



# Applying Cognitive Load Theory in Teacher Education

An Experimental Validation of the Scale by Leppink et al.

Venance Timothy<sup>1</sup>, Frank Fischer<sup>2</sup>, Bianca Watzka<sup>3</sup>, Raimund Girwidz<sup>4</sup>,  
and Matthias Stadler<sup>2,5</sup>

<sup>1</sup>Department of Educational Psychology and Curriculum Studies, Dar es Salaam University College of Education, Dar es Salaam, Tanzania

<sup>2</sup>Department of Psychology, Ludwig-Maximilians-Universität München, Munich, Germany

<sup>3</sup>Institute of Physics, Otto von Guericke University, Magdeburg, Germany

<sup>4</sup>Department of Physics Didactics, Ludwig-Maximilians-Universität München, Munich, Germany

<sup>5</sup>Institute of Medical Education, LMU University Hospital, Ludwig-Maximilians-Universität München, Munich, Germany

**Abstract:** The study investigated the validation of a rating scale to measure cognitive load in science teacher education. The rating scale was used to measure three types of cognitive load in a new learning context with 81 undergraduate students enrolled in a science education program, randomly assigned to three experimental groups: problem-solving, example-based learning, and control groups. The preservice teachers' cognitive load was measured using a rating scale during an intervention to diagnose students' misconceptions in physics. The study also assessed the effect of instructional design on cognitive load. The results showed that the three types of cognitive load can be reliably measured in science teacher education and that instructional designs that create germane cognitive load contribute to the development of preservice teachers' diagnostic competencies. Conversely, designs that create irrelevant cognitive load are detrimental to this development. These findings suggest the importance of considering cognitive load in science teacher education for effective instructional design.

**Keywords:** cognitive load, misconceptions, teacher training, example-based learning, problem-solving

Cognitive load theory (Sweller, 1988, 1994, 2011) is a theory of learning and teaching that derives instructional design implications from a model of human cognitive architecture. It suggests general instructional design principles for managing working memory load as a key issue for successful learning and performance (Ginns & Leppink, 2019; Kalyuga, 2011). Thus, it is important to understand the degree of cognitive load imposed by a particular instructional design to use it most effectively. Unfortunately, instead of a universal method for measuring cognitive load that is appropriate for different learning contexts or audiences, there is a wide variety of assessment approaches. Subjective rating scales are particularly common (Thees et al., 2021). In this study, we investigated the validity of one of the most commonly used rating scales for assessing cognitive load (Leppink et al., 2013) on a sample of Tanzanian Bachelor of Science in Education students during an intervention aimed at assessing students' misconceptions in physics. While the application of this assessment scale in science, technology, engineering, and mathematics laboratory courses is well-established (e.g., Altmeyer et al., 2021; Morrison et al.,

2015; Thees et al., 2021), these studies focused on either students or teachers, but not on student teachers (preservice teachers). Furthermore, in the studies by Leppink et al. (2013), where the rating scale was first developed and further validated in Leppink et al. (2014), the learning context was in Europe, while the participants came from health sciences, psychology, or language courses. Thus, this study adds to the existing literature on cognitive load theory by testing its applicability to this relatively understudied learning context (i.e., preservice teachers learning to diagnose student misconceptions in physics) in a non-WEIRD sample (Henrich et al., 2010). Therefore, the aim of the present study was to validate one of the most commonly used rating scales (Leppink et al., 2014) to assess cognitive load in a different learning context with preservice physics teachers and to investigate whether the rating scale could replicate the triarchic factor structure and the expected effects of instructional design on cognitive load in the context of science teacher education (Schmeck et al., 2015). Although there is ample evidence to support the validity of the instructional implications of cognitive load theory (Syring et al., 2015), there has been

no systematic investigation of the validity of the theory for facilitating knowledge and skills in science teacher education while assessing their cognitive load.

## What Is the Construct Being Measured?

Research on cognition suggests that new or novel information must be processed in working memory before it can be stored in long-term memory (Chen et al., 2018; Paas et al., 2003; Van Merriënboer & Sweller, 2010). Working memory itself is limited in terms of both the amount of information it can process and the time it takes to process it. Therefore, any learning activity that results in exceeding the capacity of working memory will always result in the learner experiencing a higher subjective cognitive load (Sweller et al., 1998). According to Paas et al. (1994), the term cognitive load refers to a multidimensional construct that represents the amount of load that learners would experience in their cognitive systems while performing specific learning tasks. Measuring cognitive load can help optimize learning and improve instructional design by providing valuable information about learners' cognitive demands and limitations. However, if unnecessary cognitive load due to inappropriate instructional procedures or interactivity of elements in learning materials is not well-managed, it can overwhelm the available cognitive capacities of preservice teachers (Moos & Pitton, 2014).

Cognitive load theory originally defined two types of cognitive load: intrinsic cognitive load, which represents the cognitive load caused by the intrinsic nature of the learning materials (i.e., the inherent difficulty of the task; Chandler & Sweller, 1991), and extraneous cognitive load, which represents information introduced into instructional designs that is not directly needed to master a given problem (Sweller et al., 1998). Later, germane cognitive load was introduced as a third category of cognitive load (Leppink, et al., 2014; Sweller et al., 1998). Germane cognitive load is generated by learning activities that support the further development of knowledge structures in long-term memory, such as the application of a learning strategy (Van Merriënboer & Sweller, 2005). The additivity hypothesis (Moreno & Park, 2010) states that intrinsic, extrinsic, and germane cognitive load contribute to total cognitive load, but only if the capacity of working memory is not exceeded. Intrinsic and extrinsic cognitive load are associated with different aspects of the learning material and are therefore assumed to be uncorrelated. On the other hand, intrinsic and germane cognitive load are assumed to share a common theoretical background and should be interdependent. Extraneous and germane cognitive load should not be correlated due to their

