EMPIRICAL ARTICLE

WILEY

Subjective cognitive load surveys lead to divergent results for interactive learning media

Alexander Skulmowski [iD] | Günter Daniel Rey

Psychology of Learning with Digital Media, Chemnitz University of Technology, Chemnitz, Germany

Correspondence
Alexander Skulmowski, Psychology of Learning with Digital Media, Chemnitz University of Technology, Straße der Nationen 12, 09111 Chemnitz, Germany.
Email: alexander.skulmowski@phil.tu-chemnitz.de

Peer Review
The peer review history for this article is available at https://publons.com/publon/10.1002/hbe2.184.

Abstract

Cognitive load theory assumes that cognitive demands that arise from the design of learning materials (known as *extraneous load*) are major obstacles in the learning process with (digital) media. Interactive digital media allow learners to utilize complex learning materials that respond to user input. However, recent research on cognitive load measurement has led to the question whether different survey instruments produce different measurements for extraneous load generated by interactive learning media. In a laboratory experiment, we investigated this question using digital visualizations. Most importantly, we found that two cognitive load questionnaires revealed divergent results regarding the extraneous load involved in learning with interactive visualizations. This finding indicates that different questionnaires may be needed for different types of tasks in technology-enhanced learning settings. A more fundamental implication is that there needs to be greater consideration of different types of extraneous load.

KEYWORDS
cognitive load, e-learning, human cognition, multimedia, psychometrics, measurement, technology-enhanced learning, human–computer interaction, interactivity, usability

Digital learning media can be enhanced with a wide variety of interactive elements, setting them apart from more static forms of instruction (for an overview, see Domagk, Schwartz, & Plass, 2010). These elements range from basic user controls that allow an influence over the presentation of instructional slides (e.g., Tabbers & de Koeijer, 2010) to immersive digitally-enhanced environments (e.g., Johnson-Glenberg, Megowan-Romanowicz, Birchfield, & Savio-Ramos, 2016). However, the literature on interactivity also contains negative results, as there are instances in which more static forms of learning media have resulted in better learning performance than versions that incorporated features such as user controls or related forms of responsive interfaces (e.g., Song et al., 2014). Results such as these have been explained in reference to an assumed cognitive overload (e.g., Skulmowski, Pradel, Kühnert, Brunnett, & Rey, 2016). Based on these results,

we intend to more closely assess how to describe and measure the cognitive load that can ensue from the use of interactive learning media. In the next sections, we will first discuss theoretical models of cognitive load, followed by an overview of issues for interactive learning.

1 | COGNITIVE LOAD THEORY

As learners only have a certain working memory capacity at their disposal, cognitive load theory presents a model aimed at maximizing the mental processing of relevant contents while minimizing the negative effects of irrelevant contents that cause cognitive load (Sweller, van Merriënboer, & Paas, 2019; for an older overview, see Sweller, van Merrienboer, & Paas, 1998). Cognitive load theory

postulates a mental capacity model in which the cognitive load types *intrinsic load* and *extraneous load* are used to describe the total mental load involved in learning processes (Sweller et al., 2019). Simply put, the intrinsic load of a learning task is associated with the task difficulty, more precisely with the complexity of learning contents and their relations (Sweller et al., 1998). However, extraneous load can be thought of as being influenced by the design parameters of learning materials (Sweller et al., 1998). The central point of cognitive load theory is that the limitations of learners' cognitive capacities need to be considered, resulting in a need for avoiding extraneous cognitive load in the design of learning materials (Sweller et al., 1998).

Importantly, there has been an effort to unite cognitive load theory with usability research (e.g., Hollender, Hofmann, Deneke, & Schmitz, 2010) and in many cases, research on technology-enhanced learning has incorporated cognitive load ratings along with usability surveys (e.g., Skulmowski et al., 2016). Since one potential source of cognitive load in the design of learning materials is interactivity (e.g., Kalyuga, 2007), our main objective is to examine extraneous load in the context of interactive learning media in order to improve its measurement.

2 | INTERACTIVE LEARNING

Interactive learning media typically involve some kind of user control over the learning materials (for an overview, see Domagk et al., 2010). This may include simple start and stop controls for animations (e.g., Song et al., 2014), letting learners manipulate items presented on-screen (e.g., Kalet et al., 2012; Song et al., 2014), or even providing learners with sophisticated simulations resembling computer games (e.g., Johnson-Glenberg et al., 2016). It needs to be noted that we exclusively refer to interactivity in the sense of user-manipulable learning environments and do not use the term to denote the concept of *element interactivity* (i.e., the complexity of relations between learning items, Sweller, 1994). Giving learners more control over the presentation of learning contents through user controls resulted both in positive and negative effects (see Scheiter & Gerjets, 2007, for an overview). While some types of interactivity (e.g., simple click-based selections) have been shown to enhance learning (e.g., Kalet et al., 2012), other studies conducted in the field of medical instruction provide support for the conclusion that static modes of presentation avoiding complicated interaction patterns may outperform more interactive learning media (e.g., Garg, Norman, Spero, & Maheshwari, 1999; Song et al., 2014). For instance, a study in which an interactive desktop-based medical training focusing on diagnostics allowed learners to click on elements of interest to use them in a task resulted in better learning performance than asking learners to perform more complex drag-and-drop interaction patterns with these items (Kalet et al., 2012). In line with the studies presented in this section, our study similarly uses a rather simple form of interactivity that lets learners switch between two visualizations by clicking on the image.

