

## Developing the Mental Effort and Load–Translingual Scale (MEL-TS) as a Foundation for Translingual Research in Self-Regulated Learning

Endres, Tino; Bender, Lisa; Sepp, Stoo; Zhang, Shirong; David, Louise; Trypke, Melanie; Lieck, Dwayne; Désiron, Juliette C.; Bohm, Johanna; More Authors

**DOI**

[10.1007/s10648-024-09978-8](https://doi.org/10.1007/s10648-024-09978-8)

**Publication date**

2025

**Document Version**

Final published version

**Published in**

Educational Psychology Review

**Citation (APA)**

Endres, T., Bender, L., Sepp, S., Zhang, S., David, L., Trypke, M., Lieck, D., Désiron, J. C., Bohm, J., & More Authors (2025). Developing the Mental Effort and Load–Translingual Scale (MEL-TS) as a Foundation for Translingual Research in Self-Regulated Learning. *Educational Psychology Review*, 37(1), Article 5. <https://doi.org/10.1007/s10648-024-09978-8>

**Important note**

To cite this publication, please use the final published version (if applicable). Please check the document version above.

**Copyright**

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

**Takedown policy**

Please contact us and provide details if you believe this document breaches copyrights. We will remove access to the work immediately and investigate your claim.



# Developing the Mental Effort and Load–Translingual Scale (MEL-TS) as a Foundation for Translingual Research in Self-Regulated Learning

Tino Endres · Lisa Bender · Stoo Sepp · Shirong Zhang · Louise David · Melanie Trypke, et al. *[full author details at the end of the article]*

Accepted: 3 December 2024  
© The Author(s) 2025

## Abstract

Assessing cognitive demand is crucial for research on self-regulated learning; however, discrepancies in translating essential concepts across languages can hinder the comparison of research findings. Different languages often emphasize various components and interpret certain constructs differently. This paper aims to develop a translingual set of items distinguishing between intentionally invested mental effort and passively perceived mental load as key differentiations of cognitive demand in a broad range of learning situations, as they occur in self-regulated learning. Using a mixed-methods approach, we evaluated the content, criterion, convergent, and incremental validity of this scale in different languages. To establish content validity, we conducted qualitative interviews with bilingual participants who discussed their understanding of mental effort and load. These participants translated and back-translated established and new items from the cognitive-demand literature into English, Dutch, Spanish, German, Chinese, and French. To establish criterion validity, we conducted preregistered experiments using the English, Chinese, and German versions of the scale. Within those experiments, we validated the translated items using established demand manipulations from the cognitive load literature with first-language participants. In a within-subjects design with eight measurements ( $N=131$ ), we demonstrated the scale's criterion validity by showing sensitivity to differences in task complexity, extraneous load manipulation, and motivation for complex tasks. We found evidence for convergent and incremental validity shown by medium-size correlations with established cognitive load measures. We offer a set of translated and validated items as a common foundation for translingual research. As best practice, we recommend four items within a reference point evaluation.

**Keywords** Mental effort · Mental load · Translingual research · Qualitative methods · Quantitative methods

Educational researchers typically adopt a monolingual approach to their research, placed within their own linguistic contexts. This approach has its merits, including an in-depth understanding of language nuances when coding data, skillful appreciation for implicit meanings, and the development of various tasks and items tailored to a particular language. On the other hand, a monolingual approach can also present challenges. It may overlook the diverse linguistic backgrounds of participants when conducting research, potentially leading to limited generalizability of findings. For example, there is a risk of misinterpretations when integrating research from multiple languages or drawing upon research conducted in the researchers' non-native language (e.g., Terry & Irving, 2010).

The same applies to research on cognitive demand in self-regulated learning (SRL). During learners' SRL, their cognitive demands such as their experienced mental load and invested mental effort play a decisive role. Various studies from different languages and cultures were combined in meta-analyses in educational science and have provided valuable insights into this role (e.g., Baars et al., 2020). Although there exist measures of cognitive load that seem culturally independent (Ayres et al., 2021), many research questions rely on the qualitative insights that can only be gained by self-rating scales of cognitive demand. Contributing to this body of inquiry, researchers from different linguistic backgrounds have provided translations of established self-rating items from English, Dutch, or German into other languages such as Chinese or French (e.g., Colliot & Jamet, 2021; Dönmez et al., 2022; Du & Zhang, 2019; Fontaine et al., 2019; Timirova, 2021).

When translating items, researchers can use either literal translation or cultural adaptation. Literal translation involves directly translating validated items verbatim but can introduce bias due to semantic nuances. Cultural adaptation, on the other hand, develops items that reflect the construct's meaning within each culture, though this can result in non-equivalent meanings across languages. Using both strategies independently to make research accessible in different languages may then complicate the interpretation of findings. This is especially true when combining results in meta-analyses or developing new theories. Subtle yet essential differences may not be adequately considered if researchers conducting these analyses do not fully grasp the linguistic nuances.

The present paper aims to enable those benefits by adopting a translingual approach in developing a measure of cognitive demand during SRL. Therefore, in the introduction, we describe the structure of cognitive demand that occurs in SRL. Afterwards, we present the existing self-ratings scales developed to assess those cognitive demands and assess their validity. Next, in a qualitative study, we interviewed bilingual participants of various languages (English, Dutch, Spanish, German, Mandarin Chinese, and French) to explore the role and understanding of mental effort and mental load within their respective cultures. This qualitative approach should prevent a biased (e.g., European or North American) view on mental effort and load, thereby increasing content validity between the cultures (Silan, 2024). This qualitative phase resulted in a set of self-rating items for every language emphasizing a shared understanding of the translingual differences. Following a rigorous methodological approach, these items were then used in a preregistered experiment focusing on criterion, convergent, and incremental validity. This process resulted in an initial

set of self-rating items for different languages that capture a translingual understanding of cognitive demand. Additionally, this work provides a research paradigm that can be used in future research to add further languages.

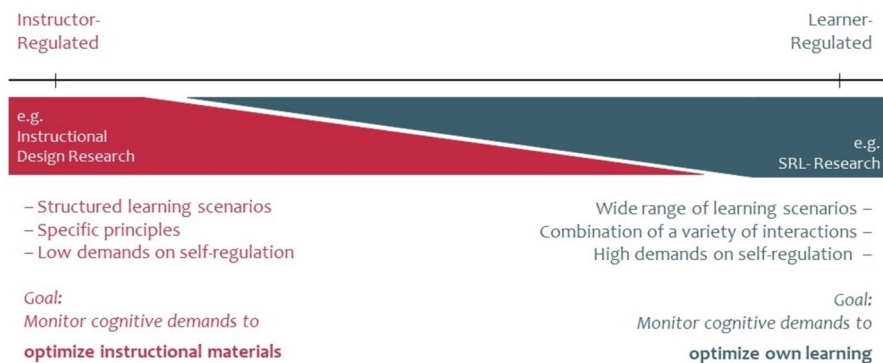
## Cognitive Demand Distinction—A Question of Educational Context?

Cognitive demand is central to learners' experiences, both shaping and being shaped by the learning process. Cognitive load theory (CLT; Sweller et al., 1998, 2019) is a widely used model that describes cognitive demand in a way that supports instructional design to optimize working memory usage, thereby enhancing learners' ability to process and retain new information. CLT highlights the importance of considering cognitive demand due to the limited capacity of working memory. By distinguishing cognitive demand into three types of load, CLT research has informed instructional design principles that align with human cognitive architecture. This research evaluates the demands learners face and examines how changes in instructional design affect learning (Paas et al., 2003a, b). While subjective measures of cognitive demand have limitations (e.g., de Jong, 2010), recent work has shown their reliability and validity (e.g., Andersen & Makransky, 2021; Klepsch & Seufert, 2021; Klepsch et al., 2017; Krieglstein et al., 2022). Studies show that these measures can account for learning differences (e.g., Coppens et al., 2020; Eitel et al., 2019; Endres & Renkl, 2015; Endres et al., 2020, 2024a; Le et al., 2021; Schneider et al., 2019).

Research in CLT mainly focuses on instructional design in structured learning environments with low demands on learner self-regulation (Seufert, 2018). Instructional design aims to improve learning by creating effective and efficient materials and procedures. Such research helps optimize instructional materials but is less applicable to SRL contexts, which involve more complex interactions and higher demands on self-regulation. In SRL scenarios, assessment tools must address a broader range of learning situations and account for overlaps between instructor-regulated and learner-regulated contexts (see Fig. 1). These differences should be considered when evaluating cognitive demand in each learning scenario.

### Instructor-Regulated Educational Settings

Over the years, various distinctions of cognitive demand components in instructor-regulated educational settings have been proposed (e.g., Kalyuga, 2011), resulting in numerous self-rating items. CLT distinguishes cognitive demand into three types of cognitive load: intrinsic, extraneous, and germane cognitive load. Intrinsic load (ICL) refers to task complexity, extraneous load (ECL) relates to how information is presented, and germane load (GCL) concerns schema construction and learning (Sweller et al., 2019). Differentiating these types is crucial for designing effective learning environments, leading to validated instruments for measuring those load types in instructional design (e.g., Andersen & Makransky, 2021; Klepsch et al., 2017; Krieglstein et al., 2022; Leppink et al., 2013). The perception of these types



**Fig. 1** Visualization of goal-driven cognitive demand in instructional design and SRL research

of load depends on the educational settings, including learners' characteristics (e.g., prior knowledge, Endres et al., 2023; Zu et al., 2021), learning techniques used (Thees et al., 2021), and the nature of the materials (Kalyuga, 2011).

### Learner-Regulated Educational Settings

In recent years, increased attention has been given to the influences of cognitive demand when investigating learners' self-regulation. This interest has led to the development of different integrative frameworks of SRL and CLT (e.g., de Bruin et al., 2020; Seufert, 2018; Seufert et al., 2024; Wang & Lajoie, 2023). Each integrative framework has recognized the links between SRL and CLT, proposing that on the one side, SRL induces cognitive load (Wirth et al., 2020), and on the other side, cognitive load influences SRL (de Bruin et al., 2020; Nugteren et al., 2018; van Gog et al., 2020). The different frameworks share the claim that CLT can benefit from incorporating the concept of dynamic changes during the learning process (e.g., Castro-Alonso et al., 2021). SRL research, on the other hand, can benefit from the focused perspective of CLT on cognitive processes by gaining a more nuanced understanding of demanding aspects of SRL (Seufert, 2018). Additionally, recent developments in CLT research have seen an increasing focus on students' self-management of cognitive load (Castro-Alonso et al., 2021; Eitel et al., 2020; Paas & Van Merriënboer, 2020; Zhang et al., 2021).

The specific cognitive load types and assessment scales proposed by CLT however seem to have less predictive value in situations in which a higher focus lies on learners' self-regulation. For example, ECL is often considered in connection to the design of the learning material, as in the often-used item from Klepsch et al. (2017): "The design of this task was inconvenient to learn something." Although this is a very important information for instructional designers, this operationalization of ECL is not essential for the learners as they might not have the chance to select a task with a different design as it was the task provided by their teacher. That is why it seems necessary to expand the understanding of SRL's cognitive demand beyond the CLT's traditional distinction.

## Why Mental Effort and Load Are Appropriate Levels of Abstraction in SRL

Due to the specific focus of some CLT scales on structured, instructor-regulated learning scenarios, researchers in the SRL community have provided a different distinction of cognitive demand that suits a wider range of learning situations and the specific affordances in SRL. One of these distinctions is that into a mental load, that learners experience during learning (e.g., Grund et al., 2024; Klepsch & Seufert, 2021), and an active component of mental effort, that learners are motivated to invest (e.g., Grund et al., 2024; van Gog et al., 2024). Similar constructs are addressed using different terminologies: the mental load (Klepsch & Seufert, 2021) is also referred to as data-driven demand (Koriat, 1997) and task-centered dimension of load (Choi et al., 2014; Paas & van Merriënboer, 1994); the active component of mental effort (Klepsch & Seufert, 2021; Krell, 2017) is also referred to as goal-driven demand (Koriat, 1997), goal-driven appraisals of cognitive load (de Bruin et al., 2020), and human-centered dimension (Choi et al., 2014; Paas & van Merriënboer, 1994). Within this paper, we will use the terminology of mental load and mental effort (Klepsch & Seufert, 2021) as those were also the most idiomatic to the participants in our qualitative interviews (see also Wolpe et al., 2024).

### Mental Load in SRL

Monitoring mental load in SRL aligns closely with learners' central goals of optimizing their own learning processes. The source of cognitive load appears to be less relevant for self-regulated learners compared to instructional designers. Instructional designers need to differentiate between the complexity of the learning content and the load evoked by its representation in the learning material. However, for self-regulated learners, the specific source of their load is less critical. Since the level of decision-making likely influences the monitored cognitive demand, it seems reasonable not to differentiate between different sources of load that are inevitable to the learner in SRL. Additionally, in SRL, there are many more sources of load than just the complexity of the cognitive representation of the task (intrinsic cognitive load) and the design and presentation of information and instruction (extraneous cognitive load) highlighted in the classical idea of CLT. For example, metacognition, the duration of learning, worry cognitions, and other factors that can contribute to mental load are considered in the literature around SRL.

**Metacognition as an Influence on Mental Load** In structured learning environments, such as they can be found in instructional design research, tasks are typically predetermined, minimizing the need for learners to choose the next task (e.g., Castro-Alonso et al., 2018). However, such decisions are critical in SRL. Indeed, SRL models (e.g., Greene & Azevedo, 2007; Panadero, 2017) emphasize the role of intentional metacognitive processes, which should therefore be considered when assessing cognitive demand in SRL. Moreover, effective SRL activities do not occur

automatically; rather, they impose a mental load on the learners who have to monitor and regulate their strategies effectively (Seufert et al., 2024).

Within metacognition, learners distinguish between mental load and mental effort. Mental effort is frequently associated with positive learning outcomes (e.g., Coppens et al., 2020; Endres & Renkl, 2015; Endres et al., 2024a, c), while mental load negatively correlates with learning outcomes (Carpenter et al., 2020; Endres et al., 2024a). These connections between cognitive demand distinctions and learning outcomes seem to be also perceived by the learners as judgments of learning negatively correlate with ratings of mental effort, while they show a positive relationship with mental effort (goal-driven effort, Baars et al., 2020). This highlights how learners differentiate between mental load and effort and their ability to assess both types of cognitive demand differently (see also Wolpe et al., 2024).

**Duration of Learning as an Influence on Mental Load** The duration of learning significantly impacts the mental load experienced by learners in SRL. Extended learning sessions require sustained cognitive effort, increasing the perceived load. Research explains that effort costs rise with both the duration and intensity of an action, necessitating that individuals assess whether the desired action is worth the required effort (Eccles & Wigfield, 2020). Furthermore, research supports the notion that as the duration of learning increases, so does the mental load experienced by learners, which leads to less sustained learning over longer learning periods (Endres et al., 2020, 2024b, c). This is particularly evident in studies examining the effects of sustained cognitive effort on learning outcomes and learners' motivation in demanding learning situations (Endres et al., 2024b, c). Another example is that prolonged engagement in learning activities can lead to mental fatigue, depleting working memory resources essential for information processing and retention (Chen et al., 2018; Lo et al., 2022). The depletion of these resources over time diminishes the learner's ability to maintain effective cognitive functioning, impacting learning outcomes and motivation.

**Worry Cognitions as an Influence on Mental Load** As a third influence, worry cognitions can increase cognitive load during learning (Moran, 2016; Plass & Kalyuga, 2019). For instance, learners with test anxiety often experience a higher mental load than those who have received treatment. Addressing these worries is crucial as learners can influence the mental load they experience. Interventions, such as short physical activity breaks, have been shown to reduce test anxiety, lower mental load, and improve performance (Mavilidi et al., 2020).