different nature, with germane cognitive load being an active process and extraneous cognitive load being a passive process. In a recent meta-analysis, Krieglstein et al. (2022) report that this triarchic theory of cognitive load (see DeLeeuw & Mayer, 2008, for more details) has been replicated across studies and across rating scales. According to this meta-analysis, reliability estimates of cognitive load rating scales were not affected by educational setting, domain of instructional material, mode of presentation, or number of rating scale points. While correlations between cognitive load types were partially inconsistent with theory-based assumptions, correlations with learning-related variables supported assumptions derived from cognitive load theory. Based on these findings, we do not expect to find substantial effects of adapting the rating scale to our context.

However, learners differ in terms of their individual learning characteristics and the cognitive load they experience in a task. The main factor that determines the amount of cognitive load that learners experience is prior knowledge (Kalyuga, 2009). With increasingly effective knowledge structures in long-term memory, the information presented in a learning situation can be structured, reducing the demand on working memory (Sweller et al., 1998). Correspondingly, the cognitive load of a given problem or task decreases as the amount of available prior knowledge relevant to the task increases. Adequate measurement of the specific cognitive load experienced by learners on a given task allows for optimization of the instructional design provided (Chandler & Sweller, 1991; Chen et al., 2018). For example, problem-based learning may impose a higher cognitive load than learning through example-based instruction because learners must direct some of their mental effort to managing the problem-solving process (e.g., dealing with problem states while relating the required solutions) during the learning process (Sweller et al., 1998). If the goal of the learning task is to acquire knowledge about concepts, while learners have experienced cognitive load due to the use of inappropriate instructional support, then this can be considered extraneous cognitive load. It is also possible for learners to experience high extraneous cognitive load when they encounter learning tasks that require them to solve structured problems that integrate many elements at once (e.g., text and diagrams; Sweller et al., 1998). This is because learners would use much of their mental effort to process multiple pieces of information, thereby increasing extraneous cognitive load (Paas et al., 1994). Similarly, if learning materials contain redundant information, learners would spend much of their mental effort on unnecessary information, thereby increasing their extraneous load (Paas et al., 1994; Sweller et al., 1998). In contrast to problem-solving, worked examples can reduce

extraneous cognitive load because there is no need to manage the problem-solving process, allowing learners to generalize solutions and focus attention on the current problem state and goals (Bichler et al., 2020; Sweller et al., 1998). The literature on cognitive load shows that worked examples are useful for learning new skills (Hoogerheide & Roelle, 2020), and especially for the acquisition of new knowledge, learners can learn best with examples before they can actually learn by solving problems (Jalani & Sern, 2015; van Harsel et al., 2020). However, according to the expertise reversal effect (Kalyuga et al., 2003), worked examples may become redundant when used with more expert learners. That is, as learners become experts in a particular learning experience, they may no longer need worked examples because they can already solve problems by applying their knowledge. In terms of learner characteristics, a recent meta-analysis (Chernikova et al., 2020) found that less advanced learners may benefit more from scaffolded instructional support with high levels of guidance (examples) than more advanced learners for whom self-regulation (problem-solving) is the best instructional strategy (cf. Chernikova et al., 2020). The use of problem-solving and example-based instructional strategies may have implications for cognitive load, while at the same time, learners' prior knowledge is a key factor in designing the best instructional strategy.

In addition to cognitive load, we assessed preservice teachers' diagnostic competence in terms of conceptual and procedural knowledge, which refers to knowledge of concepts and procedures. In teacher education, however, diagnostic competence has been defined as the ability of teachers to identify students' learning prerequisites (Barth & Henninger, 2012), students' performance, or teachers' own characteristics (Vogt & Rogalla, 2009). In science education, especially in physics, it is crucial for preservice teachers to learn how to diagnose students' misconceptions.

## What Are the Intended Uses?

Leppink et al. (2013) developed a psychometric rating scale to measure different types of cognitive load, which should be able to distinguish between different types of cognitive load in different educational contexts. To validate this scale in science teacher education, we investigated three related research questions. First, we tested whether the triarchic model of intrinsic, extrinsic, and germane load could be replicated in a novel learning environment and sample.

*Research Question 1:* Does the theoretical model of three distinct facets of cognitive load fit the data in a novel learning environment and sample?

Once a measurement model for cognitive load could be established, we investigated whether different instructional designs would have the theoretically expected effects on the extracted facets of cognitive load. Only then could we be sure that the scale was measuring the same constructs as theoretically described. One experimental group was trained to assess students' misconceptions in an intervention with an instructional design based on problem-solving, while the second experimental group was assigned to an intervention with an instructional design based on example-based instructional support. The preservice teachers in the control group did not receive any training on how to diagnose students' misconceptions in physics. However, the control group would serve as a reference point because they had not interacted with the instructional design intended for an intervention and thus would serve as a baseline for comparing the two experimental groups based on the two forms of instructional strategies in learning how to diagnose physics misconceptions. We expected to observe systematic differences between the three groups on all three facets of cognitive load. We formulated three different hypotheses for this research question.

*Hypothesis 1:* Experimental conditions vary in intrinsic cognitive load with the problem-solving group experiencing the highest load, followed by the worked example group and then finally by the control group.

*Hypothesis 2:* Experimental conditions vary in extrinsic load with the problem-solving group experiencing the highest load, followed by the worked example group and then finally by the control group.

