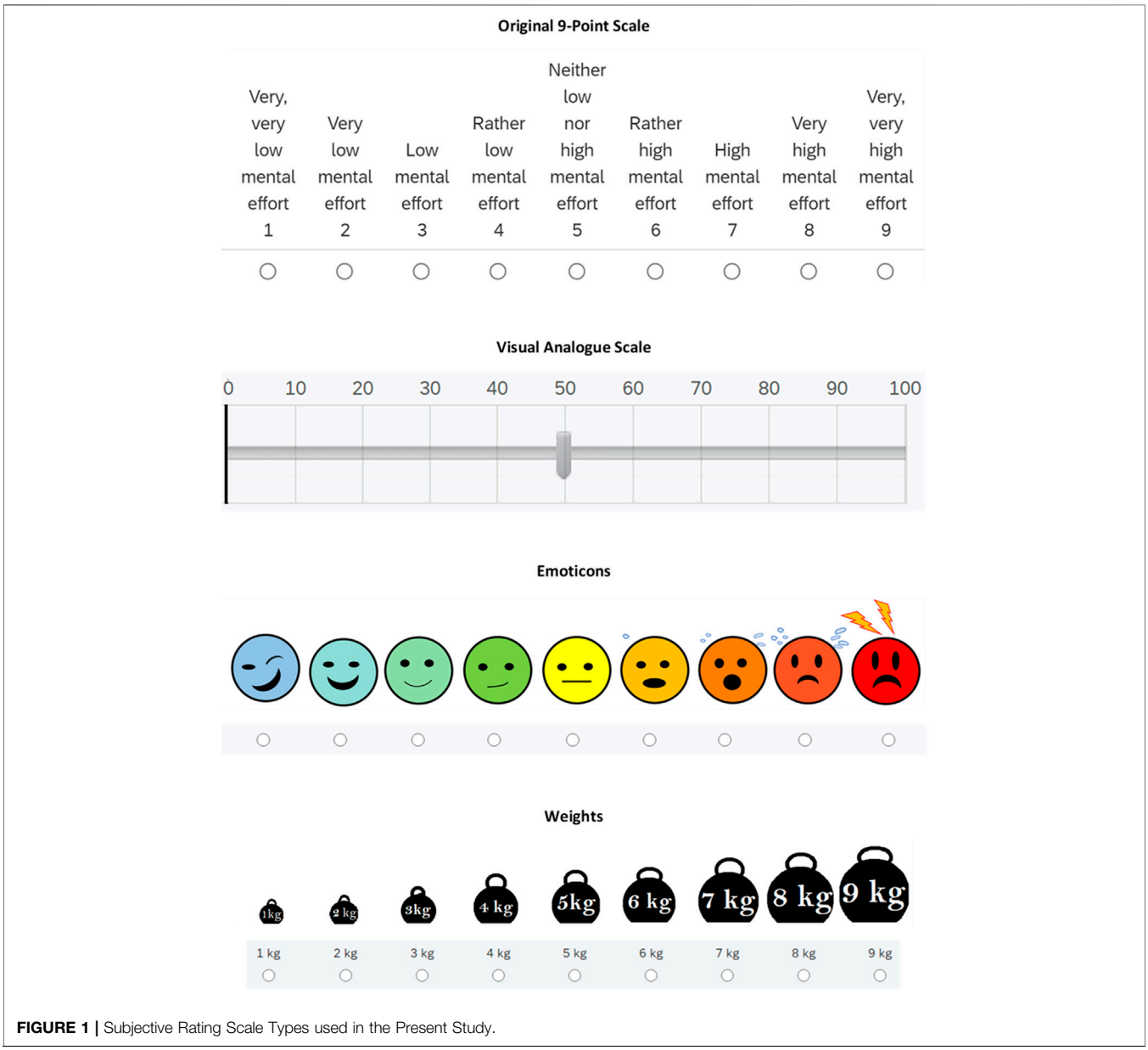

FIGURE 1 | Subjective Rating Scale Types used in the Present Study.

rating scale, VAS, emoticons and depicted weights), resulting in two mean accuracy scores for each rating scale type (i.e., one for complex problems and one for simple problems). Altogether, eight mean accuracy scores were calculated for each participant. Time on task was determined by the duration (in seconds) participants took to submit their response.

Rating (invested mental effort and perceived difficulty). To make all rating scores comparable, proportion scores were calculated by dividing the obtained scores by the maximum scores: For the original 9-point rating scale, the emoticons and depicted weights (which also had nine alternatives), proportion scores were calculated by dividing the mean ratings per scale

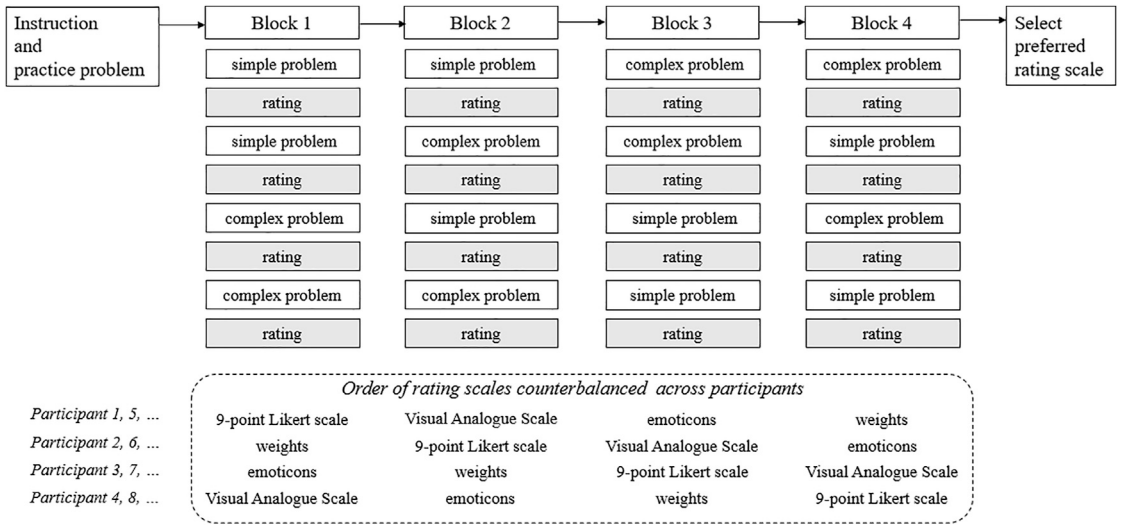

FIGURE 2 | Overview of the procedure of the experiment.

category and complexity level by 9. For the VAS, the mean percentage score was divided by 100. In this way all scores had a range between 0 and 1.

Data Analysis

Quantitative Data

Firstly, 2 (complexity; simple vs complex) x 4 (scale type; original, VAS, Emoticons, Weights) repeated measures ANOVAs were conducted for accuracy, time on task, perceived difficulty and perceived invested mental effort within subjects. Secondly, correlations were calculated between performance measures (accuracy and time on task) and the subjective ratings (perceived difficulty and mental effort) for each type of scale. To inspect validity, we tested whether significant correlations between performance measures (accuracy and time on task) and subjective ratings (perceived difficulty and mental effort) were stronger for some scales than others, significant correlations were compared by a calculation tool called cocor (Diedenhofen & Musch, 2015). This tool was also used to compare significant correlations for simple problems or complex problems per scale, to find out whether ratings on a specific type of scale was more representative of cognitive load during simple or complex problem solving. A significance level of 0.05 was used for the main analyses. On follow-up analyses a Bonferroni correction was applied. Partial eta-squared (η_p^2) was calculated as a measure of effect size for *F*-values, with values of 0.01, 0.06, and 0.14, characterizing small, medium, and large effect sizes, respectively (Cohen, 1988). Cohen's *d* was calculated as a measure of effect size for *t*-values, with values of 0.20, 0.50, and 0.80, characterizing small, medium, and large effect sizes, respectively (Cohen, 1988).

Qualitative Data

Finally, participants' indication of the type of rating scale that they liked and disliked the most was analyzed. As we had no clear expectations of the qualitative data and we wanted to explore the open-ended responses to find reasoning behind preferences for the scales, an inductive approach was taken to code these responses. This approach is appropriate for exploratory purposes in the absence of clear theory-driven hypotheses on how the data would look like (Linneberg & Korsgaard, 2019). Inductive coding is data-driven in that the responses are categorized based on the content of the responses.

Two raters independently rated and categorized the participants' responses freely (no categories were outlined before). Initially, the first rater decided on four categories (clarity, relatability, appearance, and nuance) but the second rater categorized the data in three categories (clarity, relatability, appearance). After inspection of the ratings, it seemed that the second rater initially did not distinguish between clarity and nuance. After discussion, both raters one agreed on using the four categories specified by the first rater. These answering categories were found by summarizing each given answer into keywords. The keywords were then compared to one another, and analogous keywords were combined into one category. The category "clarity" related to comments on the comprehensibility of the answering options; how unambiguous the answering options were and how easy the scale was found to be. An example of an argument classified as a remark on clarity is: 'It is most clear as to what each answer means'. The relatability category encompassed comments on how well the participants were able to relate their feelings to the scale that was used. It could also be said that this category scored answers on how intuitive the scale was found to be. A comment marked as relatable, for example, would be: 'Because it was a better illustration of how I felt'. In addition, especially the dislike comments contained a lot of remarks on how well a scale related to the questions. For example: 'Incongruent with what the scale is asking'. These kinds of comments were also scored under relatability. The third category, appearance, was a recurring aspect for a lot of the scales. Comments in this category related to the aesthetic qualities of each scale. For example: 'This scale was most

TABLE 2 | Overview of Participants' best and least Preference and Coded Explanations for the Four Types of Rating Scales.

Scales	No likes	Coded explanations for the likes				No dislikes	Coded explanations fort he dislikes				Total of the code counts	
		C	R	A	N		C	R	A	N	+	–
original scale	16	13	2	1	1	9	6	2	—	1	17	9
VAS	14	6	4	—	9	8	5	2	—	1	19	8
Emoticons	13	4	7	5	—	3	4	2	1	—	16	7
Weights	3	1	2	—	—	26	7	17	5	2	3	31

Note. *No stands for the number of participants that chose a particular scale they liked best and least. Explanations are coded under C = Clarity, R = Relatability, A = Appearance and N = Nuance. The total of the code counts represents the total number of positive and negative remarks per scale type. An explanation could fall under more than one code, therefore the No does not have to correspond to the total of code counts.*

visually pleasing'. Nuance, the final category, could be seen as the opposite of clarity, the first category. However, both were mentioned quite often as either a good or a bad quality of the scale. In addition, for both clarity and nuance a lot of arguments were given on why exactly this was a good quality of the scale. Whereas some participants praised a scale for its clearly defined and unambiguous answering options, other participants appreciated a scale for its grey areas and less well-defined answering options. An example of a nuance-category comment was: 'Because the rating is not fixed, it can give flexibility to how one perceives the task'. Subsequently, every comment was classified under one of those four categories. Some elaborations were scored under more than one category, so one elaboration could be scored more than once. For example: 'It is very ugly and not very meaningful', was scored under appearance as well as relatability.

Supplementary Appendix A shows all responses given for the scales participants indicated to like best and **B** for those they indicated to like least. In the final columns it is shown how all comments were categorized by the raters and **Table 2** shows the final categorization the raters agreed upon. In addition, this table shows how many of those comments related to a scale preference, and how many related to a disliking of the scale.

RESULTS

For accuracy, results showed a main effect of complexity, F (45, 1) = 218.75, $p < 0.001$, η_p^2 =0 .83, but not for scale, F (45, 1) = 2.2, p = 0.091, η_p^2 =0 .05, and an interaction effect, F (135, 3) = 5.41, p = 0.002, η_p^2 =0 .83. For time on task, results showed a main effect of complexity, F (45, 1) = 356.34, $p < 0.001$, η_p^2 =0 .89, and scale, F (45, 1) = 13.81, $p < 0.001$, η_p^2 =0 .24, and an interaction effect, F (135, 3) = 13.06, $p < 0.001$, η_p^2 =0 .23. For perceived difficulty, results showed a main effect of complexity, F (45, 1) = 1,306.63, $p <$ 0.001, η_p^2 =0 .97, and scale, F (45, 1) = 36.45, $p < 0.001$, η_p^2 =0 .45, and an interaction effect, F (135, 3) = 13.64, $p < 0.001$, η_p^2 =0 .23. Finally, for perceived mental effort, results showed a main effect of complexity, F (45, 1) = 1,097.30, $p < 0.001$, η_p^2 = 0.96, and scale, F (45, 1) = 30.97, $p < 0.001$, η_p^2 = 0.41, and an interaction effect, F (135, 3) = 9.50, $p < 0.001$, η_p^2 = 0.17. All means and standard deviations of the accuracy, time on task, effort and difficulty ratings are presented in **Table 3**.

Following up on the interaction effects between complexity and scale that was found for all dependent variables, we compared difference scores, i.e. instead of using a repeated measure for the performance and ratings on the simple and complex problems, we looked at Δ simple - complex. By subtracting the complex performance (accuracy and time) and rating (perceived mental effort and difficulty) scores from the simple ones, we obtained one variable for the size of the effect (instead of two) which allows for a direct comparison between effect sizes. Six paired t-tests were done on these difference scores between Original and VAS (pair 1), Original-Emoticons (pair 2), Original-Weights (pair 3), VAS-Emoticons (pair 4), VAS-Weights (pair 5), and Emoticons-Weights (pair 6). Bonferroni correction was applied by adjusting the significance level to 0.05/6 = 0.008.

For readability of the text we put all statistics in **Table 4** and report only the significant results in text. It was found that for accuracy the complexity effect was smaller for problems rated with the original scale than those with the weight scale, and smaller for problems rated with the emoticons than the weights scale. For time on task similar results were found with an additional finding that the effect of complexity was smaller for problems rated with the VAS than those with the Weight scale. For both perceived difficulty and perceived mental effort; the effect of complexity was significantly smaller using the original scale compared to the VAS or the original scale compared to the Weights scale. Also the effect of complexity was smaller when using the Emoticons compared to the Weights.

Next, correlations were calculated between performance measures (accuracy and time on task) and the ratings (difficulty and mental effort) for each scale type to examine validity of the four rating scales. For readability purposes, we present all correlations (values and significance levels) for the correlational analyses in **Table 5** and report the significant ones in text. First, for all four scale types and problem complexity levels, mental effort and perceived difficulty effort were positively correlated. For analysis on the original 9-point scale, it was found that for the complex problems (but not for the simple problems), accuracy was negatively correlated with perceived mental effort and perceived difficulty and time on task was positively correlated with perceived difficulty and invested mental effort. The analysis on the VAS showed that for complex problems, accuracy was negatively correlated with perceived mental effort and time on task was positively correlated with perceived mental effort and perceived difficulty. For the simple problems, accuracy was negatively correlated with time on task. For the emoticon scale a positive correlation between time on task and perceived difficulty was found that for the complex problems. For the simple problems,

TABLE 3 | Proportion mean scores and standard deviations of the four rating scales for accuracy, time on task, and perceived difficulty and invested mental effort.

	Accuracy (proportion)		Time on task (sec)		Perceived difficulty		Mental effort	
	Simple M (SD)	Complex M (SD)	Simple M (SD)	Complex M (SD)	Simple M (SD)	Complex M (SD)	Simple M (SD)	Complex M (SD)
Original 9	0.95 (0.16)	0.51 (0.41)	9.30 (3.93)	48.94 (19.04)	0.19 (0.09)	0.72 (0.15)	0.21 (0.12)	0.73 (0.14)
VAS	0.99 (0.07)	0.39 (0.41)	8.55 (2.86)	43.84 (17.98)	0.05 (0.05)	0.68 (0.19)	0.07 (0.06)	0.68 (0.19)
Emoticons	0.96 (0.14)	0.43 (0.37)	10.81 (3.88)	53.15 (23.48)	0.19 (0.07)	0.77 (0.16)	0.21 (0.10)	0.77 (0.16)
Weights	0.99 (0.07)	0.28 (0.33)	9.69 (3.47)	70.77 (36.41)	0.16 (0.07)	0.87 (0.14)	0.18 (0.10)	0.85 (0.13)

TABLE 4 | Statistics of the Paired t-tests Comparing Differences in Accuracy, Time on Task, Perceived Difficulty and Mental Effort Between Complex and Simple Problems Across the Four Types of Rating Scales.

Pairs	Accuracy			Time on task			Perceived difficulty			Mental effort		
	t	p	d	t	p	d	t	p	d	t	p	d
1: O vs V	−2.40	0.020	0.46	1.29	0.205	0.23	**−3.29**[a]	**0.002**[a]	**0.20**[a]	**−2.82**[a]	**0.007**[a]	**0.21**[a]
2: O vs E	−1.05	0.299	0.56	−0.93	0.355	0.20	−1.36	0.182	0.16	−1.57	0.123	0.16
3: O vs W	**−3.67**[a]	**0.001**[a]	**0.50**[a]	**−3.86**[a]	**0.000**[a]	**0.38**[a]	**−6.64**[a]	**0.000**[a]	**0.18**[a]	**−5.44**[a]	**0.000**[a]	**0.19**[a]
4: V vs E	1.12	0.267	0.46	−2.11	0.041	0.23	2.07	0.044	0.22	1.47	0.149	0.23
5: V vs W	−1.75	0.086	0.42	**−5.20**	**0.000**	**0.34**	−2.37	0.022	0.22	−1.77	0.083	0.26
6: E. vs W	**-2.85**[a]	**0.007**[a]	**0.44**[a]	**−3.29**[a]	**0.002**[a]	**0.39**[a]	**−5.00**[a]	**< 0.001**[a]	**0.20**[a]	**−4.33**[a]	**<0.001**[a]	**0.19**[a]

Note.
O = Original nine point Likert Scale, V = VAS, E = Emoticons, W = Weights.
Bonferroni correction was applied by adjusting the significance level to 0.05/6 = 0.008.
[a]*p < .05 are printed boldly.*

TABLE 5 | Correlation table of the performance (accuracy and time on task) and subjective ratings (perceived difficulty and mental effort) of the simple and complex problems.