3 | CONCEPTUALIZING AND MEASURING EXTRANEOUS LOAD IN INTERACTIVE LEARNING

The measurement of cognitive load remains a controversial aspect (e.g., de Jong, 2010). A number of subjective survey instruments aimed at measuring the cognitive load components separately have been presented (e.g., Eysink et al., 2009; Klepsch, Schmitz, & Seufert, 2017; Leppink, Paas, Van der Vleuten, van Gog, & van Merriënboer, 2013). However, some researchers have acknowledged that there may be various types of extraneous cognitive load (e.g., Schnotz & Kürschner, 2007; Skulmowski et al., 2016). Schnotz and Kürschner (2007) mention "[...] that different kinds of misalignment are associated with different kinds of extraneous load" (p. 481). Among others, they presented task complexity, the difficulty of mentally combining separate information, and irrelevant instructional components as examples of different sources of extraneous load (Schnotz & Kürschner, 2007).

Skulmowski and Rey (2017) reviewed recent results concerning which methods of measuring cognitive load are the most appropriate for more complex learning environments centered around (bodily) activity and suggested that these types of environments require different types of cognitive load measurement techniques than other types of learning scenarios. Most importantly, Skulmowski and Rey (2017) distinguished (inter-)active settings from more verbally-oriented modes of instruction and hint at the possibility that cognitive load surveys featuring items targeted at the latter forms of instruction may not be appropriate for interactive learning. Based on Skulmowski and Rey (2017), we assume that the survey by Leppink et al. (2013) may be more fitting for learning materials that primarily involve verbal contents. This can be concluded from an example item of the instrument, "The instructions and/or explanations were full of unclear language." (Leppink et al., 2013, p. 1070), with the other two extraneous load items asking similar questions. Instead of using Eysink et al.'s (2009) survey as discussed by Skulmowski and Rey (2017), we chose to utilize a more recent, but similar, survey by Klepsch et al. (2017). Among others, this survey features items targeted at learners' difficulties at accessing information, such as "During this task, it was exhausting to find the important information." (Klepsch et al., 2017, p. 10).

As the number of different instruments available for cognitive load measurement steadily increases, there have been attempts at comparing different kinds of cognitive load measures (e.g., Naismith, Cheung, Ringsted, & Cavalcanti, 2015; Szulewski, Gegenfurtner, Howes, Sivilotti, & van Merriënboer, 2017). For instance, Naismith et al. (2015) did not find significant correlations between Paas' (1992) cognitive load item and the NASA Task Load Index (Hart & Staveland, 1988) while their study revealed a significant correlation between the NASA Task Load Index and their own six-item survey. In addition to these results, Naismith and Cavalcanti (2015) draw the conclusion that several instances of using cognitive load surveys suffer from low validity. It should be noted that these results may suffer from a lack of comparability between single-item instruments and multi-item surveys. To avoid

this issue, we compare two relatively similar surveys that are both explicitly aimed at measuring the same variable (extraneous load) and mainly differ in their wording. Nevertheless, the comparisons and reviews discussed in this section still raise an interesting question, namely whether different measurement instruments can be used interchangeably in the context of interactive learning media.

4 | CHOOSING APPROPRIATE TESTS IN THE CONTEXT OF INTERACTIVE LEARNING MEDIA

While not the main focus of the article, we also investigated how multiple testing occasions and different test types can affect learning with interactive media. As outlined above, a number of interactive features have been found to hamper with learning (e.g., Song et al., 2014). Therefore, we wanted to address whether the potential negative effects of interactive learning media stemming from a higher cognitive load can be remedied by two learning phases. Since digital learning environments are usually designed for long-term usage, a second testing phase was included to shed light onto the temporal dynamics of interactivity and learning.

The so-called *testing effect* postulates increases in retention performance due to repeated occasions of retrieval (for an overview, see Roediger & Butler, 2011). We aimed to investigate the effects of repeated testing over time when using interactive learning media compared with static versions of the same learning materials. As reviewed by Rowland (2014), some studies have provided evidence for the existence of the testing effect in short-term learning situations (Rowland, 2014, cites Carpenter & DeLosh, 2006, as an example). We chose a short interval for our study as we specifically wanted to assess how information accessibility and the testing effect interact in a very controlled, small-scale learning environment. Our general research question concerning multiple rounds of testing was whether extraneous load due to interactivity (see Skulmowski et al., 2016) will have less of an impact with multiple learning phases.