*Hypothesis 3:* Experimental conditions vary in germane load with the worked example group experiencing the highest load and then followed by problem-solving group experiencing.

*Research Question 2:* Does the effect of instructional design on measured cognitive load align with theoretical assumptions?

During the learning phase, preservice teachers could experience different levels of cognitive load depending on the instructional design and elements interactivity of the learning materials (Kalyuga, 2011; Sweller et al., 1998). Higher extraneous load could be experienced due to improper instructional design, while high intrinsic load could be experienced due to elements interactivity of the learning materials. Therefore, we expected to find differences in the cognitive load experienced by preservice

teachers based on the instructional design and the nature of the learning materials.

*Hypothesis 4:* Problem-solving instruction procedure exerts higher extraneous cognitive load followed by example-based learning instruction.

Finally, we postulated that cognitive load that preservice teachers would experience during the intervention could influence the diagnostic competence as a result of instructional manipulation.

*Research Question 3:* What is the effect of instructional design on diagnostic competence while controlling cognitive load (all three facets)?

*Hypothesis 4:* The impact of experimental condition on increases in conceptual knowledge is reduced by controlling for cognitive load (all three facets).

*Hypothesis 5:* The impact of experimental condition on increases in procedural knowledge is reduced by controlling for cognitive load (all three facets).

## Methods

### Participants

Eighty-one undergraduate students pursuing a Bachelor of Science in Education with a concentration in physics participated in the study. The mean age of these preservice teachers was 25.09 years ( $SD = 2.04$ ), with a minimum age of 22 years and a maximum age of 35 years. Of these, 86.4% were men and 13.6% were women. The sample size was calculated based on the statistical power associated with testing multiple hypotheses. In our study, we were interested in detecting a medium effect size of the intervention (Cohen's  $d = 0.35$ ;  $\alpha$  level of .05 and study power of up to .95), if any, compared to the treatments of the variables. Using G\*Power statistical software, we were able to obtain an estimate of 69 participants for the repeated-measures MANOVA and study design. Therefore, the current sample ( $N = 81$ ) of available preservice teachers was sufficient to detect the required effect size.

The sample was drawn from one of the constituent colleges of education at the University of Dar es Salaam in Tanzania. Research permission was obtained from the vice chancellor of the University of Dar es Salaam before the intervention was implemented. All preservice teachers voluntarily participated in the study as part of a regular course. Participants were aware that they could end their

participation at any time without giving a reason and that the data would be anonymized. They were asked to sign a consent form. After an intervention, preservice teachers were reimbursed a small amount of money (US\$ 10) to cover their travel expenses and meals during the workshop day.

### Design

The study used an experimental research design with three independent groups. The preservice teachers were randomly assigned to two experimental groups and one control group. One experimental group ( $n = 27$ ) was trained to assess students' misconceptions in physics through an intervention with an instructional design based on problem-solving, while the second experimental group ( $n = 28$ ) was assigned to an intervention with an instructional design based on example-based instructional support. The preservice teachers in the control group ( $n = 26$ ) did not receive any training on how to diagnose students' physics misconceptions. They were asked to wait for the training later after the two experimental groups had completed the given tasks. However, they participated in the pretest (prior diagnostic knowledge), post-test (diagnostic knowledge gain), and cognitive load measures just like the experimental groups.

### Procedure

Preservice teachers were trained during a 1-day intensive training workshop; the intervention targeted preservice teachers' competence in assessing students' misconceptions in physics (Timothy et al., 2023). The preservice teachers learned how to diagnose students' misconceptions in mechanics in the first session and in electricity in the second session. The first training session lasted 2 h and 30 min, while the second training session lasted 2 h. The preservice teachers' prior diagnostic knowledge was assessed for approximately one hour prior to the training sessions. Cognitive load was measured twice: between the two training sessions and after the second training session. We assessed cognitive load by asking participants to complete a rating scale (adapted from the second study by Leppink et al., 2014) for approximately 10 to 15 min. The preservice teachers' gain in diagnostic knowledge was assessed using the same standardized test within one hour after the second training phase.

### Instruments

We measured cognitive load using a modified version of a rating scale adopted from Leppink et al. (2014) in their second study. Leppink and colleagues conducted two

studies: The first study aimed to investigate whether the psychometric instrument they developed could differentiate between three types of cognitive load, while the second study used a slightly modified version of the instrument to differentiate between intrinsic load and extrinsic load. The modified version of a rating scale in the second study consisted of 13 items, with four items measuring intrinsic load, four items measuring extraneous load, and five items measuring germane load. The preservice teachers responded to each item using a 10-point Likert scale (see Table 1). Cognitive load was assessed in the middle and at the end of the intervention phase. A mean score was calculated for each type of cognitive load (intrinsic, extrinsic, and germane) per participant. The internal consistency was good for intrinsic cognitive load ( $\alpha = .74$ ), satisfactory for extraneous cognitive load ( $\alpha = .69$ ), and excellent for germane cognitive load ( $\alpha = .91$ ). Table 1 shows the scale used to measure the cognitive load that preservice teachers would experience during the process of learning how to diagnose physics misconceptions. However, the rating scale was used to assess cognitive load in its well-established English version without further adaptation. English was used in this rating scale because it is the language of instruction in higher education institutions in Tanzania.