N **= 46**		Simple Problems			Complex Problems			Average All Problems	Difficulty	Mental Effort
		Time	Difficulty	Mental Effort	Time	Difficulty	Mental Effort	Time		
Accuracy	Original	−0.06	0.01	0.06	−0.13	**−0.37***	**−0.30***	−0.17	**−0.34***	−0.26
	VAS	**−0.45****	0.06	−0.17	−0.12	−0.20	**−0.31***	−0.15	−0.21	**−0.31***
	Emoticons	−0.01	−0.03	<0.01	−0.14	−0.28	−0.20	−0.09	−0.29	−0.26
	Weights	−0.14	**−0.57****	**−0.42****	0.05	0.18	0.10	0.05	0.06	−0.05
Time	Original		0.23	0.10		**0.35***	**0.29***		**0.30***	0.19
	VAS		0.02	0.24		**0.35***	**0.37***		**0.32***	**0.33***
	Emoticons		**0.31***	**0.35***		**0.42****	0.29		**0.48****	**0.42****
	Weights		0.17	0.25		0.22	0.28		0.15	0.17
Difficulty	Original			**0.84****			**0.95****			**0.92****
	VAS			**0.81****			**0.95****			**0.94****
	Emoticons			**0.85****			**0.84****			**0.87****
	Weights			**0.69****			**0.92****			**0.81***

**p < .05.*
***p < .01.*

time on task was positively correlated with perceived mental effort and perceived difficulty. The depicted weights scale only revealed negative correlations for simple problems between accuracy and mental effort and perceived difficulty.

Because the size of the correlation was already calculated and the values of the correlations were known, the correlational differences were tested one-sided, just as a confirmative test whether the larger (or smaller) correlation was significantly larger (or smaller). For efficiency and clarity reasons, we only report significant results here in text and present all statistics in **Table 6** and **7**.

Correlational comparisons between scale conditions showed that for the simple problems, Weight scale ratings of perceived difficulty had a significant stronger negative relation to accuracy than any other scale. Weight scale ratings of perceived mental effort during simple problem solving were also more strongly (negatively) related to accuracy compared to the ratings using the Original and Emoticon scales. For the complex problems it was

TABLE 6 | Comparisons of Correlations between the Scale types.

r		Original	VAS	Emoticons	Weights	*z*	*p*
Accuracy -PD	easy	x			x	**–3.22**[a]	**0.001**[a]
			x		x	**–3.52**[a]	**<0.001**[a]
				x	x	**–2.99**[a]	**0.001**[a]
	hard	x	x			–0.98	0.164
		x		x		–0.49	0.313
		x			x	**–3.02**[a]	**0.001**[a]
Time PD	easy	x		x		0.47	0.321
			x	x		**1.75**[a]	**0.040**[a]
				x	x	0.78	0.217
	hard	x	x			<0.01	0.500
		x		x		–0.46	0.322
		x			x	0.70	0.242
			x	x		–0.44	0.331
			x		x	0.70	0.241
				x	x	1.10	0.136
Accuracy -ME	easy	x			x	**–2.51**[a]	**0.006**[a]
			x		x	–1.33	0.091
				x	x	**–2.18**[a]	**0.015**[a]
	hard	x	x			0.05	0.479
		x		x		–0.51	0.307
		x			x	**–2.03**[a]	**0.021**[a]
			x	x		–0.61	0.273
			x		x	1.06	0.144
Time ME	easy	x		x		1.40	0.081
			x	x		0.64	0.262
				x	x	0.58	0.280
	hard	x	x			–0.48	0.317
		x		x		<0.01	1.000
		x			x	0.05	0.479
			x	x		0.47	0.320
			x		x	0.50	0.310

Note.*Example; the first correlational comparison, contrasted the correlations between Accuracy and Perceived Difficulty of the problems presented with the Original rating scale with those presented with the Weights Scale.*
[a]*p < .05 are printed boldly.*

TABLE 7 | Comparisons of Correlations between the Simple and Complex Problem Solving conditions.

	***R1* simple vs *R2* = complex**	*z*	*p*
Original	**Accuracy -PD**	**1.98**	**0.024**
	Accuracy - ME	**1.84**	**0.034**
	Time - PD	–0.65	0.257
	Time - ME	–0.99	0.161
VAS	Accuracy - ME	0.71	0.239
	time - PD	**–1.70**	**0.044**
	Time - ME	–0.71	0.240
Emoticons	time - PD	–0.64	0.261
	Time - ME	0.35	0.365
Weights	**Accuracy - PD**	**–4.21**	**<0.001**
	Accuracy - ME	**–2.79**	**0.003**

Note. *Example; the first correlational comparison, contrasted the correlations between Accuracy and Perceived Difficulty in the simple problem solving condition with those of the complex problem solving condition.*

found that ratings using the original scale for both perceived difficulty and perceived mental effort correlated stronger (negatively) with accuracy than using the Weight scale. Ratings on the the Emoticon scale for perceived difficulty on the easy problems showed a stronger (positive) correlation with time on task than the VAS scale.

Noticeable from these results is that the numeric scales (original scale and VAS) seem to better reflect effort and difficulty for the complex problems (as inferred by the correlations with accuracy), while the pictorial scales (emoticons and weights) seem to better reflect effort and difficulty in for the simple problems. To test whether correlations differed depending on complexity level within scales, each significant correlation was compared to its simple or complex counterpart one-sided. For example, for the original scale, accuracy of complex, but not simple problems was significantly correlated to perceived difficulty.

In this manner, 12 correlational pairs were tested (see **Table 7**). All comparisons not described in text had significance levels of $p > 0.16$ The significant comparisons showed that for the original scale, the relation between accuracy and perceived difficulty, $z = 1.98$, $p = 0.024$, and accuracy and perceived mental effort, $z = 1.83$, $p = 0.034$, was stronger for the complex than simple problems. For the VAS, it was found that the relation between time on task and perceived mental effort was stronger for the complex than the simple problems, $z = -1.70$, $p = 0.044$. For the Weights, it was found

that the relation between accuracy and perceived difficulty, $z = -4.21$, $p < 0.001$, and accuracy and perceived mental effort, $z = -2.79$, $p = 0.003$ was stronger for the simple than the complex problems.

Scale (dis)liking. Of the 46 participants, 16 participants preferred the original 9-point scale over the others, followed closely by the VAS with 14 votes, the emoticons with 13 votes and the weights with three votes. In response to the reversed question (which scale they disliked the most), 26 participants voted for the weights, nine participants for the original 9-point Likert scale, eight for the VAS and three for the emoticons. **Table 2** shows an overview of the best and least preference and coded explanations for the four types of rating scales.

The original 9-point scale was preferred by 16 participants, mainly for the clarity of every answering option. As one participant stated: 'There is little room for misinterpretation'. Interestingly, this clarity was exactly what nine dislikers criticized about the scale. They found that the scale contained too much text, which also made the answering options confusing. In addition, they stated that nine answering options did not leave enough room for nuance.

The 14 participants who favored the VAS reasoned that it was the most nuanced scale, leaving 'more opportunity for grey areas'. Eight participants liked the VAS the least, mainly because the scale was too unclear and could leave too much room for misinterpretation. The 13 participants who favored the emoticons scale indicated that this was because the emoticons were relatable to perceived difficulty and effort and, making the scale easiest to interpret and use. Also, the scale was found to be visually the most appealing. It was liked least by three participants who indicated that it was unclear what every option represents and because it was 'annoying'. Finally, the three participants who liked the weights-scale most indicated that this was because the weights were 'not as abstract as the other scales' and the differences between the answering options were most apparent. However, with 26 dislikes, this scale was disliked by the most participants out of all the scales predominantly because the weights were perceived as unrelated to the questions and it was 'difficult to estimate the value of the weights in relation to the answer to the question'. Also, the differences between the weights were too small. Furthermore, five participants found the scale visually unpleasant and annoying to use.

DISCUSSION

The aim of the present study was to investigate the validity of subjective rating scales measuring perceived difficulty and mental effort. More specifically, our research question was whether certain visual characteristics of a subjective rating scale intended to measure cognitive load, matter for validity (i.e. does one type of visualization elicits more valid responses than others?). By alternating the visual presentation of the scales, we compared four different types of subjective rating scales measuring perceived cognitive load (i.e., mental effort and difficulty) regarding their relation to performance (i.e., accuracy and time on task). Four scales were compared; two well-known ones, the original 9-point rating scale (Paas, 1992) and the VAS, and two specially designed for this study, using either emoticons or pictures of weights. Validity of the mental effort and difficulty ratings was estimated by correlations with performance (i.e., accuracy and time on task) and comparing correlations between difficulty levels and scale types. Also personal preference of the scales was investigated.

The results supported our first hypothesis that all scales would be able to distinguish between complexity levels of the problems for perceived difficulty and mental effort. Complex problems were rated higher than the simple ones regardless of the scale used. The second hypothesis stating that the pictorial scales might reflect cognitive load better, was partially supported. The pattern of the results of the correlations is the most striking; while numeric scales (original scale and VAS) seem to better reflect effort and difficulty for the complex problems, the pictorial scale (emoticons and weights) seems to better reflect effort and difficulty for the simple problems. More specifically, for complex problems, perceived cognitive load and difficulty as measured by the more abstract numeric scales (i.e., the original nine point Likert rating scale and the VAS) were negatively related to performance. In contrast for the simple problems, perceived cognitive load and difficulty as measured by the pictorial scales (i.e., weights and emoticons) were negatively related to performance (i.e., accuracy and time on task). This seems to suggest that for the simple problems, the pictorial scales appeared to represent experienced cognitive load better than for complex problems. Ironically, some students indicated that the differences between the weights were too small to represent the differences in difficulty they experienced. Perhaps if bigger weight increments (larger intervals, instead of 1, 2, 3 etc. 10, 20, 30 etc.) were used, the scale would be better applicable to the difficult problems. The ratings of mental effort and perceived difficulty on the affective scale were positively related with the time on task needed for the simple problems in that higher ratings were related to more time on task needed. For the complex problems, the original 9-point Likert scale was related to both the accuracy and time on task in that higher ratings were related to lower accuracy and more time on task. The VAS showed the same results as the original 9-point Likert scale, except that perceived difficulty was not (significantly) related to solution accuracy.

Therefore, it appears that the 9-point rating scale is more sensitive than the VAS scale in detecting the correlation between perceived difficulty and solution accuracy as the level of complexity in problem-solving task increases. A rating scale that is more sensitive in detecting perceived difficulty holds potential for enhancing cognitive analytics and the development of self-adaptive systems that links interaction between human and computer systems (Radanliev, et al., 2020). When comparisons between correlations was done, it was found that for the easy problems, the Weight scale provided a better reflection of perceived difficulty. However for the complex problems, the original 9-point rating scale did best.

On scale preference, it was found that the original scale was preferred most (with the VAS and emoticons closely following) and the least liked were the weights. However, likability was not a predictor for validity in terms of the association with the ratings on

the scales and performance. A point for discussion might be that students were asked about their preference for a scale after a block of four problems in the order of simple-simple -complex -complex. Having made the complex problems just before being asked about scale preference might have induced a recency effect, in that the responses recorded reflect the experience of rating mental effort and difficulty on the complex problems more than that of the simple problems. In a future studie it would be interesting to ask for participants' preference directly after they completed the simple problems and again after the complex problems.

One limitation of the present study is that the picture of the weights in the embodied scale did not fully fit onto the screen for three of the participants. Therefore, the participants had to scroll sideways to view the full length of the scale. These technical issues might have confounded the disliking of the weight scale compared to the other scales. A more theoretical limitation is that although the results showed a difference in effort and difficulty ratings for the pictorial versus numerical scales for simple versus complex problems, another factor besides the embodied cognition account may play a role in explaining these differences. In a study by Schmeck et al. (2015) that used similar week day problems, timing and topic of the ratings seemed to matter for the outcomes. Delaying effort and difficulty rating after a series or problems seemed to elicit higher scores than the average of ratings given immediately after each separate problem. The delayed ratings seemed to be better predictors of the performance on the complex problems. However, for affective components such as interest and motivation, this difference was not found. We suggest that it would be interesting to replicate the study of Schmeck et al. (2015) with the four types of scales used in this study for two reasons. First, it would be interesting to find out how the response to the affective items would differ between numeric and pictorial scales. It might be the case that the numeric scales are less sensitive to affective questions than pictorial scales and that this is the reason that these ratings were not sensitive to the timing or complexity of the problems. Second, we might find a differentiating effect depending on the timing and complexity of the problems for the affective items, using the pictorial scales.

CONCLUSION

In summary, it seems that the pictorial scales (i.e., emoticons and weights) seem to provide a more valid indication of mental effort and difficulty for simple tasks and the original 9-point Likert scale and the VAS more for complex tasks. This can be explained from the perspective of recent critics on the embodied cognition theory in that for lower-level abilities and functions, cognition may be well-grounded in sensory-motor processes, but that this may not (or to a lesser extent) be the case for higher order cognitive processes (i.e., Caramazza et al., 2014). On the other hand, it seems that numerical scales might be less suitable to reflect perceived mental effort and difficulty on simple problems. Practical implications from this finding would be to use more pictorial rating scales when assessing mental effort and perceived difficulty for simple tasks and more abstract numeric scales for more complex tasks. However, we strongly recommend future studies to replicate this setup, purely to see whether the similar results are found and to find out whether the results are reliable for other populations (i.e., other age group) and for other tasks. Note that we used university students for the present study. From the educational level and age of this rather homogeneous sample, we can expect that the learning capacity and working memory functioning are optimal compared to other populations such as older adults. For the present study, we manipulated task difficulty by increasing element interactivity and thereby intrinsic load (cognitive load elicited by task characteristics). However, populations with suboptimal cognitive functioning, such as older adults, might have more difficulty in general with the task because of an age-related decrease in working memory functioning and cognitive aging in general (e.g., Braver, & West, 2011), resulting in decreased germane cognitive load. In addition, in children, a population in which cognitive functioning and working memory are still developing, it has already been shown that use of pictures work well for clinical ratings such as pain (e.g. Keck, et al., 1996), nausea (Baxter, et al., 2011). In addition, promising results with children are also found for occupational self-assessment (e.g. Kramer, et al., 2010). Therefore, it would be interesting to replicate this study with a sample from a different age population.

In conclusion, in the present study, we made a start in the exploration of different types of subjective rating scales to self-assess invested mental effort and task difficulty. The main finding was that numerical type scales seem to better reflect cognitive processes for complex problem solving while the pictorial type scales for simple problem solving. Whether this finding applies to other forms of simple versus complex tasks, needs to be explored in the future.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics review Committee DPECS, Erasmus University Rotterdam, Netherlands. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

Authors contributed in the order presented: KO coordinated the research and had the lead in writing the manuscript. JW helped along the whole process and contributed ideas and a critical review on the drafts of KO. AK gathered the data, partly analysed the data and also reviewed the manuscript drafts. FP supervised on the background initially and critically reviewed our ideas and manuscript.

FUNDING

The fee is funded by a special fund within the Erasmus University Rotterdam, Netherlands: the Erasmus Open Access Fund.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2021.702616/full#supplementary-material

REFERENCES

Alghadir, A. H., Anwer, S., Iqbal, A., and Iqbal, Z. A. (2018). Test-retest Reliability, Validity, and Minimum Detectable Change of Visual Analog, Numerical Rating, and Verbal Rating Scales for Measurement of Osteoarthritic Knee Pain. *J. Pain Res.* 11, 851–856. doi:10.2147/JPR.S158847

Ayres, P. (2018). "Subjective Measures of Cognitive Load: What Can They Reliably Measure?" in *Cognitive Load Measurement and Application: A Theoretical Framework for Meaningful Research and Practice*. Editor R. Z. Zheng (Routledge/Taylor & Francis Group), 9–28.

Ayres, P. (2006). Using Subjective Measures to Detect Variations of Intrinsic Cognitive Load within Problems. *Learn. Instruction* 16 (5), 389–400. doi:10.1016/j.learninstruc.2006.09.001

Baddeley, A. (2000). The Episodic Buffer: a New Component of Working Memory? *Trends Cogn. Sci.* 4 (11), 417–423. doi:10.1016/S1364-6613(00)01538-2

Baddeley, A. (1992). Working Memory. *Science* 255 (5044), 556–559. doi:10.1126/science.1736359

Barsalou, L. W. (2008). Grounded Cognition. *Annu. Rev. Psychol.* 59, 617–645. doi:10.1146/annurev.psych.59.103006.093639

Barsalou, L. W. (2016). On Staying Grounded and Avoiding Quixotic Dead Ends. *Psychon. Bull. Rev.* 23 (4), 1122–1142. doi:10.3758/s13423-016-1028-3

Baxter, A. L., Watcha, M. F., Baxter, W. V., Leong, T., and Wyatt, M. M. (2011). Development and Validation of a Pictorial Nausea Rating Scale for Children. *Pediatrics* 127 (6), e1542–9. doi:10.1542/peds.2011-231210.1542/peds.2010-1410

Braver, T. S., and West, R. (2011). "Working Memory, Executive Control, and Aging," in *The Handbook of Aging and Cognition*. Editors F. I. M. Craik and T. A. Salthouse (van der, New York: Psychology Press), 311–372.

Caramazza, A., Anzellotti, S., Strnad, L., and Lingnau, A. (2014). Embodied Cognition and Mirror Neurons: a Critical Assessment. *Annu. Rev. Neurosci.* 37, 1–15. doi:10.1146/annurev-neuro-071013-013950

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Erlbaum.