In summary, metacognition, duration of learning, worry cognitions, and other potential influences on mental load are usually not included in classical assessments of CLT. However, they seem essential when analyzing a broader range of learning scenarios, especially in ecological settings of SRL. The broader scope of mental load as a cognitive construct allows learners and SRL researchers to analyze a wide range of learning situations without sacrificing predictive validity for future learning.

## Mental Effort in SRL

The monitoring of mental effort as a motivational component of cognitive demand that is actively invested also aligns closely with learners' central goals of optimizing their own learning processes in SRL (Grund et al., 2024). For example, learning processes such as retrieval practice, spacing, and interleaving, known as “desirable difficulties,” (Bjork & Bjork, 2020) require significant cognitive engagement to enhance knowledge consolidation (Richter et al., 2022; Roelle et al., 2022). Ensuring learners' active engagement with the material is essential for lasting learning (Bjork & Bjork, 2020). For self-regulated learners, monitoring their mental effort allows them to adjust their learning strategies to meet their individual needs and maintain an optimal challenge level (Roediger & Butler, 2011). Additionally, understanding the link between mental effort and long-term learning outcomes can motivate learners by encouraging persistence with challenging tasks (Dunlosky et al., 2013). Learners can foresee subtle differences when judging their own learning processes and recognize that effortful strategies like retrieval practice improve their memory (Rivers, 2021).

Mental effort is linked to the GCL (e.g., Klepsch & Seufert, 2021). However, the advantage of the concept of mental effort over GCL is its broader applicability to various scenarios. While most CLT scales typically focus on comprehension-oriented strategies, active investment of mental effort can also be encouraged by choosing to work on more complex tasks or engaging in desirable difficulties. These strategies may contribute to long-term retention without directly enhancing immediate learning outcomes (e.g., Roelle et al., 2023).

## The Basis for Validation: The Overlap of Mental Load and Effort with CLT in Instructional Design

As discussed, SRL-specific educational situations require a different distinction of cognitive demand compared to that used in CLT-based research. Nonetheless, a validated scale for SRL educational situations should identify demand differences in contexts where both research areas overlap. Since an SRL-specific cognitive demand scale aims to encompass a broad range of learning situations, it should also predict differences typically examined in CLT research. This overlapping shared area of interest allows us to implement a validation approach that uses established experimental procedures from CLT research. Specifically, we intend to investigate educational settings where CLT-based research and SRL-based research overlap. With these settings, we will apply both the newly developed mental effort and load scale, alongside an instructional design-based cognitive load scale, to demonstrate convergent and incremental validity.

This overlap was previously demonstrated by Klepsch and Seufert (2021). In their study, ICL and ECL were strongly related to passive mental load, whereas GCL was associated with active mental effort through instructional design. In the following,



we describe the overlap between mental load and mental effort with classical ICL, ECL, and GCL situations. We will leverage this overlap in our quantitative validation experiments.

## Influences on Mental Load in SRL

**ICL** The basis of all cognitive demand assessment is the intrinsic load. Intrinsic load refers to the complexity of the learning task and is often quantified by the degree of element interactivity (e.g., Chen et al., 2023; Haji et al., 2015; Huang, 2018; Larmuseau et al., 2020; Sweller, 2010). Element interactivity comprises both the structure of the information being processed and the knowledge activated from memory that is processed to solve a task (Endres et al., 2023). Element interactivity represents the interconnectedness between essential elements of information that needs to be considered in working memory simultaneously to be able to solve a task (Kalyuga, 2011; Sweller, 2010).

Importantly, intrinsic load is relevant in different aspects of SRL. Research on the goal specificity effect examines situations in which learners must work on a task and simultaneously monitor whether the specific learning goal has been reached or not (Locke & Latham, 2002; Sweller, 1988). This research highlights the role of task complexity and the additional elements that must be processed in educational settings, in which metacognitive requirements are more pronounced (Sweller & Levine, 1982). Those different influences of task complexity and metacognition in SRL could be measured separately. The decisive question for learners to reach their learning goal seems to be the perceived load of both complexity and metacognition together, rather than distinguishing between the two.

**ECL** ECL is highly prevalent in SRL, where learners are required to ignore external distractions and persist in their learning activity. Such external distractions can be background noises in a library, advertisements on information websites, or attractive and easily accessible social media websites (see cyber-slacking, e.g., Flanigan & Kiewra, 2018). One well-investigated phenomenon that is related to such distractions is seductive details. Seductive details are interesting, but irrelevant elements in learning materials that have been shown to influence learners' processing focus, increase ECL, and hamper learning performance (e.g., Colliot & Boucheix, 2024; Bender et al., 2021a; Eitel et al., 2022; González et al., 2019; Harp & Mayer, 1998; Rey, 2012; Tsai et al., 2019; Wang & Adesope, 2016b). Against this background, the likelihood of increased mental load due to interesting or appealing extraneous content might be quite high in SRL. Apart from competing for cognitive resources, such content (especially when somehow related to the learning task) has also been assumed to improve the learning process through enhanced effect and motivation (e.g., Lenzner et al., 2013; Magner et al., 2014; Wang & Adesope, 2016a). However, such effects were observed very seldomly and have not yet been replicated (see Bender et al., 2021b for an overview).

This line of research highlights the role of ECL that must be managed when engaging in SRL situations in a natural learning environment. As any given SRL

situation has some type of extraneous demand, the mental load imposed by different learning settings on the learners must be considered together with the task complexity and the metacognitive demand. To monitor their current learning process, learners seem to use this aggregation of ICL and ECL as mental load (Baars et al., 2020).

### Influences on Mental Effort

**GCL** Given that learning involves the construction of schemas, this type of load refers to the working memory's resources required for deep learning (Sweller et al., 2019). This concept seems to be highly correlated with mental effort in research (Klepsch & Seufert, 2021).

GCL has been manipulated in previous research by several interventions, all of which led to an increase in learners' motivation to engage in deep learning processes (e.g., Klepsch & Seufert, 2020; Klepsch et al., 2017). One of these interventions is the implementation of the imagination principle (Cooper et al., 2001; Leopold, 2021), where learners receive instructions to engage in mental imagery such as "Please imagine the steps in the nervous system when the brain sends a signal to the diaphragm and rib muscles" (e.g., Krieglstein et al., 2022; Leopold & Mayer, 2015; Leopold et al., 2019). When imagination is triggered, learners invest more GCL, which leads to better learning. Similarly, when validating their cognitive load scale, Klepsch and colleagues (2017) let the students imagine hypothetical learning situations which differed in the GCL they should induce. All interventions led to a higher GCL. Another intervention that influences GCL is game-based learning (e.g., Huang, 2011; Woo, 2014), which has consistently resulted in more mental effort investment in learning and additionally in an increase in learning performance (Woo, 2014). This research shows how GCL and mental effort could be increased. In SRL, similarly, the mental effort learners invested is key to increase learning performance (desirable difficulties; Bjork & Bjork, 2020).

**ICL** Intrinsic load can significantly impact not only the mental load but also the mental effort learners actively invested in a task. More complex tasks prompt learners to engage more actively, leading to increased mental effort and greater engagement. For example, a study by van Merriënboer and Sweller (2005) indicates that tasks with high ICL increase the cognitive investment required to understand and integrate information. This increased effort can result in deeper processing and better retention of the material (Paas et al., 2003a, b).

Moreover, complex tasks can stimulate learners' intrinsic motivation, as they may find such tasks more challenging and rewarding (Deci & Ryan, 1985). The engagement required to manage a high intrinsic load can foster deeper cognitive investment, enhancing learning outcomes. Furthermore, learners employing SRL strategies are better equipped to handle ICL. For instance, Greene and Azevedo (2009) found that learners who effectively plan, monitor, and regulate their cognition can manage intrinsic load more efficiently, resulting in improved learning outcomes. Additionally, Sirock et al. (2023) found that motivated learners put in extra effort to

compensate for task complexity or deficits in instructional design. This effort helps in forming better mental representations, which supports the findings that motivated learners to increase their effort in challenging tasks.

Thus, while intrinsic load directly relates to the complexity and structure of the content, it also impacts learners' invested mental effort, beyond mere difficulty. Indeed, it influences both the cognitive and the motivational aspects of learning, driving learners to invest more mental effort and engage more deeply with the material.

## Cultural Understanding of Effort and Load

Research in cognitive science has shown that language profoundly influences thought processes (e.g., Boroditsky, 2011; Casasanto & Boroditsky, 2008; Lupyan et al., 2020). Speakers of different languages conceptualize time, space, and even colors differently due to linguistic structures (Boroditsky, 2011; Lupyan et al., 2020). This phenomenon is also important in educational psychology, especially in translanguaging research, as it suggests that learners with different linguistic backgrounds may experience and assess educational learning scenarios uniquely based on their linguistic backgrounds. Usually, researchers only translate items from a given language into the language they want to use in their study. The process of translating these items may follow different approaches. For example, one could use (almost) literal translations of the validated items. So far, the load and effort scales used in the extant studies worldwide mostly originated from Dutch (Paas, 1992), German (e.g., Klepsch et al., 2017; Krell, 2017), and English (e.g., Leppink et al., 2013). While this approach may seem reasonable, it can introduce bias and limit comparability due to translanguaging effects (Boroditsky, 2011; Silan, 2024). These effects occur when words encompass cultural, contextual, or semantic nuances that cannot be fully captured or easily conveyed through a literal or direct translation in another language. For instance, in German, the two constructs mental effort and mental load are translated by a subtle change in phrasing “es war anstrengend” (it has been strenuous) vs. “ich habe mich angestrengt” (I have made an effort) (see also Klepsch & Seufert, 2021). These formulations are used in validated tasks but have no direct translation that truly preserves this meaning.

Another approach might be not to translate the items directly, but to develop comparable items that the authors assume best reflect the meaning of the construct in the respective culture. For instance, asking learners in Chinese if they “gave their heart and soul,” in French if they “invested themselves,” or in English if they “invested mental effort” may be representative within their respective cultures (see qualitative interviews). However, such translations may not have equivalent meanings across different languages and may therefore assess different constructs.

In recent educational research, the impact of cultural factors on learners' perceptions has received significant attention. It is crucial to acknowledge that cognitive constructs, such as mental effort and load, are interpreted and valued differently across cultural contexts (e.g., Chen, 2023). This understanding is pivotal for our

study, aiming to develop items that are culturally sensitive and valid across various linguistic settings. For instance, Chen (2023) claims that effort is given higher relevance in many east Asian cultures than in Western cultures. This difference is explained by Confucian-influenced cultures more strongly conceptualizing effort as a central social duty and moral virtue. Effort is additionally ascribed more explanatory power over personal success or failure than in many Western cultures. This trend is evident in studies showing a pronounced tendency in east Asian cultures to attribute academic success to effort rather than to innate ability (e.g., Chen et al., 2009, 2018).

Recognizing these cultural nuances is essential in our research. The items we develop to measure cognitive demand must not only be linguistically accurate translations but also culturally congruent. For example, the higher valuation of effort in Asian cultures may lead to a greater tendency to report higher efforts due to socially desirable responding (Paulhus, 1984). Therefore, to assess cognitive demand effectively, item stems must be capable of evaluating a higher relative level of cognitive demand while still validly differentiating between various interventions. Achieving this goal will enable a more comprehensive and accurate understanding of cognitive demand in SRL across different languages and cultures.

## Translingual Research and Validity

When developing items for use across different languages and cultures, it is essential to consider the quality criteria specific to each language and culture while ensuring a shared understanding across them. The translingual assessment of self-rating items for cognitive demand is primarily a matter of construct validity. Construct validity has multiple facets, all of which must be evaluated holistically for a rigorous methodological procedure.

**Content Validity** The initial consideration in construct validity is content validity. Content validity examines whether a scale adequately represents all aspects of a given construct through its items. In translingual research on cognitive demand, it is crucial to ensure consistent understanding of the concepts assessed, that are mental effort and mental load, across different languages. To achieve this, we will conduct qualitative interviews to ensure linguistically appropriate translation and cultural representation.

**Criterion Validity** Another aspect of construct validity to address is criterion validity. This type of validity assesses whether a scale can predict a specific behavior. In the context of cognitive demand, a newly developed scale should be capable of measuring changes due to task aspects that may arise in SRL situations. When manipulating different aspects of a task (e.g., in an experimental design), the resulting differences should be consistently measurable among different participants within the same language and across different languages and cultures. To ensure translingual criterion validity, manipulations should yield comparable effect sizes across tasks that differ in a similar manner.

**Convergent and Incremental Validity** The third aspect of construct validity to consider is convergent and incremental validity. These types of validity explore constructs that should reasonably correspond to or differ from the construct under assessment and examine whether they overlap.

On the one hand, a newly developed scale should show convergent validity by aligning with constructs that assess similar aspects in a similar learning situation. For our scale, this means that when implemented in a controlled learning scenario manipulating CLT variations, we should find results similar to established scales or aggregates as explained earlier (e.g., Klepsch et al., 2017).

On the other hand, the objective is for each construct to contribute a unique aspect, demonstrating incremental validity. In the context of cognitive demand, one such construct is difficulty appraisal (Hoch et al., 2023). Difficulty appraisal typically measures the perceived difficulty of a task for the learner. In contrast to mental effort and load, it does not account for the learners' actual effort expended on the task or their perceived inner load while working on the task. For instance, when learners work on a task that requires demands besides complexity, such as metacognition or other external demands, they may assess the difficulty based on their understanding of these complexities instead of considering their overall load. During an easy task, the difficulty might be low, but additional ECL might lead to a higher overall mental load. Consequently, a scale assessing mental load should demonstrate incremental validity in relation to difficulty appraisals. The same is true for established CLT scales. The newly developed scale should be able to contribute a unique aspect of description to the established scales, which for example could be shown by substantial but limited correlations.

## Development of Items

**Content Validity** Single items often fail to capture the full scope of constructs like cognitive demand, being sensitive to biases and representing an incomplete picture. Employing multiple items increases measurement reliability by reducing the impact of random errors or outliers, thus improving the signal-to-noise ratio (Cronbach, 1951; Rouder et al., 2019). Additionally, a diverse set of items enhances content validity in cognitive demand assessments across various learning scenarios by encompassing different dimensions and nuances of the construct (DeVellis, 2017). Including both positively and negatively worded items mitigates response biases, such as acquiescence bias, and balances individual response styles, resulting in more accurate and thoughtful responses (Porst, 2014).

To ensure content validity, we selected and extended established items from the literature. We compiled ten self-rating items: five for mental effort and five for mental load. The item construction was inspired by established measures (Klepsch & Seufert, 2021; Krell, 2017; Leppink et al., 2013; Paas, 1992). We adapted two items from well-known scales such as those developed by Paas (1992) and by Klepsch and Seufert (2021). We also incorporated items used in Chinese studies to broaden the

international scope. These sources provided a foundational set of items proven effective in instructional design.

The collected items were refined following Porst (2014) item construction rules, emphasizing the creation of both positive and negative items, avoiding quantifiers in item stems, and abstracting items to comprehensively cover various aspects of SRL. Each construct comprised five items, with two items formulated in a reversed manner to mitigate response bias. The items were iteratively created in both German and English, serving as base languages for subsequent translations into Mandarin Chinese, Dutch, Spanish, and French. This multilingual approach ensured the scale's applicability across diverse linguistic contexts. The final set of items developed for this study is detailed in Table 1 and available on the Open Science Framework.