### Diagnostic Competence

The diagnostic competence of preservice teachers was assessed using an objective test that is well-described in the recently published article (Timothy et al., 2023). The test was piloted with a similar sample of preservice teachers from another constituent university college of education. The test consisted of two sections (scales): the first section of 32 multiple-choice items measuring conceptual diagnostic knowledge and a second section of 14 items measuring procedural diagnostic knowledge. A section measuring conceptual diagnostic knowledge consisted of items derived from similar diagnostic cases used in the learning phase of an intervention. An item in the first section of this knowledge test was worth either one point or zero depending on whether a preservice teacher chose a correct or incorrect answer. Some items were later removed during scale construction to increase the internal consistency of the instrument. The overall internal consistency of all items in the first scale was  $\alpha = .72$ , but the reliability analysis suggested that if certain items were removed, the internal consistency would increase to  $\alpha = .74$ , a value close to the recommended value of  $\alpha = .75$  (Field, 2013). Next, a total of nine items were progressively removed from the first scale to increase its internal consistency to at least the recommended value for objective

**Table 1.** A rating scale to measure three types of cognitive load

Number	Item	Response												
		0	1	2	3	4	5	6	7	8	9	10		
1	The content of this activity was very complex.													
2	The problem/s covered in this activity was/were very complex.													
3	In this activity, very complex terms were mentioned.													
4	I invested a very high mental effort in the complexity of this activity.													
5	The explanations and instructions in this activity were very unclear.													
6	The explanations and instructions in this activity were full of unclear language.													
7	The explanations and instructions in this activity were, in terms of learning, very ineffective.													
8	I invested a very high mental effort in unclear and ineffective explanations and instructions in this activity.													
9	This activity really enhanced my understanding of the content that was covered.													
10	This activity really enhanced my understanding of the problem/s that was/were covered.													
11	This activity really enhanced my knowledge of the terms that were mentioned.													
12	This activity really enhanced my knowledge and understanding of how to deal with the problem/s covered													
13	I invested a very high mental effort during this activity in enhancing my knowledge and understanding													

Note. Item categories: intrinsic load (1–4), extraneous load (5–8), and germane load (9–13). Scale: 0 = *not at all*, 10 = *completely the case*.

measures. The final score was  $\alpha = .81$ . For the second scale, which was used to measure procedural knowledge, the overall internal consistency was  $\alpha = .68$ , while the reliability analysis suggested that the value would never be higher than .68 if any items were removed from the scale. Although this value was lower than the recommended value of  $\alpha = .75$ , we decided to retain all 15 items to assess procedural diagnostic knowledge.

### Data Processing and Analyses

We applied confirmatory factor analysis using the *lavaan* package (Rosseel, 2012) in R 4.0.3. The aim was to test whether a theoretical model about the rating scale developed by Leppink and colleagues to measure the three forms of cognitive load could still hold in another learning context of teacher education in physics (RQ1). The model with three correlated factors was tested for both measures. In addition, we tested for measurement invariance across time points. All models were estimated using diagonal weighted least squares. Model fit was considered acceptable with a comparative fit index (CFI)  $> .95$  and a root-mean-square error of approximation (RMSEA)  $< .05$ . A decrease in CFI of less than .01 was considered acceptable to assume invariance (Cheung & Rensvold, 2002). If strong invariance could be assumed, we could average both measures for all further analyses to increase the reliability of our measures.

To analyze the main effect of instructional design on cognitive load (RQ2), we conducted a MANOVA with ANOVAs and Tukey-corrected post hoc tests to test for significant effects. To analyze the effect of instructional design on diagnostic competence (RQ3), we calculated knowledge gain scores (post-test–pretest) for both facets of diagnostic competence. Based on these scores, we performed MANOVAs to test the hypotheses without

covariates (H4 and H5) and MANCOVAs to test the hypotheses with covariates (H4 and H5). Both types of omnibus tests are followed by ANOVAs to test for significant effects. The  $\alpha$  level was set at 5% for all analyses.

## Results

### Descriptive Statistics

Table 2 provides descriptive statistics for all manifest variables used in the analyses. There were no variables that digressed substantially from the expected distributions, and all correlations were in the expected directions.

#### Factorial Validation of the Cognitive Load Rating Scales (RQ1)

The first research question addressed whether the theoretical model of three correlated cognitive load factors still held in the context of teacher education. The assumed model fit the data well at both time points (see Table 3). All factor loadings were significant and positive. There was a strong positive latent correlation between intrinsic load and extrinsic load ( $r_{t1} = .74, p < .001; r_{t2} = .58, p < .001$ ) but only a small negative latent correlation between intrinsic and germane load ( $r_{t1} = -.14, p = .043; r_{t2} = -.01, p = .804$ ). The negative latent correlation between extrinsic and germane load was not statistically significant either ( $r_{t1} = -.11, p = .144; r_{t2} = -.04, p = .303$ ). There were no substantial residual covariances.

The measurement model showed metric and scalar invariance across the two time points (Table 3). The two measures were, therefore, aggregated into a single measure for all further analyses.

**Table 2.** Means, standard deviations, and correlations for all manifest variables

Variable	M	SD	1	2	3	4	5	6	7	8	9
1. Intrinsic load t1	2.98	2.24	—								
2. Extrinsic load t1	1.53	1.95	.49**	—							
3. Germane load t1	7.36	2.36	-.13	-.17	—						
4. Intrinsic load t2	2.12	2.38	.58**	.41**	.05	—					
5. Extrinsic load t2	1.63	2.13	.38**	.43**	-.31**	.39**	—				
6. Germane load t2	7.11	3.18	-.19	-.23*	.76**	-.04	-.04	—			
7. Conceptual know t1	0.43	0.18	-.07	-.11	.26*	-.01	-.00	.16	—		
8. Procedural know t1	0.46	0.20	-.05	-.16	.20	-.14	-.14	.14	.19	—	
9. Conceptual know t2	0.52	0.21	-.10	-.18	.26*	-.08	-.08	.19	.62**	.09	—
10. Procedural know t2	0.54	0.21	-.03	-.10	.43**	-.03	-.16	.25*	.31**	.45**	.40**

Note. M = mean; SD = standard deviation; Know = knowledge.