DeLeeuw, K. E., and Mayer, R. E. (2008). A Comparison of Three Measures of Cognitive Load: Evidence for Separable Measures of Intrinsic, Extraneous, and Germane Load. *J. Educ. Psychol.* 100 (1), 223–234. doi:10.1037/0022-0663.100.1.223

Diedenhofen, B., and Musch, J. (2015). Cocor: A Comprehensive Solution for the Statistical Comparison of Correlations. *PLoS ONE* 10 (4), e0121945. doi:10.1371/journal.pone.0121945

Faul, F., Erdfelder, E., Buchner, A., and Lang, A. G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods* 41, 1149–1160. doi:10.3758/BRM.41.4.1149

Gordan, M., and Amutan, K. I. (2014). A Review of B. F. Skinnerâ€s â€Reinforcement Theory of Motivationâ€. *Ijrem* 5 (3), 680–688. doi:10.24297/ijrem.v5i3.3892

Hadie, S. N. H., and Yusoff, M. S. B. (2016). Assessing the Validity of the Cognitive Load Scale in a Problem-Based Learning Setting. *J. Taibah Univ. Med. Sci.* 11 (3), 194–202. doi:10.1016/j.jtumed.2016.04.001

Keck, J. F., Gerkensmeyer, J. E., Joyce, B. A., and Schade, J. G. (1996). Reliability and Validity of the Faces and Word Descriptor Scales to Measure Procedural Pain. *J. Pediatr. Nurs.* 11 (6), 368–374. doi:10.1016/S0882-5963(96)80081-9

Kramer, J. M., Kielhofner, G., and Smith, E. V. (2010). Validity Evidence for the Child Occupational Self Assessment. *Am. J. Occup. Ther.* 64 (4), 621–632. doi:10.5014/ajot.2010.08142

Naismith, L. M., and Cavalcanti, R. B. (2015). Validity of Cognitive Load Measures in Simulation-Based Training: a Systematic Review. *Acad. Med.* 90 (11), S24–S35. doi:10.1097/ACM.0000000000000893

O'Doherty, J. P., Cockburn, J., and Pauli, W. M. (2017). Learning, Reward, and Decision Making. *Annu. Rev. Psychol.* 68, 73–100. doi:10.1146/annurev-psych-010416-044216

Ouwehand, K., van Gog, T., and Paas, F. (2014). Effects of Gestures on Older Adults' Learning from Video-Based Models. *Appl. Cognit. Psychol.* 29, 115–128. doi:10.1002/acp.3097

Paas, F. G. W. C. (1992). Training Strategies for Attaining Transfer of Problem-Solving Skill in Statistics: A Cognitive-Load Approach. *J. Educ. Psychol.* 84, 429–434. doi:10.1037/0022-0663.84.4.429

Paas, F., Tuovinen, J. E., Tabbers, H., and van Gerven, P. W. M. (2003). Cognitive Load Measurement as a Means to advance Cognitive Load Theory. *Educ. Psychol.* 38 (1), 63–71. doi:10.1207/S15326985EP3801_8

Radanliev, P., De Roure, D., Van Kleek, M., Santos, O., and Ani, U. (2020). Artificial Intelligence in Cyber Physical Systems. *AI Soc.*, 1–14. doi:10.1007/s00146-020-01049-0

Rhodes, M. G. (2015). *Judgments of Learning: Methods, Data, and Theory*. Oxford University Press. doi:10.1093/oxfordhb/9780199336746.013.4

Schmeck, A., Opfermann, M., van Gog, T., Paas, F., and Leutner, D. (2015). Measuring Cognitive Load with Subjective Rating Scales during Problem Solving: Differences between Immediate and Delayed Ratings. *Instr. Sci.* 43 (1), 93–114. doi:10.1007/s11251-014-9328-3

Shenhav, A., Musslick, S., Lieder, F., Kool, W., Griffiths, T. L., Cohen, J. D., et al. (2017). Toward a Rational and Mechanistic Account of Mental Effort. *Annu. Rev. Neurosci.* 40, 99–124. doi:10.1146/annurev-neuro-072116-031526

Sitzmann, T., Ely, K., Brown, K. G., and Bauer, K. N. (2010). Self-assessment of Knowledge: A Cognitive Learning or Affective Measure? *Amle* 9 (2), 169–191. doi:10.5465/amle.9.2.zqr169

Skinner, B. F. (1963). Operant Behavior. *Am. Psychol.* 18 (8), 503–515. doi:10.1037/h0045185

Skjott Linneberg, M., and Korsgaard, S. (2019). Coding Qualitative Data: A Synthesis Guiding the Novice. *Qrj* 19 (3), 259–270. doi:10.1108/QRJ-12-2018-0012

Staddon, J. E., and Cerutti, D. T. (2003). Operant Conditioning. *Annu. Rev. Psychol.* 54 (1), 115–144. doi:10.1146/annurev.psych.54.101601.145124

Sung, Y. T., and Wu, J. S. (2018). The Visual Analogue Scale for Rating, Ranking and Paired-Comparison (VAS-RRP): a New Technique for Psychological Measurement. *Behav. Res. Methods* 50 (4), 1694–1715. doi:10.3758/s13428-018-1041-8

Sweller, J., Ayres, P., and Kalyuga, S. (2011). "Measuring Cognitive Load," in *Cognitive Load Theory* (New York, NY: Springer), 71–85. doi:10.1007/978-1-4419-8126-4_6

Sweller, J. (2010). Element Interactivity and Intrinsic, Extraneous, and Germane Cognitive Load. *Educ. Psychol. Rev.* 22, 123–138. doi:10.1007/s10648-010-9128-5

Sweller, J., and Paas, F. (2017). Should Self-Regulated Learning Be Integrated with Cognitive Load Theory? A Commentary. *Learn. Instruction* 51, 85–89. doi:10.1016/j.learninstruc.2017.05.005

Sweller, J. (1993). Some Cognitive Processes and Their Consequences for the Organisation and Presentation of Information. *Aust. J. Psychol.* 45, 1–8. doi:10.1080/00049539308259112

Sweller, J., van Merriënboer, J. J. G., and Paas, F. (2019). Cognitive Architecture and Instructional Design: 20 Years Later. *Educ. Psychol. Rev.* 31 (2), 261–292. doi:10.1007/s10648-019-09465-5

Sweller, J., van Merriënboer, J. J. G., and Paas, F. G. W. C. (1998). Cognitive Architecture and Instructional Design. *Educ. Psychol. Rev.* 10, 251–296. doi:10.1023/A:1022193728205

van Gog, T., Kirschner, F., Kester, L., and Paas, F. (2012). Timing and Frequency of Mental Effort Measurement: Evidence in Favour of Repeated Measures. *Appl. Cognit. Psychol.* 26, 833–839. doi:10.1002/acp.2883

van Merriënboer, J. J. G., and Sweller, J. (2005). Cognitive Load Theory and Complex Learning: Recent Developments and Future Directions. *Educ. Psychol. Rev.* 17, 147–177. doi:10.1007/s10648-005-3951-0

Zanolie, K., Dantzig, Sv., Boot, I., Wijnen, J., Schubert, T. W., Giessner, S. R., et al. (2012). Mighty Metaphors: Behavioral and ERP Evidence that Power Shifts Attention on a Vertical Dimension. *Brain Cogn.* 78 (1), 50–58. doi:10.1016/j.bandc.2011.10.006

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Education

ORIGINAL RESEARCH
published: 10 January 2022
doi: 10.3389/feduc.2021.648319

Analysing the Relationship Between Mental Load or Mental Effort and Metacomprehension Under Different Conditions of Multimedia Design

Lenka Schnaubert[1]* *and Sascha Schneider*[2]

[1]*Media-Based Knowledge Construction—Research Methods in Psychology, Computer Science and Applied Cognitive Science, Faculty of Engineering, University of Duisburg-Essen, Duisburg, Germany,* [2]*Educational Technology, Institute of Education, Faculty of Arts and Social Sciences, University of Zurich, Zurich, Switzerland*

OPEN ACCESS

Edited by:
Fred Paas,
Erasmus University Rotterdam, Netherlands

Reviewed by:
Tina Seufert,
University of Ulm, Germany
Lisette Wijnia,
Open University of the Netherlands, Netherlands

***Correspondence:**
Lenka Schnaubert
lenka.schnaubert@uni-due.de

Specialty section:
This article was submitted to Educational Psychology, a section of the journal Frontiers in Education

Received: *31 December 2020*
Accepted: *10 December 2021*
Published: *10 January 2022*

Citation:
Schnaubert L and Schneider S (2022) Analysing the Relationship Between Mental Load or Mental Effort and Metacomprehension Under Different Conditions of Multimedia Design. Front. Educ. 6:648319. doi: 10.3389/feduc.2021.648319

Cognitive load theory assumes effort may only lead to comprehension if the material-induced load leaves enough resources for learning processes. Therefore, multimedia materials should induce as little non-relevant load as possible. Metacognition research assumes that learners tap into their memory processes to generate a mental representation of their comprehension to regulate learning. However, when judging their comprehension, learners need to make inferences about actual understanding using cues such as their experienced mental load and effort during learning. Theoretical assumptions would assume both to affect understanding and its metacognitive representation (metacomprehension). However, the question remains how perceived effort and load are related to metacomprehension judgments while learning with multimedia learning material. Additionally, it remains unclear if this varies under different conditions of multimedia design. To better understand the relationship between perceived mental load and effort and comprehension and metacomprehension under different design conditions of multimedia material, we conducted a randomised between-subjects study (N = 156) varying the design of the learning material (text-picture integrated, split attention, active integration). Mediation analyses testing for both direct and indirect effects of mental load and effort on metacomprehension judgments showed various effects. Beyond indirect effects via comprehension, both mental load and effort were directly related to metacomprehension, however, this seems to vary under different conditions of multimedia design, at least for mental effort. As the direction of effect can only be theoretically assumed, but was not empirically tested, follow-up research needs to identify ways to manipulate effort and load perceptions without tinkering with metacognitive processes directly. Despite the limitations due to the correlative design, this research has implications for our understanding of cognitive and metacognitive processes during learning with multimedia.

Keywords: metacomprehension judgments, multimedia learning, mental load, mental effort, cue utilization

1 INTRODUCTION

Research on multimedia learning aims at examining the influence of differences in the design of learning materials on learning outcomes (for an overview, see Li et al., 2019). For example, the inclusion of signaling cues (i.e., the highlighting of learning-relevant information or the structure of a learning material; Schneider et al., 2018) or the segmentation of learning materials into meaning-related sections (Rey et al., 2019) were found to foster learning by highlighting relevant and coherent concepts of a learning material. However, not all materials are well designed or activate learners to build a coherent mental model leading to huge learning differences. One major explanation for learning differences is based on the cognitive load theory (Sweller, 1994; Sweller et al., 2011). According to this theory, learners experience a cognitive load when processing the information presented in a multimedia learning material. Cognitive load, also often referred to as mental load is, thus, said to be task-related and reflects the cognitive resources needed to cope with the complexity of the learning material.

Cognitive load can be distinguished into two groups of processes: learning-relevant (productive) and learning-irrelevant (unproductive) cognitive load processes (Kalyuga and Singh, 2016). While productive cognitive load refers to all cognitive processes that are needed to reach a learning goal, unproductive cognitive load refers to all cognitive processes that occur due to design-induced information search processes. For example, decorative (learning-irrelevant) pictures integrated into multimedia learning material may distract learners' attention away from learning-relevant information (for an overview, see Schneider et al., 2016). When both textual and pictorial information are learning relevant and relate to each other, their spatial distance is of major importance for the amount of cognitive load since spatially distant representation lead to a learning-hindering split-attention effect (for an overview, see Schroeder and Cenkci, 2018). In this case, the integration of textual information into the pictorial information source is found to enhance complex learning (i.e., the spatial contiguity principle; Mayer and Moreno, 2003; for an overview, see Schroeder and Cenkci, 2018) by a reduction of unproductive cognitive load (Schroeder and Cenkci, 2020). Another possibility to foster learning is not to reduce learning-irrelevant processes, but to induce learning-relevant processes. Similar to a generation effect (see Bertsch et al., 2007), asking learners to actively integrate pictorial and symbolic sources of information in multimedia learning material seems to support them in building coherent mental representations and fosters learning (Bodemer et al., 2004; Bodemer et al., 2005). Thus, within such an active-integration procedure, learners are asked to integrate disintegrated material assuming the load induced by the split attention is gradually reduced by actively integrating the material. During this process it gets gradually replaced by the additional productive cognitive load that supports building coherent mental models. Thus, while the load imposed is supposed to be quite high, such "desirable difficulties" (see Bjork and Bjork, 2011) should ultimately be beneficial for learning as has been found for other instructions designed to induce germane cognitive processes, for example self-explanation prompts (e.g., Berthold and Renkl, 2009; Renkl et al., 2009).

While the terms "mental load" and "cognitive load" are sometimes used interchangeably, other conceptualizations differentiate the concept. In these, cognitive load is not seen as a unidimensional construct based on task-induced affordances, but also includes the effort learners assign to task-processing (Paas and van Merriënboer, 1994). As described with the above examples, cognitive load is imposed by the demands from the learning environment to perform a certain learning task. However, not all learners will achieve the same learning under the same task conditions. Learners allocate a different amount of cognitive resources for a given task demand (i.e., their cognitive engagement in the task). This allocation of cognitive resources is known as mental effort (Orru and Longo, 2019). Mental effort reflects a second dimension of possible assessment factors besides the measurement of (task-driven) mental load and the actual learning performance and can be described as a second indicator for possible learning differences by addressing the human-centered dimension of cognitive load (Scheiter et al., 2020). Thus, mental effort refers to learners' actually invested cognitive resources while processing information of a learning material (Paas and van Merriënboer, 1994).

The relationship between mental load and mental effort and their relation to learning processes has been examined to a minor degree. Some studies propose that mental effort and mental load are different concepts with unique consequences for the measurement of learning processes (e.g., Ayres and Youssef, 2008; Schmeck et al., 2015). The experienced mental load and invested mental effort are most often measured with subjective rating scales (e.g., Naismith et al., 2015; Anmarkrud et al., 2019). When using such a subjective measurement, researchers assume that learners are able to access and assess their own invested cognitive resources and load imposed by a learning task (Paas et al., 2008).

If insights into mental processes and resources like mental load and mental effort are available to learners, we must assume that these learners are also able to use this information for metacognitive regulation purposes. While there clearly is more to (successful) regulation than metacognitive monitoring (e.g., using monitoring outcomes to control learning; Schnaubert and Bodemer, 2017), from a learner's perspective, internal experiences may be used to guide regulation attempts (independent of their successfulness). Metacognitive processes are deemed vital for understanding how learners approach and process learning material (Schraw et al., 2006). While this may apply for memory processes like word-pair or vocabulary learning, this also extends to comprehending complex expository material (Wiley et al., 2005). Not only do learners need to plan what to study, but also divert attention towards relevant sources and invest effort in processing and integrating the content, for example when integrating texts and graphics during multimedia learning (e.g., Burkett and Azevedo, 2012). To do so, according to metacognition theories, learners monitor their learning-related cognitive processes and outcomes (e.g., their comprehension) and—by comparing it to standards—evaluate the need for further studying or a change of tactics or strategies

(e.g., Nelson and Narens, 1990; Winne and Hadwin, 1998) and may thus actively steer their learning processes to successfully foster learning (e.g., Thiede, 1999; Metcalfe, 2009).

Thus, monitoring learning and comprehension is a crucial part of learning. To support learners in doing so, it is critical to understand what affects their metacognitive monitoring judgments, which is the evaluation of own learning, memory and comprehension. It is widely assumed that learners have no direct access to their memory or comprehension strength, but that metacognitive judgments rely on cues and are inferential in nature. Cue utilization theory (Koriat, 1997) assumes that learners use a number of relevant cues to judge their learning and comprehension. A crucial part of these are mnemonic cues, that is cues relating to perceiving memory processes such as the ease with which information is processed, encoded or retrieved from memory (e.g., Begg et al., 1989; Benjamin et al., 1998). Such experience-based cues inform metacognitive judgments and are in turn influenced by intrinsic and extrinsic factors, for example, material complexity or prior exposure (Koriat, 1997; Koriat et al., 2008). However, as not all cues are equally diagnostic to learning in all circumstances, their usage may result in low monitoring accuracy (i.e., the relationship between actual comprehension and metacomprehension), which is often found to be quite poor for metacomprehension (Dunlosky et al., 2005; Dunlosky and Lipko, 2007). While there are various aspects of memory and cognition learners may monitor, our research is focused on the monitoring of own levels of comprehension. In this context, metacomprehension is defined as the metacognitive representation of own comprehension as a result of metacognitive monitoring (Maki and Berry, 1984; Wiley et al., 2005) rather than a valid understanding of own comprehension (metacomprehension accuracy). Thus, while a lot of multimedia-focused research focusses rather on the accuracy of metacognitive judgments, this is a rather instructional design perspective (as it relates it to learning outcomes from an outside perspective). From a metacognition perspective, it is of equal importance to understand what affects a learner's judgment itself as this forms the basis for self-regulatory processes (e.g., Koriat, 1997). The subjective experience and evaluation of own learning (not their accuracy) is key in regulating behavior and form the basis of control decisions (Nelson and Narens, 1990; Son and Schwartz, 2002), for example regarding study time allocation (e.g., Son and Metcalfe, 2000). While both the subjective monitoring judgments and their accuracy ultimately contribute to learning and are worthwhile studying, the learner-centred view aims at understanding the learners' experience. This is the perspective taken within this study. From a metacognitive self-regulation perspective, it is of vital importance to understand how learners judge their comprehension and form metacomprehension judgments and how this relates to cognitive processing. From a multimedia-learning perspective, it is important to understand how this is affected by the design of (multimedia) learning material.