## The Present Project

Our project aimed at providing a translingual, validated scale of cognitive demand during SRL. To ensure *content validation* in the different languages, we first implemented qualitative interviews. Before translating the scales, bilingual participants discussed and reflected on their understanding of the concepts of mental effort and load. Afterwards, they translated a scale of established, emerging, and newly developed items aimed at assessing those concepts in English, Dutch, Spanish, German, Mandarin Chinese, and French. A second bilingual participant, who was fluent in the same languages, back-translated these items and reflected on possible discrepancies between the two versions. Second, to ensure *criterion* as well as *convergent and incremental validity*, we conducted a quantitative experiment. We validated the translated items using established demand manipulations from the cognitive load literature in samples of first-language participants to identify potential differences within the languages.

## Qualitative Study—Content Validity

To establish content validation across languages, we conducted qualitative semi-structured interviews with bilingual participants. The first goal of those interviews was to investigate the interviewees' understanding of mental effort and load within their respective cultures and to identify any cross-cultural differences in these conceptions. The second goal was to obtain a set of translated items that consider these differences, thereby ensuring translingual equivalence of the content.

## Methods

### Participants and Study Plan

The interviews were conducted in three stages. A first stage focused on the refinement of the German and English items originally selected and adapted by the

**Table 1** Selected established and newly generated items as a basis for translation

Construct	Item number	German	English
Mental effort	1	Ich habe mich beim Lösen dieser Aufgabe bemüht. <sup>P</sup>	I invested effort while working on this task. <sup>P</sup>
	2	Ich habe mich beim Lösen dieser Aufgabe angestrengt. <sup>K</sup>	I mentally strained myself to solve this task. <sup>K</sup>
	3	Ich habe keine Anstrengung in das Lösen dieser Aufgabe investiert. <sup>R</sup>	I refrained from putting effort into this task. <sup>R</sup>
	4	Ich habe keinen Aufwand in das Lösen dieser Aufgabe gesteckt. <sup>R</sup>	I did not invest effort into solving this task. <sup>R</sup>
	5	Ich habe Herzblut in das Lösen dieser Aufgabe investiert	I put my heart and soul into solving this task
Mental load	6	Das Lösen dieser Aufgabe hat mich Mühe gekostet	Solving this task required effort
	7	Das Lösen dieser Aufgabe war anstrengend. <sup>K</sup>	Solving this task was mentally demanding. <sup>K</sup>
	8	Es war mühelos, diese Aufgabe zu lösen. <sup>R</sup>	It was effortless to solve this task. <sup>R</sup>
	9	Das Lösen dieser Aufgabe hat Aufwand von mir gefordert	Solving this task required effort on my part
	10	Es war einfach, diese Aufgabe zu lösen. <sup>R</sup>	It was easy to solve this task. <sup>R</sup>

*R* reversed items, *P* adapted from Paas (1992), *K* adapted from Klepsch and Seufert (2021)

authors. In a second stage, the interviews were conducted with two bilingual interviewees per target language (bilingual in either German or English and the target language). Both interviews followed the same three phases and were followed by a comparison phase. Overall, the interviews comprised five language combinations (English–German, English–Mandarin Chinese, English–Spanish, English–Dutch, and German–French) resulting in a sample of 10 participants. The participants were recruited by the authors and received a predetermined payment (~ 18 €), which we adjusted if the interviews lasted longer than the allotted time of 1.5 h.

During the first phase, in the reflection part, participants were asked some predefined questions to reflect on their understanding of the concepts of effort and load. For each target language, one interviewee was assigned to the base language (German or English) and the other to the target language. In the second phase, the translation part, the participants were asked to translate some predefined items while considering the understanding of effort and load discussed in phase one. In a third phase, participants were asked to back-translate the items developed by the other interviewee. Finally, in a fourth phase, both bilingual participants discussed the two translations with the language coordinator in the comparison phase.

### **Phase 1: Reflection**

The interview followed a semi-structured protocol created to maintain consistency across all sessions. First, the interviewer informed the bilingual participants about the goal and duration (about 1.5 h) of the interview and asked for consent to audio record their responses (all recordings were transcribed and stored anonymously according to local data protection laws). Then, the interview started with discussing the concepts of load and effort according to several predefined questions and prompts. Following transparency criteria, an overview of those questions is available on OSF.

### **Phase 2: Translation**

After conceptual discussion of load and effort, the item translation part of the interview started. The translation process was based on the guidelines of the Psychological Science Accelerator (Psychological Science Accelerator, [n.d.](#)). Before translation, the interviewer emphasized the mental effort or mental load aspect of the items to be translated. One interviewee per pair (Y) was assigned to translating the items from the base language (either German or English) to the target language (English, Dutch, Spanish, German, Mandarin Chinese, and French), resulting in the initial translated Version A1. The other interviewee (Z) was assigned to translating the items from the target language (English, Dutch, Spanish, German, Mandarin Chinese, and French) to the base language (either German or English), resulting in the initial translated Version B1.



### Phase 3: Back-translation

Afterwards, interviewee Z independently back-translated the items of Version A1 to the base language, resulting in Version A2, while interviewee Y independently back-translated the items of the Version B1 to the target language, resulting in Version B2. The participants also reflected on potential differences regarding their back-translated items (Version 2) and the Version 1 of the items.

### Phase 4: Comparison

Finally, the language coordinator together with both bilingual participants discussed similarities or differences and any further necessary language-specific cultural adjustments in phrasings, resulting in the final items, Version C.

## Results

Our analysis of the interviews followed a two-step approach. First, the researcher (interviewer) examined their interview recordings and transcripts with respect to two predefined questions: (1) What is the participants' cultural understanding of effort and load in the learning context? (2) How did culture and language influence the understanding and translation of the items? A detailed overview of the results of the interview examinations is available on OSF and discussed in the next section.

Second, the language expert decided on the final set of effort and load items based on the results of the interviews. The final sets consisted of four items for the mental effort scale and four items for the mental load scale to be used in our quantitative experiment. Moreover, the language experts also decided on one item presumably reflecting a more intense view of effort ("put heart and soul"), which was inserted as a result of the Chinese understanding of effort. Additionally, one item served to assess perceived "difficulty" representing a discriminant construct. In making those decisions, the language experts considered whether the items were in line with (a) the translations and accompanying discussions regarding cultural understanding in the interviews, (b) the scientific understanding of the two constructs, and (c) the methodological criteria of item construction. Table 2 shows an overview of those items in our six languages: English, Mandarin Chinese, Spanish, Dutch, German, and French.

## Discussion

The goals of our interviews were (a) to get some deeper insights into how different cultures perceive and understand mental effort and load and (b) to translate established self-rating items of effort and load into different languages while considering those cultural aspects. With respect to both goals, our interviews

**Table 2** Proposed items for the six languages as a result of the translation process

English	Mandarin Chinese	Spanish	Dutch	German	French
<b>Mental effort</b>					
I invested effort while working on this activity	我投入了努力去完成这个任务。	Puse esfuerzo en hacer esta actividad	Ik heb moeite gestoken in het werken aan deze taak	Ich habe mich bei der Bearbeitung der Aufgabe bemüht	J'ai fait des efforts pour réaliser cette tâche
I did not mentally strain myself while working on this activity	我很努力地去完成这个任务。	No puse esfuerzo mental en hacer esta actividad	Ik heb mezelf niet mentaal ingespannen tijdens het werken aan deze taak	Ich habe mich bei der Bearbeitung der Aufgabe nicht angestrengt	Je ne me suis pas fatigué-e pour réaliser cette tâche
I refrained from putting effort into this activity	我没有尽力去完成这个任务。	Evité ponerle esfuerzo a esta actividad	Ik heb geen moeite gestoken in deze taak	Ich habe keine Anstrengung in diese Aufgabe investiert	Je n'ai pas fait d'effort pour réaliser cette tâche
I worked hard on this activity	我没有为这个任务付出任何努力。	Me esforcé para esta actividad	Ik heb hard gewerkt aan deze taak	Ich habe hart an dieser Aufgabe gearbeitet	Je me suis donné-e de la peine pour réaliser cette tâche
<b>Mental load</b>					
Working on this activity required effort	完成这个任务需要投入努力。	Hacer esta actividad requirió de esfuerzo	Het werken aan deze taak kostte moeite	Das Bearbeiten der Aufgabe hat mich Mühe gekostet	Réaliser cette tâche m'a demandé des efforts
Working on this activity was mentally demanding	完成这个任务很消耗脑力。	Hacer esta actividad fue mentalmente demandante	Het werken aan deze taak was mentaal veeleisend	Das Bearbeiten der Aufgabe war anstrengend	Réaliser cette tâche était mentalement exigeant
Working on this activity was effortless	完成这个任务很轻松。	Hacer esta actividad fue fácil	Het werken aan deze taak kostte geen moeite	Das Bearbeiten der Aufgabe war mühelos	Réaliser cette tâche ne m'a pas coûté d'effort
Working on this activity didn't mentally strain me	完成这个任务并不费力。	Hacer esta actividad no requirió de mi esfuerzo mental	Het werken aan deze taak heeft me mentaal niet belast	Das Bearbeiten der Aufgabe hat keinen Aufwand von mir gefordert	Réaliser cette tâche était fatigant

Table 2 (continued)

English	Mandarin Chinese	Spanish	Dutch	German	French
Lifeblood					
I put my heart and soul into working on this activity	我全身心投入地去完成这个任务。	Puse todo mi corazón en hacer esta actividad	Ik heb mijn ziel en zaligheid gestoken in het werken aan deze taak	Ich habe die Aufgabe mit Herzblut bearbeitet	J'ai mis tout mon cœur dans la réalisation de cette tâche
Difficulty					
Working on this activity was difficult	完成这个任务是困难的。	Hacer esta actividad fue difícil	Het was moeilijk om aan deze taak te werken	Die Bearbeitung der Aufgabe war schwierig	Réaliser cette tâche était difficile

These items were developed ensuring content validity within their respective languages and are not expected to have literal matching translations between languages. Stoo Sepp functioned as an English language expert, Shirong Zhang as a Chinese language expert, and Louise David functioned as a Dutch language expert. Lisa Bender and Juliette Désiorin functioned as French language experts, while Melanie Trypke functioned as a German language expert

revealed some interesting findings that might provide some new perspectives on the interpretation of established effort and load scales in different cultures and on the translation of such scales, in general.

### **Cultural Understanding of Mental Effort and Load**

Participants in all our six languages could relate to the distinction between mental effort and load in their respective cultures, with Chinese participants perceiving only a subtle difference of the two constructs and Dutch participants not being sure about how common a conscious distinction would be in daily life. Overall, effort was described as an active investment that is made willfully and controllably and depends on motivation. Participants from all cultures acknowledged that effort depends on the benefits associated with working on a task (e.g., personal value, grades, recognition, and goal-pursuit). Another aspect mentioned by some participants was the association of effort with complex, non-routine tasks (English, French) and concentration (French, Mandarin Chinese). Chinese-speaking participants in particular mentioned that invested effort also depends on the sense of responsibility. In all cultures, effort had an additional physical component (e.g., described via sports analogies).

The participants described mental load as something imposed or even forced that cannot be controlled easily (English and French). While Spanish and Chinese participants felt that mental load encompasses all aspects of working on a task (i.e., also dealing with unclear instructions or distractions), French participants specifically emphasized the feelings of exhaustion in long, uninteresting tasks and mentioned a negative connotation of the construct. The Mandarin Chinese participants' responses revealed a particularly strong association with difficulty (i.e., mental load is high when a task is too difficult).

Taken together, although in all cultures the distinction between effort and load was present, we noticed some differences in how clear this distinction was made and what aspects were emphasized. Those differences could also influence how self-rating scales of these constructs are perceived and answered. For example, while effort ratings in all of our investigated cultures might depend on participants' motivation during the task (e.g., interest, benefits), Chinese participants specifically discussed their sense of responsibility for completing the task when answering such items. Moreover, Mandarin Chinese (and Spanish and Dutch) participants may associate the established load items more strongly with the difficulty of the task or aggravating external factors while French participants might rather rely on how long and therefore exhausting the task was for them. In future studies, it should also be considered that the distinction of effort and load might not be as clear in the ratings of Chinese participants making it more difficult to interpret those measures. Moreover, load was not consistently perceived as negative across cultures, but also had positive aspects for some participants. It is unclear whether those are reflected in the established items or not.

## Translation of Effort and Load Items

In our interviews, we aimed at translating established effort and load items while considering the above-described aspects. While many items could be translated rather directly into different languages (as supposedly done by many previous translations for research projects), some translations were discussed more intensely and needed to be adjusted. A more general difficulty faced by many translations was the intended intensity of the items. According to the participants, some items differed in intensity from the original or from the other items in the subscale—especially those expressing mental effort. Such differences should be considered when interpreting the scales, as they could lead to floor or ceiling effects. Moreover, as mentioned above, mental effort has a physical component for some of our participants and, therefore, the items should specifically address *mental* effort in some languages. For Mandarin Chinese, however, this is not possible without making the items sound superficial. This problem could perhaps be solved by giving very clear instructions before the rating. Another aspect of discussion that came up in multiple interviews was the phrasing of the object to be rated (i.e., the activity participants engaged in before rating). In many studies, the phrasing is adjusted to the specific task. Future studies should also consider linguistic and cultural particularities for this adjustment (e.g., “solving this task” is inappropriate in Spanish and Mandarin Chinese).

For some languages, we noticed interesting deviations from the literal translation. For example, instead of using specific effort and load expressions, French-, Spanish-, and Chinese-speaking participants used expressions that refer to fatigue (French and Spanish) or energy (Mandarin Chinese). Moreover, the French-speaking participants decided on an effort expression that would appear rather extreme in other languages (“I invested myself,” “I gave pain”). Finally, our test of the “lifeblood”-item revealed that this expression of effort is suitable for learning contexts in Mandarin Chinese and Spanish, but not in the other cultures interviewed.

Taken together, our interviews revealed that a translation of mental effort and load self-rating scales should be conducted under multiple considerations. Moreover, although we were able to address many of the above-described issues by lingual adjustments, some difficulties remained unresolved (e.g., with respect to the item intensity or slight differences in meaning). We therefore conclude that while a culturally appropriate translation of effort and load scales is valuable and necessary, researchers should nevertheless interpret such scales with cultural context in mind.

Having obtained these qualitative insights into the cultural understanding of effort and load scales and the tentative item translation, we aimed to assess the reliability and validity of these translations in the quantitative part of this project.

## Quantitative Study—Criterion Validity

To ensure criterion validity, we conducted a quantitative experiment. The translated items were validated using established demand manipulations from the cognitive load literature. We implement three factors to increase specific load types

prevalent in SRL: degree of ICL manipulated by task complexity, degree of ECL manipulated by seductive details, and degree of GCL manipulated by an imaginary scenario and incentives. We validated the English, Mandarin Chinese, and German versions of the scale with first-language participants. Our research aimed to investigate the validity of the self-rating scales across cultures and investigate potential differences between cultures. More specifically, we investigated the following hypotheses:

### **Expectations About Mental Load Items**

*Complexity-increases-load* hypothesis: We expect that participants will rate their mental load higher after completing tasks with higher complexity than after completing tasks with lower complexity (intrinsic-load factor).

*Seductive-details-increase-load* hypothesis: We expect that participants will rate their mental load higher after completing tasks with additional seductive details than after completing tasks without seductive details (extraneous-load factor).

*Scenario-does-not-affect-load* hypothesis: We expect that participants will not differ on their mental load after completing tasks that are embedded in an imaginary scenario with financial incentive and tasks without a scenario and incentive (germane-load factor).

### **Expectations About Mental Effort Items**

*Explorative complexity-effort* hypothesis: We will explore potential differences in mental effort after completing tasks with higher complexity compared to completing tasks with lower complexity (bidirectional hypothesis, intrinsic-load factor).

*Seductive-details-do-not-affect-effort* hypothesis: We expect that participants will not differ on their mental effort after completing tasks with seductive details and after completing tasks without seductive details (extraneous-load, Bayesian null-hypothesis).

*Scenario-increases-effort* hypothesis: We expect that participants will rate their mental effort higher after completing tasks that are embedded in an imaginary scenario than after completing tasks that are neutrally designed (germane-load factor).

## **Methods**

### **Participants and Design**

We first conducted the quantitative study in English, followed by Mandarin Chinese, and German. An a priori power analysis with the software G\*Power (Faul et al., 2007) revealed that for the assumed medium effect size, our experimental

design would require a sample size of 33 participants per language (Cohen's  $f = .25$ ,  $\alpha$ -level,  $p = .5$ , power 80%). To perform additional analyses regarding the item order for the English version of the scale, we decided to increase the size of the English-first-language sample to 70 participants. Participants were recruited and accessed the study online via Prolific. Each volunteer received a financial reward of 4.5£ for their participation. Depending on their performance in some of the tasks, participants could gain up to 1£ of bonus reward (see materials section). The study was programmed and displayed with the software *labvanced* (scicoverly GmbH). To prevent bots and individuals who do not take the study seriously, participants had to successfully complete a captcha and two response quality checks to participate and receive their financial reward. Moreover, a built-in webcam-based eye-tracking function reminded participants to focus on the screen while working on the study (without actually recording eye-tracking-data). Participants using a phone or tablet were prohibited from proceeding with the study.

A total of 142 participants completed the study in the three languages. We excluded three participants who did not indicate the requested mother tongue, two participants who failed at least one out of our two response quality checks, and four participants who answered less than 25% of our tasks correctly indicating that they did not work on the tasks seriously. For another two participants, technical issues led to missing data. We included a final sample of  $N = 131$  ( $n_{\text{English}} = 66$ ,  $n_{\text{Mandarin Chinese}} = 34$ ,  $n_{\text{German}} = 31$ ). Our sample had a mean age of  $M = 36.92$  ( $SD = 11.35$ ). Of the participants, 60.3% indicated their gender as female, with 38.2% indicating male and 1.5% indicating non-binary/third gender.

The experiment followed a  $2 \times 2 \times 2$ -within-subjects design. Participants worked on eight math problems that varied with respect to the three factors: (1) degree of ICL (low complexity vs. high complexity), (2) degree of ECL (without seductive details vs. with seductive details), and (3) degree of GCL (not embedded in an imaginary scenario with feedback and monetary incentive vs. embedded in an imaginary scenario with feedback and monetary incentive). Order and task condition assignments were randomized. Our translated scale from the qualitative study served as a dependent measure.

## Material and Experimental Manipulation

A detailed view of the materials is also available on OSF ([https://osf.io/c7bjx/?view\\_only=daf17411033641d7be1ef231e5a8b3aa](https://osf.io/c7bjx/?view_only=daf17411033641d7be1ef231e5a8b3aa)). Each participant worked on two different types of math problems (see Fig. 2): four arithmetic problems and four word problems (“Boolean algebra”). The *arithmetic problems* consisted of three angled lines meeting at the center that had to be compared in lengths. For this comparison, participants needed to engage in multiple arithmetic operations (addition and/or division). The *Boolean word problems* consisted of short texts about different overlapping sets of items. From those texts, participants were

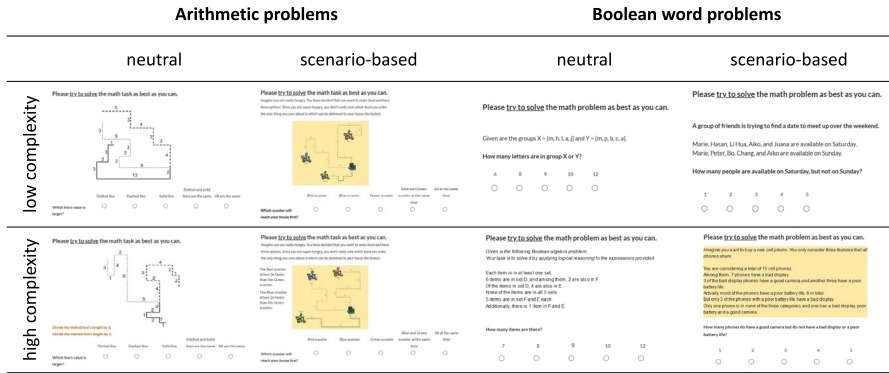


Fig. 2 Problem-solving tasks used in the study

asked to deduce the correct number of items that met certain criteria. For each problem (arithmetic and word), participants were asked to choose the correct answer out of five options.

Following methodological rigor, our materials and procedure were developed and evaluated in an iterative process via three pilot studies with English first language participants ( $n_1=33, n_2=26, n_3=29$ ). Throughout this process, we revised the material according to the feedback and ratings from the participants.

### ICL Manipulation

We manipulated ICL by varying the degree of element interactivity (see Chen et al., 2023 for an overview).

**Arithmetic Problems** For the low intrinsic load version of the arithmetic problem, participants needed to perform multiple additions in order to solve the task. For the high intrinsic load version, participants had to additionally divide some of the lines by specific numbers (see Fig. 2).

**Boolean Word Problems** Compared to the low intrinsic load version, the high intrinsic load word problems had higher overlap and interaction between the sets of items (see Fig. 2).

### ECL Manipulation

Our ECL manipulation relied on the coherence principle indicating that inserting additional irrelevant information into a task increases extraneous cognitive load (Fiorella & Mayer, 2021; cf. *seductive details effect*, e.g., Eitel et al., 2019; Harp & Mayer, 1998). In situations of SRL outside from formal classroom settings (e.g., online learning), such information may consist of personalized advertisements or easily accessible social media sites.



To increase extraneous cognitive load in four of our tasks, we inserted videos displaying interesting facts on famous TV shows on the side of those tasks. Pictures accompanied the facts and were each visible for five seconds before they slowly slid up across the screen to make way for the next fact. We personalized this information by letting participants choose one out of six TV shows at the beginning of the study. Participants only saw the videos on their favored TV show while completing the tasks. The selection of TV shows available to participants was based on official viewer numbers, as well as diversity of genre and provider. Four of the shows are known especially in the Western culture and two are known especially in the Chinese culture. The selection consisted of the following shows: *Stranger Things* (science fiction/mystery, *Netflix*), *Game of Thrones* (fantasy, *HBO*), *The Mandalorian* (space western, *Disney+*), *Bridgerton* (historical romance, *Netflix*), *My Own Swordsman* (comedy, *China Central Television*), *Empresses in the Palace* (historical fiction, *Shaoxing News*). For each TV show, we produced four videos (one for each high ECL task) containing different kinds of facts (i.e., interesting numbers, special effects, stories from the set, facts about actors). All videos were aligned with respect to the number of words (117 to 125 words), idea units (12 to 13 idea units), and pictures (5 pictures). The assignment of the videos to the tasks was randomized for each participant (see Fig. 3).

### GCL Manipulation

To manipulate participants' invested effort, we presented the tasks either in a neutral version or in a (supposedly) motivating version (see Fig. 2). Two measures served to create the motivating version: First, the motivating tasks were accompanied by performance feedback, as well as a monetary incentive (.25 pounds)

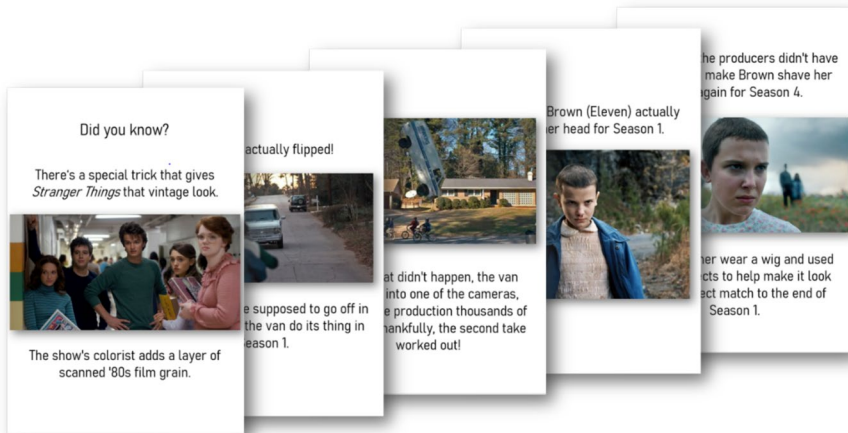


Fig. 3 Screenshots of seductive details videos presented on the side of the tasks

for correctly solving the task (Endres & Eitel, 2024; Kang & Pashler, 2014). For this purpose, we presented the following instruction in advance of the respective tasks: “Now it’s time to shine! Please get ready for the following math task. This task will be followed by feedback. If you are successful, you will receive an extra .25 £.” After completing the task, participants received feedback on whether their answer was right or wrong and whether they earned the additional money or not. Moreover, we presented them with the correct solution.

As a second measure to create a motivating version of the tasks, we additionally embedded them into a scenario. We derived this approach from previous research that successfully increased invested mental effort by designing the learning material in an appealing, game-like way (Woo, 2014), or in an imaginary situation (e.g., Klepsch et al., 2017; Kriegelstein et al., 2022). The scenario manipulation for our two task types was as follows:

**Arithmetic Problem** The neutral version of the arithmetic problems was black and white line drawings depicting the lines to be compared (see Fig. 2). In the motivating version, the task additionally included drawings of a delivery scooter at the beginning of each line and a building at the center where all the lines meet. A short text on the top of the page introduced participants into a scenario where they should imagine that they are really hungry and want to order some food. The task was to determine which of the different delivery scooters would be the first to reach their house (the building in the center) by following its respective line. This version of the tasks was presented on a yellow background and used colored drawings of the scooters and the house (see Fig. 2).

**Boolean Word Problem** The neutral version of the word problems consisted of a short text defining the sets of items and the question to be answered. For the motivating version, the texts were altered so that they would embed the sets of items and question into an everyday problem (see Fig. 2). One scenario described a group of friends trying to find a date to meet for the weekend. In the other scenario, the aim was to determine the best choice for a new cell phone based on different features. Hence, the neutral and motivating versions of the problem were similar with respect to required mental operations and design but differed in wording and number of words (see Fig. 2).

Taken together, participants received four of the tasks in a neutral version without any feedback or incentive (low GCL), and the other four tasks were embedded in a scenario with feedback and incentive instruction.

## Measures

**Control and Demographic Variables** We collected data about participants’ age, gender, occupational background, level of education, mother tongue, and country of residence. Moreover, after completing all tasks, participants were asked to retrospectively rate their prior knowledge, familiarity, and experience with arithmetic and

Boolean algebra tasks on a slider from 0 to 100% (e.g., “How much prior knowledge did you have with arithmetic tasks/Boolean algebra tasks before this study?”; Cronbach’s  $\alpha = .92$ ).

**Our Developed Mental Effort and Load Scale** The mental effort and load scales included the items from the qualitative part of our study. Participants responded to these items once after each task. They received the following instruction: “Please rate your experience while working on the math problem by indicating the extent to which the following statements apply to you.”

All items had to be rated on a 9-point Likert scale from 1 *strongly disagree* to 9 *strongly agree*. The scales of passively perceived mental load (e.g., “Working on this activity was mentally demanding”) and actively invested mental effort (e.g., “I invested effort while working in this activity”) each consisted of four statements (see Table 2).

We tested an additional statement that also represents mental effort, but presumably is rather specific for the Chinese culture (“I put my heart and soul into working on this activity”). Moreover, one item served to measure the perceived difficulty of working on the task—primarily to test whether participants differentiate between the two constructs difficulty and load (“Working on this activity was difficult”).

In the English study, which we conducted first, the item order was block-randomized per subscale. That is, participants were randomly assigned to one of six item orders that stayed the same for the whole experiment (i.e., all eight measuring points). After observing no (interaction) effect of the item randomization (all  $BF_{01} > 3.98$ ), we decided to only use one item order for the other languages to reduce the technical and computational requirements of our study. The detailed analyses are available on OSF.

**Established Cognitive Load Scale** To compare our results to an already established cognitive load scale for purposes of validity, participants completed the items from the scale by Klepsch et al. (2017) after each task. On a 9-point Likert scale from 1 *strongly disagree* to 9 *strongly agree* participants indicated their ICL (2 items plus one item added by the authors; e.g., “This task was complex”; Cronbachs  $\alpha = .89$ ), their ECL (3 items; e.g., “During this task, it was difficult to recognize and link the crucial information”; Cronbachs  $\alpha = .84$ ), and their GCL (3 items; e.g., “My point while dealing with the task was to understand everything correctly.”; Cronbachs  $\alpha = .64$ ).

## Procedure

Participants could access our study via the subject pool Prolific and participated online. They were first informed about the study’s goal, conditions of participation, and data processing. After giving their informed consent to participate in the study, participants answered questions on their demographics. They then got the instruction that they would work on eight short problem-solving tasks, and

after each task, they were to rate their experience while working on these tasks. Moreover, they were informed that some tasks would allow them to earn additional money. A multiple-choice attention check served to confirm that all participants understood the instructions correctly. Subsequently, participants indicated the TV show they would favor (see ECL manipulation). Participants then worked on the eight tasks in their own speed in randomized order. After each task, participants indicated their perceived mental effort and load on our translated scales. Moreover, they filled out the cognitive load scale by Klepsch et al. (2017). Once participants had completed all tasks, they were asked to indicate their prior knowledge and to type in their age again (for quality check purposes). Moreover, they were asked to indicate whether they had been distracted or disrupted while working on the tasks and also asked to give feedback on the overall study and the clarity of instructions.

## Results

We took a rigorous three-step approach to analyzing our data. First, we explored the psychometric properties of our scales and decided upon a long and short version of our scale for each language. Second, we explored criterion validity via hypotheses tests with three-factorial repeated-measures ANOVAs (within-subjects). Additional Bayesian analysis (Bayes Factor; BF) served to test null hypotheses. Third, we explored the convergent validity of our scales by conducting correlation analyses with established scales measuring cognitive load types (Klepsch et al., 2017). We conducted the frequentist analyses with the software IBM SPSS Statistics (alpha level of .05) and used two-sided tests for all statistical analyses.  $\eta^2$  served as the effect size index, with values of .01, .06, and .14 considered small, medium, and large effects, respectively. The Bayesian analyses were conducted with the software JASP (JZS prior), with BF values of 1–3, 3–10, 10–30, 30–100, and over 100 considered anecdotal, moderate, strong, very strong, and extreme evidence, respectively (Lee & Wagenmakers, 2014).