\* $p < .050$ ; \*\* $p < .010$ .

### Effect of Instructional Design on Cognitive Load (RQ2)

In line with Hypotheses 1–3, a MANOVA analyzing mean differences within any of the groups across the three facets of cognitive load showed a significant main effect ( $F_{6,154} = 3.23, p = .005$ ). Follow-up ANOVAs revealed significant main effects for intrinsic load ( $F_{2,78} = 5.30, p = .007, \eta^2 = .12$ ), extrinsic load ( $F_{2,78} = 6.02, p = .004, \eta^2 = .14$ ), and germane load ( $F_{2,78} = 3.74, p = .028, \eta^2 = .09$ ).

The example-based learning condition showed the highest intrinsic load, followed by the problem-solving and control groups. However, only the difference between worked examples and the control group was statistically significant ( $p_{\text{Tukey}} = .006, d = 0.87$ ). The same pattern was found for extraneous load, with the worked examples condition showing the highest load, followed by the problem-solving condition and the control group. Again, only the difference between the worked examples condition and the control group was statistically significant ( $p_{\text{Tukey}} = .003, d = 0.92$ ). Finally, the problem-solving condition showed the highest germane load, followed by the control group and the worked examples condition. Only the difference between problem-solving and worked examples condition was statistically significant ( $p_{\text{Tukey}} = .025, d = 0.72$ ). These results partially contradict Hypotheses 1–3 and are visually illustrated in Figure 1.

### Effect of Instructional Design on Diagnostic Competency While Controlling Cognitive Load (RQ3)

A MANOVA analyzing mean differences within any of the groups across the gains on the two facets of diagnostic competencies showed a significant main effect ( $F_{4,156} = 6.40, p < .001$ ). Follow-up ANOVAs revealed significant main effects of the experimental condition on conceptual knowledge ( $F_{2,78} = 10.41, p < .001, \eta^2 = .21$ ) and procedural knowledge ( $F_{2,78} = 6.28, p = .003, \eta^2 = .14$ ). For conceptual knowledge, participants in the problem-solving condition increased their scores significantly more than participants in the control condition ( $p_{\text{Tukey}} < .001, d = 1.24$ ) and the worked examples condition ( $p_{\text{Tukey}} = .012, d = 0.79$ ). There was no significant difference in gain scores between the

control and worked example conditions ( $p_{\text{Tukey}} = .236, d = -0.45$ ).

A similar pattern was observed for procedural knowledge. Participants in the problem-solving condition increased their scores significantly more than participants in the control condition ( $p_{\text{Tukey}} < .001, d = 0.95$ ) but not the worked examples condition ( $p_{\text{Tukey}} = .056, d = 0.63$ ). There was no significant difference in gain scores between the control and worked example conditions ( $p_{\text{Tukey}} = .469, d = -0.32$ ). Adding the three cognitive load variables as covariates in a MANCOVA analyzing mean differences within any of the groups across the gains on the two facets of diagnostic competencies still showed a significant main effect ( $F_{4,150} = 6.27, p < .001$ ). Follow-up ANCOVAs revealed significant main effects of the experimental condition on conceptual knowledge ( $F_{2,75} = 11.01, p < .001, \eta^2 = .22$ ) and procedural knowledge ( $F_{2,75} = 5.40, p = .006, \eta^2 = .12$ ). Adjusting for cognitive load, participants in the problem-solving condition still increased their conceptual knowledge scores significantly more than participants in the control condition ( $p_{\text{Tukey}} < .001, d = 1.31$ ) and the worked examples condition ( $p_{\text{Tukey}} = .049, d = 0.70$ ). Also, participants in the problem-solving condition increased their procedural knowledge significantly more than participants in the control condition ( $p_{\text{Tukey}} = .005, d = 0.90$ ) but not the worked examples condition ( $p_{\text{Tukey}} = .137, d = 0.56$ ). These results contradict Hypotheses 4 and 5.

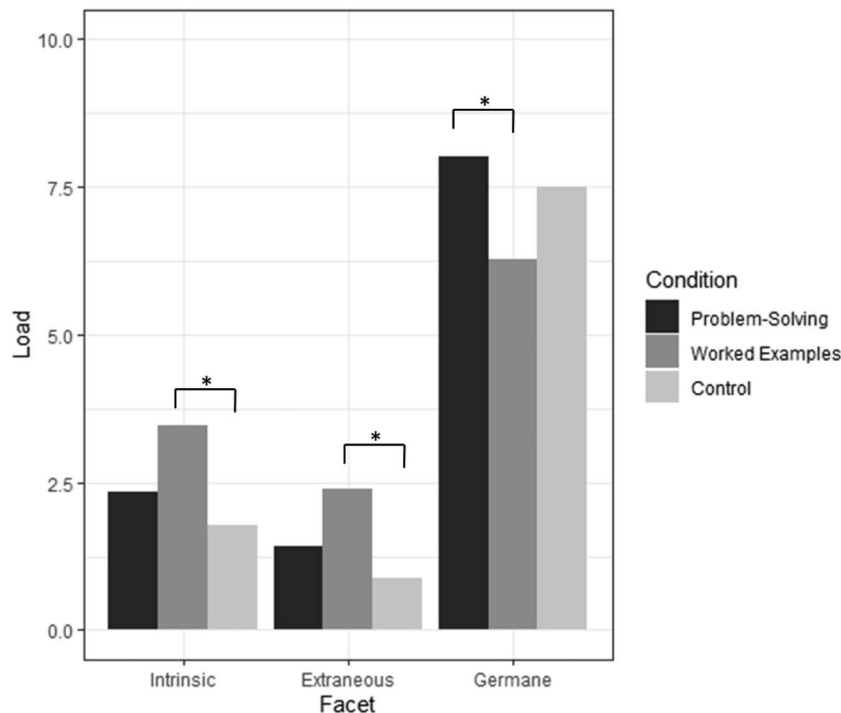
## Discussion

Several researchers have described the importance of having reliable and valid measures of cognitive load (Korbach et al., 2018). Subjective measures of cognitive load are easy to administer, and learners can use standardized rating scales to rate their perceived task difficulty, engagement, or effort in performing a given learning task. In addition, subjective measures can be used in different learning environments with different learning contexts and participants. To validate one of the most

**Table 3.** Model fit for all measurement models

Model	$\chi^2$	<i>df</i>	<i>p</i>	CFI	RMSEA [CI]	$\Delta$ CFI	Decision
Measurement models							
Measurement model t1	60.13	62	.544	1.00	.00 [.00–.06]		Good fit
Measurement model t2	58.56	62	.600	1.00	.00 [.00–.06]		Good fit
Invariance testing							
Configural model	118.69	128	.618	1.00	.00 [.00–.05]		Good fit
Metric model	135.19	134	.455	1.00	.01 [.00–.06]	0.00	Accept invariance
Scalar model	138.49	144	.614	1.00	.00 [.00–.05]	0.00	Accept invariance

Note. *df* = degrees of freedom; CFI = comparative fit index; CI = confidence interval; RMSEA = root-mean-square error of approximation.



**Figure 1.** Average cognitive load by facet and experimental condition. \* $p < .050$ .

commonly used subjective measures of cognitive load (Leppink et al., 2013) in a novel context, we designed an experimental study to measure the cognitive load of preservice teachers in the context of science teacher education in Tanzania. Participants were undergraduate students on the Bachelor of Science Education program, studying physics in the third year of the program.

The first research question in this study was whether the theoretical model of three correlated cognitive load factors still holds in the context of teacher education. We examined the factorial validity of the instrument in science education. Our results support the three-factor solution. The pattern of correlations is consistent with meta-analytic findings (Kriegelstein et al., 2022) on rating scales based on Leppink et al. (2013). We found a substantial correlation between internal and external cognitive load, but hardly any correlation between the two facets and germane load. These findings support the argument that both sources of cognitive load may be difficult for learners to assess in a differentiated way, as complex learning content cannot be presented in a simple way, and thus, extrinsic cognitive load increases due to the complex presentation. Intrinsic and germane load, on the other hand, showed no correlation despite their common conceptual background (Kalyuga, 2011). However, intrinsic load results from the complexity of the learning material and is experienced passively by the learner, whereas germane load refers to the allocation of cognitive

resources and is therefore active in nature (Klepsch & Seufert, 2021). Finally, since germane load refers to the allocation of cognitive resources to learning-relevant activities (Bannert, 2002), its active nature is obvious. In contrast, learners experience extrinsic load as a result of the passive presentation of learning materials. Consistent with this distinction, the two types of load were not correlated. These results are also consistent with previous findings on the cognitive load of preservice teachers in comparable tasks (e.g., Syring et al., 2015).

The second research question concerned the effect of instructional design on cognitive load. Our findings partially support H1, H2, and H3, which postulated that subjective cognitive load should vary depending on whether participants learned in a problem-solving condition, a worked example condition, or an active control condition. The rating scale is appropriate to differentiate the three types of cognitive load in a higher education context, where preservice teachers learned in a simulation-based learning environment to diagnose students' physics misconceptions. However, the results indicated that preservice teachers experienced higher intrinsic and extrinsic cognitive load when learning to diagnose with example-based learning instructions than with problem-solving. This may be due to the nature of the learning material and the instructional strategies used in the training intervention. In the example-based condition, preservice teachers were expected to study examples and



consider any student misconceptions. Also, preservice teachers may already have enough knowledge to easily solve the given problems, but still need to study the examples. The findings are consistent with Chernikova et al. (2020) who found that less advanced learners might benefit more from scaffolding support with a high level of guidance (examples) than more advanced learners for whom self-regulated learning (problem-solving) is the best instructional strategy. Surprisingly, preservice teachers who learned through a problem-solving instructional strategy experienced higher germane cognitive load than those who learned through examples. Again, the type of instructional support (problem-solving) may explain this observation.

Finally, we investigated whether our participants would differ in their learning gains and whether these differences could be attributed to the different cognitive load imposed by the instructional design. While we found that participants benefited most from the problem-solving condition, adjusting for cognitive load did not eliminate this difference. The results are identical for both conceptual and procedural knowledge. This finding contrasts with the general notion that experienced cognitive load should influence learning gains through reduced working memory capacity. However, similar findings were reported by Schwaighofer et al. (2016) and replicated by Bichler et al. (2020). Both papers found that individual differences in shifting, rather than working memory capacity, explained differences in worked examples and problem-based learning.

## Limitations

There are several limitations that should be taken into account when interpreting the results. First, the sample size was relatively small, which may limit the generalizability of the findings. Future studies should aim to replicate these findings with a larger sample size to increase the generalizability of the results. Second, although this study provided initial evidence for the validity of the rating scale in a novel context, further replication studies are needed to establish the generalizability of the scale beyond the physics domain. Future studies should examine the validity of the scale in other domains to determine whether the findings generalize to other domains.

Another limitation of the present study is the lack of an active control group, which may limit our ability to attribute observed effects solely to the intervention under investigation. By using a passive control group instead, we acknowledge the potential for confounding factors or natural changes over time to influence the observed results. Future research should consider including an active

control group to better isolate the specific effects of the intervention and to increase the internal validity of the results. Investigating the potential differential effects between the passive and active control conditions could provide valuable insights into the comparative effectiveness and mechanisms of action, further strengthening the evidence base in this area. Finally, the current study only quantitatively examined a limited ontological network of construct validity. Future research should examine the validation of the rating scale through interviews with preservice teachers in a new setting to test how they perceive the items.

## Conclusions

The current study provided evidence that the theoretical model underlying the rating scales developed by Leppink et al. (2014) is still valid in a different learning context. The scales of Leppink and colleagues used in this study were found to be reliable and valid for science teacher education. They can be easily used in science teacher education to assess the cognitive load caused by learning activities and materials across the curriculum. Therefore, the results of this study validate the use of a recently established psychometric self-report rating scale to assess three types of cognitive load in a different learning context and with learners of different races and characteristics.

## References

- Altmeyer, K., Malone, S., Kapp, S., Barz, M., Lauer, L., Thees, M., Kuhn, J., Peschel, M., Sonntag, D., & Brünken, R. (2021). The effect of augmented reality on global coherence formation processes during STEM laboratory work in elementary school children. In *Proceedings of the International cognitive load theory conference* (pp. 20–22).
- Bannert, M. (2002). Managing cognitive load – Recent trends in cognitive load theory. *Learning and Instruction*, 12(1), 139–146. [https://doi.org/10.1016/S0959-4752\(01\)00021-4](https://doi.org/10.1016/S0959-4752(01)00021-4)
- Barth, C., & Henninger, M. (2012). Fostering the diagnostic competence of teachers with multimedia training – A promising approach? In I. Deliyannis (Ed.), *Interactive multimedia* (pp. 49–66). InTech.
- Bichler, S., Schwaighofer, M., Stadler, M., Bühner, M., Greiff, S., & Fischer, F. (2020). How working memory capacity and shifting matter for learning with worked examples – A replication study. *Journal of Educational Psychology*, 112(7), 1320–1337. <https://doi.org/10.1037/edu0000433>
- Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction*, 8(4), 293–332. [https://doi.org/10.1207/s1532690xci0804\\_2](https://doi.org/10.1207/s1532690xci0804_2)
- Chen, O., Castro-Alonso, J., Paas, F., & Sweller, J. (2018). Extending cognitive load theory to incorporate working memory resource depletion: Evidence from the spacing effect. *Educational*

- Psychology Review*, 30(2), 483–501. <https://doi.org/10.1007/s10648-017-9426-2>
- Chernikova, O., Heitzmann, N., Fink, M. C., Timothy, V., Seidel, T., & Fischer, F., & DFG Research Group COSIMA (2020). Facilitating diagnostic competences in higher education – A meta-analysis in medical and teacher education. *Educational Psychology Review*, 32(1), 157–196. <https://doi.org/10.1007/s10648-019-09492-2>
- Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, 9(2), 233–255. [https://doi.org/10.1207/S15328007SEM0902\\_5](https://doi.org/10.1207/S15328007SEM0902_5)
- DeLeeuw, K. E., & Mayer, R. E. (2008). A comparison of three measures of cognitive load: Evidence for separable measures of intrinsic, extraneous, and germane load. *Journal of Educational Psychology*, 100(1), 223–234. <https://doi.org/10.1037/0022-0663.100.1.223>
- Field, A. P. (2013). *Discovering statistics using IBM SPSS statistics* (4th ed.). SAGE Publications.
- Gianns, P., & Leppink, J. (2019). Special issue on cognitive load theory. *Educational Psychology Review*, 31(2), 255–259. <https://doi.org/10.1007/s10648-019-09474-4>
- Henrich, J., Heine, S., & Norenzayan, A. (2010). The weirdest people in the world?. *Behavioral and Brain Sciences*, 33(2-3), 61–83. <https://doi.org/10.1017/S0140525X0999152X>
- Hoogerheide, V., & Roelle, J. (2020). Example-based learning: New theoretical perspectives and use-inspired advances to a contemporary instructional approach. *Applied Cognitive Psychology*, 34(4), 787–792. <https://doi.org/10.1002/acp.3706>
- Jalani, N. H., & Sern, L. C. (2015). The example-problem-based learning model: Applying cognitive load theory. *Procedia – Social and Behavioral Sciences*, 195, 872–880. <https://doi.org/10.1016/j.sbspro.2015.06.366>
- Kalyuga, S. (2009). Knowledge elaboration: A cognitive load perspective. *Learning and Instruction*, 19(5), 402–410. <https://doi.org/10.1016/j.learninstruc.2009.02.003>
- 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>
- Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. *Journal of Educational Psychologist*, 38(1), 23–31. [https://doi.org/10.1207/S15326985EP3801\\_4](https://doi.org/10.1207/S15326985EP3801_4)
- Klepsch, M., & Seufert, T. (2021). Making an effort versus experiencing load. *Frontiers in Education*, 6, Article 645284. <https://doi.org/10.3389/educ.2021.645284>
- Korbach, A., Brünken, R., & Park, B. (2018). Differentiating different types of cognitive load: A comparison of different measures. *Educational Psychology Review*, 30(2), 503–529. <https://doi.org/10.1007/s10648-017-9404-8>
- Krieglstein, F., Beege, M., Rey, G. D., Gianns, 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>
- Leppink, J., Paas, F., Van der Vleuten, C., Van Gog, T., & Van Merriënboer, J. (2013). Development of an instrument for measuring different types of cognitive load. *Journal of Behavioral Research*, 45(4), 1058–1072. <https://doi.org/10.3758/s13428-013-0334-1>
- Leppink, J., Paas, F., Van Gog, T., van Der Vleuten, C. P., & Van Merriënboer, J. J. (2014). Effects of pairs of problems and examples on task performance and different types of cognitive load. *Learning and Instruction*, 30, 32–42. <https://doi.org/10.1016/j.learninstruc.2013.12.001>
- Moos, D. C., & Pitton, D. (2014). Student teacher challenges: Using the cognitive load theory as an explanatory lens. *Teaching Education*, 25(2), 127–141. <https://doi.org/10.1080/10476210.2012.754869>
- Moreno, R., & Park, B. (2010). Cognitive load theory: Historical development and relation to other theories. In J. L. Plass, R. Moreno, & R. Brünken (Eds.), *Cognitive load theory*, (pp. 9–28). Cambridge University Press. <https://doi.org/10.1017/CBO9780511844744.003>
- Morrison, J., Roth McDuffie, A., & French, B. (2015). Identifying key components of teaching and learning in a STEM school. *School Science and Mathematics*, 115(5), 244–255. <https://doi.org/10.1111/ssm.12126>
- Paas, F., Renkl, A., & Sweller, J. (2003). 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)
- Paas, F. G., Van Merriënboer, J. J., & Adam, J. J. (1994). Measurement of cognitive load in instructional research. *Perceptual and Motor Skills*, 79(1), 419–430. <https://doi.org/10.2466/pms.1994.79.1.419>
- Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2), 1–36. <https://doi.org/10.18637/jss.v048.i02>
- Schmeck, A., Opfermann, M., Van Gog, T., Paas, F., & Leutner, D. (2015). Measuring cognitive load with subjective rating scales during problem solving: Differences between immediate and delayed ratings. *Instructional Science*, 43(1), 93–114. <https://doi.org/10.1007/s11251-014-9328-3>
- Schwaighofer, M., Bühner, M., & Fischer, F. (2016). Executive functions as moderators of the worked example effect: When shifting is more important than working memory capacity. *Journal of Educational Psychology*, 108(7), 982–1000. <https://doi.org/10.1037/edu0000115>
- Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2), 257–285. [https://doi.org/10.1016/0364-0213\(88\)90023-7](https://doi.org/10.1016/0364-0213(88)90023-7)
- Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, 4(4), 295–312. [https://doi.org/10.1016/0959-4752\(94\)90003-5](https://doi.org/10.1016/0959-4752(94)90003-5)
- Sweller, J. (2011). Cognitive load theory. In J. P. Mestre, & B. H. Ross (Eds.), *The psychology of learning and motivation: Cognition in education* (pp. 37–76). Elsevier Academic Press. <https://doi.org/10.1016/B978-0-12-387691-1.00002-8>
- Sweller, J., Van Merriënboer, J. J., & Paas, F. G. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, 10(3), 251–296. <https://doi.org/10.1023/A:1022193728205>
- Syring, M., Kleinknecht, M., Bohl, T., Kuntze, S., Rehm, M., & Schneider, J. (2015). How problem-based or direct instructional case-based learning environments influence secondary school pre-service teachers' cognitive load, motivation and emotions: A quasi-experimental intervention study in teacher education. *Journal of Education and Human Development*, 4(4), 115–129. <https://doi.org/10.15640/jehd.v4n4a14>
- 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, Article 705551. <https://doi.org/10.3389/educ.2021.705551>
- Timothy, V., Watzka, B., Stadler, M., Girwidz, R., & Fischer, F. (2023). Fostering preservice teachers' diagnostic competence in identifying students' misconceptions in physics. *International Journal of Science and Mathematics Education*, 21(5), 1685–1702. <https://doi.org/10.1007/s10763-022-10311-4>
- van Harsel, M., Hoogerheide, V., Verkoeijen, P., & van Gog, T. (2020). Examples, practice problems, or both? Effects on motivation and learning in shorter and longer sequences. *Applied Cognitive Psychology*, 34(4), 793–812. <https://doi.org/10.1002/acp.3649>
- Van Merriënboer, J. J., & 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>

Van Merriënboer, J. J., & Sweller, J. (2010). Cognitive load theory in health professional education: Design principles and strategies. *Medical Education*, 44(1), 85–93. <https://doi.org/10.1111/j.1365-2923.2009.03498.x>

Vogt, F., & Rogalla, M. (2009). Developing adaptive teaching competency through coaching. *Teaching and Teacher Education*, 25(8), 1051–1060. <https://doi.org/10.1016/j.tate.2009.04.002>

### History

Received November 14, 2022

Revision received July 12, 2023

Accepted July 17, 2023

Published online September 13, 2023

Section: Educational Psychology

### Acknowledgments

We thank the students for their interest and participation as well as the anonymous reviewers who supplied many comments, which enhanced the quality of the paper.

### Conflict of Interest

The authors have no conflict of interest to report.

### Funding


A grant of the German Academic Exchange Service (DAAD) supported the contribution of Venance Timothy (grant number 91584993).

### ORCID

Bianca Watzka

 <https://orcid.org/0000-0002-3867-9683>

Raimund Girwidz

 <https://orcid.org/0000-0003-2551-4449>

Matthias Stadler

 <https://orcid.org/0000-0001-8241-8723>

### Matthias Stadler

Institute of Medical Education

LMU University Hospital

Pettenkoferstr 8a

80336 München

Germany

[matthias.stadler@med.uni-muenchen.de](mailto:matthias.stadler@med.uni-muenchen.de)