De Bruin and van Merriënboer (2017) made a first explicit attempt to connect self-regulated learning and cognitive load theory. While they focused on differences and similarities between the theories, concepts and measurement rather than the actual relationship between the specific cognitive and metacognitive constructs in question, within their special issue, there are some studies relating cognitive load and metacognitive judgments empirically. For example, Schleinschok and colleagues (2017) found a strong negative correlation between prospective judgments-of-learning (JOLs) and cognitive load. They further performed regression analyses finding cognitive load not contributing to explaining test performance when JOLs were included in the analyses. Baars and colleagues (2017), on the other hand, assume that under various task conditions varying in cognitive load, learners may use their invested mental effort as cue for forming metacognitive judgments. While this hypothesis needs empirical validation, the authors previously found strong correlations between mental effort and metacognitive judgments, although the impact of actual learning was not included in these calculations (Baars et al., 2013). A current meta-analysis found a substantial negative overall relationship between perceived mental effort and monitoring judgments (Baars et al., 2020). This negative relationship, however, defused for self-agent ratings (i.e., ratings stressing the effort put willingly into learning rather than the effort necessary to solve a task; Koriat, 2018). Thus, it can be assumed that mental effort only negatively relates to monitoring judgments, when it reflects a need rather than a choice to invest effort. Our study relates to the latter conceptualisation, as it distinguishes task-induced load and effort invested willingly.

While cue utilization has been researched extensively within metacognition research, multimedia research provides insights into cognitive processes relevant for processing complex learning material. Within cognitive load research, it is commonly assumed that learners are able to access information about and thus validly judge their mental load and effort (Paas et al., 2008). However, judging own mental effort may also be viewed from a metacognitive perspective as it entails monitoring own cognitive processes (Scheiter et al., 2020). Thus, it seems logical that learners may use these insights into their cognitive processes as cues to form metacognitive judgments about their learning; perceived mental load and effort may therefore affect metacognitive judgments (Baars et al., 2017). Consequently, in our study, we want to investigate if learners' mental effort and mental load are related to metacomprehension judgments beyond their effect on the learner's actual level of comprehension.

De Bruin and colleagues related cues, monitoring judgments, and learning by cue utilization, diagnosticity of the cues for learning, and monitoring accuracy (de Bruin et al., 2017). Their (metacognitive) model resembles a mediation model predicting learning or performance based on monitoring judgments. This is in line with other studies, for example, Schleinschok and colleagues (2017), who regressed JOLs and cognitive load on test performance. While we understand the reasons for predicting performance based on prospective JOLs, when directly targeting metacomprehension, i.e., learners' mental representation of their comprehension, learners judge their current state of learning not their later task performance. Such presumably subtle difference in assessment may have severe impact on the outcome of a metacognitive judgment (Kelemen, 2000). Thus, in our model, metacomprehension is

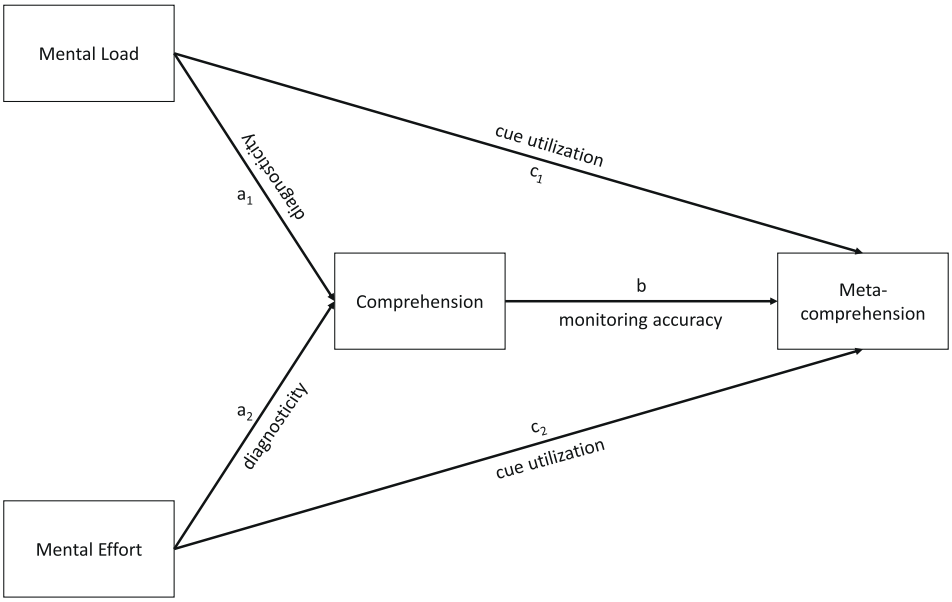

FIGURE 1 | Assumed mediation model.

following comprehension not preceding it (although it may precede its assessment if it can be assumed that comprehension is relatively stable and not affected by the process of judging one's comprehension on a metacomprehension scale; see **Section 2.3**). We additionally argue that while the value of mental load and mental effort for predicting learning and resulting comprehension may diminish when metacognitive judgments are involved, that does not mean that they are not affecting the judgments themselves and the strong intercorrelation found in Schleinschok and colleagues' study (2017) indicates there might be more to this relationship. To sum it up, while we are aware of the different approaches to these relationships, we assume mental processes are not only predictive for actual comprehension, but learners' awareness of these processes may be used as mnemonic cues for metacognitive judgments as well. Thus, we assume that learning and comprehension precedes monitoring said comprehension and consequently, in our model, comprehension is the mediator while metacomprehension is the criterion with (perceived) mental load and mental effort as predictors (**Figure 1**).

Based on multimedia research, we assume high diagnosticity for perceived mental effort (positive) and perceived mental load (negative) for predicting comprehension (a paths). Additionally, based on metacognition research, we assume a high positive relationship between comprehension and metacomprehension judgments (b path). While this relationship is termed "monitoring accuracy" and can theoretically be modelled as such, please note that within the statistical model used in our empirical study (see **Section 2**), a between-subject relationship between comprehension and metacomprehension judgments does not reflect monitoring accuracy as it does not reflect how individual learners monitor their cognitions and differentiate between high or low comprehension (see Schraw, 2009 and Schraw et al., 2013 for measures of monitoring accuracy). It remains unclear, whether perceived mental load and mental effort are predictive for metacomprehension judgments (c paths) and especially whether they thus affect metacomprehension judgments beyond their actual effects on comprehension (c' paths). Based on de Bruin's model (de Bruin et al., 2017), these paths are termed "cue utilization" describing a theoretical relationship between those concepts. Considering the above discussed literature (e.g., Koriat, 2018; Baars et al., 2020) as well as the above assumed relationships (a and b paths), we assume (perceived) mental effort to be rather positively and (perceived) mental load to be negatively related to metacomprehension judgments. The positive relationship of mental effort hereby refers to a conceptualization of effort that includes elements of self-agency and thus a choice to invest effort rather than a need to invest.

We know from multimedia and cognitive load theory and research that while load and effort contribute to comprehension, their value for judging comprehension depends strongly on the target processes involved and if they ultimately are relevant or not relevant for learning and comprehension. As described above, within multimedia learning, this may strongly depend on the design of the learning material. For example, learning-irrelevant load may be induced by a split attention format whereby an active integration format may induce load positively related to learning. This begs the question if learners—who may use mnemonic cues like the perceived mental load put upon them by the learning material and the effort applied to process the content to judge their comprehension (Baars et al., 2017)—are aware whether the processes conducted are directly related to learning or not. Thus, we further want to know if these mechanisms apply to different design versions of multimedia material similarly or if designs that evoke unproductive load (like a split attention format) or

additional productive learning processes (like active integration) or reduce load (like an integrated material) have differential effects. For example, the ease with which integrated material may be processed may lead to learners judging their comprehension to be quite high while processing that requires more effort may be judged as less understood (ease-of-processing hypothesis; Undorf and Erdfelder, 2011). While this may be reasonable when the load imposed is based on unfortunate design of the multimedia material (e.g., split attention format), this may not hold true while actively integrating material, which may induce a high load, but may ultimately be rewarding (comparable to desirable difficulties, Bjork and Bjork, 2011). Thus, the actual diagnosticity of (perceived) mental load and effort for comprehension may vary under various load-inducing conditions and it remains unclear, if cue utilization differs accordingly. Thus, we will use a variety of different design mechanisms to find out if the relationship between mental load or effort and metacomprehension differs for these conditions and if—by affecting diagnosticity—they may hamper the relationship between comprehension and metacomprehension judgments. Thus, although effects on mental load and mental effort caused by the multimedia design are assumed, the main target of the paper is not to assess the effect of multimedia design on mental load or effort as this has been done extensively in the past (e.g., Xie et al., 2017; Mutle-Bayraktar et al., 2019), but to investigate if the multimedia design and its potential effects on the relationship between load, effort and learning effectiveness (i.e., comprehension) affect metacomprehension judgments and their relationship with mental load or effort as well. Consequently, the study aims to investigate if a potential direct and indirect relationship between perceived load and effort and metacomprehension judgments differs between conditions of multimedia design.

Taken together, we assume that overall, (perceived) mental effort is positively related to metacomprehension judgments (hypothesis 1) while (perceived) mental load is negatively related to metacomprehension judgments (hypothesis 2). With regard to the precise relationship and also the effect of multimedia design, our research questions are: 1) Is learners' perceived mental effort and load related to learners' judgments of comprehension (beyond actual effects on comprehension), and 2) does this vary under different load-imposing multimedia conditions (i.e., split attention, integrated, active integration condition)? While in line with Baars and colleagues (2017) we assume learners to use mental effort and load as cues for their metacognitive judgments, please be advised that intentional cue usage is not in the focus of our research, but the relationship between the constructs.

While the research basis with regard to the effect of the multimedia design on metacomprehension judgments (vs. metacomprehension accuracy) is too scarce to form explicit hypotheses and we thus chose a more exploratory approach to that respect, some effects seem more likely than others. For example, based on research on generative activities and its relation to monitoring accuracy (e.g., Prinz et al., 2020; van Gog et al., 2020), it may be assumed that effort and load within active integration affect metacomprehension judgments rather indirectly *via* comprehension rather than directly (without relating to comprehension). With regard to split attention, one may assume that a possible effort-metacomprehension-relationship may not be mediated by comprehension as the effort-comprehension relationship may be hampered by investing effort in overcoming the split-attention rather than germane learning activities (e.g., Beege and Colleagues, 2019). However, as the research basis for these assumptions is not yet solid enough to form distinct hypotheses (apart from the direction of the general relationship between perceived mental load or effort and metacomprehension judgments), we chose a more exploratory approach to take a first step towards understanding how (perceived) mental load and mental effort are related to metacognitive monitoring of learning under varying conditions of multimedia design.

While our study focusses on assumptions about the relationship between mental effort, mental load, comprehension, and metacomprehension under varying conditions of multimedia design, we further assume to replicate effects frequently found in multimedia research. Based on the literature on multimedia learning, we assume the multimedia design to affect comprehension with a split attention format rather hampering learning (unnecessary effort needs to be invested into integrating text and graphics; e.g., Schroeder and Cenkci, 2018) and active integration rather fostering learning (generative activity; e.g., Bodemer et al., 2004). Further, we assume multimedia design to affect especially mental load, with the potentially load inducing conditions (split attention and active integration) resulting in higher perceived mental load (e.g., Schroeder and Cenkci, 2020). However, as these issues are not in the focus of the study, we did not include them in the formal hypotheses.

2 MATERIALS AND METHODS

The study (study-ID: psychmeth_2019_MMMC_66) was conducted November 2019 to December 2019 at the University of Duisburg-Essen and approved by the local ethics committee (ethics votum-ID: 1910PFYE114). It was not officially pre-registered.

2.1 Sample

The sample consisted of 156 university students, most of them (145) studying applied cognitive and media science or loosely related subjects (seven psychology, one applied informatics, one engineering, one applied language science, one did not provide a course of study). They had a mean age of 21.00 years (SD = 3.35). Most of them (121) identified as female, 34 as male, and one as non-binary. Altogether, the sample can be described as quite homogeneous: predominantly females in their early 20 s studying applied cognitive and media science at a German university. The participants were randomly assigned to one of three conditions with a different multimedia design (see **Section 2.2**): integrated format (n = 51), split attention format (n = 52) and active integration format (n = 53). Due to an assignment error, not all pre-test data to describe sample characteristics could be

TABLE 1 | Motivational questionnaire: descriptive data and group differences.

Variable	Active integration		Split attention		Integrated		Cronbach	Welch test		
	(*n* = 53)		(*n* = 52)		(*n* = 51)		α	*F*	*df1/df2*	*p*
	M	*SD*	*M*	*SD*	*M*	*SD*				
Interest/Enjoyment	3.92	1.36	3.79	1.46	3.46	1.53	0.910	1.34	2/101.45	0.266
Perceived competence	3.60	1.35	3.49	1.40	3.77	1.32	0.917	0.57	2/101.93	0.566
Pressure/Tension	3.77	1.50	4.05	1.53	3.89	1.49	0.802	0.46	2/101.95	0.633

confidently traced back, but a rigorously cleaned minimal dataset with 137 participants showed a medium interest in the topic of information transmission within the nervous system ($M = 4.09$, $SD = 1.59$ on a scale from 1 = "very low" to 7 = "very high") and rather low self-assessed prior knowledge on the nervous system ($M = 3.09$, $SD = 1.37$ on a scale from 1 = "very poor" to 7 = "very well") and ability to explain the difference between a excited und inhibited synapse ($M = 2.47$, $SD = 1.52$ on a scale from 1 = "very poor" to 7 = "very well").

2.2 Design

Within this study, we assessed all variables (predictors, mediator, criterion) throughout the sample. Additionally, we randomly assigned each participant to one of three multimedia design conditions. The experimentally varied factor (design of the multimedia learning material) thus had three factor levels varying the design of the learning material to induce various types of cognitive load. On level one (integrated format), the material, consisting of text and picture, was presented in an integrated format with textual annotations attached to the pictorial content to decrease cognitive load. On level two, the text was presented below the picture with numerical indicators of what each text described (split attention format) to induce search processes irrelevant for learning. On level three, the text and picture were presented like in level two, but without the numbers indicating where each text belonged and learners could drag and drop the text blocks into blanks within the picture, thereby actively generating an integrated format (active integration format). For more details see **Section 2.4**.

2.3 Procedure

The study took place in research laboratories that seated up to three participants simultaneously. After welcoming, participants were seated on a computer screen each with a 24″ monitor. The setup was identical for all participants. After an introduction into the study and giving informed consent, participants received a short pre-study questionnaire assessing rough indicators of self-assessed prior knowledge and topic-specific interest (see **Section 2.1**).

Afterwards, participants received multimedia learning material on stimulus transduction at a synapse (text and graphic) from Florax and Plötzner (2010). It was presented in an integrated format and explained basic functions and processes of synaptic transduction (see **Section 2.4**). Participants had 4 minutes to familiarise themselves with the process. Afterwards they were redirected to their respective learning material according to condition. The learning material was identical except for the integration of the material (see **Section 2.4**) and consisted of a multimedia representation of an inactive, excited and inhibited synapse. Learners had 7 minutes to study the information before they were redirected to another questionnaire page.

For questionnaires, they first filled out a short motivational questionnaire based on Wilde and colleagues et al. (2009) assessing interest/enjoyment, perceived competence and pressure/tension. As the information is not relevant for the purpose of this study, results are reported in **Table 1** but not discussed further. Afterwards, learners were asked to provide various metacognitive judgments. They were asked how many items they expected to answer correctly on a 30-item test. This is consistent with a judgment-of-learning as used by Wiley and colleagues (Wiley et al., 2008). Further and on the same questionnaire page, they were asked to provide metacomprehension judgments (comparable to Thiede et al., 2003), first on the overall topic (information transmission within the nervous system) and then separately for processes regarding an inactive, an excited and an inhibited synapse. The last three were later combined to assess metacomprehension (see **Section 2.5.2**). As we were interested in how the learners judged their current level of comprehension of the material rather than how they predicted their future performance, we used the metacomprehension judgments for our analyses. Both measures (judgment-of-learning and metacomprehension judgments) were highly interrelated ($\rho = 0.716$, $p < 0.001$). Data on the judgments-of-learning are included in **Table 2** for reference, but please note that the measures (including monitoring accuracy) only provide very rough indicators as they are based on one prediction of a performance in an unknown test. Thus, this data may provide background information, but will not be considered for further analyses. After providing metacomprehension ratings, participants were asked to fill out the StuMMBE-Q by Krell (2015), a 12-item instrument to assess mental effort and mental load (see **Section 2.5.1**). Learners were then asked to conduct a 30-item comprehension test adapted from Beege and Colleagues (2019). Finally, learners filled out a demographic's questionnaire, assessing age, gender and course of study. Afterwards, they were thanked and were able to receive course credit for participation. The procedure is depicted in **Figure 2**.

While at first glance, the procedure may seem inappropriate for the assumed mediation model as it is widely agreed upon (and usually a good strategy) that in mediation and regression

TABLE 2 | Judgments-of-learning: descriptive data and group differences.