### Analysis of the Scale Properties

To decide upon the final versions of our scales, we first explored internal consistencies via Cronbach's  $\alpha$  using the ratings of all measuring time points. In an iterative process, we then explored different scale versions considering reliability and validity. The goal of this process was to achieve a good internal consistency while at the same time maintaining the content and criterion validity of the scale. Table 3 shows a proposed 4-item (long) and 3-item version (short) of the two subscales per language, as well as the corresponding Cronbach's alphas. Overall, our scales showed good to excellent internal consistencies (all  $\alpha > .81$ ). Interestingly, in all languages, the effort scale showed

**Table 3** Final long and short versions of our effort and load scales (with corresponding Cronbach's  $\alpha$ s) per language

Overall* (N = 131)	English (N = 65)	Mandarin Chinese (N = 34)	German (N = 31)
<b>Mental load</b>			
Load Item 1 <sup>L,S</sup>	Working on this activity required effort. <sup>L,S</sup>	完成这个任务需要投入努力。 <sup>L</sup>	Das Bearbeiten der Aufgabe hat mich Mühe gekostet. <sup>L,S</sup>
Load Item 2 <sup>L,S</sup>	Working on this activity was mentally demanding. <sup>L,S</sup>	完成这个任务很消耗脑力。 <sup>L,S</sup>	Das Bearbeiten der Aufgabe war anstrengend. <sup>L,S</sup>
Load Item 3 <sup>L,S</sup>	Working on this activity was effortless. <sup>L,S</sup>	完成这个任务是毫不费力的。 <sup>L,S</sup>	Das Bearbeiten der Aufgabe war mühelos. <sup>L,S</sup>
Load Item 4 <sup>L</sup>	Working on this activity did not mentally strain me. <sup>L</sup>	完成这个任务不需要我绞尽脑汁。 <sup>L,S</sup>	Das Bearbeiten der Aufgabe hat keinen Aufwand von mir gefordert. <sup>L</sup>
4-item-scale <sup>L</sup> ; $\alpha = .87$	4-item-scale <sup>L</sup> ; $\alpha = .88$	4-item-scale <sup>L</sup> ; $\alpha = .85$	-item-scale <sup>L</sup> ; $\alpha = .91$
3-item-scale <sup>S</sup> ; $\alpha = .85$	3-item-scale <sup>S</sup> ; $\alpha = .86$	3-item-scale <sup>S</sup> ; $\alpha = .87$	3-item-scale <sup>S</sup> ; $\alpha = .92$
<b>Mental effort</b>			
Effort Item 1 <sup>L,S</sup>	I invested effort while working on this activity. <sup>L,S</sup>	我投入了努力去完成这个任务。 <sup>L,S</sup>	Ich habe mich bei der Bearbeitung der Aufgabe bemüht. <sup>L,S</sup>
Effort Item 2	I did not mentally strain myself while working on this activity	我没有让自己绞尽脑汁地去完成这个任务。	Ich habe mich bei der Bearbeitung der Aufgabe nicht angestrengt. <sup>L</sup>
Effort Item 3 <sup>L</sup>	I refrained from putting effort into this activity. <sup>L</sup>	我没有为这个任务付出任何努力。 <sup>L</sup>	Ich habe keine Anstrengung in diese Aufgabe investiert
Effort Item 4 <sup>L,S</sup>	I worked hard on this activity. <sup>L,S</sup>	我很努力地地去完成这个任务。 <sup>L,S</sup>	Ich habe hart an dieser Aufgabe gearbeitet. <sup>L,S</sup>
Lifeload Item <sup>L,S</sup>	I put my heart and soul into working on this activity. <sup>L,S</sup>	我全身心投入到了这个任务中。 <sup>L,S</sup>	Ich habe die Aufgabe mit Herzblut bearbeitet. <sup>L,S</sup>
4-item-scale <sup>L</sup> ; $\alpha = .81$	4-item-scale <sup>L</sup> ; $\alpha = .83$	4-item-scale <sup>L</sup> ; $\alpha = .81$	4-item-scale <sup>L</sup> ; $\alpha = .82$
3-item-scale <sup>S</sup> ; $\alpha = .88$	3-item-scale <sup>S</sup> ; $\alpha = .90$	3-item-scale <sup>S</sup> ; $\alpha = .87$	3-item-scale <sup>S</sup> ; $\alpha = .82$

Scale versions are the result of an iterative process considering reliability and validity

<sup>L</sup>Proposed long version (4 items)

<sup>S</sup>Proposed short version (3 items)

\*The overall scale versions are based on a sample with unequal language distribution (50% English participants) and therefore do not represent a general recommendation for the scales. We recommend adapting the scale version to the respective language (see other columns)

better psychometric quality (reliability and validity) when the presumably more extreme item referring to “heart and soul” (lifeblood-item) was included instead of one of the other effort items (either item 2 or 3, see Table 3). This increase in psychometric quality was more obvious for the samples with English- and Chinese-speaking participants than in the sample with German-speaking participants.

For the following analyses, we computed the means of the two scales (effort and load, long version) per condition. Detailed means and standard deviations are available on OSF. The mean over all conditions was  $M = 5.58$  ( $SD = 1.39$ ) for the load ratings and  $M = 6.95$  ( $SD = 1.37$ ) for the effort ratings.

### Analysis of Criterion Validity of Our Scale

We used a frequentist repeated measures ANOVA to test our hypotheses regarding the criterion validity of our scales. The three task manipulations (ICL, ECL, and GCL) served as factors, and the effort or load ratings (long-scale versions) served as dependent variables. Additionally, we performed the same analysis with Bayesian methods to allow for evidence supporting the null hypothesis. Table 4 gives an overview of the results of these tests. In the following section, we describe the hypothesis tests for the overall sample (all three languages).

### Effects on Mental Load

**Complexity-Increases-Load Hypothesis** In line with our hypothesis that mental load ratings would be higher after tasks with higher complexity compared to tasks with lower element interactivity (ICL manipulation), the frequentist ANOVA revealed a significant effect of the ICL manipulation on load ratings in the expected direction,  $F(1, 130) = 243.70$ ,  $p < 0.001$ ,  $\eta^2 = .65$ . This pattern of results was present in all three languages.

**Seductive-Details-Increase-Load Hypothesis** In line with our hypothesis that mental load ratings would be higher after tasks with additional interesting elements (Seductive Details; ECL manipulation) compared to tasks without such additional elements, the frequentist ANOVA revealed a significant effect of the ECL manipulation on load ratings in the expected direction,  $F(1,130) = 4.52$ ,  $p = .035$ ,  $\eta^2 = .03$ . Surprisingly, we observed this pattern of results for the English-speaking sample, but not for the German- and Mandarin Chinese-speaking samples.

**Scenario-Does-Not-Affect-Load Hypothesis** The Bayesian ANOVA revealed anecdotal evidence for our null hypothesis that participants would not differ on their mental load ratings after completing tasks that are embedded in an imaginary scenario with financial incentive and tasks without a scenario and incentive (GCL manipulation, H3),  $BF_{01} = 3.41$ . We observed moderate evidence in favor

**Table 4** Results of hypotheses tests per language subsample and for the overall sample

	Overall (N = 130)		English (N = 65)		Mandarin Chinese (N = 34)		German (N = 31)	
	Frequentist	Bayesian	Frequentist	Bayesian	Frequentist	Bayesian	Frequentist	Bayesian
<b>Effects on mental load (4-item scale version)</b>								
<i>Complexity-increases-load</i> hypothesis	$p < .001, \eta^2 = .65$	-	$p < .001, \eta^2 = .60$	-	$p < .001, \eta^2 = .71$	-	$p < .001, \eta^2 = .70$	-
<i>Seductive-details-increase-load</i> hypothesis	$p = .035, \eta^2 = .03$	-	$p = .014, \eta^2 = .09$	-	$p = .875, \eta^2 < .01$	-	$p = .515, \eta^2 < .01$	-
<i>Scenario-does-not-affect-load</i> hypothesis	$p = .054, \eta^2 = .03$	<b>BF<sub>01</sub> = 3.41</b>	$p < .008, \eta^2 = .04$	BF <sub>01</sub> = .62	$p = .654, \eta^2 = .01$	<b>BF<sub>01</sub> = 7.82</b>	$p = .779, \eta^2 = .01$	<b>BF<sub>01</sub> = 6.51</b>
<b>Effects on mental effort (4-item scale version)</b>								
<i>Explorative complexity-effort</i> hypothesis	$p < 0.001, \eta^2 = .21$	BF <sub>01</sub> < 1/100	$p < 0.001, \eta^2 = .35$	BF <sub>01</sub> < 1/100	$p = .032, \eta^2 = .13$	BF <sub>01</sub> = .17	$p = .133, \eta^2 = .07$	BF <sub>01</sub> = 2.73
<i>Seductive-details-do-not-affect-effort</i> hypothesis	$p = .135, \eta^2 = .02$	<b>BF<sub>01</sub> = 6.32</b>	$p = .155, \eta^2 = .03$	<b>BF<sub>01</sub> = 5.60</b>	$p = .139, \eta^2 = .07$	<b>BF<sub>01</sub> = 3.89</b>	$p = .375, \eta^2 = .03$	<b>BF<sub>01</sub> = 6.99</b>
<i>Scenario-increases-effort</i> hypothesis	$p = .488, \eta^2 < 0.01$	-	$p = .974, \eta^2 < 0.01$	-	$p = .350, \eta^2 = .03$	-	$p = .146, \eta^2 = .07$	-

Frequentist as well as Bayesian analyses rely on three-factorial within-subjects ANOVAs with the ICL, GCL, and ECL manipulations as factors and mental load or effort ratings as dependent variables. Bold printed values represent evidence in favor of our hypotheses

of a null effect in the Mandarin Chinese and German subsample, while there was anecdotal evidence in favor of an effect in the English subsample (i.e., a difference between conditions).

### Effects of Mental Effort

**Explorative Complexity-Effort Hypothesis** The frequentist ANOVA revealed a significant effect of the ICL manipulation (i.e., task complexity) on effort ratings,  $F(1,130)=35.40$ ,  $p<0.001$ ,  $\eta^2=.21$ . Participants showed higher effort ratings after tasks with higher complexity compared to tasks with lower complexity. This was supported by strong evidence for an effect in the Bayesian ANOVA,  $BF_{10}>100$ . This pattern of results was present in the English and Mandarin Chinese, but not in the German subsample.

**Seductive-details-Do-Not-Affect-Effort Hypothesis** The Bayesian ANOVA revealed moderate evidence for our null hypothesis that participants would not differ on their mental effort ratings after completing tasks with additional interesting elements (seductive details, ECL manipulation) and tasks without such elements,  $BF_{01}=6.32$ . We observed this pattern of results in all subsamples with the Mandarin Chinese study showing only anecdotal evidence.

**Scenario-Increases-Effort Hypothesis** Contrary to our hypothesis that mental effort ratings would be higher after tasks that are embedded in an imaginary scenario with a financial incentive compared to tasks without a scenario and an incentive (GCL manipulation), the frequentist ANOVA revealed no main effect of the GCL manipulation on effort ratings,  $F(1,130)=.49$ ,  $p=.488$ ,  $\eta^2=.01$ . Interestingly, further analyses revealed a significant interaction effect between the GCL and ICL Factor in the overall sample ( $p=.003$ ,  $\eta^2=.06$ ), indicating that the GCL manipulation might have the assumed effect on effort only in conditions with high ICL (i.e., tasks of higher complexity). A subsequent two-factorial ANOVA with the GCL and ECL manipulation as factors that we performed only for the high ICL conditions supported this observation,  $F(1,130)=6.66$ ,  $p=.011$ ,  $\eta^2=.05$ . In the tasks with high complexity, embedding the tasks in a scenario with incentives leads to higher effort ratings compared to tasks without such measures. This pattern of results was also present for the English and Chinese subsample but failed to reach the level of statistical significance (English:  $p=.109$ ,  $\eta^2=.04$ ; Mandarin Chinese:  $p=.127$ ,  $\eta^2=.07$ ).

### Analysis of Convergent and Incremental Validity of Our Scale

To investigate the convergent and incremental validity of our scale, we analyzed its relation to the conceptualization of cognitive demand proposed by CLT. As stated in the theory section, we expected that our conceptualization of cognitive demand in SRL—namely mental load and mental effort—would overlap with



the three constructs of CLT (ICL, ECL, and GCL), but still represent broader constructs. We examined whether our scales would show specific correlation patterns with the three CLT constructs measured by the established scale by Klepsch et al. (2017). We hypothesized that our mental load scale would show higher correlations with the ICL and ECL scales than with the GCL scale. Similarly, we expected higher correlations between our mental effort scale and the GCL scale than with the ICL and ECL scales. These expectations were based on our scale's conceptualization and a previous study by Klepsch and Seufert (2021) showing similar correlation patterns. As shown in Table 5, our data generally exhibited the expected correlation patterns. Unexpectedly, we found a moderate, statistically significant correlation between mental load and GCL. Additionally, consistent with previous findings (e.g., Klepsch & Seufert, 2021), mental effort also correlated significantly with ICL. When comparing our results to those of Klepsch and Seufert (2021), we observed similar correlation strengths: mental load with ICL ( $r = .63$ ) and ECL ( $r = .44$ ). Especially for mental effort, our multi-item scale showed a slightly different pattern. This different pattern is desirable as our items cover a wider range of learning scenarios as compared to single-item scales.

To further investigate the incremental validity of our scale, we compared our load scale with the difficulty item ("Working on this activity was difficult") that we also included in our study. As mentioned in the theory section, we expected that the construct of difficulty would overlap with the broader load construct but would represent only specific aspects of load. Specifically, while our load construct should be sensitive to task complexity (i.e., ICL manipulation) and other aggravating factors such as seductive details (i.e., ECL manipulation), the difficulty item should only be sensitive to task complexity, not to other aggravating factors.

Overall, we observed a significant correlation between the difficulty and load constructs ( $r = .77$ ,  $p < 0.001$ ). To test our expectations regarding the effect of the manipulations on difficulty appraisal, we performed both a frequentist and a Bayesian three-factor ANOVA with the three manipulations (ICL, ECL, and GCL) as factors and the load ratings as dependent variables. There was a significant effect of the ICL condition on difficulty ratings,  $F(1,130) = 265.11$ ,  $p < 0.001$ ,  $\eta^2 = .67$ ,  $BF_{01} < 1/100$ , with higher difficulty ratings after the high ICL tasks. As expected, such an effect could not be observed for the ECL

**Table 5** Correlations of our mental effort and load scale (long versions) with the CLT scales by Klepsch et al. (2017)

Our scale	Scale by Klepsch et al. (2017)		
	ICL	ECL	GCL
Mental load	0.76**	0.46**	0.29* <sup>a</sup>
Mental effort	0.44**	0.01	0.59*** <sup>a</sup>

\*\* $p < 0.001$ ; \* $p < 0.01$ ; <sup>a</sup> $N = 102$ : smaller sample due to technical issues. For an item-to-item correlation table, please see Appendix

manipulation,  $F(1,130) = .95$ ,  $p = .331$ ,  $\eta^2 < 0.01$ , with  $BF_{01} = 9.83$  with the data providing moderate evidence for a null effect and showing incremental validity of our scale.

## General Discussion

The goal of this study was to develop and validate a cognitive demand scale that distinguishes between two components of cognitive demand, namely mental effort and mental load, across different languages. With methodological rigor, we conducted various steps to ensure different types of validity, including content validity, criterion validity, and convergent and incremental validity. These efforts were aimed at establishing a robust, reliable, and language-adaptable scale. We found evidence for all investigated types of validity.

### Content Validity

To establish content validity, we conducted qualitative interviews with bilingual individuals to develop translations of the items, which were then evaluated and adapted by our language experts. This process revealed that while the core concept of cognitive demand was universally shared, slight differences in interpretation existed across languages. To address these differences, we focused on translating items to preserve their intended meaning rather than relying on literal translations. This approach ensured that the items retained their conceptual integrity across languages and showed the importance of a language-sensitive translation, beyond literal word-by-word translations.

Additionally, we used insights from these qualitative interviews to inform variations in our quantitative study. Although we employed a classical manipulation of task complexity, we also introduced a monetary incentive as a motivational manipulation, reflecting themes that emerged in discussions about mental effort. Participants indicated that financial rewards were a significant motivator which might influence their mental effort, highlighting our dual approach of qualitative and quantitative methods as being essential to create a robust scale that accounts for linguistic nuances.

### Criterion Validity

Criterion validity was assessed through a quantitative study to determine whether the scale could predict specific behaviors. This type of validity ensures that the scale accurately measures changes in task aspects, in SRL situations. Consistent measurability of task differences among participants within the same

language and across different languages was a key requirement. Our comparative analysis revealed that the scale produced comparable effects between most languages, indicating its effectiveness in diverse linguistic contexts. Translingual criterion validity was achieved as manipulations yielded comparable effect sizes across similarly differing tasks.

**Complexity Criterion (ICL)—Mental Load** Criterion validity was established for our complexity manipulation in the mental load scale. This validity was confirmed across all three languages—English, German, and Mandarin Chinese—and held true even when analyzed separately. These consistent results across different linguistic groups demonstrate that our central affordance was effectively met. The robustness of our scale in assessing mental load related to task complexity was evident, supporting its use in diverse linguistic contexts. By confirming criterion validity across multiple languages, we ensured that the scale can be confidently used in international research, facilitating cross-language comparisons and collaborations for future empirical contributions.

**Complexity Criterion (ICL)—Mental Effort** Our complexity manipulation also influenced the investment of mental effort, showing that learners invested higher mental effort when confronted with a more complex task. This effect is consistent with the literature, especially in an SRL context, where learners tend to be more engaged when challenged (van Merriënboer & Sweller, 2005; Paas et al., 2003b; Deci & Ryan, 1985; Greene & Azevedo, 2009; Sirock et al., 2023). However, this effect was not consistent across all languages. In German, the effect was in the same direction but failed to reach statistical significance.

This inconsistency may have multiple reasons. One possibility is that the motivational effects of complex tasks are limited in controlled learning situations. In our experiment, we chose this manipulation to compare our interventions with established CLT research. However, as learners were not able to choose whether they wanted to work on a more complex task, the effect might have been smaller than in an authentic context. A smaller effect size of complexity as a source of motivation would explain our pattern of results, as the German sample was not doubled (as the English sample was) and smaller effects were therefore less likely to reach statistical significance. Further research should investigate the role of choosing more complex tasks and its impact on motivation. It is challenging to distinguish between the motivation evoked by the autonomy of choice and the higher complexity as a motivational factor (Deci & Ryan, 1985). Understanding these dynamics will enhance our comprehension of how task complexity and autonomy interact to influence learners' motivation to invest mental effort.

**Seductive Details Criterion (ECL)** Our ECL criterion for mental load was effectively met. The manipulation using seductive details elicited a higher overall mental load, as hypothesized. There was no effect on mental effort, which aligns with

our expectations. However, the effects in the German and Mandarin Chinese samples were slightly smaller than in the English sample. This discrepancy could be attributed to several factors.

Firstly, we designed our seductive details to align with learners' individual interests (e.g., favorite series) rather than directly relating them to the learning material. It is possible that a more significant ECL effect would occur if the seductive details were more closely tied to the content being learned (Bender et al., 2024; Harp & Mayer, 1998).

The limited ECL effects on mental load in the German and Mandarin Chinese samples could also be due to our selection of TV shows. To account for cultural differences, we chose four internationally renowned series and two series famous in China. This intervention seemed successful, as most participants in the Mandarin Chinese sample chose the series popular in their cultural background. However, the Chinese sample had only two series options compared to four in the English sample, potentially not meeting all participants' preferences and reducing the effect size. In the German sample, the smaller effect size of the manipulation might be due to the potential difference in the popularity of the selected series. Although we picked well-known series based on international data, they might still be more popular in English-speaking countries. These potential explanations could be interesting for research on seductive details. Further research could investigate the importance of the relevance or adaptability of seductive details and their influence on mental load.

**Motivational Scenario Criterion (GCL)** In our motivational manipulation, we combined multiple interventions (monetary goal, scenario-based learning, and game-based learning). We observed an increase in mental effort only when more complex tasks were used. This limitation can be explained by the fact that the easy tasks were very straightforward, as evidenced by the lack of mistakes made by hardly any participant. Therefore, it appears unnecessary to invest mental effort in very easy tasks. Given that the effect was robust in high-complexity conditions, we still consider our criterion validity to be met.

Our results highlight that motivation and mental effort are primarily necessary for challenging tasks. Future research validating translations of the introduced items should include slightly more complex tasks, even in the easier versions, to determine if this interaction holds true at higher levels of complexity. This approach will help ensure that the motivational effects and mental effort are adequately captured across different task difficulties.

## Convergent and Incremental Validity

Our investigation of our scale's convergent and incremental validity also appears to have been successful. First, convergent validity was met as our cognitive demand scale correlated with the expected facets of the established CLT scales (Klepsch

et al., 2017). We observed similar correlations to those found by Klepsch and Seufert (2021) regarding our mental load scales. This pattern of results highlights the usability of our scale in contexts comparable to CLT-based research.

More importantly, our scale also showed substantial incremental validity. Our multi-item scale is more accurately aligned with the motivational variations implemented to increase mental effort than single-item scales. Although we found a significant correlation with GCL in established load scales, the GCL scale was not sensitive enough to identify the differences evoked by our motivational criterion variation. This finding demonstrates the incremental focus of a broader understanding of learning-related mental effort. Our broader focus on mental effort encompasses comprehension and other aspects beneficial for learning (e.g., knowledge consolidation by desirable difficulties). Mental effort covers a wider range of learning scenarios than classical GCL scales.

Additionally, our mental load sub-scale showed incremental validity, especially in contrast to difficulty appraisals. Our analysis indicated that the ECL manipulation influenced our mental load scale while having no effect on difficulty appraisals. This pattern of results highlights a distinction between mental load and difficulty appraisal. Our mental load scale is sensitive to more than just difficulty and thus covers a wide range of load aspects important in SRL.

As explained in our literature review, influences on mental effort can include multiple factors such as metacognition, length of learning session, or other cognitive processes such as worry cognitions. Our mental load scale can be seen as a broad-range scale that encompasses these aspects in SRL. Specific research fields could use it to understand interactions between these effects, such as the interaction between higher metacognitive demand and the length of learning scenarios. This might call for a finer distinction of mental load in SRL, focusing on specific cognitive processes.

## Limitations

### Base language

One potential limitation of our study lies in the approach of starting our translation process from only one base language per item set. We translated the separate items either from German or English. We started with a set of English items as the international composition of our research team and the widespread use of English as the language of science made the collaboration the easiest. Further, most of the research tradition of CLT (and SLR) is based upon research conducted with English-based participants and materials. The materials were then translated to German, in a strict procedure ensuring idiomatic accuracy and cultural appropriateness. From there on, we also considered German as a base language and translated other languages to facilitate the recruitment of bilingual interview partners for our qualitative study phase.

Despite our rigorous translation process, a possibility of discrepancies when translating from different base languages remains. Specifically, translating Dutch from English might result in nuanced differences that could affect the scales' comparability across languages. Although we aimed to ensure linguistic accuracy and cultural relevance, employing both English and German may introduce discrepancies. This potential bias might be less pronounced for the translation from German to French, as the language experts were multilingual and could consider both the English and German materials in their review. Future research could involve more multilingual translators and triangulate their translations to identify potential biases. Using the MEL-TS scales in future studies will be crucial for evaluating these issues and ensuring the robustness of the translations.

### **Factorial Validity**

One potential avenue to further demonstrate the robustness of our scale is by addressing factorial validity. In our study, we focused on elements of construct validity: content validity, criterion validity, convergent validity, and incremental validity. Although our within-subjects design effectively addressed these types of validity, it did not encompass factorial validity. Factorial validity typically requires between-subjects data and a fully representative sample across all languages involved in the study. Addressing this limitation was beyond the logistical constraints of our current research.

Future research, particularly those using international large-scale assessments, could address this gap. Such studies often have access to between-subjects data and diverse, representative samples, making them well-suited for conducting confirmatory factor analyses (CFAs) to test for measurement invariance across languages and cultures. These efforts would significantly contribute to validating and refining our scales for cognitive demand on a broader, more generalizable level. Pursuing factorial validity in future studies will ensure that constructs are consistently and accurately interpreted across different cultural contexts, enhancing the robustness and applicability of the developed scales.

### **Potential Ceiling Effects in the Effort Scale in Mandarin Chinese**

Another limitation is the potential ceiling effect in the Mandarin Chinese scale for mental effort. As identified in our literature review, investing effort is seen as a highly favorable behavior in many Asian countries, which may lead to socially desirable response patterns in our scale. Ceiling effects, where many respondents score at the upper limit, challenge the validity of the rating dimensions. These effects hinder differentiation in effort and load ratings.

Although the ceiling effects do not seem problematic, the Mandarin Chinese scale could be improved. One way to address this is by including more extreme quantifiers in the item sets. For example, the item "I invested effort

while working on this activity" could be modified to "I invested much effort while working on this activity" (我投入了大量的努力去完成这个任务). Such an adaptation might better capture differences in effort in Asian countries. Additionally, we could revise the term used for effort. In Chinese culture, "effort" (努力) can imply a moral judgment, prompting respondents to inflate their ratings. To address this, we could avoid terms with moral implications and emphasize "mental effort" (认知努力) in all items, reducing ambiguity. In conclusion, addressing ceiling effects in the effort scale should involve refining item specificity and mitigating social desirability bias. These steps will enhance the validity of effort ratings in cross-cultural research.

### Practical Considerations of MEL-TS

In our criterion validation studies, we used a repeated measures design. We think that the positive outcomes of our study might be influenced by the multiple ratings involved in such designs. This idea provides us with some practical application guidelines on how to assess cognitive demand with a higher quality. The repetition of items might help (a) learners' metacognition and (b) our statistical analysis. From the perspective of learners' metacognition, the initial ratings of mental effort and mental load can help learners as a point of internal reference for subsequent ratings. This point of internal reference makes it easier for learners to judge the differences between tasks and come to a more calibrated understanding of their perceived cognitive demand. From a statistical perspective, this methodological approach allows us to control for between-subject variability by focusing on within-subject variance, thereby increasing the validity of our measurements. Similar procedures have been used in experimental between-subject designs (e.g., Endres et al., 2024a). The authors implemented a reference point evaluation by first asking learners to rate mental effort and mental load after completing a task with unvaried learning material as a reference point. Learners provided a second rating after engaging in a varied learning task that was of actual interest to the specific research question. This design allowed the researchers to use the initial ratings as an internal reference point in their statistical analysis, which may have also supported learners' metacognition. Both points likely increased the credibility of their assessed self-rated data. The between-subjects design led to robust mediation results, demonstrating a clear positive effect of mental effort investment on learning and a negative effect of cognitive load on learning.

A second practical consideration is the number of items to include in future studies. Many studies only assess mental effort using a single item. This is very plausible as single-item measures offer advantages, particularly in terms of practicality and participant satisfaction (Allen et al., 2022). For narrow or highly homogeneous constructs, the validity of single-item measures can also be acceptable (Allen et al., 2022). However, as demonstrated in our criterion validation, a higher number of items generally led to more favorable statistical outcomes (e.g., Andersen & Makransky, 2021; Leppink et al., 2013).

Furthermore, our findings indicated that selecting different single items could have resulted in varying interpretations by participants, suggesting that mental effort and cognitive load ratings are not as narrow constructs as sometimes assumed. We assume that cognitive demand is better captured with multiple items, especially when, as in our scales, both positive and negative item formulations were included. Additionally, our results show, that ceiling effects of single-item measures can be avoided when using multiple items with different formulations (notably observed in our studies with the Chinese sample).

A third consideration is that each item was formulated broadly to encompass various learning situations. If researchers wish to examine different aspects of cognitive demand, we recommend tailoring the item's to-be-rated object to the specific learning context. For example, when a task requires both retrieval and elaboration of content, researchers may want to assess the effort invested in each task separately. In such cases, we recommend adapting the item to reflect the distinct demands of each task.

To summarize, as a practical guideline, we recommend using the MEL-TS with an internal point of reference evaluation to enhance the credibility of the measure (e.g. assessed after a neutral first task). As best practice, we suggest selecting as many items as feasible, given practical constraints of the individual study, with a range of 1 to 4 items per construct.

### **Advancing Translingual Research: A Call for Collaboration and Knowledge Sharing**

The methodological framework presented in this paper has successfully achieved a validated mental demand scale across different languages. This framework has the potential to inspire further validation in additional languages. Expanding our research into more diverse linguistic and cultural contexts will enhance the generalizability of our findings in educational science, contributing to a broader and more nuanced understanding of SRL and its translingual differences. This could lead to important empirical contributions.

Showing our commitment to methodological rigor, transparency, and supporting translingual research, we have shared all materials and resources on the Open Science Framework. Our aim is to facilitate collaboration and encourage the dissemination of knowledge across language barriers. We welcome collaboration with researchers interested in enriching the linguistic diversity of this research, offering qualitative expertise and experience in support. We hope our research marks the beginning of a journey, paving the way for comprehensive and profound explorations in translingual research within educational science.



Appendix

Variable	M	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1.Lead1	6.14	1.72																		
2.Lead2	5.01	1.83	.85																	
3.Lead3	6.29	1.45	-.43	-.58																
4.Lead4	5.01	1.83	-.41	-.39	-.78															
5.Outbody	6.65	1.68	.71	.59	-.43	-.56														
6.Outm1	6.94	1.77	.71	.59	-.45	-.57	.19													
7.Outm2	6.57	1.79	-.51	-.29	.79	.54	-.35	-.22												
8.Outm3	2.54	1.17	-.21	-.29	.59	.52	-.15	-.46	.29											
9.Outm4	6.90	1.84	.71	.51	-.41	-.48	-.14	-.18	.29	.29										
10.BodyBlood	5.64	2.47	.51	.44	-.13	-.18	-.14	-.14	.52	.52	.29									
11.OSL1	3.76	1.47	.39	.44	-.27	-.28	-.65	.28	-.24	.26	.27	.19								
12.OSL2	4.13	1.58	.17	.29	-.69	-.46	.38	-.69	.10	.33	-.10	-.24	.66							
13.OSL3	4.09	1.39	.38	.31	-.49	-.18	-.61	.18	-.14	.30	.17	.06	.84	.65						
14.OSL4	6.41	2.01	.66	.46	-.14	.05	.38	.35	.00	-.43	.30	.79	.29	-.87	.08					
15.OSL2	7.25	1.56	.51	.37	.87	.14	.22	.79	.01	-.40	.68	.61	.62	-.14	-.08	.38				
16.OSL3	5.92	1.86	.53	.37	.18	.07	.24	.08	-.05	-.23	.67	.75	.20	-.14	.09	.84	.66			
17.KL1	5.90	1.76	.75	.72	-.17	-.22	.59	.71	-.26	-.18	-.18	.72	.61	.48	.13	.35	.00	.27	.08	
18.KL2	4.60	1.75	.76	.80	-.47	-.35	.89	.49	-.33	-.09	-.53	.44	.63	.29	.55	.44	.30	.40	.21	.71
19.KL3	5.34	1.75	.85	.80	-.43	-.46	.83	.61	-.46	-.12	.64	.55	.62	.20	.48	.65	.45	.54	.84	.90

**Funding** Open Access funding enabled and organized by Projekt DEAL. This project was supported by the colleges of the EARLI Emerging Field Group (EFG) “Unifying Cognitive Load and Self-Regulated Learning Research: Monitoring and Regulation of Effort (MRE).”

Financial support was provided by the Chairs of the universities of Freiburg, Maastricht, Kassel, Rotterdam, and Osnabrück.

**Data Availability** Material and data of the project are available on OSF: [https://osf.io/c7bjx/?view\\_only=daf17411033641d7be1ef231e5a8b3aa](https://osf.io/c7bjx/?view_only=daf17411033641d7be1ef231e5a8b3aa).

## Declarations

**Ethics Approval** This study was approved by the ethics board of the Ruhr University Bochum, Germany (registration number: EPE-2022–029, October 2022).

**Conflict of Interest** The authors declare no competing interests.

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

## References

- Allen, M. S., Iliescu, D., & Greiff, S. (2022). Single item measures in psychological science. *European Journal of Psychological Assessment*, *38*(1), 1–5. <https://doi.org/10.1027/1015-5759/a000699>
- Andersen, M. S., & Makransky, G. (2021). The validation and further development of a multi-dimensional cognitive load scale for virtual environments. *Journal of Computer Assisted Learning*, *37*(1), 183–196. <https://doi.org/10.1111/jcal.12478>
- Ayres, P., Lee, J. Y., Paas, F., & van Merriënboer, J. J. G. (2021). The validity of physiological measures to identify differences in intrinsic cognitive load. *Frontiers in Psychology*, *12*, 702538. <https://doi.org/10.3389/fpsyg.2021.702538>
- Baars, M., Wijnia, L., de Bruin, A., & Paas, F. (2020). The relation between students’ effort and monitoring judgments during learning: A meta-analysis. *Educational Psychology Review*, *32*(4), 979–1002. <https://doi.org/10.1007/s10648-020-09569-3>
- Bender, L., Renkl, A., & Eitel, A. (2021a). When and how seductive details harm learning. A study using cued retrospective reporting. *Applied Cognitive Psychology*, *35*(4), 948–959. <https://doi.org/10.1002/acp.3822>
- Bender, L., Renkl, A., & Eitel, A. (2021b). Seductive details do their damage also in longer learning sessions - When the details are perceived as relevant. *Journal of Computer Assisted Learning*, *37*(5), 1248–1262. <https://doi.org/10.1111/jcal.12560>
- Bender, L., Brosemer, K., Renkl, A., Endres, T., Eitel, A. (2024). Keep some distance: Seductive details are only harmful when closely related to the learning content. 10.17605/OSF.IO/9FKRW
- Bjork, R. A., & Bjork, E. L. (2020). Desirable difficulties in theory and practice. *Journal of Applied Research in Memory and Cognition*, *9*(4), 475–479. <https://doi.org/10.1016/j.jarmac.2020.09.003>
- Boroditsky, L. (2011). How language shapes thought. *Scientific American*, *304*(2), 62–65.

- Carpenter, S. K., Endres, T., & Hui, L. (2020). Students' use of retrieval in self-regulated learning: Implications for monitoring and regulating effortful learning experiences. *Educational Psychology Review*, 32(4), 1029–1054. <https://doi.org/10.1007/s10648-020-09562-w>
- Casasanto, D., & Boroditsky, L. (2008). Time in the mind: Using space to think about time. *Cognition*, 106(2), 579–593. <https://doi.org/10.1016/j.cognition.2007.03.004>
- Castro-Alonso, J. C., Ayres, P., Wong, M., & Paas, F. (2018). Learning symbols from permanent and transient visual presentations: Don't overlap the hand. *Computers & Education*, 116, 1–13. <https://doi.org/10.1016/j.compedu.2017.08.011>
- Castro-Alonso, J. C., de Koning, B. B., Fiorella, L., & Paas, F. (2021). Five strategies for optimizing instructional materials: Instructor- and learner-managed cognitive load. *Educational Psychology Review*, 33(4), 1379–1407. <https://doi.org/10.1007/s10648>
- Chen, S. W. (2023). Learning motivations and effort beliefs in Confucian cultural context: A dual-mode theoretical framework of achievement goal. *Frontiers in Psychology*, 14, 1058456. <https://doi.org/10.3389/fpsyg.2023.1058456>
- Chen, S.-W., Wang, H.-H., Wei, C.-F., Fwu, B.-J., & Hwang, K.-K. (2009). Taiwanese students' self-attributions for two types of achievement goals. *The Journal of Social Psychology*, 149(2), 179–194. <https://doi.org/10.3200/SOCP.149.2.179-94021-09606-9>
- Chen, O., Paas, F., & Sweller, J. (2023). A cognitive load theory approach to defining and measuring task complexity through element interactivity. *Educational Psychology Review*, 35(2), 63. <https://doi.org/10.1007/s10648-023-09782-w>
- Chen, S.-W., Fwu, B.-J., Wei, C.-F., & Wang, H.-H. (2018). Effort beliefs count: The predictive effects of effort beliefs on students' emotion, attribution, and behavior toward academic failure in a Confucian cultural context. In Liem, G. A. D. & Tan, S.H. (eds.) (2018). *Asian Education Miracles: In Search of Sociocultural and Psychological Explanation*. Routledge. <https://doi.org/10.4324/9781315180625>
- Choi, H.-H., van Merriënboer, J. J. G., & Paas, F. (2014). Effects of the physical environment on cognitive load and learning: Towards a new model of cognitive load. *Educational Psychology Review*, 26(2), 225–244. <https://doi.org/10.1007/s10648-014-9262-6>
- Colliot, T., & Boucheix, J. M. (2024). Exploring the effects of seductive details and illustration dynamics on young children's performance in an origami task. *Journal of Computer Assisted Learning*, 40(2), 437–451. <https://doi.org/10.1111/jcal.12879>
- Colliot, T., & Jamet, É. (2021). Improving students' learning by providing a graphic organizer after a multimedia document. *British Journal of Educational Technology*, 52(1), 252–265. <https://doi.org/10.1111/bjjet.12980>
- Cooper, G., Tindall-Ford, S., Chandler, P., & Sweller, J. (2001). Learning by imagining. *Journal of Experimental Psychology: Applied*, 7(1), 68–82. <https://doi.org/10.1037/1076-898X.7.1.68>
- Coppens, L., de Jonge, M., van Gog, T., & Kester, L. (2020). The effect of practice test modality on perceived mental effort and delayed final test performance. *Journal of Cognitive Psychology*, 32(8), 764–770. <https://doi.org/10.1080/20445911.2020.1822366>
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334. <https://doi.org/10.1007/BF02310555>
- de Jong, T. (2010). Cognitive load theory, educational research, and instructional design: Some food for thought. *Instructional Science*, 38(2), 105–134. <https://doi.org/10.1007/s11251-009-9110-0>
- de Bruin, A. B. H., Roelle, J., Carpenter, S. K., Baars, M., & EFG-MRE. (2020). Synthesizing cognitive load and self-regulation theory: A theoretical framework and research agenda. *Educational Psychology Review*, 32(4), 903–915. <https://doi.org/10.1007/s10648-020-09576-4>
- Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. Plenum. <https://doi.org/10.1007/978-1-4899-2271-7>
- DeVellis, R. F. (2017). *Scale development: theory and applications* (4th ed.). Sage.
- Dönmez, O., Akbulut, Y., Telli, E., Kaptan, M., Özdemir, İH., & Erdem, M. (2022). In search of a measure to address different sources of cognitive load in computer-based learning environments. *Education and Information Technologies*, 27(7), 10013–10034. <https://doi.org/10.1007/s10639-022-11035-2>
- Du, X., & Zhang, Q. (2019). Tracing worked examples: Effects on learning in geometry. *Educational Psychology*, 39(2), 169–187. <https://doi.org/10.1080/01443410.2018.1536256>
- Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest: A Journal of the American Psychological Society*, 14(1), 4–58. <https://doi.org/10.1177/1529100612453266>

- Eccles, J. S., & Wigfield, A. (2020). From expectancy-value theory to situated expectancy-value theory: A developmental, social cognitive, and sociocultural perspective on motivation. *Contemporary Educational Psychology*, *61*, 101859. <https://doi.org/10.1016/j.cedpsych.2020.101859>
- Eitel, A., Bender, L., & Renkl, A. (2019). Are seductive details seductive only when you think they are relevant? An experimental test of the moderating role of perceived relevance. *Applied Cognitive Psychology*, *33*(1), 20–30. <https://doi.org/10.1002/acp.3479>
- Eitel, A., Endres, T., & Renkl, A. (2020). Self-management as a bridge between cognitive load and self-regulated learning: The illustrative case of seductive details. *Educational Psychology Review*, *32*(4), 1073–1087. <https://doi.org/10.1007/s10648-020-09559-5>
- Eitel, A., Endres, T., & Renkl, A. (2022). Specific questions during retrieval practice are better for texts containing seductive details. *Applied Cognitive Psychology*, *36*(5), 996–1008. <https://doi.org/10.1002/acp.3984>
- Endres, T., & Eitel, A. (2024). Motivation brought to the test: Successful retrieval practice is modulated by mastery goal orientation and external rewards. *Applied Cognitive Psychology*, *38*(1), e4160. <https://doi.org/10.1002/acp.4160>
- Endres, T., & Renkl, A. (2015). Mechanisms behind the testing effect: An empirical investigation of retrieval practice in meaningful learning. *Frontiers in Psychology*, *6*, 1054. <https://doi.org/10.3389/fpsyg.2015.01054>
- Endres, T., Weyreter, S., Renkl, A., & Eitel, A. (2020). When and why does emotional design foster learning? Evidence for situational interest as a mediator of increased persistence. *Journal of Computer Assisted Learning*, *36*(4), 514–525. <https://doi.org/10.1111/jcal.12418>
- Endres, T., Lovell, O., Morkunas, D., Rieß, W., & Renkl, A. (2023). Can prior knowledge increase task complexity? - Cases in which higher prior knowledge leads to higher intrinsic cognitive load. *The British Journal of Educational Psychology*, *93*(Suppl. 2), 305–317. <https://doi.org/10.1111/bjep.12563>
- Endres, T., Carpenter, S., & Renkl, A. (2024a). Constructive retrieval: Benefits for learning, motivation, and metacognitive monitoring. *Learning and Instruction*, *94*, 101974. <https://doi.org/10.1016/j.learninstruc.2024.101974>
- Endres, T., Eitel, A., Renninger, K. A., Vössing, C., & Renkl, A. (2024b). Why narrative frames matter for instructional videos: A value-evoking narrative frame is essential to foster sustained learning with emotional design videos. *Learning & Instruction*, *94*, 101962. <https://doi.org/10.1016/j.learninstruc.2024.101962>
- Endres, T., Vössing, C., Renninger, K. A., Eitel, A., & Renkl, A. (2024c). Learning contexts shape the effect of emotional design—Facilitating sustained learning in distraction-prone situations. *SSRN*. <https://doi.org/10.2139/ssrn.5002709>
- Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G\*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191. <https://doi.org/10.3758/BF03193146>
- Fiorella, L., & Mayer, R. E. (2021). Principles for reducing extraneous processing in multimedia learning. In R. E. Mayer & L. Fiorella (Eds.), *The Cambridge Handbook of Multimedia Learning* (pp. 185–198). Cambridge University Press. <https://doi.org/10.1017/9781108894333.019>
- Flanigan, A. E., & Kiewra, K. A. (2018). What college instructors can do about student cyber-slacking. *Educational Psychology Review*, *30*(2), 585–597. <https://doi.org/10.1007/s10648-017-9418-2>
- Fontaine, G., Cossette, S., Maheu-Cadotte, M.-A., Mailhot, T., Lavoie, P., Gagnon, M.-P., Dubé, V., & Côté, J. (2019). Traduction, adaptation et évaluation psychométrique préliminaire d'une mesure d'engagement et d'une mesure de charge cognitive en contexte d'apprentissage numérique. *Pédagogie Médicale*, *20*(2), 79–90. <https://doi.org/10.1051/pmed/2020009>
- González, F. M., Saux, G., & Burin, D. (2019). The decorative images' seductive effect in e-learning depends on attentional inhibition. *Australasian Journal of Educational Technology*, *35*(3). <https://doi.org/10.14742/ajet.4577>
- Greene, A. M., & Azevedo, R. (2009). A theoretical review of Winne and Hadwin's SRL model: New perspectives and directions. *Educational Psychologist*, *44*(1), 5–15. <http://www.jstor.org/stable/4624902?> Accessed 1.8.2024.
- Greene, J. A., & Azevedo, R. (2007). A theoretical review of Winne and Hadwin's model of self-regulated learning: New perspectives and directions. *Review of Educational Research*, *77*(3), 334–372. <https://doi.org/10.3102/003465430303953>
- Grund, A., Fries, S., Nückles, M., Renkl, A., & Roelle, J. (2024). When is learning “effortful”? Scrutinizing the concept of mental effort in cognitively oriented research from a motivational perspective. *Educational Psychology Review*, *36*, 11. <https://doi.org/10.1007/s10648-024-09852-7>

- Haji, F. A., Rojas, D., Childs, R., de Ribaupierre, S., & Dubrowski, A. (2015). Measuring cognitive load: Performance, mental effort and simulation task complexity. *Medical Education*, 49(8), 815–827. <https://doi.org/10.1111/medu.12773>
- Harp, S. F., & Mayer, R. E. (1998). How seductive details do their damage: A theory of cognitive interest in science learning. *Journal of Educational Psychology*, 90(3), 414–434. <https://doi.org/10.1037/0022-0663.90.3.414>
- Hoch, E., Sidi, Y., Ackerman, R., Hoogerheide, V., & Scheiter, K. (2023). Comparing mental effort, difficulty, and confidence appraisals in problem-solving: A metacognitive perspective. *Educational Psychology Review*, 35, 61. <https://doi.org/10.1007/s10648-023-09779-5>
- Huang, Y. H. (2018). Influence of instructional design to manage intrinsic cognitive load on learning effectiveness. *Eurasia Journal of Mathematics, Science and Technology Education*, 14(6), 2653–2668. <https://doi.org/10.29333/ejmste/90264>
- Huang, W.-H. (2011). Evaluating learners' motivational and cognitive processing in an online game-based learning environment. *Computers in Human Behavior*, 27(2), 694–704. <https://doi.org/10.1016/j.chb.2010.07.021>
- Kalyuga, S. (2011). Cognitive Load Theory: How many types of load does it really need? *Educational Psychology Review*, 23(1), 1–19. <https://doi.org/10.1007/s10648-010-9150-7>
- Kang, S. H., & Pashler, H. (2014). Is the benefit of retrieval practice modulated by motivation? *Journal of Applied Research in Memory and Cognition*, 3(3), 183–188. <https://doi.org/10.1016/j.jarmac.2014.05.006>
- Klepsch, M., & Seufert, T. (2020). Understanding instructional design effects by differentiated measurement of intrinsic, extraneous, and germane cognitive load. *Instructional Science*, 48(1), 45–77. <https://doi.org/10.1007/s11251-020-09502-9>
- Klepsch, M., Schmitz, F., & Seufert, T. (2017). Development and validation of two instruments measuring intrinsic, extraneous, and germane cognitive load. *Frontiers in Psychology*, 8, 1997. <https://doi.org/10.3389/fpsyg.2017.01997>
- Klepsch, M., & Seufert, T. (2021). Making an effort versus experiencing load. *Frontiers in Education*, 6. <https://doi.org/10.3389/educ.2021.645284>
- Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, 126(4), 349–370. <https://doi.org/10.1037/0096-3445.126.4.349>
- Krell, M. (2017). Evaluating an instrument to measure mental load and mental effort considering different sources of validity evidence. *Cogent Education*, 4(1), 1280256. <https://doi.org/10.1080/2331186X.2017.1280256>
- Kriegelstein, F., Beege, M., Rey, G. D., Ginns, P., Krell, M., & Schneider, S. (2022). A systematic meta-analysis of the reliability and validity of subjective cognitive load questionnaires in experimental multimedia learning research. *Educational Psychology Review*, 34(4), 2485–2541. <https://doi.org/10.1007/s10648-022-09683-4>
- Larmuseau, C., Cornelis, J., Lancieri, L., Desmet, P., & Depaepe, F. (2020). Multimodal learning analytics to investigate cognitive load during online problem solving. *British Journal of Educational Technology*, 51(5), 1548–1562. <https://doi.org/10.1111/bjet.12958>
- Le, Y., Chen, Z., Liu, S., Pang, W., & Deng, C. (2021). Investigating the effectiveness of emotional design principle to attenuate ego depletion effect. *Computers & Education*, 174, 104311. <https://doi.org/10.1016/j.compedu.2021.104311>
- Lee, M. D., & Wagenmakers, E.-J. (2014). Bayesian cognitive modeling. *Cambridge University Press*. <https://doi.org/10.1017/CBO9781139087759>
- Lenzner, A., Schnotz, W., & Müller, A. (2013). The role of decorative pictures in learning. *Instructional Science*, 41(5), 811–831. <https://doi.org/10.1007/s11251-012-9256-z>
- Leopold, C., & Mayer, R. E. (2015). An imagination effect in learning from scientific text. *Journal of Educational Psychology*, 107(1), 47–63. <https://doi.org/10.1037/a0037142>
- Leopold, C., Mayer, R. E., & Dutke, S. (2019). The power of imagination and perspective in learning from science text. *Journal of Educational Psychology*, 111(5), 793–808. <https://doi.org/10.1037/edu0000310>
- Leopold, C. (2021). The imagination principle in multimedia learning. In R. E. Mayer & L. Fiorella (Eds.), *The Cambridge Handbook of Multimedia Learning* (pp. 370–380). Cambridge University Press. <https://doi.org/10.1017/9781108894333.039>

- Leppink, J., Paas, F., van der Vleuten, C. P. M., van Gog, T., & van Merriënboer, J. J. G. (2013). Development of an instrument for measuring different types of cognitive load. *Behavior Research Methods*, 45(4), 1058–1072. <https://doi.org/10.3758/s13428-013-0334-1>
- Lo, K. W., Ngai, G., Chan, S. C., & Kwan, K. P. (2022). How students' motivation and learning experience affect their service-learning outcomes: A structural equation modeling analysis. *Frontiers in Psychology*, 13, 825902. <https://doi.org/10.3389/fpsyg.2022.825902>
- Locke, E. A., & Latham, G. P. (2002). Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American Psychologist*, 57(9), 705–717. <https://doi.org/10.1037/0003-066X.57.9.705>
- Lupyan, G., Abdel Rahman, R., Boroditsky, L., & Clark, A. (2020). Effects of language on visual perception. *Trends in Cognitive Sciences*, 24(11), 930–944. <https://doi.org/10.1016/j.tics.2020.08.005>
- Magner, U. I., Schwonke, R., Alevin, V., Popescu, O., & Renkl, A. (2014). Triggering situational interest by decorative illustrations both fosters and hinders learning in computer-based learning environments. *Learning and Instruction*, 29, 141–152. <https://doi.org/10.1016/j.learninstruc.2012.07.002>
- Mavilidi, M. F., Ouweland, K., Riley, N., Chandler, P., & Paas, F. (2020). Effects of an acute physical activity break on test anxiety and math test performance. *International Journal of Environmental Research and Public Health*, 17(5), 1523. <https://doi.org/10.3390/ijerph17051523>
- Moran, T. P. (2016). Anxiety and working memory capacity: A meta-analysis and narrative review. *Psychological Bulletin*, 142, 831–864. <https://doi.org/10.1037/bul0000051>
- Nugteren, M. L., Jarodzka, H., Kester, L., & van Merriënboer, J. J. G. (2018). Self-regulation of secondary school students: Self-assessments are inaccurate and insufficiently used for learning-task selection. *Instructional Science*, 46(3), 357–381. <https://doi.org/10.1007/s11251-018-9448-2>
- Paas, F. G. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology*, 84(4), 429–434. <https://doi.org/10.1037/0022-0663.84.4.429>
- Paas, F. G. W. C., & van Merriënboer, J. J. G. (1994). Instructional control of cognitive load in the training of complex cognitive tasks. *Educational Psychology Review*, 6(4), 351–371. <https://doi.org/10.1007/BF02213420>
- Paas, F., & van Merriënboer, J. J. G. (2020). Cognitive-load theory: Methods to manage working memory load in the learning of complex tasks. *Current Directions in Psychological Science*, 29(4), 394–398. <https://doi.org/10.1177/0963721420922183>
- Paas, F., Tuovinen, J. E., Tabbers, H., & van Gerven, P. W. M. (2003a). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist*, 38(1), 63–71. [https://doi.org/10.1207/S15326985EP3801\\_8](https://doi.org/10.1207/S15326985EP3801_8)
- Paas, F., Renkl, A., & Sweller, J. (2003b). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist*, 38(1), 1–4. [https://doi.org/10.1207/S15326985EP3801\\_1](https://doi.org/10.1207/S15326985EP3801_1)
- Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology*, 8, 422. <https://doi.org/10.3389/fpsyg.2017.00422>
- Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology*, 46(3), 598–609. <https://doi.org/10.1037/0022-3514.46.3.598>
- Plass, J. L., & Kalyuga, S. (2019). Four ways of considering emotion in cognitive load theory. *Educational Psychology Review*, 31(2), 339–359. <https://doi.org/10.1007/s10648-019-09473-5>
- Porst, R. (2014). Frageformulierung. *Handbuch Methoden der empirischen Sozialforschung*, 687–699. [https://doi.org/10.1007/978-3-531-18939-0\\_50](https://doi.org/10.1007/978-3-531-18939-0_50)
- Psychological Science Accelerator (n.d.). PSA COVID-Rapid (PSACR) Translation process. Retrieved October 17, 2024, from [https://docs.google.com/document/d/12G8aPocfn2KxIJvASbjMUzzTTjf\\_bTIVZiiFgrveRgs/](https://docs.google.com/document/d/12G8aPocfn2KxIJvASbjMUzzTTjf_bTIVZiiFgrveRgs/)
- Rey, G. D. (2012). A review of research and a meta-analysis of the seductive detail effect. *Educational Research Review*, 7(3), 216–237. <https://doi.org/10.1016/j.edurev.2012.05.003>
- Richter, T., Berger, R., Ebersbach, M., Eitel, A., Endres, T., Ferri, R. B., Hänze, M., Lachner, A., Leutner, D., Lipowsky, F., Nemeth, L., Renkl, A., Roelle, J., Rummer, R., Scheiter, K., Schweppe, J., von Aufschneider, C., & Vorholzer, A. (2022). How to promote lasting learning in schools. *Zeitschrift Für Entwicklungspsychologie und Pädagogische Psychologie*, 54(4), 135–141. <https://doi.org/10.1026/0049-8637/a000258>
- Rivers, M. L. (2021). Metacognition about practice testing: A review of learners' beliefs, monitoring, and control of test-enhanced learning. *Educational Psychology Review*, 33(3), 823–862. <https://doi.org/10.1007/s10648-020-09578-2>

- Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, 15(1), 20–27. <https://doi.org/10.1016/j.tics.2010.09.003>
- Roelle, J., Schweppe, J., Endres, T., Lachner, A., Aufschnaiter, C. V., Renkl, A., Eitel, A., Leutner, D., Rummer, R., Scheiter, K., & Vorholzer, A. (2022). Combining retrieval practice and generative learning in educational contexts: Promises and challenges. *Zeitschrift Für Entwicklungspsychologie und Pädagogische Psychologie*, 54(4), 142–150. <https://doi.org/10.1026/0049-8637/a000261>
- Roelle, J., Endres, T., Abel, R., Obergassel, N., Nückles, M., & Renkl, A. (2023). Happy together? On the relationship between research on retrieval practice and generative learning using the case of follow-up learning tasks. *Educational Psychology Review*, 35(4), 102. <https://doi.org/10.1007/s10648-023-09810-9>
- Rouder, J. N., Kumar, A., & Haaf, J. M. (2019). *Why most studies of individual differences with inhibition tasks are bound to fail*. OSF.
- Schneider, S., Häbler, A., Habermeyer, T., Beege, M., & Rey, G. D. (2019). The more human, the higher the performance? Examining the effects of anthropomorphism on learning with media. *Journal of Educational Psychology*, 111(1), 57–72. <https://doi.org/10.1037/edu0000273>
- Seufert, T. (2018). The interplay between self-regulation in learning and cognitive load. *Educational Research Review*, 24, 116–129. <https://doi.org/10.1016/j.edurev.2018.03.004>
- Seufert, T., Hamm, V., Vogt, A., & Riemer, V. (2024). The interplay of cognitive load, learners' resources and self-regulation. *Educational Psychology Review*, 36, 50. <https://doi.org/10.1007/s10648-024-09890-1>
- Silan, M. (2024). Rethinking multi-site studies in social and personality psychology: Can the cross-indigenous approach remedy common cross-cultural vulnerabilities? *Social and Personality Psychology Compass*, 18(10), e70007. <https://doi.org/10.1111/spc3.70007>
- Sirock, J., Vogel, M., & Seufert, T. (2023). Analyzing and supporting mental representations and strategies in solving Bayesian problems. *Frontiers in Psychology*, 14, 1085470. <https://doi.org/10.3389/fpsyg.2023.1085470>
- Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2), 257–285. [https://doi.org/10.1207/s15516709cog1202\\_4](https://doi.org/10.1207/s15516709cog1202_4)
- Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational Psychology Review*, 22(2), 123–138. <https://doi.org/10.1007/s10648-010-9128-5>
- Sweller, J., & Levine, M. (1982). Effects of goal specificity on means-ends analysis and learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 8(5), 463–474. <https://doi.org/10.1037/0278-7393.8.5.463>
- Sweller, J., van Merriënboer, J. J. G., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, 10(3), 251–296. <https://doi.org/10.1023/a:1022193728205>
- Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review*, 31(2), 261–292. <https://doi.org/10.1007/s10648-019-09465-5>
- Terry, N. P., & Irving, M. A. (2010). Cultural and linguistic diversity: Issues in education. *Special Education for All Teachers*, 5, 109–132.
- Thees, M., Kapp, S., Altmeyer, K., Malone, S., Brünken, R., & Kuhn, J. (2021). Comparing two subjective rating scales assessing cognitive load during technology-enhanced STEM laboratory courses. *Frontiers in Education*, 6, 705551. <https://doi.org/10.3389/educ.2021.705551>
- Timirova, A. M. (2021). Cognitively adapted multimedia educational instructions for teaching information technologies in a multilingual university environment. *Informatics and Education*, (4), 47–53. <https://doi.org/10.32517/0234-0453-2021-36-4-47-53>
- Tsai, M. J., Wu, A. H., & Chen, Y. (2019). Static and dynamic seductive illustration effects on text-and-graphic learning processes, perceptions, and outcomes: Evidence from eye tracking. *Applied Cognitive Psychology*, 33(1), 109–123. <https://doi.org/10.1002/acp.3514>
- van Gog, T., Hoogerheide, V., & van Harsel, M. (2020). The role of mental effort in fostering self-regulated learning with problem-solving tasks. *Educational Psychology Review*, 32(4), 1055–1072. <https://doi.org/10.1007/s10648-020-09544-y>
- van Gog, T., Janssen, E., Lucas, F., & Taheij, M. (2024). A motivational perspective on (anticipated) mental effort investment: The biopsychosocial model of challenge and threat. *Educational Psychology Review*, 36, 54. <https://doi.org/10.1007/s10648-024-09861-6>













- van Merriënboer, J. J. G., & Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. *Educational Psychology Review*, 17(2), 147–177. <https://doi.org/10.1007/s10648-005-3951-0>
- Wang, Z., & Adesope, O. (2016a). Exploring the effects of seductive details with the 4-phase model of interest. *Learning and Motivation*, 55, 65–77. <https://doi.org/10.1016/j.lmot.2016.06.003>
- Wang, Z., & Adesope, O. (2016b). Does learners' prior knowledge moderate the detrimental effects of seductive details in reading from text? A 2 by 3 study. *International Journal of Instruction*, 9(2), 35–50. <https://doi.org/10.12973/iji.2016.923a>
- Wang, T., & Lajoie, S. P. (2023). How does cognitive load interact with self-regulated learning? A dynamic and integrative model. *Educational Psychology Review*, 35, 69. <https://doi.org/10.1007/s10648-023-09794-6>
- Wirth, J., Stebner, F., Trypke, M., Schuster, C., & Leutner, D. (2020). An interactive layers model of self-regulated learning and cognitive load. *Educational Psychology Review*, 32(4), 1127–1149. <https://doi.org/10.1007/s10648-020-09568-4>
- Wolpe, N., Holton, R., & Fletcher, P. C. (2024). What is mental effort: A clinical perspective. *Biological Psychiatry*, 95(11), 1030–1037. <https://doi.org/10.1016/j.biopsych.2024.01.022>
- Woo, J.-C. (2014). Digital game-based learning supports student motivation, cognitive success, and performance outcomes. *Educational Technology & Society*, 17(3), 291–307.
- Zhang, S., Koning, B. de, Agostinho, S., Tindall-Ford, S., Chandler, P., & Paas, F. (2021). The cognitive load self-management principle in multimedia learning. In R. E. Mayer & L. Fiorella (Eds.), *The Cambridge Handbook of Multimedia Learning* (pp. 430–436). Cambridge University Press. <https://doi.org/10.1017/9781108894333.044>
- Zu, T., Munsell, J., & Rebello, N. S. (2021). Subjective measure of cognitive load depends on participants' content knowledge level. *Frontiers in Education*, 6, 647097. <https://doi.org/10.3389/educ.2021.647097>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This is a preregistered report: Hypotheses and experimental plans for the quantitative part of our study were evaluated by the initial reviewers of this manuscript. Data collection started after approval from the reviewers. The preregistration is available on asPredicted: <https://aspredicted.org/c3vt-wpjz.pdf>.



## Authors and Affiliations

Tino Endres<sup>1,2</sup>  · Lisa Bender<sup>1</sup>  · Stoo Sepp<sup>3</sup>  · Shirong Zhang<sup>4,5</sup>  ·  
Louise David<sup>6</sup>  · Melanie Trypke<sup>7</sup>  · Dwayne Lieck<sup>1</sup>  · Juliette C. Désiron<sup>1</sup>  ·  
Johanna Bohm<sup>1</sup>  · Sophia Weissgerber<sup>8</sup>  · Juan Cristobal Castro-Alonso<sup>9</sup>  ·  
Fred Paas<sup>4,10</sup> 

✉ Tino Endres  
tino.endres@psychologie.uni-freiburg.de

<sup>1</sup> Department of Psychology, University of Freiburg, Freiburg, Germany

<sup>2</sup> Educational Technology, University of Zürich, Zurich, Switzerland

<sup>3</sup> School of Education, University of New England, Armidale, Australia

<sup>4</sup> Department of Psychology, Erasmus University Rotterdam, Education, and Child Studies, Rotterdam, The Netherlands

<sup>5</sup> Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Delft, The Netherlands

<sup>6</sup> Department of Educational Development and Research, School of Health Professions Education (SHE), Maastricht University, Maastricht, The Netherlands

<sup>7</sup> Institute of Educational Science, University of Osnabrück, Osnabrück, Germany

<sup>8</sup> Department of Psychology, University of Kassel, Kassel, Germany

<sup>9</sup> School of Education, University of Birmingham, Birmingham, UK

<sup>10</sup> School of Education, University of New South Wales, Kensington, Australia