5 | THE PRESENT EXPERIMENT

We conducted the experiment to assess how different cognitive load surveys measure the extraneous load generated by interactive elements. It is known that learning performance can be put in jeopardy by requiring learners to relate separated information due to increased attentional demands (known as the *split-attention effect*, Sweller, Chandler, Tierney, & Cooper, 1990; Chandler & Sweller, 1991, 1992; for a meta-analysis, see Schroeder & Cenkci, 2018). Thus, we conducted our study using an interactive function that allows learners to view two layers of an anatomical diagram of the human back. The interactive condition lets learners switch between drawings of two muscle layers, while the static version shows these layers as one integrated image. Implementations of user controls allowing learners to view different layers are used in the field of medical education (see Yue, Kim, Ogawa, Stark, &

Kim, 2013, for an overview). However, previous research suggests that having to keep track of dynamic presentations may be a cause of lower learning performance (e.g., Lowe, 1999).

We hypothesized that Klepsch et al.'s (2017) survey would measure a larger difference in extraneous load than Leppink et al.'s (2013) survey in the interactive version; with a smaller difference between the survey scores in the static version (interaction effect: H_1). Furthermore, we intended to investigate the effect of repeated testing by repeating the learning phase and presenting the retention test twice as well. As interactive features sometimes resulted in lower learning performance in previous research (e.g., Song et al., 2014), we were interested to see whether an interactive feature can lead to a stronger rise in retention scores when learners are given a second opportunity to learn with the interactive version compared to a static version. Skulmowski et al. (2016) explain the lowered learning performance in the interactive versions of Song et al. (2014) in terms of heightened demands imposed by the user interface. Therefore, we assumed that giving learners a second chance to learn with an interactive version might be less strongly affected by the demands arising through the need to learn the interface (interaction effect: H_2).

6 | METHOD

6.1 | Participants and design

The study used a 2×2 mixed design with the between-subject factor interactivity (static vs. switchable) and, depending on the dependent variable, a within-subjects factor with two levels. Only 18–30-year-old native speakers of German with little or no knowledge of back muscle anatomy were eligible for participation. As we were not aware of similar research comparing cognitive load surveys in this particular context, we decided to be conservative in our power estimation and assumed a rather small effect of $\eta_p^2 = .04$. A power analysis using G*Power (Version 3.1.9.2; Faul, Erdfelder, Buchner, & Lang, 2009) revealed that 50 participants were sufficient to detect a within-between interaction with an estimated effect of $\eta_p^2 = .04$ (power = 0.80, α error probability = .05, correlation between repeated measures = 0.5). The data of 50 participants were collected, but the data of participants who did not interact with the learning materials in one or both learning phases was not included in the analyses. Therefore, the data of 42 participants (33 female, 9 male) were analyzed. We used block randomization to achieve an almost balanced distribution between the two between-subjects groups (n_{static} = 26, $n_{switchable}$ = 24), resulting in n_{static} = 26 and $n_{switchable}$ = 16 after removing participants who did not interact with the learning materials. Two additional incomplete datasets resulted from restarting the experiment (before beginning the learning phase). Our participants were students of Media Communication, Computer Science and Communication Science, or Media and Instructional Psychology and took part in partial fulfillment of course requirements. We conducted the web-based experiment using SoSci Survey (Version 2.6.00-i; https://www.soscisurvey.de).

6.2 | Learning materials

Schematic line diagrams of the human back based on information conveyed in medical illustrations (Bammes, 2009; Gray, 1918; Tillmann, 2016) with nine labeled muscles were presented to the participants (see Figure 1). In the static version, the superficial layer of muscles was presented on the left half and deeper muscles were shown on the right half of the image (see Figure 1a). In the interactive version, participants swapped between these two layers by clicking on the image (see Figure 1b,c). The JavaScript-based presentation logged the count of switches between the layers in order to exclude participants who did not engage with the interactive display as a means to enhance data quality. Participants were not re-assigned to a different learning setting in the second learning phase but were presented with the same version (static or interactive) that they had used in the first phase.

6.3 | Retention tests

Retention performance was tested using a two-page labeling task using versions of the two pictures of muscles presented in Figure 1b,c (without color and with letters instead of the muscle labels). Participants responded to a test concerning the superficial muscles first (four items) and, on the second test page, completed a test about the deeper muscles (five items). The first page informed participants that the two-page test was designed to test their knowledge concerning the previously learned contents and that there was no time limit. They were asked not to use any additional assistance. As these two test pages were repeated after the second learning phase, the first page of the second round of learning tests included a note that the following learning test was identical to the first one. McDonald's ω (McDonald, 1999) for the nine items of the first round of the retention test was .76; with ω = .79 for the second round of testing (one question item result had to be removed for the latter analysis due to a lack of variance). We used McDonald's ω for learning tests due to numerous advantages of this reliability measure as outlined by Dunn, Baguley, and Brunsden (2014). Scores were determined by awarding participants one point for every correctly assigned muscle. The maximum score was nine points on each of the two rounds of testing.

6.4 | Extraneous load surveys

For our study, we chose to use the three items measuring extraneous load in the survey developed by Klepsch et al. (2017) and adapted German translations of three of the four extraneous load questions presented in Table 1 of Leppink and van den Heuvel (2015). The three items from Leppink and van den Heuvel (2015) are largely identical to the three items presented in appendix 1 of Leppink et al. (2013) and were chosen due to license considerations. However, we will treat both versions of the survey as identical throughout this paper and usually cite Leppink et al. (2013) when discussing this instrument. For these three items, the adaption only consisted in a change of the word "activity" to "task" in our

translation in order to be consistent with the items of Klepsch et al. (2017). For all six extraneous load questions we used 7-point Likert scales with endpoints labeled "absolutely wrong" and "absolutely right" in line with Study 2 by Klepsch et al. (2017, p. 10). Reliability analyses resulted in Cronbach's α = 0.88 for the survey by Klepsch et al. (2017) and Cronbach's α = 0.79 for the survey by Leppink et al. (2013).

6.5 | Procedure

Participants provided informed consent and were displayed a survey asking them if they were within our targeted age range of 18–30 years, were native speakers of German, have little or no knowledge of the muscle anatomy of the back, and had not previously participated in the study. Additionally, we asked them to select their course of study and to specify their gender. Next, participants received the instructions for the first learning phase, namely that they would learn the superficial and deep anatomy of the back using a labeled diagram on the following page. The instructions stated that the names and positions of the muscles should be learned within 45 s. For those participants who were assigned to the interactive condition, an additional instruction stating that they could alternate between figures of the superficial and deep muscle layers by clicking on the image was displayed. As soon as the learning phase of 45 s had elapsed, the cognitive load survey page was displayed. The first retention test was presented on the subsequent two pages. After this first block, the instructions, learning phase (with a learning time of 45 s), and the two-page retention test were repeated. Participants were thanked and offered further information regarding the study.
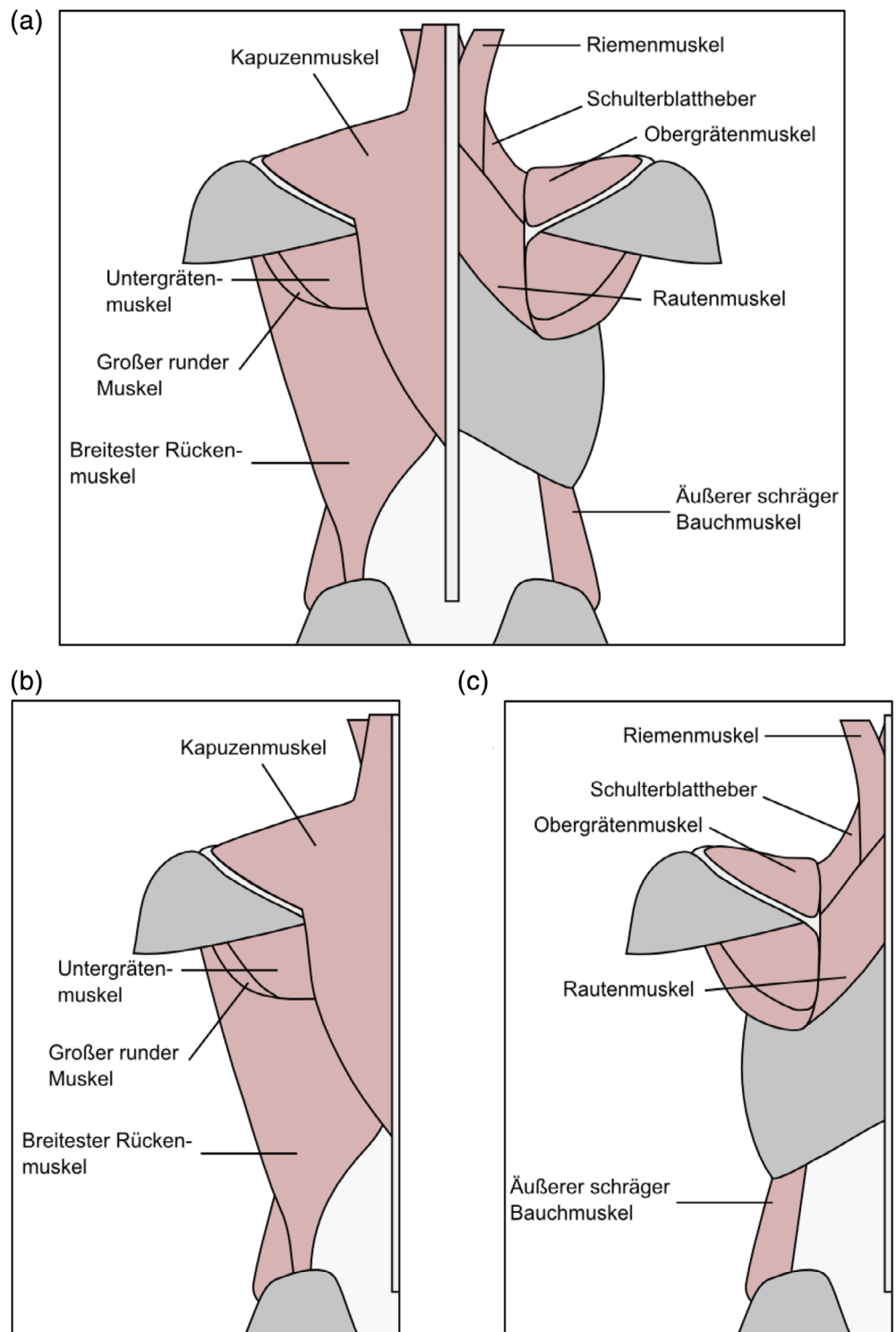
7 | RESULTS

Based on simulation studies concerning the power of the Shapiro–Wilk test, the test does not have adequate power to detect deviations from the normal distribution when used on sample sizes comparable to ours (Razali & Wah, 2011). Therefore, we used a nonparametric version of the analysis of variance (ANOVA) procedure based on aligned rank transformations (Fawcett & Salter, 1984) for the analyses in this paper. As the three tests of a 2×2 ANOVA (i.e., two main effects and one interaction effect) increase the Type I error (Cramer et al., 2016) and since we used two different dependent variables, we used Holm's sequential Bonferroni procedure (Holm, 1979) with the two hypotheses listed above to control the Type I error.

7.1 | Extraneous load

The analysis was conducted using the aligned rank transformation ANOVA procedure on averaged data. The hypothesized interaction effect between the between-subjects factor and the within-subjects factor extraneous load test with two levels (Klepsch et al., 2017 vs. Leppink et al., 2013) reached significance, $F(1, 40)$ = 8.65, p = .005, η_p^2 = .18. This

FIGURE 1 Learning materials used in the experiment (based on information conveyed in Bammes, 2009; Gray, 1918; Tillmann, 2016). (a) Static version. Panels (b) and (c) show the two images of the interactive version in which participants could switch between (b) and (c) by clicking on the image

result indicates that the difference between extraneous load scores in the two between-subjects groups is much stronger when measured with Klepsch et al.'s (2017) survey compared to Leppink et al.'s (2013) survey (see Figure 2a for the untransformed, averaged data). H_1 was thus confirmed by the data.

Besides our hypothesized result, there were two significant main effects. The interactive version significantly raised extraneous load compared to the static version, $F(1, 40) = 8.71$, $p = .005$, $\eta_p^2 = .18$. The Klepsch et al. (2017) survey resulted in significantly higher extraneous load scores overall compared to the Leppink et al. (2013) survey, $F(1, 40) = 67.93$, $p < .001$, $\eta_p^2 = .63$. The significant effect of interactivity confirms that interactive learning settings can induce extraneous load.

7.2 | Retention

We again used aligned rank-transformed ANOVAs to analyze retention data, this time with the within-subjects factor learning phase

TABLE 1 Spearman correlations between retention and extraneous load scores

		Retention	Klepsch et al. (2017)	Leppink et al. (2013)
Retention	Spearman's rho	—		
	p	—		
Klepsch et al. (2017)	Spearman's rho	−.331	—	
	p	.032	—	
Leppink et al. (2013)	Spearman's rho	−.297	.641	—
	p	.056	<.001	—
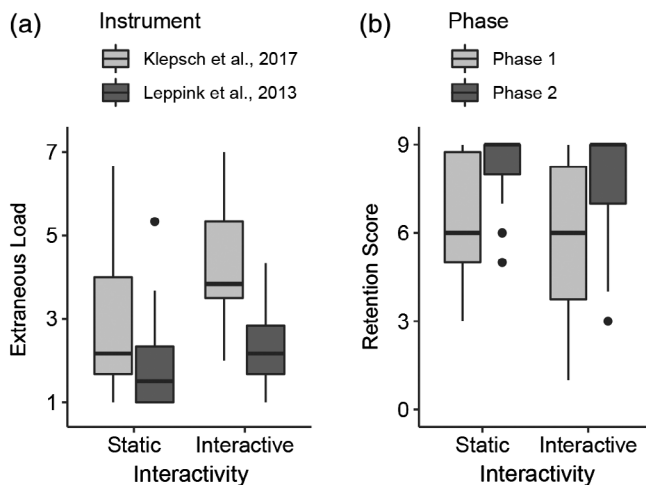
FIGURE 2 Descriptive results of the experiment. (a) Boxplot of the averaged extraneous load ratings (minimum = 1, maximum = 7). (b) Boxplot of retention scores (minimum = 0, maximum = 9)

(Phase 1 vs. Phase 2) and the between-subjects factor. Our hypothesized interaction effect between the between-subjects factor and the within-subjects factor did not result in a significant interaction, *p* = .855. The untransformed data (see Figure 2b) show that the second learning phase did not induce a stronger rise in retention performance for the interactive group compared to the static group. In addition to this result, there was a significant main effect of the factor learning phase with higher retention results in the second learning phase, $F(1, 40) = 19.92$, $p < .001$, $\eta_p^2 = .33$. There was no significant effect of the factor interactivity (*p* = .776).

7.3 | Correlations

We computed the Spearman correlations between the retention scores of the first learning phase and the two extraneous load measures (see Table 1). Both extraneous load surveys were significantly positively correlated, but only the Klepsch et al. (2017) survey reached a significant negative correlation with the retention scores of the first learning phase (in line with the assumption that a higher extraneous load is associated with a lower retention performance). This result underlines the importance of choosing a suitable cognitive load survey.

8 | DISCUSSION

We assessed the effects of the design of interactive learning environments on learning and cognitive load. As expected, we found that different extraneous load surveys deliver different results when used to assess the mental demands of interactive learning media. While one extraneous load survey (Klepsch et al., 2017) measured increased extraneous load when participants used an interactive feature during learning compared to a static version, the other survey (Leppink et al., 2013) did not detect a large difference in extraneous load. These results have far-reaching implications for cognitive load measurement and cognitive load theory as a whole.

9 | IMPLICATIONS OF THE MEASUREMENT OF EXTRANEOUS LOAD IN INTERACTIVE LEARNING MEDIA

The results indicate that the current practices of cognitive load measurement in the field of technology-enhanced learning are in dire need for revision. We demonstrated that different surveys indicate vastly different levels of extraneous load depending on whether the learning environment included interactive elements or not. In light of our data concerning the variability in the measurement of extraneous load, we propose two major conclusions.

Generally speaking, the results support the claim that the language used in different surveys should match the learning task (as discussed by Skulmowski & Rey, 2017). In the context of activity-based and interactive learning media, Skulmowski and Rey (2018) suggested a more task-oriented approach based on Wilson and Golonka (2013) that recommends the use of task analyses. A conceptualization of cognitive load involving a learner, a learning task, the physical environment, and relations between these factors was introduced by Choi, van Merriënboer, and Paas (2014). Based on these models, we emphasize the importance of task analyses for the conceptualization of extraneous load. Before selecting a cognitive load survey, researchers should check their appropriateness for the learning task (Skulmowski & Rey, 2017). Our results demonstrate that measuring extraneous cognitive load as it manifests itself in interactive learning media requires a survey featuring question items focused on the cognitive demands that potentially arise from the interaction design (see also Skulmowski & Rey, 2017). However, more research is

needed to develop clear criteria and guidelines concerning the use of different cognitive load surveys in different contexts.

A second conclusion that can be drawn from our results is that there may be multiple types of extraneous load (as previously suggested, among others, by Schnotz & Kürschner, 2007). In the case of our studies, the extraneous load could be thought of as consisting of the mental demands resulting from understanding the visuospatial arrangement of the anatomical parts on one hand and the demands created by interactive controls on the other hand. Being instructed to memorize these anatomical structures most likely generates very little extraneous load–yet the measurement of this constituent of learning tasks is a major aspect of Leppink et al.'s (2013) survey (as discussed by Skulmowski & Rey, 2017). While our results indicate that certain surveys may be more appropriate than others in specific contexts, we stress that this does not necessarily mean that one survey will be superior across several contexts. Rather, the results support the claim that the usefulness of a survey depends on contextual factors. Therefore, our study provides evidence for the claim that cognitive load surveys cannot be used interchangeably in all instructional settings.

Another important point is that Klepsch et al. (2017) emphasize that their survey can be used in more diverse and short learning settings while mentioning that Leppink et al. (2013) devised their instrument for the evaluation of more extensive settings such as entire courses. This aspect may have contributed to our results, but our results still suggest that the wordings used by Klepsch et al. (2017) make their survey more compatible with research on interactive learning media regardless of the duration or scope of a learning task. However, the aspect of learning duration and the suitability of surveys should be empirically investigated in future studies.

10 | LIMITATIONS AND OUTLOOK

An important limitation of the study is the use of a small-scale learning environment. It will be interesting to see whether our result will transfer to other forms of technology-enhanced learning such as augmented reality and virtual reality. Future research should compare additional surveys across a wide variety of learning contexts. However, the use of a limited and controlled learning environment was, in our opinion, important for establishing a foundation for the divergent effects of cognitive load surveys in interactive settings. Lastly, our investigation was focused on the effects of interactivity on cognitive load and survey measurement. Using the approach presented in this paper, further comparisons of cognitive load measurement in other contexts should be conducted to see whether similar divergent effects emerge.

It should be noted that our study focused on the extraneous load induced by the learning materials themselves and not a specific method of instruction (such as using worked examples, Paas, 1992). Hence, our approach is more aligned with the idea of using cognitive load theory as a tool for measuring usability in the context of digital learning media (see Hollender et al., 2010). Furthermore, more research is needed to determine the suitability of germane load and intrinsic load surveys for different learning settings, as our study only included extraneous load ratings.

11 | CONCLUSION

Cognitive load measurement has been described as a challenging task (e.g., de Jong, 2010). In line with perspectives suggesting that cognitive load theory itself can be improved through new findings related to cognitive load measurement (e.g., Paas, Tuovinen, Tabbers, & van Gerven, 2003), we conducted a study focusing on the survey-based measurement of extraneous load in interactive learning settings. Most importantly, our results demonstrate that different cognitive load surveys can lead to varying outcomes depending on the instructional design of the learning task. As cognitive load is not only highly relevant for the design of instructional materials, but also has been identified as a critical component in the field of usability research (e.g., Hollender et al., 2010), In sum, our results offer some of the first cues for the demanding task of choosing extraneous load measures for interactive learning. Further research will be necessary to establish more precise guidelines.

AUTHOR CONTRIBUTIONS
A. S. designed the studies and materials with critical input from G. D. R. A. S. conducted the experiments and analyzed the data. G. D. R. supervised the analyses and provided feedback and analysis tools. A. S. wrote the initial draft; G. D. R. made critical revisions. AS and GDR have read and approved the manuscript.

ORCID
Alexander Skulmowski [ID] https://orcid.org/0000-0002-1682-021X

REFERENCES

Bammes, G. (2009). *Die Gestalt des Menschen*. Freiburg, Germany: Christophorus Verlag.

Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition*, 34, 268–276.

Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction*, 8, 293–332.

Chandler, P., & Sweller, J. (1992). The split-attention effect as a factor in the design of instruction. *British Journal of Educational Psychology*, 62, 233–246.

Choi, H. H., van Merriënboer, J. J., & Paas, F. (2014). Effects of the physical environment on cognitive load and learning: Towards a new model of cognitive load. *Educational Psychology Review*, 26, 225–244.

Cramer, A. O., van Ravenzwaaij, D., Matzke, D., Steingroever, H., Wetzels, R., Grasman, R. P., ... Wagenmakers, E. J. (2016). Hidden multiplicity in exploratory multiway ANOVA: Prevalence and remedies. *Psychonomic Bulletin & Review*, 23, 640–647.

de Jong, T. (2010). Cognitive load theory, educational research, and instructional design: Some food for thought. *Instructional Science*, 38, 105–134.

Domagk, S., Schwartz, R. N., & Plass, J. L. (2010). Interactivity in multimedia learning: An integrated model. *Computers in Human Behavior*, 26, 1024–1033.

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, 105, 399–412.

Eysink, T. H., de Jong, T., Berthold, K., Kolloffel, B., Opfermann, M., & Wouters, P. (2009). Learner performance in multimedia learning arrangements: An analysis across instructional approaches. *American Educational Research Journal*, 46, 1107–1149.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41, 1149–1160.

Fawcett, R. F., & Salter, K. C. (1984). A Monte Carlo study of the F test and three tests based on ranks of treatment effects in randomized block designs. *Communications in Statistics-Simulation and Computation*, 13, 213–225.

Garg, A., Norman, G. R., Spero, L., & Maheshwari, P. (1999). Do virtual computer models hinder anatomy learning? *Academic Medicine*, 74, S87–S89.

Gray, H. (1918). *Anatomy of the human body* (20th ed.). Philadelphia, PA: Lea & Febiger Retrieved from https://hdl.handle.net/2027/uc2.ark:/13960/t9m32qk0z

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in Psychology*, 52, 139–183.

Hollender, N., Hofmann, C., Deneke, M., & Schmitz, B. (2010). Integrating cognitive load theory and concepts of human–computer interaction. *Computers in Human Behavior*, 26, 1278–1288.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6, 65–70.

Johnson-Glenberg, M. C., Megowan-Romanowicz, C., Birchfield, D. A., & Savio-Ramos, C. (2016). Effects of embodied learning and digital platform on the retention of physics content: Centripetal force. *Frontiers in Psychology*, 7, 1819. https://doi.org/10.3389/fpsyg.2016.01819

Kalet, A. L., Song, H. S., Sarpel, U., Schwartz, R., Brenner, J., Ark, T. K., & Plass, J. (2012). Just enough, but not too much interactivity leads to better clinical skills performance after a computer assisted learning module. *Medical Teacher*, 34, 833–839.

Kalyuga, S. (2007). Enhancing instructional efficiency of interactive e-learning environments: A cognitive load perspective. *Educational Psychology Review*, 19, 387–399.

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

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, 1058–1072.

Leppink, J., & van den Heuvel, A. (2015). The evolution of cognitive load theory and its application to medical education. *Perspectives on Medical Education*, 4, 119–127. https://doi.org/10.1007/s40037-015-0192-x

Lowe, R. K. (1999). Extracting information from an animation during complex visual learning. *European Journal of Psychology of Education*, 14, 225–244.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Naismith, L. M., & Cavalcanti, R. B. (2015). Validity of cognitive load measures in simulation-based training: A systematic review. *Academic Medicine*, 90, S24–S35.

Naismith, L. M., Cheung, J. J., Ringsted, C., & Cavalcanti, R. B. (2015). Limitations of subjective cognitive load measures in simulation-based procedural training. *Medical Education*, 49, 805–814.

Paas, F., Tuovinen, J. E., Tabbers, H., & van Gerven, P. W. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist*, 38, 63–71.

Paas, F. G. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology*, 84, 429–434.

Razali, N. M., & Wah, Y. B. (2011). Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics*, 2, 21–33.

Roediger, H. L., III, & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, 15, 20–27.

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, 140, 1432–1463.

Scheiter, K., & Gerjets, P. (2007). Learner control in hypermedia environments. *Educational Psychology Review*, 19, 285–307.

Schnotz, W., & Kürschner, C. (2007). A reconsideration of cognitive load theory. *Educational Psychology Review*, 19, 469–508.

Schroeder, N. L., & Cenkci, A. T. (2018). Spatial contiguity and spatial split-attention effects in multimedia learning environments: A meta-analysis. *Educational Psychology Review*, 30, 679–701. https://doi.org/10.1007/s10648-018-9435-9

Skulmowski, A. (2019). Bridging the gap between embodied cognition and cognitive load theory [Doctoral dissertation]. Retrieved from https://nbn-resolving.org/urn:nbn:de:bsz:ch1-qucosa2-353589

Skulmowski, A., Pradel, S., Kühnert, T., Brunnett, G., & Rey, G. D. (2016). Embodied learning using a tangible user interface: The effects of haptic perception and selective pointing on a spatial learning task. *Computers & Education*, 92-93, 64–75.

Skulmowski, A., & Rey, G. D. (2017). Measuring cognitive load in embodied learning settings. *Frontiers in Psychology*, 8, 1191. https://doi.org/10.3389/fpsyg.2017.01191

Skulmowski, A., & Rey, G. D. (2018). Embodied learning: Introducing a taxonomy based on bodily engagement and task integration. *Cognitive Research: Principles and Implications*, 3, 6. https://doi.org/10.1186/s41235-018-0092-9

Song, H. S., Pusic, M., Nick, M. W., Sarpel, U., Plass, J. L., & Kalet, A. L. (2014). The cognitive impact of interactive design features for learning complex materials in medical education. *Computers & Education*, 71, 198–205.

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, 4, 295–312.

Sweller, J., Chandler, P., Tierney, P., & Cooper, M. (1990). Cognitive load as a factor in the structuring of technical material. *Journal of Experimental Psychology: General*, 119, 176–192.

Sweller, J., van Merriënboer, J. J., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review* Advance online publication, 31, 261–292. https://doi.org/10.1007/s10648-019-09465-5

Sweller, J., van Merrienboer, J. J., & Paas, F. G. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, 10, 251–296.

Szulewski, A., Gegenfurtner, A., Howes, D. W., Sivilotti, M. L., & van Merriënboer, J. J. (2017). Measuring physician cognitive load: validity evidence for a physiologic and a psychometric tool. *Advances in Health Sciences Education*, 22, 951–968.

Tabbers, H. K., & de Koeijer, B. (2010). Learner control in animated multimedia instructions. *Instructional Science*, 38, 441–453.

Tillmann, B. N. (2016). *Atlas der Anatomie des Menschen* (3rd ed.). Berlin: Springer-Verlag.

Wilson, A. D., & Golonka, S. (2013). Embodied cognition is not what you think it is. *Frontiers in Psychology*, 4, 58. https://doi.org/10.3389/fpsyg.2013.00058

Yue, C., Kim, J., Ogawa, R., Stark, E., & Kim, S. (2013). Applying the cognitive theory of multimedia learning: An analysis of medical animations. *Medical Education*, 47, 375–387.

AUTHOR BIOGRAPHIES

Dr. Alexander Skulmowski is a postdoc at the Chair of Psychology of Learning with Digital Media, Chemnitz University of Technology, Chemnitz, Germany. He teaches courses on the design of multimedia learning environments and interactive media. His research interests include interactive learning media, realism, and virtual reality.

Prof. Dr. Günter Daniel Rey is a professor of Psychology of Learning with Digital Media at Chemnitz University of Technology, Chemnitz, Germany. His courses focus on instructional psychology and he holds lectures on learning with media. His research encompasses cognitive, emotional, social, and motivational facets of learning with digital media.