Variable	Active integration		Split attention		Integrated		Welch test		
	(*n* = 51[a])		(*n* = 48[a])		(*n* = 46[a])		*F*	*df1/df2*	*p*
	M	*SD*	*M*	*SD*	*M*	*SD*			
Judgment of learning	11.8	6.29	13.5	6.13	14.7	7.03	2.39	2/93.5	0.097
absolute accuracy[b]	4.55	3.43	3.89	3.69	4.76	3.43	0.76	2/94.1	0.470
bias[c]	1.14	5.62	1.41	5.19	1.37	5.75	0.03	2/94.0	0.966

[a]n *differs due to missing or illogical answers (i.e., answers > 30).*
[b]*Absolute distance between judgment and test performance.*
[c]*Positive values indicate overestimation of performance (negative values underestimation).*

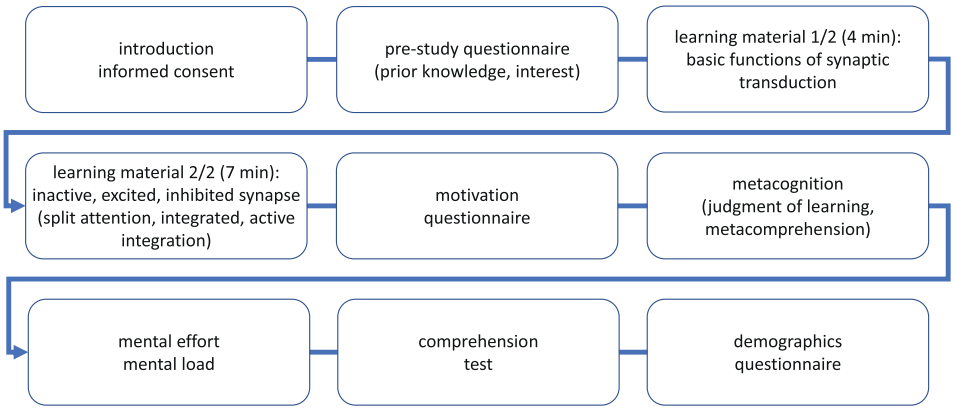

FIGURE 2 | Procedure.

analyses, predictors should be assessed prior to mediators and criterion variables, we chose to change the order of assessment. This has various reasons. First, the order of assessment is rather less relevant to theoretical assumptions as overtly sometimes implied, but it is the actual order of occurrence of the true processes that matter. While assessing predictors before a criterion (or mediator) ensures the order of assessment matches the order of occurrence, more stable predictors not fazed by timely changes do not need to be assessed prior to a criterion variable if it can be validly argued that they preceded the criterion. While we not necessarily expect comprehension or (self-assessed) mental load or effort to be stable over a longer period of time, no relevant learning and rehearsal processes will take place after the learning phase and especially comprehension should thus be relatively stable for the short amount of time it takes to fill out a few short metacomprehension questions. Second, some variables are much more susceptible to influence than others and the order of assessment may severely affect results and mask true relationships between constructs. While asking for metacognitive judgments may act as metacognitive prompts during or before learning (e.g., Schnaubert and Bodemer, 2017), these are unlikely to affect either mental load or mental effort (or their assessment) or comprehension when there is no further study phase. On the other hand, metacomprehension judgments are assumed to be highly susceptible to assessing other variables. For example, testing for comprehension may severely alter metacognitive judgments as learners may use their experience during testing as indicators for their actual comprehension (e.g., Maki, 1998). If testing is just a means to an end (to assess comprehension) rather than part of the studied scenario, assessing metacomprehension after conducting a comprehension test would have blurred all prior existing connections. Similarly, the assessment of mental effort or mental load may impact metacognitive judgments. As we assume those to be used as indicators during learning, assessing them before metacomprehension might act as a self-fulfilling prophecy with no meaning for real-life processes. Thus, we opted to assess the criterion before the predictors and mediator.

2.4 Learning Material and Independent Variable

The learning material consisted of two webpages. On the first webpage, an introduction to the learning topic "Functioning of a synapse" was given. This introduction consisted of definitions and explanations of the nervous system and its components obtained from Florax and Plötzner (2010). The second webpage included a graphic of a synapse with three segments. This graphic was also obtained from Florax and Plötzner (2010). The second webpage was prepared with three different versions of the graphic. The first version of the graphic showed a synapse that explained the processes at a non-active synapse, the processes at

an excited synapse, and the processes at an inhibited synapse. Overall, 21 synaptic sub-processes were displayed within the graphic, and every subprocess had an associated text label. In this version of the graphic, all verbal explanations of processes were displayed in rectangular boxes close to the place that stands for the process in the graphic. This version of the graphic is further called "Integrated format". A second version of the graphic showed the same graphic but the verbal explanations were exchanged by numbers. Then, under the graphic, all the (numbered) explanations were listed one by one. This version is further called "Split-Attention format". The third version of the graphic was designed similar to the graphic in the first version. In contrast, the verbal explanation boxes did not contain any information but a white box. Similar to the second version, under the graphic, all the explanations were listed one by one. In contrast to the second version, learners were able to drag the verbal explanations to the appropriate boxes. All boxes were programmed to accept the placement of correct explanations only so that learners did not combine an explanation box with a false box place. When learners dragged an explanation onto an incorrect place, the explanation automatically returned to the list under the graphic. This third interactive version of the graphic is further called "Active Integration format". In all versions of the second webpage, a timer was integrated. This timer was set to 7 minutes in order to regulate the learning time of students. After this timer expired, learners were directed to the next webpage.

2.5 Dependent Variables

2.5.1 Mental Load and Mental Effort

Mental load and mental effort were measured with the StuMMBE-Q, a questionnaire developed by Krell (2015), Krell (2017). The questionnaire consists of 12 items. Six items assessed mental load (Cronbach's $\alpha = 0.839$; e.g., "The contents of the tasks were complicated"). Another six items assessed mental effort (Cronbach's $\alpha = 0.780$; e.g., "I have given my best to complete the tasks"). Please note that the items for mental effort include asking participants to judge whether they tried hard to solve the task, and the measure thus contains elements of self-agency which is a somewhat different conceptualization than the one-item scale by Paas (1992). Students had to rate these items using a 7-point equidistant response format as within the original conceptualization ranging from "not at all" to "totally". While we are aware of the research indicating a 3-point format may ease distinction between categories (Krell, 2015), for the models assumed we feared a 3-point scale may not be adequately used to assume more than ordinal level of information (Leppink, et al., 2013). We used mean answers as estimates for (perceived) mental load and mental effort. Intercorrelation between both scales was low and non-significant ($r = -0.028$, $p = 0.730$).

2.5.2 Metacomprehension

In line with a method used by Thiede and colleagues (e.g., Thiede et al., 2003) and based on previous work by Glenberg and Epstein (1985), metacomprehension was measured by asking learners to rate their understanding of the learning content using a 7-point equidistant response format from "very poorly" (1) to "very well" (7). We asked them separately for their understanding with regard to the processes on an inactive, an excited and an inhibited synapse and used their mean metacomprehensionrating as indicator for the level of metacomprehension. Cronbach's α was at 0.909.

2.5.3 Comprehension

Comprehension was measured with 30 multiple-choice questions (Cronbach's $\alpha = 0.739$) adapted from Beege and Colleagues (2019). These questions were 5-answer single-choice questions with "I don't know" as one option. Some questions contained verbal answer options, some graphical answer options, some a combination of both. In order to answer these questions, participants needed to remember and understand the verbal explanations and the processes displayed in the graphic. If participants marked the correct answer of one question, one point was given. We used the number of correctly answered questions as estimator for comprehension. Overall, learners could achieve a maximum of 30 points.

2.6 Statistical Models

To answer our research questions, we conducted a number of mediation models with (perceived) mental load and mental effort as predictors, comprehension as the mediator and metacomprehension as criterion (all standardised; **Figure 1**). We used the jamovi 1.1.9.0 JAMM package and 10,000 percentile bootstrapping and 95% CI to estimate beta. We first used all data for an overall model and then computed one model for each multimedia condition. α was set at 5%.

3 RESULTS

Before we report the model results, we report relevant descriptive statistics and the results of testing for differences in the variables between the conditions.

3.1 Descriptive Statistics and Group Differences

Table 3 shows the descriptive statistics for the relevant variables by condition. We can see that—in contrast to prior assumptions assuming active integration fosters comprehension—students in the active integration condition performed worst at the comprehension test and accordingly also judged their comprehension lowest.

We conducted Welch's ANOVA to test for differences between the groups in terms of (perceived) mental effort, (perceived) mental load, comprehension or metacomprehension (cf. **Table 3**). We found mental effort and mental load not to differ significantly. Thus, although the conditions were meant to induce various levels of cognitive load, the overall level did not differ. However, this was not the case for comprehension [F (2,101) = 5.75, $p = 0.004$] or metacomprehension [F (2,101) = 3.19, $p = 0.045$]. Games Howell post-hoc test confirmed a difference between the learners working with the active integration material and those using integrated material [t (98.1) = 3.24, $p = 0.005$] for

TABLE 3 | Descriptive statistics by experimental condition and group differences.

Variable	Active integration		Split attention		Integrated		Group differences	
	(*n* = 53)		(*n* = 52)		(*n* = 51)		*F* (2,101)	*p*
	M	*SD*	*M*	*SD*	*M*	*SD*		
Comprehension	10.47	4.07	12.50	5.06	13.29	4.79	5.75	0.004
Metacomprehension	3.82	1.27	4.09	1.27	4.50	1.43	3.19	0.045
Mental effort	4.81	1.01	5.01	0.85	4.94	0.88	0.62	0.541
Mental load	4.21	1.17	4.10	1.02	4.15	0.83	0.12	0.885

TABLE 4 | Games Howell post-hoc test for group differences.

Group comparison	Comprehension			Metacomprehension		
	t	*df*	*p*	*t*	*df*	*p*
Active integration vs. Integrated	3.24	98.1	0.005	2.53	99.7	0.034
Active integration vs. Split attention	2.26	97.7	0.066	1.07	103.0	0.535
Integrated vs. Split attention	0.82	100.9	0.693	1.53	99.2	0.283

TABLE 5 | Intercorrelations between variables.

	Comprehension			Metacomprehension			Mental effort		
	r	95% CI	*p*	*r*	95% CI	*p*	*r*	95% CI	*p*
C	—	—	—	—	—	—	—	—	—
MC	0.571	(0.455, 0.668)	<0.001	—	—	—	—	—	—
ME	0.260	(0.107, 0.401)	0.001	0.267	(0.114, 0.407)	<0.001	—	—	—
ML	–0.291	(–0.428, –0.140)	<0.001	-0.512	(–0.620, –0.386)	<0.001	–0.028	(–0.184, 0.139)	0.730

C = Comprehension; MC = Metacomprehension; ME = Mental effort; ML = Mental load.

comprehension with learners with the active integration material scoring significantly worse in the comprehension test, but not between learners working with the active integration material and those learning with split-attention material or between those with split-attention and integrated material. For metacomprehension, the picture was similar. Again, learners with the active integration material had significantly lower ratings than learners with integrated material [t (99.7) = 2.53, p = 0.034], but not than learners with split attention material and the latter groups did also not differ. See **Table 4** for the full results of the post-hoc tests.

We further conducted correlation analyses (cf. **Table 5** for full results). As expected, comprehension and metacomprehension showed a rather high correlation (r = 0.571). Mental effort (r = 0.260) and mental load (r = −0.291) rather weakly correlated with comprehension. They did not correlate with each other significantly (r = −0.028). Metacomprehension, however, showed a rather weak correlation with mental effort (r = 0.267) but a considerably stronger, albeit negative, association with mental load (r = −0.512).

As the active integration condition was not informationally equivalent to the other conditions due to the missing link between text boxes and correct placements, we additionally checked how many of the students managed to correctly connect textual and pictorial information. Results showed that by the end of the learning phase, 39 (74%) participants had correctly placed at least 20 text boxes in the respective fields (meaning all assignments were ascertained), 11 (21%) had not, and 3 (6%) could not be confidently analysed due to a logging failure.

3.2 Mediation Analyses

The mediation analyses were conducted using jamovi's JAMM package. We used z-standardised variables and used 10,000 percentile bootstrapping to estimate beta-coefficients.

3.2.1 Mediation Model (all Conditions)

To assess if and how perceived mental effort and load affected metacomprehension judgments (beyond their actual effects on comprehension), we first computed one mediation model over all conditions, disregarding possible differences. We found all paths to be highly significant and both direct and indirect effects of mental effort (positive) and mental load (negative) on metacomprehension (**Table 6** and **Figure 3**). While the indirect effects of effort and load seem to be comparable in size (although not direction) due to the comparable effects of load and effort on comprehension, the direct effect of load seems considerably larger than the effect of effort, indicating that learners' metacognitive judgments seem to be more sensitive to (perceiving) mental load than mental effort. In general, perceived mental effort affected metacomprehension positively (β =

TABLE 6 | Full mediation model (all conditions).

Effects[b]		β	*SE*	95% CI[a]		*z*	*p*
				Lower bound	Upper bound		
Indirect	ME → C → MC (a_2*b)	0.106	0.036	0.040	0.181	2.96	0.003
	ML → C → MC (a_1*b)	–0.119	0.034	–0.189	–0.056	–3.51	<0.001
Component	ME → C (path a_2)	0.252	0.079	0.096	0.406	3.19	0.001
	C → MC (path b)	0.421	0.064	0.296	0.546	6.59	<0.001
	ML → C (path a_1)	–0.284	0.071	–0.416	–0.137	–3.98	<0.001
Direct	ME → MC (path c_2')	0.147	0.058	0.038	0.266	2.54	0.011
	ML → MC (path c_1')	–0.386	0.061	–0.509	–0.267	–6.29	<0.001
Total	ME → MC (path c_2)	0.253	0.066	0.123	0.382	3.83	<0.001
	ML → MC (path c_1)	–0.505	0.066	–0.635	–0.376	–7.66	<0.001

[a]*CIs based on 10,000 percentile bootstrapping samples.*
[b]*ME = (perceived) Mental Effort; ML = (perceived) Mental Load; C = Comprehension; MC = Metacomprehension.*

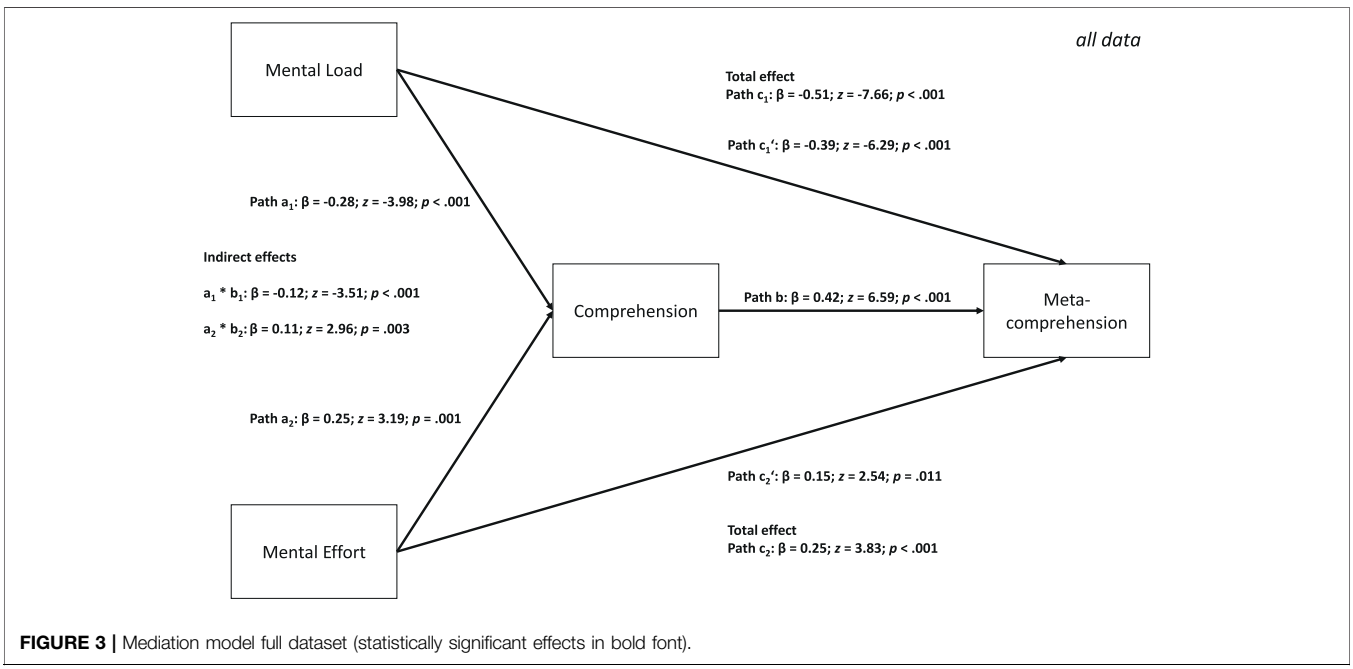

FIGURE 3 | Mediation model full dataset (statistically significant effects in bold font).

0.25, $p < 0.001$) and perceived mental load negatively ($\beta = -0.51$, $p < 0.001$), thus confirming hypotheses 1 and 2.

3.2.2 Mediation Models per Condition

To see if the effects vary under different load-imposing multimedia conditions, we then computed mediation models for each condition separately. An overview of the results can be seen in **Figure 4** (active integration format), **Figure 5** (split attention format), and **Figure 6** (integrated format), the full results in **Table 7**. As can be seen, within all conditions, there was a clear effect of comprehension on metacomprehension (b) and although not exactly equal, the size of the effect is roughly comparable, albeit descriptively a bit larger for learners using integrated material. There was also a direct negative effect of mental load on metacomprehension (c1), descriptively larger for learners using active integration material, even when indirect effects were eliminated (c1'; cue utilization). However, as mental load seems to not be very indicative of comprehension for learners with the split attention and integrated format (a1; diagnosticity), indirect effects of mental load on metacomprehension via comprehension were only confirmed for learners within the active integration condition (partial mediation).

For mental effort, this looked somewhat different. Here, mental effort only seems to be indicative of comprehension for learners provided with the integrated material (a2) and thus, the indirect effect was statistically significant only for those. However, while the indirect effect fully explained the effect of mental effort on metacomprehension for learners with integrated material, learners with split attention material showed a significant direct effect of mental effort on metacomprehension not explained by comprehension (c2'). Thus, even lacking diagnosticity for the latter, for learners with

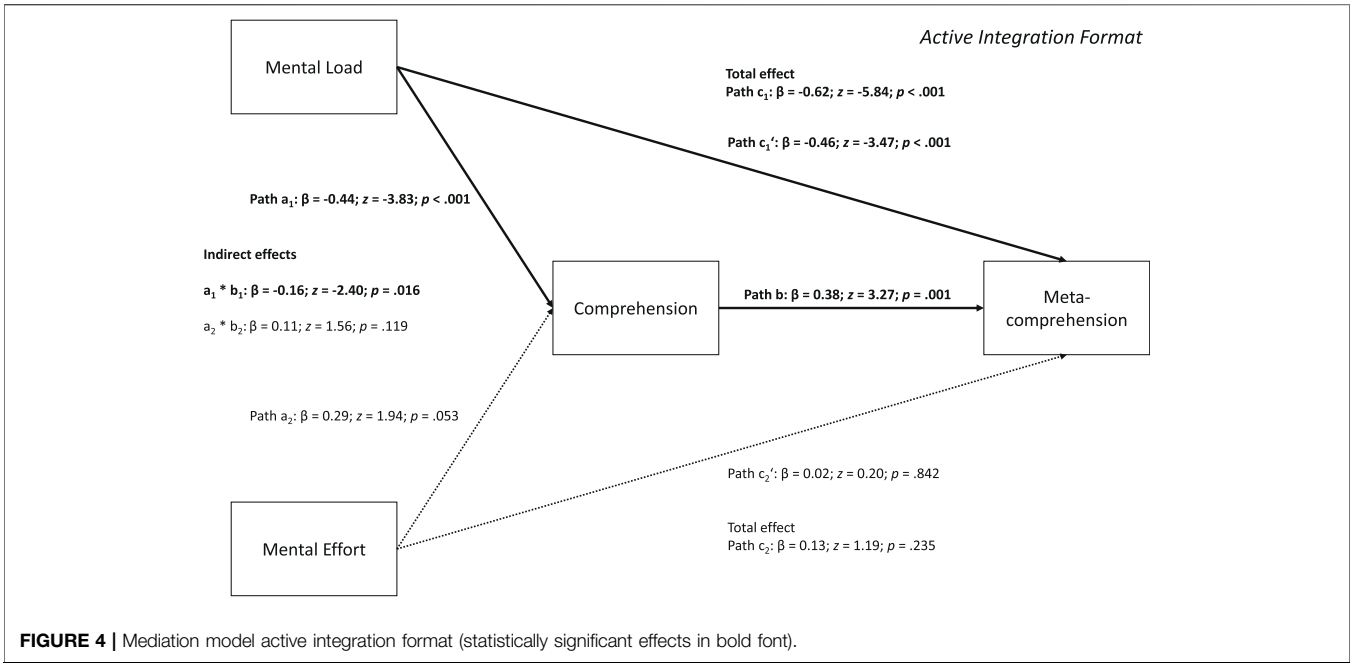

FIGURE 4 | Mediation model active integration format (statistically significant effects in bold font).

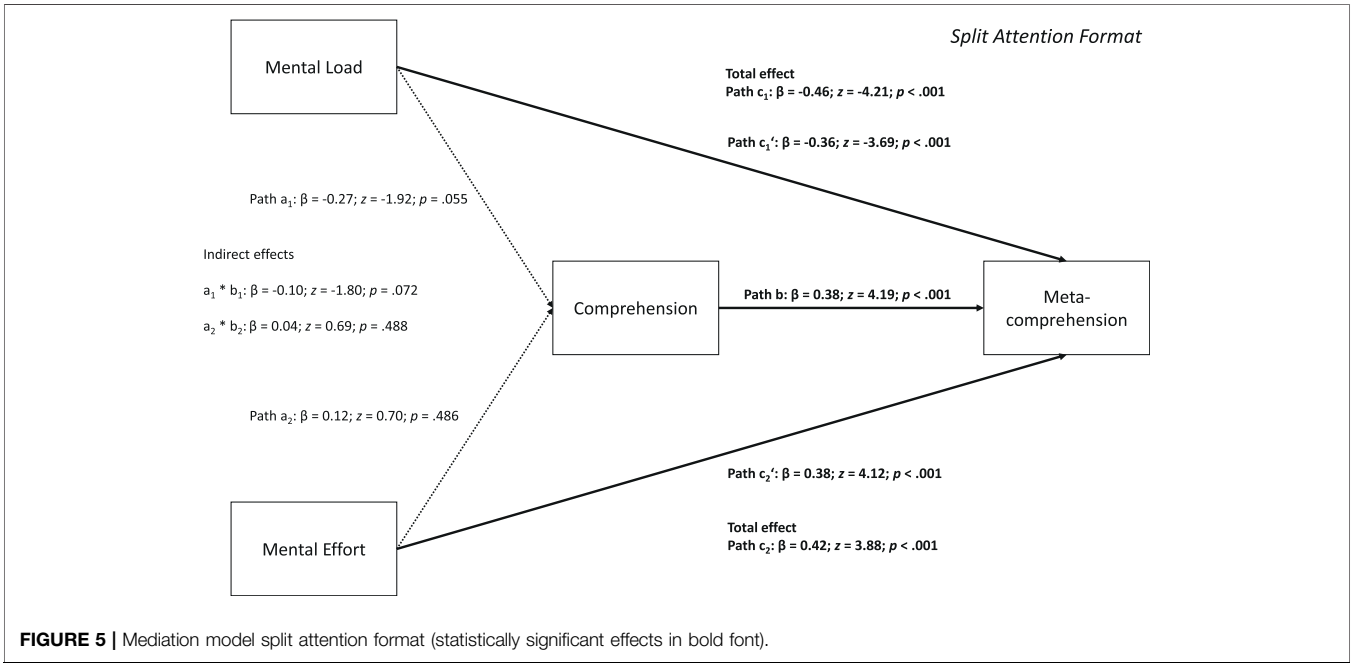

FIGURE 5 | Mediation model split attention format (statistically significant effects in bold font).

split attention material mental effort seems to be a relevant factor for metacomprehension judgments. For learners with active integration material, mental effort seems to be less relevant.

As statistically securing smaller effects may have been hampered by the lower power of the separate mediation models and the described differences may not be generalised beyond the sample, we conducted further post-hoc analyses (see **Section 3.2.3**).

3.2.3 Post-Hoc Comparisons of the Models

To test if the (direct) effects between mental load and effort on metacomprehension differed significantly between the conditions, we integrated condition as a moderator on the c' path using 10,000 percentile bootstrapping. We contrasted all conditions separately and used bonferroni corrected alpha-levels (0.017) to adjust for alpha-error inflation. Results for the moderations showed a significant moderation effect only for the split attention versus active integration contrast (for full

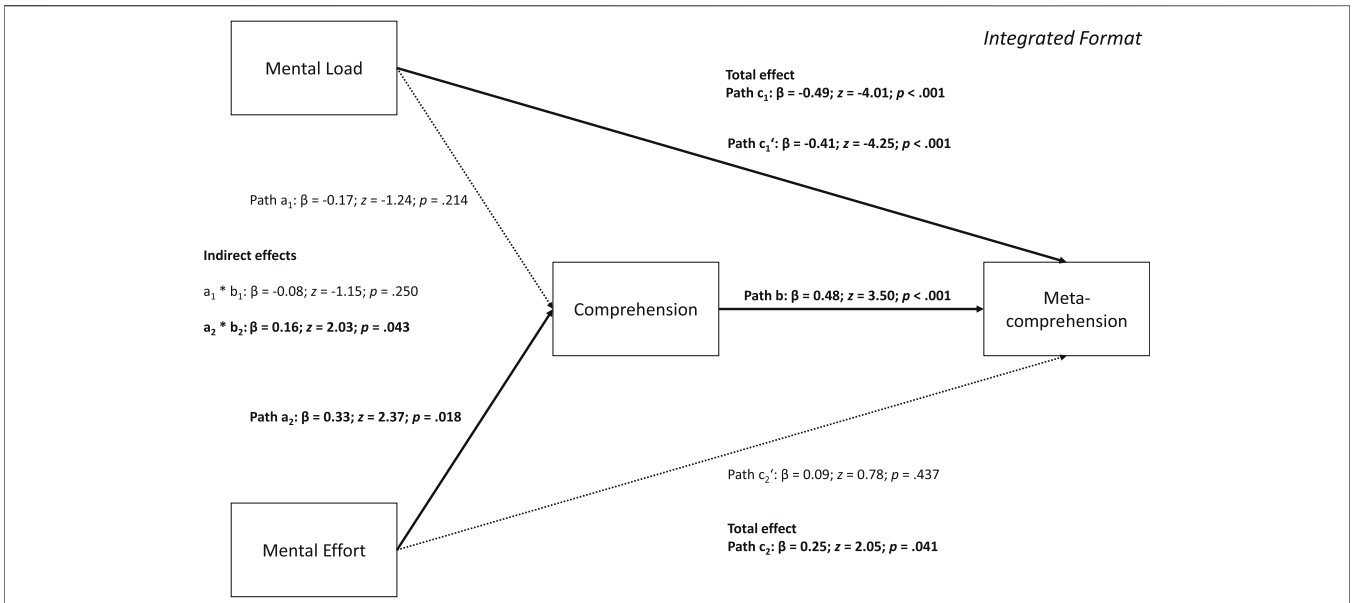

FIGURE 6 | Mediation model integrated format (statistically significant effects in bold font).

TABLE 7 | Mediation model by experimental condition.

Effects[b]		β	*SE*	95% CI[a]		*z*	*p*
				Lower bound	Upper bound		
Active integration format							
Indirect	ME → C → MC (a_2*b)	0.110	0.070	0.010	0.279	1.56	0.119
	ML → C → MC (a_1*b)	−0.164	0.068	−0.311	−0.043	−2.40	0.016
Component	ME → C (path a_2)	0.292	0.151	0.047	0.623	1.94	0.053
	C → MC (path b)	0.376	0.115	0.125	0.573	3.27	0.001
	ML → C (path a_1)	−0.436	0.114	−0.658	−0.206	−3.83	<0.001
Direct	ME → MC (path c_2')	0.017	0.084	−0.176	0.162	0.20	0.842
	ML → MC (path c_1')	−0.457	0.132	−0.705	−0.190	−3.47	<0.001
Total	ME → MC (path c_2)	0.126	0.106	−0.082	0.335	1.19	0.235
	ML → MC (path c_1)	−0.621	0.106	−0.830	−0.413	−5.84	<0.001
Split attention format							
Indirect	ME → C → MC (a_2*b)	0.042	0.061	−0.095	0.148	0.69	0.488
	ML → C → MC (a_1*b)	−0.100	0.056	−0.202	0.023	−1.80	0.072
Component	ME → C (path a_2)	0.112	0.160	−0.228	0.397	0.70	0.486
	C → MC (path b)	0.376	0.090	0.191	0.549	4.19	<0.001
	ML → C (path a_1)	−0.266	0.139	−0.488	0.058	−1.92	0.055
Direct	ME → MC (path c_2')	0.379	0.092	0.212	0.576	4.12	<0.001
	ML → MC (path c_1')	−0.357	0.097	−0.567	−0.188	−3.69	<0.001
Total	ME → MC (path c_2)	0.421	0.109	0.209	0.634	3.88	<0.001
	ML → MC (path c_1)	−0.457	0.109	−0.670	−0.244	−4.21	<0.001
Integrated format							
Indirect	ME → C → MC (a_2*b)	0.155	0.077	0.007	0.307	2.03	0.043
	ML → C → MC (a_1*b)	−0.081	0.070	−0.240	0.038	−1.15	0.250
Component	ME → C (path a_2)	0.327	0.138	0.019	0.565	2.36	0.018
	C → MC (path b)	0.476	0.136	0.195	0.733	3.50	<0.001
	ML → C (path a_1)	−0.169	0.137	−0.457	0.080	−1.24	0.214
Direct	ME → MC (path c_2')	0.094	0.120	−0.122	0.348	0.78	0.437
	ML → MC (path c_1')	−0.408	0.096	−0.608	−0.228	−4.25	<0.001
Total	ME → MC (path c_2)	0.250	0.122	0.010	0.487	2.05	0.041
	ML → MC (path c_1)	−0.488	0.122	−0.726	−0.250	−4.01	<0.001

[a]*CIs based on 10,000 percentile bootstrapping samples.*
[b]*ME = (perceived) Mental Effort; (perceived) ML = Mental load; C = Comprehension; MC = Metacomprehension.*

TABLE 8 | Post-hoc moderation effects on direct effect (c') on metacomprehension.

Contrast	IV: Mental load			IV: Mental effort		
	β	*z*	*p*[a]	β	*z*	*p*[a]
Active integration vs. Integrated	–0.085	–1.14	0.255	0.056	0.77	0.444
Integrated vs. Split attention	–0.103	–1.42	0.155	–0.114	–1.58	0.114
Active integration vs. Split attention	0.037	0.50	0.616	0.190	3.04	0.002

[a]p *is not corrected, we used adjusted alpha-levels of .017 to account for alpha-error inflation.*

results **Table 8**) with the direct effect of mental effort on metacomprehension being highly significant within the split attention condition and substantially smaller and non-significant within the active integration condition (moderation effect: $\beta = 0.19$, $z = 3.04$, $p = 0.002$). While the integrated conditions' effect descriptively ranged somewhere in between (**Table 7**), the difference was not statistically significant between the integrated and any of the other conditions. There were no significant moderation effects on the c' path between mental load and metacomprehension.

Please note that even though mental load and mental effort did not differ significantly between our experimental conditions in the study and the predictors showed no sign of multicollinearity, possible influences of the conditions (moderator) on the predictors cannot be completely discarded and the results of the moderation models thus need to be treated with caution and may only provide a first indicator on possible moderation effects.

4 DISCUSSION

4.1 Effects on Comprehension

In general, as expected, we found the relationships between (perceived) mental effort or mental load and metacomprehension to be positive for the former and negative for the latter (hypotheses 1 and 2) and a positive relationship between comprehension and metacomprehension. Comparable to other research, mental effort and load were not related (e.g., Minkley et al., 2021). Although the effects of mental load and mental effort on comprehension were not large, the relationship for both but especially for mental load seems to be comparable to findings in the literature (e.g., Krell, 2017; Minkley et al., 2021). This confirms basic research on cognitive load and multimedia learning, but also highlights the relevance of mental effort exerted by learners as a distinct construct apart from mental load (e.g., Paas and van Merriënboer, 1994). Thus, it is important to separately measure and include both when studying (multimedia) learning. While this did not hold true for each experimental condition separately and thus may vary under various load-inducing designs of learning material, a lack of statistical power may have contributed to these observed differences.

In the two load-inducing conditions (split attention and active integration), mental effort seems to have less effect on comprehension (less diagnosticity) and although smaller effects might have been statistically secured with more power to the analyses, especially for the split attention condition, effort seems to be rather futile. This is especially interesting as the level of mental effort and mental load did not differ between the condition, so all groups reported to have exerted similar effort, but to different avail. It seems that if learning material is designed to reduce cognitive load, the effort learners put into the learning process shows the greatest effect, while it is less effective when learners are asked to actively integrate the information and least effective under a split attention format, presumably because most of the effort exerted is wasted on processes not directly relevant for learning (unproductive).

For mental load, the picture was different. Here, (negative) effects on comprehension were especially strong within the active integration condition and were smaller with the split attention format and smallest in the integrated format. It seems that if actively integrating text and graphics put extra load on the learner, comprehension suffered. This is interesting, as it was assumed that actively integrating information would rather induce productive learning processes and successful conduction would thus foster comprehension (cf. Bodemer et al., 2004; Bodemer et al., 2005). However, the active integration condition showed lower comprehension after the learning phase (a descriptive difference which was also statistically confirmed in comparison with the integrated format) and it is thus possible, that due to the complexity of the material, they never overcame the split attention format (without proper assignment) and just did not manage to benefit from the design. Although most participants managed to integrate the information (at least physically), a considerable percentage did not manage to integrate the information within the given timeframe. This gave these learners a further disadvantage as without correct assignments, the conditions were not informationally equivalent.

4.2 Effects on Metacognition

As for metacomprehension, we did confirm effects of comprehension on metacomprehension and although somewhat stronger in the integrated format, the relationship between comprehension and metacomprehension was in general consistent with effects found in metacognition research (e.g., Schleinschok et al., 2017).

However, apart from comprehension, other factors contribute to explaining metacomprehension variance, for example mental effort. The first important observation is that the relationship between mental effort and metacomprehension is positive. This would be expected assuming learners are aware of the positive influence invested mental effort has on comprehension and learning. However, Baars and colleagues (2013) found (strong)

negative correlations between judgments of learning and mental effort using the one-item scale by Paas (1992). Explicitly differentiating between mental load imposed and mental effort invested by using the instrument by Krell (2015) may have led to a more refined picture, shifting the focus more clearly on effort as a voluntary activity ["I have given my best (. . .)"] as opposed to task difficulty ["The tasks were easy (. . .)"]. However, this also means that a differential view is necessary to understand how learners form metacognitive judgments. A recent meta-analysis by Baars et al. (2020) comes to a similar conclusion: They find that the usually found negative relationship between metacognitive judgment and effort vanished when considering rating scales promoting self-agency. Thus, when effort regulation is goal-driven, said effort may be interpreted positive, while this may revert when it is data-driven (Koriat, 2018; de Bruin et al., 2020). The scale by Krell (2015) used in our study arguably conceptually aligns with a self-agent view and thus, the found relationship seems to align with findings in literature.

Learners in the integrated format condition seem to have (rightfully) used their mental effort as an indicator for comprehension. While our research design did not allow to ascertain cue usage in an intentional way, the effect of mental effort on metacomprehension is mediated by actual comprehension. However, for learners learning with material in a split attention format effort also seems to influence metacomprehension, although it does seem less diagnostic for comprehension. Thus, while these learners judge their comprehension higher when exerting more effort, actual comprehension was rather unfazed. Such mechanisms could potentially lead to misjudgments and lower monitoring accuracy, which may severely hamper self-regulated learning attempts when making study decisions (Thiede et al., 2003; Schnaubert and Bodemer, 2017). For learners using an active integration format, mental effort was not significantly related to metacomprehension. Due to the design providing feedback during learning (explanation did only stick to correct places), learners may have experienced their efforts being more or less fruitful during the learning process. While this warrants further investigation, feedback has previously been found to not only correct faulty assumptions, but also metacognitive errors (e.g., Butler et al., 2008). The difference between the split attention and active integration condition with regard to the effect of mental effort was also confirmed by the post-hoc moderation analyses which showed a significant effect of condition on the direct relationship between mental effort and metacomprehension (when the mediation effect was excluded). While these results have to be treated with caution, it seems that multimedia design may affect the relationship between mental effort invested and metacomprehension reported.

Mental load is strongly connected to metacomprehension under all conditions of learning material provided, and this relationship is even higher than the one between comprehension and metacomprehension. This may be an indicator for learners using mental load as a cue for judging their comprehension. While at least for active integration, a significant part of this is mediated by comprehension, a large portion of (negative) effect is unwarranted. Learners may overestimate the negative effect of load on comprehension, especially when learning with integrated, theoretically less load-inducing multimedia learning material. Again, this could lead to severe misjudgments and hamper self-regulated learning.

4.3 Mental Load, Mental Effort and Metacognition

Overall, these results show that mental load and mental effort are distinct concepts related differently to metacomprehension and should be studied in line with other cognitive experiences regarded as cues for metacognitive judgments like ease-of-processing or ease-of-retrieval (e.g., Benjamin and Bjork, 2014 see also Koriat and Ma'ayan, 2005). While the causal relationship cautiously assumed in our models are based on theoretical assumptions rather than experimental design, there is merit in combining research on metacognition with the vast amount of research regarding cognitive load and multimedia learning. However, there are some conceptual issues that may need specific attention when doing so (e.g., Sweller and Paas, 2017). One of these is the perspective taken on the constructs involved. Cognitive load research is mostly concerned with (working) memory processes and resources. Thereby, cued self-report is a means to gather information about cognitive processes assuming these are not only accessible, but also transformable to a given scale (for more detail, see Paas et al., 2008). This is inherently different for metacognition research. Here, the subjective view on cognitive constructs (like comprehension) is not a measurement issue, but inherent in the concept (Nelson and Narens, 1990; Nelson and Narens, 1994). Within metacognition research, metacomprehension scales are not meant to measure comprehension but an individual's unique perception of their own comprehension. What would a metacognitive view on cognitive load look like? If we assume mental load and mental effort to be directly accessible for learners (as assumed when using self-report to measure it), is awareness of this as a metacognitive construct just an epiphenomenon with no impact or does it actually affect learning processes? Our research as well as other current approaches viewing mental effort through a metacognitive lens (e.g., Scheiter et al., 2020) suggest that monitoring mental load and mental effort is more than just a fall-out, but may actually be relevant for forming metacognitive judgments. Although requiring further empirical research, in line with the cyclic model of metacognitive regulation assumed in metacognition theory (e.g., Nelson and Narens, 1990), this means that learners may use this information to regulate their learning processes, for example by exerting effort, allocating study time and diverting attention. This brings us to another understudied yet highly discussed question of how learners regulate mental effort itself and how this may depend on their perception of cognitive load (de Bruin and van Merriënboer, 2017; see also de Bruin et al., 2020). Including multimedia research and building on well-established multimedia design effects may help to strategically design further studies to investigate not only how perceived mental load and effort affect metacognitive judgments, but how sensitive these relationships are to differences in

diagnosticity of perceived mental load and effort for actual learning gains.

While our research showed first connections between perceived mental load, perceived mental effort, comprehension and metacomprehension under various load-inducing design conditions of multimedia learning material, there are several open questions that need further discussion. First, while the various effects under specific design conditions all need replications with different material and substantial statistical power to validly generalise the effects for more or less demanding content, especially the results of active integration seem puzzling. While the lack of beneficial effects of active integration may have been due to learners not performing mental activities while behaviorally integrating the material, active integration as a means to induce productive learning processes may also be more or less effective under various conditions relating to the content or the learners. Although within our sample, learners did not provide ratings indicative of "overload", it should be taken into consideration that these productive learning processes that were attempted to be triggered did not take place or were even hindered due to the complexity of the task or low prior knowledge (prior knowledge was rated quite low within our sample, see **Section 2.1**). As with other multimedia effects (e.g., imagination effect; Lin et al., 2017), there could be reversed effects depending on prior knowledge (i.e., expertise reversal; Kalyuga, 2007) and thus, prior knowledge should be taken into consideration in further studies. Second, other interindividual differences may need to be considered as well. Cue utilization theory does not make specific assumptions about interindividual differences, but research has found that not all learners use the same cues. Although self-reported cue use needs to be viewed with caution, Thiede and colleagues (2010), for example, found especially high-risk readers reported using more surface level cues (i.e., relating to surface properties of a text) than typical readers. This strengthens the notion that not only do learners vary in their cue use, but also that this may be trainable (Wiley et al., 2016). A current meta-analysis found that interventions supporting learners' use of situation-model cues (i.e., cues relating to the learners' situation model constructed during text comprehension; cf. Kintsch, 1994) positively and considerably affect metacomprehension accuracy (Prinz et al., 2020). While we were rather interested in how perceived mental load and effort affected metacomprehension judgments under various load-inducing conditions, a next step would be to assess how this affects metacomprehension accuracy within students.

4.4 Limitations

As with all research, there are limitations to this study. First of all, our sample size especially for the single mediation models was too small to confirm smaller effects. While we can draw some conclusions about the processes involved, a final verdict needs careful replications of the found differences between the conditions to confirm and potentially generalise the effects. Additionally, we could not confirm differences in (perceived) mental effort or mental load between conditions, which could be due to incorrect assumptions about the relationship between treatment and cognitive load (which is improbable for split attention and the integrated format, but may hold true for the less studied active integration format). However, it could also be due to the assessment of load and effort, which may have been invariant to more subtle changes, especially since it was measured with a delay. However, the instrument had shown sensitivity to instructional variations before (Krell, 2017) and we did find expected (albeit rather small) relationships with comprehension giving at least small indications of valid assessment. Additionally, differences in how learners' effort and load was related to metacomprehension between conditions indicate some sensitivity of the (perceived) mental load and mental effort measurement to changes. While there are other measures of cognitive load, we opted for the questionnaire by Krell (2015), as it differentiates between mental load imposed and effort deliberately invested, which are both central when studying self-regulatory processes in multimedia learning. Due to the intricate relationship between the requirements of a task and regulatory processes by the learner, the relationship between cognitive load and metacognition is a source for some debate (e.g., de Bruin and van Merriënboer, 2017; Seufert, 2020). Thus, it seems vital to differentiate between material-induced mental load and invested mental effort (see also Seufert, 2020) as the latter may be regulated by learners (de Bruin et al., 2020) while the former much more relies on the instructional design (although both may affect each other). Further studies may put more focus on the type of load imposed by the material and the reasons for investing mental effort in the learning task to distinguish more or less beneficial load conditions [see also discussion by Seufert (2020)], possibly linking more active (effort) and more passive (load) aspects of cognitive load to the assumed tripartite nature of the cognitive load concept (see Klepsch and Seufert, 2021). As the assessment (and structure) of cognitive load is subject to an ongoing debate (see for example Kirschner et al., 2011), a combination of multiple measures may prove useful, and while objective measures may provide reliable information for instructional design and multimedia research (e.g., Korbach et al., 2017), subjective measures additionally provide insights about subjective judgments of effort or load (Scheiter et al., 2020) that are relevant when the aim is to understand how learners themselves view their learning process and regulate their behaviour. Additionally, various subjective measures focus on different aspects of load and effort that may lead to very different empirical results (e.g., elements of self-agency, Baars et al., 2020).

Rather than a limitation, a further open question concerns the role of restricting study time on the results found. It is not uncommon that multimedia effects are more pronounced under conditions of system-paced time pressure (e.g., Rey et al., 2019). While a defined time frame allowed us to minimise effects of self-regulatory processes (i.e., study time decisions) to affect comprehension and thus fostered experimental scrutiny within our setup, Baars and colleagues' (2020) meta-analysis found the (negative) relationship between mental effort and monitoring judgments to diffuse under time pressure. Following de Bruin and colleagues (2020) argumentation, our setup could have reinforced a positive interpretation of effort by fostering a more goal-driven approach. Coupled with our use of a mental effort scale

including elements of self-agency (see above), this may have reinforced the positive association and may not be generalisable to settings allowing for self-paced study.

Another limitation to be discussed are the potentially incorrect assumptions about the use of the active integration format. While we did assume learners to actively integrate the material, which had previously been shown to be beneficial for learning (e.g., Bodemer et al., 2004) and roughly 3/4 of learners did manage to correctly assign the boxes, learners may have just actively dragged and dropped the boxes without mentally integrating the content (trial and error). Such mere behavioural activity however (correctly placing the boxes) is not in itself conducive to learning, but has to be accompanied by cognitive activity (mentally connecting concepts and building coherent mental representations) to show the assumed positive effects (Mayer, 2001). Thus, apart from participants simply not managing to integrate the information (See **Section 4.1**), participants focussing on behavioural (versus cognitive) activities may have contributed to the findings in our study and explain why test performance was worst for the active integration condition.

One further and central limitation of the study is that we cannot rule out that further variables affect the found relationships or that the direction of effect may be different than assumed. While we based our model on theoretical assumptions about (causal) processes, our empirical model can only confirm the relationship rather than the mechanisms involved. While we randomised participants' allocation to experimental conditions and found differences between metacomprehension ratings as a result, that does not exclude the possibility that learners perceiving themselves as more potent and therefore provide higher metacomprehension ratings may have exerted more mental effort during learning rather than using effort as a cue to judge their comprehension. Further, additional variables may affect the found results. For example, Zu and colleagues found that prior knowledge may influence how learners respond to cognitive load items affecting the factor structure (Zu et al., 2021), which may simultaneously affect metacomprehension. Although the authors used a different survey that was designed to measure the tripartite nature of cognitive load, prior knowledge may have affected our measures of cognitive and metacognitive processes as well. While load imposed by the design of learning material may be manipulated in experimental settings, disentangling metacognitive monitoring and invested effort is harder to accomplish as it is assumed learners actively regulate their mental effort during learning (de Bruin and van Merriënboer, 2017), but would give further insights into the relationship between cognitive processes during multimedia learning and metacognitive regulation. Directly manipulating mental load and effort without running the risk of the intervention affecting metacomprehension simultaneously may be a hard to accomplish goal, but would be pertinent to ascertain causality as implied by the theoretical model assumptions underlying the cue utilization model proposed. Additionally, while we decided upon the order of assessing our variables with possible sequence effects in mind, we cannot fully exclude possible interferences. For example, the motivational questionnaire may have had unexpected effects when learners judge their perceived competence or interest. Allowing for more scrutiny, follow-up research may use within-subject variations of multimedia-material, on-time measurements of effort and load and additional metacognitive measures and explicitly test for sequence effects or randomise the order of assessment to provide further insights into how learners take variations of multimedia design into account when monitoring their comprehension during multimedia learning.

5 CONCLUSION

Our research suggests that the subjective perceptions of mental load and mental effort are not epiphenomenons only to be considered for assessment purposes, but that these perceptions may impact learners' self-regulatory processes. Although it would be inappropriate to conclude an intentional cue usage from the data collected in this study, independent of the validity of the subjective measures (i.e., their relation to actual mental load and mental effort), experiences of mental load and invested effort may inform learners and may be used as cues when making monitoring judgments. As put by Nelson and Narens (1990, p .128): "A system that monitors itself (even imperfectly) may use its own introspections as input to alter the system's behavior". We argue that while the validity of subjective measures may be a major concern for research on cognitive load, the value of information about the subjective experience of those processes is underrated and warrants further empirical (and possibly experimental) research. Thus, with validity of assessments referring to interpretations and usage of scores (Kane, 2013), a metacognitive perspective on cognitive load shifts the focus from assessing cognitive processes to assessing their subjective experience by learners (please note that the term "experience" in this context refers to conscious perceptions, but is also being used to differentiate passive and active forms of load; e.g., Klepsch and Seufert, 2021). The found relationships between perceived mental load, perceived mental effort and metacomprehension indicate not only that perceived mental load and effort are distinct concepts, but that they may play an additional role for learning which hinges on their subjective experience by learners. Although the direction of effect cannot conclusively be established within this study, a metacognitive view on cognitive load has implications for the interpretation of subjective measures of cognitive load. While studying the relationship between subjective and objective measures of cognitive load may be a matter of validating assessment strategies (e.g., Minkley et al., 2021), it also establishes a relationship between cognitive processes and their idiosyncratic experience. Applying Nelson and Narens' view on metacognition (Nelson and Narens, 1994), misalignment between the two is a distortion providing insights into how learners perceive their mental processes and may thus be studied in terms of metacognitive accuracy. Thus, validly

interpreting subjective measures needs to consider their subjectivity explicitly while studying possible distortions. Validity hinges on the interpretation of a score rather than the score itself (e.g., Kane, 2013). Consequently, when studying self-regulatory processes during (multimedia) learning, carefully implementing subjective rating scales may be more appropriate to capture the subjective experience than physiological measures. While this does not mean subjective rating scales may validly assess experiences of effort and load per se and for example scale characteristics (Ouwehand et al., 2021) or the framing with regard to self-agency (Koriat, 2018) may influence results, it stresses the need to differentiate between cognitive load and their impact on learning processes and its subjective experience, which may impact metacognitive regulation. Our study provided some indications for the relevance of experiences of mental load and effort for learning and thus calls for a conceptual rather than methodological differentiation between cognitive processing and their idiosyncratic experience when research targets learning regulation rather than working memory capacity.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee of the division of Computer Science and Applied Cognitive Sciences at the Faculty of Engineering at the University of Duisburg-Essen, Germany. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

LS: planned and supervised experiment, conducted majority of statistical analyses, wrote majority of the manuscript (focus on metacognition). SS: technical realisation of experiment, supported data analyses, wrote parts of the manuscript (focus on multimedia), proofreading.

ACKNOWLEDGMENTS

We want to thank Ebru Yilmaz and Sonja Glantz for their support in planning the study and collecting data, Peter Bellstedt for programming especially the active integration website, and Mareike Florax and Rolf Plötzner for the permission to use and adapt the learning material. We acknowledge support by the Open Access Publication Fund of the University of Duisburg-Essen.

REFERENCES

Anmarkrud, Ø., Andresen, A., and Bråten, I. (2019). Cognitive Load and Working Memory in Multimedia Learning: Conceptual and Measurement Issues. *Educ. Psychol.* 54, 61–83. doi:10.1080/00461520.2018.1554484

Ayres, P., and Youssef, A. (2008). "Investigating the Influence of Transistory Information and Motivation during Instructional Animations," in *Proceedings of the 8th International Conference of the Learning Sciences*. Editors P. A. Kirschner, F. Prins, V. Jonker, and G. Kanselaaer (Utrecht, Netherlands: International Society of the Learning Sciences), Vol. 1, 68–75.

Baars, M., van Gog, T., de Bruin, A., and Paas, F. (2017). Effects of Problem Solving after Worked Example Study on Secondary School Children's Monitoring Accuracy. *Educ. Psychol.* 37 (7), 810–834. doi:10.1080/01443410.2016.1150419

Baars, M., Visser, S., Gog, T. v., Bruin, A. d., and Paas, F. (2013). Completion of Partially Worked-Out Examples as a Generation Strategy for Improving Monitoring Accuracy. *Contemp. Educ. Psychol.* 38 (4), 395–406. doi:10.1016/j.cedpsych.2013.09.001

Baars, M., Wijnia, L., de Bruin, A., and Paas, F. (2020). The Relation between Students' Effort and Monitoring Judgments during Learning: A Meta-Analysis. *Educ. Psychol. Rev.* 32 (4), 979–1002. doi:10.1007/s10648-020-09569-3

Beege, M.; Colleagues (2019). Spatial Continuity Effect vs. Spatial Contiguity Failure. Revising the Effects of Spatial Proximity between Related and Unrelated Representations. *Front. Educ.* 4. doi:10.3389/feduc.2019.00086

Begg, I., Duft, S., Lalonde, P., Melnick, R., and Sanvito, J. (1989). Memory Predictions Are Based on Ease of Processing. *J. Mem. Lang.* 28 (5), 610–632. doi:10.1016/0749-596X(89)90016-8

Benjamin, A. S., Bjork, R. A., and Schwartz, B. L. (1998). The Mismeasure of Memory: When Retrieval Fluency Is Misleading as a Metamnemonic index. *J. Exp. Psychol. Gen.* 127 (1), 55–68. doi:10.1037//0096-3445.127.1.55

Benjamin, A. S., and Bjork, R. A. (2014). "Retrieval Fluency as a Metacognitive Index," in *Implicit Memory and Metacognition*. Editor L. M. Reder (New York, NY: Psychology Press), 321–350. doi:10.4324/9781315806136-19

Berthold, K., and Renkl, A. (2009). Instructional Aids to Support a Conceptual Understanding of Multiple Representations. *J. Educ. Psychol.* 101 (1), 70–87. doi:10.1037/a0013247

Bertsch, S., Pesta, B. J., Wiscott, R., and McDaniel, M. A. (2007). The Generation Effect: A Meta-Analytic Review. *Mem. Cognit* 35 (2), 201–210. doi:10.3758/BF03193441

Bjork, E. L., and Bjork, R. A. (2011). "Making Things Hard on Yourself, but in a Good Way: Creating Desirable Difficulties to Enhance Learning," in *Psychology and the Real World: Essays Illustrating Fundamental Contributions to Society* (New York, NY, US: Worth Publishers), 56–64.

Bodemer, D., Ploetzner, R., Bruchmüller, K., and Häcker, S. (2005). Supporting Learning with Interactive Multimedia through Active Integration of Representations. *Instr. Sci.* 33 (1), 73–95. doi:10.1007/s11251-004-7685-z

Bodemer, D., Ploetzner, R., Feuerlein, I., and Spada, H. (2004). The Active Integration of Information during Learning with Dynamic and Interactive Visualisations. *Learn. Instruction* 14 (3), 325–341. doi:10.1016/j.learninstruc.2004.06.006

Burkett, C., and Azevedo, R. (2012). The Effect of Multimedia Discrepancies on Metacognitive Judgments. *Comput. Hum. Behav.* 28 (4), 1276–1285. doi:10.1016/j.chb.2012.02.011

Butler, A. C., Karpicke, J. D., and Roediger, H. L., III (2008). Correcting a Metacognitive Error: Feedback Increases Retention of Low-Confidence Correct Responses. *J. Exp. Psychol. Learn. Mem. Cogn.* 34 (4), 918–928. doi:10.1037/0278-7393.34.4.918

De Bruin, A. B. H., Dunlosky, J., and Cavalcanti, R. B. (2017). Monitoring and Regulation of Learning in Medical Education: The Need for Predictive Cues. *Med. Educ.* 51 (6), 575–584. doi:10.1111/medu.13267

de Bruin, A. B. H., Roelle, J., Carpenter, S. K., and Baars, M. (2020). Synthesizing Cognitive Load and Self-Regulation Theory: A Theoretical Framework and Research Agenda. *Educ. Psychol. Rev.* 32 (4), 903–915. doi:10.1007/s10648-020-09576-4

de Bruin, A. B. H., and van Merriënboer, J. J. G. (2017). Bridging Cognitive Load and Self-Regulated Learning Research: A Complementary Approach to Contemporary Issues in Educational Research. *Learn. Instruction* 51, 1–9. doi:10.1016/j.learninstruc.2017.06.001

Dunlosky, J., and Lipko, A. R. (2007). Metacomprehension. *Curr. Dir. Psychol. Sci.* 16 (4), 228–232. doi:10.1111/j.1467-8721.2007.00509.x

Dunlosky, J., Rawson, K. A., and Middleton, E. L. (2005). What Constrains the Accuracy of Metacomprehension Judgments? Testing the Transfer-Appropriate-Monitoring and Accessibility Hypotheses. *J. Mem. Lang.* 52 (4), 551–565. doi:10.1016/j.jml.2005.01.011

Florax, M., and Ploetzner, R. (2010). What Contributes to the Split-Attention Effect? the Role of Text Segmentation, Picture Labelling, and Spatial Proximity. *Learn. Instruction* 20, 216–224. doi:10.1016/j.learninstruc.2009.02.021

Glenberg, A. M., and Epstein, W. (1985). Calibration of Comprehension. *J. Exp. Psychol. Learn. Mem. Cogn.* 11 (4), 702–718. doi:10.1037/0278-7393.11.1-4.702

Kalyuga, S. (2007). Expertise Reversal Effect and its Implications for Learner-Tailored Instruction. *Educ. Psychol. Rev.* 19 (4), 509–539. doi:10.1007/s10648-007-9054-3

Kalyuga, S., and Singh, A.-M. (2016). Rethinking the Boundaries of Cognitive Load Theory in Complex Learning. *Educ. Psychol. Rev.* 28, 831–852. doi:10.1007/s10648-015-9352-0

Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores. *J. Educ. Meas.* 50 (1), 1–73. doi:10.1111/jedm.12000

Kelemen, W. L. (2000). Metamemory Cues and Monitoring Accuracy: Judging what You Know and what You Will Know. *J. Educ. Psychol.* 92 (4), 800–810. doi:10.1037/0022-0663.92.4.800

Kintsch, W. (1994). Text Comprehension, Memory, and Learning. *Am. Psychol.* 49 (4), 294–303. doi:10.1037/0003-066X.49.4.29410.1037//0003-066x.49.4.294

Kirschner, P. A., Ayres, P., and Chandler, P. (2011). Contemporary Cognitive Load Theory Research: The Good, the Bad and the Ugly. *Comput. Hum. Behav.* 27 (1), 99–105. doi:10.1016/j.chb.2010.06.025

Klepsch, M., and Seufert, T. (2021). Making an Effort versus Experiencing Load. *Front. Educ.* 6, 56. doi:10.3389/feduc.2021.645284

Korbach, A., Brünken, R., and Park, B. (2017). Measurement of Cognitive Load in Multimedia Learning: a Comparison of Different Objective Measures. *Instr. Sci.* 45 (4), 515–536. doi:10.1007/s11251-017-9413-5

Koriat, A. (2018). Agency Attributions of Mental Effort during Self-Regulated Learning. *Mem. Cognit* 46 (3), 370–383. doi:10.3758/s13421-017-0771-7

Koriat, A., and Ma'ayan, H. (2005). The Effects of Encoding Fluency and Retrieval Fluency on Judgments of Learning. *J. Mem. Lang.* 52, 478–492. doi:10.1016/J.JML.2005.01.001

Koriat, A. (1997). Monitoring One's Own Knowledge during Study: A Cue-Utilization Approach to Judgments of Learning. *J. Exp. Psychol. Gen.* 126 (4), 349–370. doi:10.1037/0096-3445.126.4.349

Koriat, A., Nussinson, R., Bless, H., and Shaked, N. (2008). "Information-Based and Experience-Based Metacognitive Judgments" in *Handbook of Metamemory and Memory*. Editors J. Dunlosky and R. A. Bjork (New York, NY: Psychology Press), 117–135. doi:10.4324/9780203805503.ch7

Krell, M. (2017). Evaluating an Instrument to Measure Mental Load and Mental Effort Considering Different Sources of Validity Evidence. *Cogent Edu.* 4 (1), 1280256. doi:10.1080/2331186X.2017.1280256

Krell, P. (2015). Evaluating an Instrument to Measure Mental Load and Mental Effort Using Item Response Theory. *Sci. Edu. Rev. Lett.*, 1–16. doi:10.5771/9783845263991-1

Leppink, J., Paas, F., Van der Vleuten, C. P., Van Gog, T., and Van Merriënboer, J. J. (2013). Development of an Instrument for Measuring Different Types of Cognitive Load. *Behav. Res. Methods* 45 (4), 1058–1072. doi:10.3758/s13428-013-0334-1

Li, J., Antonenko, P. D., and Wang, J. (2019). Trends and Issues in Multimedia Learning Research in 1996-2016: A Bibliometric Analysis. *Educ. Res. Rev.* 28, 100282. doi:10.1016/j.edurev.2019.100282

Lin, L., Lee, C. H., Kalyuga, S., Wang, Y., Guan, S., and Wu, H. (2017). The Effect of Learner-Generated Drawing and Imagination in Comprehending a Science Text. *J. Exp. Edu.* 85 (1), 142–154. doi:10.1080/00220973.2016.1143796

Maki, R. H., and Berry, S. L. (1984). Metacomprehension of Text Material. *J. Exp. Psychol. Learn. Mem. Cogn.* 10 (4), 663–679. doi:10.1037/0278-7393.10.4.66310.1037//0278-7393.10.4.663

Maki, R. H. (1998). "Test Predictions over Text Material," in *Metacognition in Educational Theory and Practice*. Editors D. J. Hacker, J. Dunlosky, and A. C. Graesser (Lawrence Erlbaum Associates Publishers), 117–144.

Mayer, R. E. (2001). "A Cognitive Theory of Multimedia Learning," in *Multimedia Learning*. Editor R. E. Mayer (Cambridge, UK: Cambridge University Press), 41–62.

Mayer, R. E., and Moreno, R. (2003). Nine Ways to Reduce Cognitive Load in Multimedia Learning. *Educ. Psychol.* 38 (1), 43–52. doi:10.1207/S15326985EP3801_6

Metcalfe, J. (2009). Metacognitive Judgments and Control of Study. *Curr. Dir. Psychol. Sci.* 18 (3), 159–163. doi:10.1111/j.1467-8721.2009.01628.x

Minkley, N., Xu, K. M., and Krell, M. (2021). Analyzing Relationships between Causal and Assessment Factors of Cognitive Load: Associations between Objective and Subjective Measures of Cognitive Load, Stress, Interest, and Self-Concept. *Front. Educ.* 6, 53. doi:10.3389/feduc.2021.632907

Mutlu-Bayraktar, D., Cosgun, V., and Altan, T. (2019). Cognitive Load in Multimedia Learning Environments: A Systematic Review. *Comput. Edu.* 141, 103618. doi:10.1016/j.compedu.2019.103618

Naismith, L. M., Cheung, J. J., Ringsted, C., and Cavalcanti, R. B. (2015). Limitations of Subjective Cognitive Load Measures in Simulation-Based Procedural Training. *Med. Educ.* 49, 805–814. doi:10.1111/medu.12732

Nelson, T. O., and Narens, L. (1990). "Metamemory: A Theoretical Framework and New Findings," in *Psychology of Learning & Motivation*. Editor G. H. Bower (Academic Press), 26, 125–173. doi:10.1016/s0079-7421(08)60053-5

Nelson, T. O., and Narens, L. (1994). "Why Investigate Metacognition? *Metacognition: Knowing about Knowing*. Editors J. Metcalfe and A. Shimamura (MIT Press), 1–25.

Orru, G., and Longo, L. (2019). "The Evolution of Cognitive Load Theory and the Measurement of its Intrinsic, Extraneous and Germane Loads: A Review," in *Human Mental Workload: Models and Applications*. Editors L. Longo and M. C. Leva (Springer International Publishing), 23–48. doi:10.1007/978-3-030-14273-5_3

Ouwehand, K., Kroef, A. v. d., Wong, J., and Paas, F. (2021). Measuring Cognitive Load: Are There More Valid Alternatives to Likert Rating Scales? *Front. Educ.* 6, 370. doi:10.3389/feduc.2021.702616

Paas, F., Ayres, P., and Pachman, M. (2008). "Assessment of Cognitive Load in Multimedia Learning: Theory, Methods and Applications," in *Recent Innovations in Educational Technology that Facilitate Student Learning*. Editors D. H. Robinson and G. Schraw (Charlotte, NC: Information Age Publishing, Inc), 11–35.

Paas, F. G. W. C. (1992). Training Strategies for Attaining Transfer of Problem-Solving Skill in Statistics: A Cognitive-Load Approach. *J. Educ. Psychol.* 84 (4), 429–434. doi:10.1037/0022-0663.84.4.429

Paas, F. G. W. C., and Van Merriënboer, J. J. G. (1994). Instructional Control of Cognitive Load in the Training of Complex Cognitive Tasks. *Educ. Psychol. Rev.* 6 (4), 351–371. doi:10.1007/BF02213420

Prinz, A., Golke, S., and Wittwer, J. (2020). To what Extent Do Situation-Model-Approach Interventions Improve Relative Metacomprehension Accuracy? Meta-Analytic Insights. *Educ. Psychol. Rev.* 32 (4), 917–949. doi:10.1007/s10648-020-09558-6

Renkl, A., Hilbert, T., and Schworm, S. (2009). Example-Based Learning in Heuristic Domains: A Cognitive Load Theory Account. *Educ. Psychol. Rev.* 21 (1), 67–78. doi:10.1007/s10648-008-9093-4

Rey, G. D., Beege, M., Nebel, S., Wirzberger, M., Schmitt, T. H., and Schneider, S. (2019). A Meta-Analysis of the Segmenting Effect. *Educ. Psychol. Rev.* 31, 389–419. doi:10.1007/s10648-018-9456-4

Scheiter, K., Ackerman, R., and Hoogerheide, V. (2020). Looking at Mental Effort Appraisals through a Metacognitive Lens: Are They Biased? *Educ. Psychol. Rev.* 32 (4), 1003–1027. doi:10.1007/s10648-020-09555-9

Schleinschok, K., Eitel, A., and Scheiter, K. (2017). Do drawing Tasks Improve Monitoring and Control during Learning from Text? *Learn. Instruction* 51, 10–25. doi:10.1016/j.learninstruc.2017.02.002

Schmeck, A., Opfermann, M., van Gog, T., Paas, F., and Leutner, D. (2015). Measuring Cognitive Load with Subjective Rating Scales during Problem Solving: Differences between Immediate and Delayed Ratings. *Instr. Sci.* 43, 93–114. doi:10.1007/s11251-014-9328-3

Schnaubert, L., and Bodemer, D. (2017). Prompting and Visualising Monitoring Outcomes: Guiding Self-Regulatory Processes with Confidence Judgments. *Learn. Instruction* 49, 251–262. doi:10.1016/j.learninstruc.2017.03.004

Schneider, S., Beege, M., Nebel, S., and Rey, G. D. (2018). A Meta-Analysis of How Signaling Affects Learning with media. *Educ. Res. Rev.* 23, 1–24. doi:10.1016/j.edurev.2017.11.001

Schneider, S., Nebel, S., and Rey, G. D. (2016). Decorative Pictures and Emotional Design in Multimedia Learning. *Learn. Instruction* 44, 65–73. doi:10.1016/j.learninstruc.2016.03.002

Schraw, G. (2009). A Conceptual Analysis of Five Measures of Metacognitive Monitoring. *Metacognition Learn.* 4 (1), 33–45. doi:10.1007/s11409-008-9031-3

Schraw, G., Crippen, K. J., and Hartley, K. (2006). Promoting Self-Regulation in Science Education: Metacognition as Part of a Broader Perspective on Learning. *Res. Sci. Educ.* 36 (1–2), 111–139. doi:10.1007/s11165-005-3917-8

Schraw, G., Kuch, F., and Gutierrez, A. P. (2013). Measure for Measure: Calibrating Ten Commonly Used Calibration Scores. *Learn. Instruction* 24, 48–57. doi:10.1016/j.learninstruc.2012.08.007

Schroeder, N. L., and Cenkci, A. T. (2020). Do measures of Cognitive Load Explain the Spatial Split-Attention Principle in Multimedia Learning Environments? A Systematic Review. *J. Educ. Psychol.* 112, 254–270. doi:10.1037/edu0000372

Schroeder, N. L., and Cenkci, A. T. (2018). Spatial Contiguity and Spatial Split-Attention Effects in Multimedia Learning Environments: A Meta-Analysis. *Educ. Psychol. Rev.* 30, 679–701. doi:10.1007/s10648-018-9435-9

Seufert, T. (2020). Building Bridges between Self-Regulation and Cognitive Load-An Invitation for a Broad and Differentiated Attempt. *Educ. Psychol. Rev.* 32 (4), 1151–1162. doi:10.1007/s10648-020-09574-6

Son, L. K., and Metcalfe, J. (2000). Metacognitive and Control Strategies in Study-Time Allocation. *J. Exp. Psychol. Learn. Mem. Cognlearning, Memory, Cogn.* 26 (1), 204–221. doi:10.1037/0278-7393.1.20410.1037//0278-7393.26.1.204

Son, L. K., and Schwartz, B. L. (2002). "The Relation between Metacognitive Monitoring and Control," in *Applied Metacognition*. Editors T. J. Perfect and B. L. Schwartz (Cambridge University Press), 15–38. doi:10.1017/CBO9780511489976.003

Sweller, J., Ayres, P., and Kalyuga, S. (2011). *Cognitive Load Theory*. Springer.

Sweller, J. (1994). Cognitive Load Theory, Learning Difficulty, and Instructional Design. *Learn. Instruction* 4, 295–312. doi:10.1016/0959-4752(94)90003-5

Sweller, J., and Paas, F. (2017). Should Self-Regulated Learning Be Integrated with Cognitive Load Theory? A Commentary. *Learn. Instruction* 51, 85–89. doi:10.1016/j.learninstruc.2017.05.005

Thiede, K. W. (1999). The Importance of Monitoring and Self-Regulation during Multitrial Learning. *Psychon. Bull. Rev.* 6 (4), 662–667. doi:10.3758/BF03212976

Thiede, K. W., Anderson, M. C. M., and Therriault, D. (2003). Accuracy of Metacognitive Monitoring Affects Learning of Texts. *J. Educ. Psychol.* 95 (1), 66–73. doi:10.1037/0022-0663.95.1.66

Thiede, K. W.; colleagues (2010). Poor Metacomprehension Accuracy as a Result of Inappropriate Cue Use. *Discourse Process.* 47 (4), 331–362. doi:10.1080/01638530902959927

Undorf, M., and Erdfelder, E. (2011). Judgments of Learning Reflect Encoding Fluency: Conclusive Evidence for the Ease-Of-Processing Hypothesis. *J. Exp. Psychol. Learn. Mem. Cogn.* 37 (5), 1264–1269. doi:10.1037/a0023719

van Gog, T., Hoogerheide, V., and van Harsel, M. (2020). The Role of Mental Effort in Fostering Self-Regulated Learning with Problem-Solving Tasks. *Educ. Psychol. Rev.* 32 (4), 1055–1072. doi:10.1007/s10648-020-09544-y

Wilde, G., Bätz, K., Kovaleva, A., and Urhahne, D. (2009). Überprüfung einer Kurzskala intrinsischer Motivation (KIM) [Testing a short scale of intrinsic motivation]. *Z. Für Didaktik Der Naturwissenschaften* 15 (15/2009), 31–45.

Wiley, J., Griffin, T. D., Jaeger, A. J., Jarosz, A. F., Cushen, P. J., and Thiede, K. W. (2016). Improving Metacomprehension Accuracy in an Undergraduate Course Context. *J. Exp. Psychol. Appl.* 22 (4), 393–405. doi:10.1037/xap0000096

Wiley, J., Griffin, T. D., and Thiede, K. W. (2005). Putting the Comprehension in Metacomprehension. *J. Gen. Psychol.* 132 (4), 408–428. doi:10.3200/GENP.132.4.408-428

Wiley, J., Griffin, T. D., and Thiede, K. W. (2008). To Understand Your Understanding You Must Understand what Understanding Means. *Proc. Cogn. Sci. Soc.* 30. , 2008 Available at: http://csjarchive.cogsci.rpi.edu/Proceedings/2008/pdfs/p817.pdf.

Winne, P. H., and Hadwin, A. F. (1998). "Studying as Self-Regulated Learning," in *Metacognition in Educational Theory and Practice*. Editors D. J. Hacker, J. Dunlosky, and A. C. Graesser (Lawrence Erlbaum), 277–304.

Xie, H., Wang, F., Hao, Y., Chen, J., An, J., Wang, Y., et al. (2017). The More Total Cognitive Load Is Reduced by Cues, the Better Retention and Transfer of Multimedia Learning: A Meta-Analysis and Two Meta-Regression Analyses. *PLOS ONE* 12 (8), e0183884. doi:10.1371/journal.pone.0183884

Zu, T., Munsell, J., and Rebello, N. S. (2021). Subjective Measure of Cognitive Load Depends on Participants' Content Knowledge Level. *Front. Educ.* 6, 144. doi:10.3389/feduc.2021.647097

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership