Psychometric Properties of a Cognitive Load Measure When Assessing the Load Associated with a Course

Dan J Miller

Prisca Noble

Sue Medlen

Karina Jones

Suzy Munns

James Cook University, Townsville, Queensland, Australia

Correspondence to: Dan J. Miller, Department of Psychology, College of Healthcare Sciences, Division of Tropical Health and Medicine, James Cook University, Townsville, Qld 4811, Australia. Email: daniel.miller1@jcu.edu.au

The Version of Record of this manuscript has been published and is available in *Journal of Experimental Education*, 20 July, 2021, https://doi.org/10.1080/00220973.2021.1947763

Embargo this document until 20 January, 2023

Abstract

The cognitive load imposed by instruction is an important consideration for instructional designers. Theoretical models have traditionally divided total cognitive load into intrinsic, extrinsic, and germane load. The 10-item Cognitive Load Inventory (CLI-10) is designed to measure these three types of cognitive load. It is typically administered immediately following a discrete learning activity (e.g., a lecture). This study assesses the properties of the CLI-10 when used to measure the "long-term" cognitive load experienced in a course, over a semester. To do this, the instrument was given to a group of students enrolled in a veterinary anatomy course (N = 94), toward the end of a 13-week semester. Students were asked to indicate the cognitive load they experienced across this course. Confirmatory factor analysis supported the instrument's three-factor structure when used in this way. Further, the instrument's three subscales performed well in terms of internal reliability and convergent and discriminant validity.

Keywords: Cognitive Processes/Development; Factor Analysis; Higher Education; Instructional Design/Development; Metacognition

Introduction

In the context of instructional design research, cognitive load refers to the mental effort required to learn new material. Leppink et al.'s (2013) ten-item Cognitive Load Inventory (CLI-10) is a short measure designed to assess three types of cognitive load: intrinsic load (IL), extrinsic load (EL), and germane load (GL). Previous research has focused on assessing the psychometric properties of the instrument when administered immediately following an educational activity (e.g., immediately following a lecture). This study adds to the literature by assessing the properties of the instrument when used as a long-term measure of cognitive load. The CLI-10 being valid when used to assess the cumulative cognitive load experienced over a semester would allow for more flexibility in the way educational designers assess the cognitive load associated with a course.[1]

Cognitive load theory (see Leppink & van den Heuvel, 2015) posits that a) for learning to occur information needs to be encoded into long-term memory via working memory, b) there are inherent limitations to working memory (in terms of both capacity and duration), and c) these limitations should be considered in instructional design in order to maximize learning. In cognitive load theory knowledge is assumed to be stored in long-term memory in the form of schemas: cognitive frameworks for organizing interrelated information elements (Sweller et al., 1998; van Merriënboer & Sweller, 2010). Schema construction brings information elements together (i.e., chunking) into something that can be treated as a single unit in working memory (thereby reducing load). Schemas which are frequently used become automated, such that they do not need to be consciously processed in working memory (e.g., learning to read).

As mentioned above, total cognitive load has traditionally been divided into EL, IL, and GL (Sweller et al., 1998). IL refers to working memory load imposed as a result of the natural complexity of the material being studied. IL is also reflective of the prior knowledge of the learner. EL refers to mental effort resulting from the way material is organized and presented, with suboptimal instructional design imposing greater EL.

Whereas IL and EL primarily concern the characteristics and presentation of the material to be learned, GL concerns the cognitive resources the learner applies to this material in order to create and automate schemas. This would include deliberate strategies that the learner applies to learning material. Originally, cognitive load theory distinguished

[1] By "course" we mean the equivalent to the Australian and UK term "unit", that is, a unit of teaching on a particular topic area that lasts one academic semester (e.g., *Introductory Psychology 101*).

between IL and EL only, with GL being added by Sweller et al. (1998) at a later point. There is currently debate as to whether GL should be considered a distinct form of cognitive load or merely a subcomponent of IL (see Kalyuga 2011; Leppink & van den Heuvel, 2015; Young & Sewell, 2015).

Improving instructional design involves reducing EL, encouraging GL, and managing IL (van Merriënboer & Sweller, 2010). For example, EL can be minimized through the use of worked examples and the avoidance of split-attention (e.g., avoiding having the information needed to solve a problem dispersed over multiple documents). Although the inherent difficulty of material is thought to be largely immutable, strategies can also be applied to optimize IL (e.g., gradually increasing complexity of tasks, and ensuring material difficulty matches the prior knowledge of the learner). GL can by promoted by encouraging students to use meta-cognitive skills which promote schema construction (e.g., self-explaining information).

Of course, assessing the quality of instructional design from a cognitive load perspective requires a valid measure of cognitive load. Approaches to the measurement of cognitive load include the collection of performance data (e.g., the accuracy of detecting an auditory signal), physiological indices (e.g., brain activity, pupil dilation), and subjective, self-report measures (Paas et al., 2003).

One issue with older self-report measures of cognitive load (e.g., Paas, 1992)—and also physiological and performance-based measures—is that they produce an index of overall cognitive load only. As such, it is not possible to assess how changes to instructional design differently impact the types of cognitive load. Accordingly, Leppink et al. (2013) developed the CLI-10 to produce scores for IL, EL, and GL.

Using factor analytic techniques, Leppink et al. (2013) demonstrated the CLI-10's three-factor structure (representing IL, EL, and GL) and internal reliability, across multiple samples of social and health science students attending lectures on statistics and/or research methodology. Zukić et al. (2016) confirmed the instrument's three-factor structure among their own sample of undergraduate psychology students learning statistics. The CLI-10's three-factor structure has also been demonstrated outside of statistical instruction, among medical students engaging in a problem-based learning activity (Hadie & Yusoff, 2016), high-school students undergoing a medical simulation activity (Cook et al., 2017), and university students taking language classes (Leppink et al., 2014). The CLI-10 has also been successfully adapted and expanded (to 19 items) to assess cognitive load associated with

performing colonoscopies among surgical residents (Sewell et al., 2016), while maintaining a clear three-factor structure.

This said, a two-factor solution (consisting of EL and IL) was found to be more appropriate than a three-factor solution when measuring the cognitive load associated with e-textbook use, among a large sample of undergraduate biology and anatomy/physiology students (Novak et al., 2018). Furthermore, attempts to adapt the CLI-10 to assess the cognitive load associated with patient-handover simulations among medical students failed to produce a clear three-factor solution (Young et al., 2016; Young et al., 2017). However, Young and colleagues' *Cognitive Load Inventory for Handovers* used CLI-10 items as a starting point only. Thus, these findings may not necessarily reflect problems with the factor structure of the CLI-10 itself.

In demonstration of the instrument's predictive validity, CLI-10 scores have been found to correlate with various learning outcomes; showing moderate correlations with performance on post-lecture quizzes (Đapo, & Husremović, 2016), small-to-moderate correlations with course exam performance (Leppink et al., 2014), small correlations with course letter grades (Novak et al., 2018), and a moderate correlation with accuracy during patient handover simulations (Young et al., 2017).

Current Study

As can be seen, there are some inconsistencies within this literature as to whether items on the CLI-10 form three distinct factors (which one would expect if the CLI-10 really does index three distinct forms of cognitive load). Another limitation of this literature is the timing of the measurement of cognitive load. Typically, the CLI-10 (or one of its derivative instruments) is administered directly following a discrete learning activity (e.g., immediately after a lecture or directly following participation in a problem-based learning activity). This lack of delay between the learning activity and measurement of cognitive load has been identified as problematic by the instrument's lead author (Leppink, 2017) and others (Young & Sewell, 2015), as schema construction and automation would be unlikely to occur immediately. This problem is especially relevant to the measurement of GL (which more directly relates to schema construction and automation).

In the current study, a slightly modified version of the CLI-10 was administered to a sample of undergraduate veterinary science students in a veterinary anatomy course—a population to which, to the authors' best knowledge, the CLI-10 has not previously been administered. The instrument was given toward the end of the 13-week semester, with

questions being modified to assess the cumulative cognitive load experienced over the semester. This article aims to assess the psychometric properties of the CLI-10 among this population, when used as a long-term measure of cognitive load in this way. The instrument is assessed in terms of its factor structure, the convergent and discriminant validity of subscales, and the internal reliability of subscales.

Method

Procedure and Participants

Data were collected as part of a larger longitudinal study tracking several cohorts of undergraduate veterinary students across their study of anatomy. The data for this study were collected in the first of three anatomy courses. At the time of data collection (the final week of a 13-week teaching semester), the course had covered locomotor anatomy (anatomy of the locomotor system, e.g., bones, joints, and muscles) and nervous anatomy (anatomy of the sensory organs and nervous system) in relation to multiple species. Locomotor anatomy was covered is Weeks 1-10 (with an additional lecture-recess week during this period), whereas nervous anatomy was covered in Weeks 11-12.

Participation in the study involved completing a short, pen-and-paper questionnaire. Of the 119 students enrolled in the subject, 99 were present at the lecture in which data were collected. Ninety-six of these students consented to complete the questionnaire, resulting in a response rate of 80.7%. The sample had a mean age of 21.22 years (*SD* = 4.91, range = 17–54). The majority of participants identified as female (80.2%) and Australian (94.8%).

Measures

The questionnaire assessed demographic information and the cognitive load associated with the course. As mentioned above, cognitive load was assessed using Leppink et al.'s (2013) instrument. The instrument consists of 10 11-point rating scales, anchored by 0 = *not at all the case* and 10 = *completely the case*. The instrument is designed to produce scores for IL (measured by items 1, 2, and 3), EL (items 4, 5, and 6), and GL (items 7, 8, 9, and 10; see Leppink et al., 2013, for original item wording). The instrument was presented twice (but at the same point in time): once in relation to the teaching of locomotor anatomy and once in relation to nervous anatomy.

As suggested by the instrument's authors, items were modified to make them specific to the content areas being taught. For example, Item 8 reads "This activity really enhanced my knowledge and understanding of statistics." Here *statistics* was altered to *functional*

locomotor anatomy and *functional nervous anatomy* respectively. Items 2 and 9 of the CLI-10 relate to understanding around formulas. These items were modified to ask about the clinical applications of anatomical concepts (a focus of the subject). Students were instructed that the questions related to the teaching activities across the semester (lectures and tutorials) associated with the respective content blocks.

Data Analysis

Given our *a priori* expectations of the scale's factor structure, confirmatory factor analysis (CFA) was utilized over exploratory factor analysis (EFA; Fabrigar et al., 1999). CFA was performed using maximum likelihood estimation.

Data cleaning (see Supplementary Material) left a final *N* of 94 and 93 for the analysis of the locomotor and nervous anatomy questions respectively. Larger sample sizes are preferred for factor analytic techniques. This said, smaller samples are permissible for simpler models, especially when factors are overdetermined (at least 3-4 items per factor) and communalities are high (an average of >.70; Fabrigar et al., 1999). Similarly, Hair et al. (2014) suggest a sample of around 100 is adequate in situations where measurement models contain fewer than five factors, each factor has more than three indicators and communalities are above .60. We note that the measurement models tested in the current study have three factors, with three to four indicators per factor and commonalities exceeding .70.

Results

CFA

Two CFA models (one for the nervous anatomy data and one for the locomotor anatomy data) were specified. The models consisted of three first-order factors (to represent IL, EL, and GL), with items being treated as reflective indicators of their proposed factors (Items 1–3 for the IL factor, Items 4–6 for the EL factor, and Items 7–10 for the GL factor). Items were not permitted to cross-load on non-salient factors and item error terms were not permitted to covary. The three factors were permitted to covary (see Figure S1, Supplementary Material).

Model fit was assessed based on the following indices: the model χ^2 test, the comparative fit index (CFI), the Tucker-Lewis index (TLI), the standardized root mean square residual (SRMR), and the root mean square error of approximation (RMSEA). The RMSEA is reported along with its 90% CIs and the $p_{\text{close-fit}}$. Kline's (2011) method of using the RMSEA's 90% CI to inspect model fit was utilized. Correlation residuals and standardized covariance residuals were also inspected. In both models, very few correlational

residuals had an absolute value greater than .10, and no standardized covariance residuals had an absolute value greater than 2.

Nervous Anatomy. In terms of the nervous anatomy model, the χ^2 test was significant, $\chi^2(32) = 50.02$, $p = .022$, so the exact-fit hypothesis was rejected. However, the hypothesis that the model exactly fits the data (i.e., all residual values are zero) is overly stringent. This, and other problems with this index, have resulted in many authors de-emphasizing this fit index (Hoyle, 2011). Traditionally .90 has been suggested as the cut-off value for the CFI and TLI, however many authors now recommend a more stringent cut-off value of .95 (Hoyle, 2011). The CFI and TLI were both well above .95 (CFI = .98, TLI = .97), indicating good fit. It is recommended that the SRMR be below .08 (Hu & Bentler, 1999), which was the case (SRMR = .050). The RMSEA was .078 [.030, .118]. As the lower bound of the RMSEA's 90% CI was less than .05, the close-fit hypothesis (the hypothesis that the model closely fits the data) could not be rejected (this is supported by the non-significant $p_{\text{close-fit}}$ of .136). However, the upper bound of the 90% CI exceeded .10. Thus, the poor-fit hypotheses (the hypothesis that the model poorly fits the data) also could not be rejected. These RMSEA values are interpreted with caution, as the RMSEA is biased toward indicating poor fit for models with low degrees of freedom and small sample sizes (Kenny et al., 2015).

In some of the CFA models outlined in Leppink et al. (2013), covariances were freed between error terms for Items 7 (which related to "understanding of the topics covered") and 9 ("understanding of the formulas covered") and Items 9 and 10 ("understanding of concepts and definitions"), based on the theoretical rationale that, depending on the focus of the particular lecture, there could be significant overlap between "topics" and "formulas" or "topics" and "concepts and definitions." Zukić et al. (2016) also found that freeing the errors for Items 9 and 10 to covary improved model fit. Modification indices were inspected to determine if freeing these parameters would improve model fit. This was not the case.

Locomotor Anatomy. In terms of the model assessing the locomotor anatomy data, the model χ^2 test was not significant, $\chi^2(32) = 43.26$, $p = .088$, so the exact-fit hypothesis was not rejected. Again, the CFI, TLI and SRMR all indicated good fit (CFI = .98, TLI = .98, SRMR = .051). The RMSEA was .062 [.000, .105]. Accordingly, the close-fit hypothesis could not be rejected (which, again, was supported by the non-significant $p_{\text{close-fit}}$ of .322). However, once again the poor-fit hypothesis also could not be rejected.

Inspection of modification indices indicated that freeing errors for Items 9 and 10 to covary would improve model fit (as was done in Zukić et al., 2016, and Leppink et al., 2013).

The fit statistics for this model indicated very good fit, $\chi^2(31) = 31.35$, p = .499; CFI > .99; TLI > .99; SRMR = .048, RMSEA = .011 [.000, .079], $p_{\text{close-fit}} = .759$, with the upper-bound of the RMSEA's 90% CI even dropping below .10.

Correlations between within-factor errors indicate that the correlated items share variance that cannot be explained by their factor (GL in the case of Items 9 and 10). There are many potential causes for correlated within-factor errors (Gerbing & Anderson, 1984; Lucke, 2005), the most problematic (from the perspective of trying to establish the factor structure of an instrument) being that the items are measuring an additional factor which was not, but should have been, included in the model.

It is recommended (e.g., Schmitt, 2011) that researchers be open to performing follow-up EFA should a CFA model show a pattern of problematic fit. Given the covaried error terms observed here, and previously, a follow-up EFA was performed on the locomotor anatomy data (see Supplementary Material) to verify the three-factor solution proposed by Leppink et al. (2013). This analysis strongly supported the three-factor solution.

Reliability and Validity

Factor loadings for each model are presented in Table 1. All tested factor loadings where significant at $p < .001$. Squared multiple correlation (R^2) values (also referred to as *communalities* in factor analysis) are also given in Table 1. These values indicate the proportion of variance in the item that is explained by that item's factor. It is recommended that all items have large and statistically-significant standardized factor loadings (e.g., Hair et al., 2014, recommend standardized factor loadings be at least .50, and ideally > .70). All items had good factor loadings in both models apart from Item 9, which showed a marginal factor loading (.56) in the locomotor anatomy model. Given that the factors were freed to covary, structure coefficients are also reported.

Table 2 reports correlations between factors, and McDonald's omega (ω) and average variance extracted (AVE) values for each subscale (see Supplementary Materials for information on the calculation of ω and AVE values).

ω is reported over Cronbach's alpha, as not all of the assumptions of alpha—that factor loadings are equal across items (e.g., essential tau equivalency), that the test is unidimensional, and that error terms for items are uncorrelated (Flora, 2020)—would be tenable for all subscales. ω was greater than .88 across all constructs, in both models, indicating good reliability for all subscales. Hair et al. (2014) give the following recommendations for determining the convergent validity of a subscale (the degree to which all items on the subscale converge, i.e., measure the same construct): AVE > .5, reliability >

.7, and standardized factor loadings for all items > .5. This was the case for all three constructs in both models.

Leppink et al. (2014) suggests that if the three types of cognitive load are independent and additive, we might expect the correlation between them to be near zero. As can be seen from the table, correlations between IL and EL scores and IL and GL scores were close to zero. Significant negative correlations were observed between EL and GL scores (although this is consistent with previous research; Hadie et al., 2016; Leppink et al., 2014). However, these correlations were still in the small-to-medium range. The relatively small correlations would support the discriminant validity of factors. Additionally, all subscales across the two models meet Hair et al.'s (2014, p. 620) rule of thumb for establishing discriminant validity between two constructs: AVE values > the square of the correlation between the two constructs.

Discussion

This study investigated the psychometric properties of the CLI-10 when used to assess the cumulative cognitive load experienced in a course over a semester. CFA was used to evaluate the three-factor structure of the instrument in relation to the two areas covered in the course: nervous and locomotor anatomy. Although the CFA models did not meet all fit criteria, it is not unusual for goodness of model fit to differ across fit indices (Kenny et al., 2015). We note that for indices which are biased toward indicating poor model fit in smaller samples (e.g., the upper-bound 90% CI of the RMSEA; Kenny et al., 2015) the models performed worse than on indices which are not bound to sample size (e.g., the CFI and TLI). Accordingly, we believe the data can be said to fit the proposed three-factor model in relation to both the nervous and locomotor anatomy learning blocks. Longer delays between a phenomenon and its measurement are known to impact response accuracy on surveys (Tourangeau, 1999). Accordingly, the performance of the instrument in relation to the locomotor anatomy content is especially noteworthy (given that there was a three-week delay between the end of the locomotor anatomy learning block and administration of the instrument).

It should be noted that, in the locomotor anatomy model, model fit improved once error terms for Items 9 and 10 were freed to covary. Given that items were presented in sequential order, it is proposed that this may be the result of *item bundling*. Lucke (2005) describes an item bundle as "a cluster of items that share a common stimulus, contain common content, or possess a common structure" (Lucke, 2005, p. 110). We note that Items

9 and 10 do start with the same stem. Furthermore, Items 9 and 10 were presented sequentially, potentially exacerbating this bundling effect. Other studies have similarly found that freeing error terms for Items 9 and 10 to covary improves model fit (Zuzik et al., 2016). Follow-up EFA (see Supplementary Material) did not indicate that Items 9 and 10 were loading on a fourth, previously-unspecified factor, giving us confidence in the three-factor solution. Future studies may simply want to separate these items to minimize potential bundling effects.

Subscale items showed good internal reliability across both areas of anatomy. Furthermore, evidence was found for the convergent and discriminant validity of the three subscales. That is, the results were consistent with the notion that items on each subscale are measuring the same construct (convergence) and that these constructs are distinct (discrimination). However, it should be kept in mind that the convergent validity of a subscale does not necessarily mean that this subscale is measuring the construct we believe it to measure (i.e., IL, EL, or GL). This is a possibility that Leppink et al. (2014) investigated. The authors argue that it may be that those sampled in their earlier validation study of the CLI-10 (Leppink et al. 2013) responded to items to indicate their *estimations* of the three types of cognitive load required to learn the material, rather than the actual level of mental effort they invested in learning the material. Accordingly, the authors added three items to the CLI-10 which more closely assess the mental effort invested around each type of cognitive load. The authors then assessed if these new items were internally consistent with the existing items on each subscale, as it was argued that this would confirm that the items on the subscale measure the same latent construct, confirming the original interpretation of the subscale as measuring actual cognitive load. This was found to be the case for IL and EL (i.e., items on these subscales do appear to be assessing IL and EL, not just participants' estimations of required IL and EL), but not GL. These findings have resulted in Leppink et al. (2014) endorsing the more recent conceptualizations of GL as a subcomponent of IL (see Kalyuga, 2011), rather than a distinct form of cognitive load. Accordingly, Leppink et al. (2014) and others (Hadie & Yusoff, 2016; Leppink & van den Heuvel, 2015) now refer to this subscale as a measure of "self-perceived learning", as opposed to GL.

As mentioned above, whether GL represents a distinct cognitive load type is currently an active area of debate. Although some authors have called for a reinterpretation of GL as a subcomponent of IL, not all agree. Young and Sewell (2015) contend that until further research indicates otherwise, there is still value in conceptualizing IL and GL as distinct entities. They note that multiple previous studies finding a three-factor solution for cognitive

load type provides strong evidence for there being three distinct forms of cognitive load. This study adds to the literature which supports cognitive load having three separate components.

In summary, the current study provides instructional designers interested in the measurement of cognitive load further evidence as to the psychometric properties of the CLI-10 in terms of its three-factor structure, and the internal reliability, convergent validity, and discriminant validity of its three subscales. Importantly, the study also provides evidence that the CLI-10 performs well when used as a long-term cognitive load measure. In this way, the study opens the use of the CLI-10 in different kinds of instructional assessment scenarios (e.g., assessing the cognitive load experienced across a semester, as opposed to assessing the cognitive load associated with a particular task only), assisting instructional designers and researchers to address questions like *Does including self-explaining activities in tutorials have a noticeable impact on GL? Does structuring lectures to gradually increase the complexity of material optimize IL across the semester? Is online lecture delivery associated with greater EL compared to in-person delivery?* and *Do these changes to my subject actually work to reduce students' cognitive load, and does this positively impact students' learning and/or satisfaction?*

Data Disclosure Statement

The data that support the findings of this study are available from the corresponding author, DJM, upon reasonable request.

Declaration of Interest Statement

No potential conflict of interest was reported by the authors.

References

Cook, D. A., Castillo, R. M., Gas, B., & Artino Jr, A. R. (2017). Measuring achievement goal motivation, mindsets and cognitive load: Validation of three instruments' scores. *Medical Education*, *51*, 1061–1074. https://doi.org/10.1111/medu.13405

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*, 272–299. http://dx.doi.org/10.1037/1082-989X.4.3.272

Flora, D. B. (2020). Your coefficient alpha is probably wrong, but which coefficient omega is right? A tutorial on using R to obtain better reliability estimates. *Advances in Methods and Practices in Psychological Science, 3*, 484–501. https://doi.org/10.1177/2515245920951747

Gerbing, D. W., & Anderson, J. C. (1984). On the meaning of within-factor correlated measurement errors. *Journal of Consumer Research*, *11*, 572–580. https://doi.org/10.1086/208993

Hadie, S. N., & Yusoff, M. S. (2016). Assessing the validity of the cognitive load scale in a problem-based learning setting. *Journal of Taibah University Medical Sciences*, *11*, 194–202. http://dx.doi.org/10.1016/j.jtumed.2016.04.001

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2014). *Multivariate data analysis* (7th ed.). Harlow, UK: Pearson.

Hoyle, R. H. (2011). *Structural equation modeling for social and personality psychology*. Thousand Oaks, CA: SAGE.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*, 1–55. https://doi.org/10.1080/10705519909540118

Kalyuga, S. (2011). Cognitive load theory: How many types of load does it really need? *Educational Psychology Review*, *23*, 1–19. https://doi.org/10.1007/s10648-010-9150-7

Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2015). The performance of RMSEA in models with small degrees of freedom. *Sociological Methods & Research*, *44*, 486–507. https://doi.org/10.1177/0049124114543236

Kline, R. B. (2011). *Principles and practices of structural equation modeling* (3rd ed.). New York, NY: Guilford.

Leppink, J. (2017). Cognitive load theory: Practical implications and an important challenge. *Journal of Taibah University Medical Sciences, 12*, 385–391. http://dx.doi.org/10.1016/j.jtumed.2017.05.003

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. http://dx.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. http://dx.doi.org/10.1016/j.learninstruc.2013.12.001

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

Lucke, J. F. (2005). "Rassling the hog": The influence of correlated item error on internal consistency, classical reliability, and congeneric reliability. *Applied Psychological Measurement, 29*(2), 106-125. https://doi.org/10.1177/0146621604272739

Novak, E., Daday, J., & McDaniel, K. (2018). Assessing intrinsic and extraneous cognitive complexity of E-textbook learning. *Interacting with Computers, 30*, 150–161. https://doi.org/10.1093/iwc/iwy001

Paas, F. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology, 84*, 429–434. http://dx.doi.org/10.1037/0022-0663.84.4.429

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. https://doi.org/10.1207/S15326985EP3801_8

Schmitt, T. A. (2011). Current methodological considerations in exploratory and confirmatory factor analysis. *Journal of Psychoeducational Assessment, 29*, 304–321. https://doi.org/10.1177/0734282911406653

Sewell, J. L., Boscardin, C. K., Young, J. Q., ten Cate, O., & O'Sullivan, P. S. (2016). Measuring cognitive load during procedural skills training with colonoscopy as an exemplar. *Medical Education, 50*, 682–692. https://doi.org/10.1111/medu.12965

Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational Psychology Review, 22*, 123–138. https://doi.org/10.1007/s10648-010-9128-5

Sweller, J., van Merriënboer, J. J., & Paas, F. G. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*, 251–296. https://doi.org/10.1023/A:1022193728205

Tourangeau, R. (1999). Remembering what happened: Memory errors and survey reports. In A. A. Stone, J. S. Turkkan, C. A. Bachrach, J. B. Jobe, H. S. Kurtzman, & V. S. Cain (Eds.), *The science of self-report: Implications for research and practice* (pp. 41-60). Psychology Press.

Young, J. Q., Boscardin, C. K., van Dijk, S. M., Abdullah, R., Irby, D. M., Sewell, J. L., ... & O'Sullivan, P. S. (2017). Performance of a cognitive load inventory during simulated handoffs: Evidence for validity. *SAGE Open Medicine, 4*, 1–7. https://doi.org/10.1177/2050312116682254

Young, J. Q., Irby, D. M., Barilla-LaBarca, M. L., ten Cate, O., & O'Sullivan, P. S. (2016). Measuring cognitive load: Mixed results from a handover simulation for medical students. *Perspectives on Medical Education, 5*, 24–32. https://doi.org/10.1007/s40037-015-0240-6

Young, J. Q., & Sewell, J. L. (2015). Applying cognitive load theory to medical education: construct and measurement challenges. *Perspectives on Medical Education, 4*(3), 107–109. https://doi.org/10.1007/s40037-015-0193-9

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

Zukić, M., Đapo, N., & Husremović, D. (2016). Construct and predictive validity of an instrument for measuring intrinsic, extraneous and germane cognitive load. *Universal Journal of Psychology, 4*, 242–248.https://doi.org/10.13189/ujp.2016.040505

Table 1
Means (Standard Deviations), Factors Loadings, and R^2 for Items Across Models

Model	Mean (Standard Deviation)	Unstandardized Factor Loading	Standard Error	Standardized Factor Loading	R^2	Structure Coefficients		
						IL	EL	GL
Nervous Anatomy								
Factor: IL								
Item 1	8.38 (1.74)	1[a]		.86	.73	.86	.03	.02
Item 2	7.73 (1.98)	1.24	.10	.93	.87	.93	.04	.02
Item 3	7.69 (1.92)	1.22	.10	.94	.89	.94	.04	.02
Factor: EL								
Item 4	4.88 (2.75)	1[a]		.86	.74	.03	.86	-.23
Item 5	4.00 (2.62)	1.02	.08	.93	.86	.04	.93	-.25
Item 6	4.48 (2.89)	1.12	.09	.92	.85	.04	.92	-.24
Factor: GL								
Item 7	6.98 (1.92)	1[a]		.88	.78	.02	-.23	.88
Item 8	7.09 (1.83)	1.02	.08	.94	.88	.02	-.25	.94
Item 9	7.19 (2.00)	0.95	.10	.81	.65	.02	-.21	.81
Item 10	8.38 (1.95)	0.98	.09	.85	.73	.02	-.23	.85
Locomotor Anatomy								
Factor: IL								
Item 1	7.80 (1.76)	1[a]		.84	.70	.84	.09	.00
Item 2	7.15 (1.92)	1.22	.10	.93	.87	.93	.10	.00
Item 3	7.13 (2.16)	1.35	.12	.92	.84	.92	.10	.00
Factor: EL								
Item 4	4.46 (2.59)	1[a]		.87	.75	.09	.87	-.25
Item 5	3.39 (2.47)	1.03	.09	.94	.88	.10	.94	-.27
Item 6	4.22 (2.87)	1.09	.10	.85	.73	.09	.85	-.24
Factor: GL								
Item 7	7.69 (1.68)	1[a]		.90	.81	.01	-.26	.90
Item 8	7.84 (1.83)	1.17	.08	.96	.92	.01	-.27	.96
Item 9	7.66 (1.97)	.73	.12	.56	.31	.01	-.16	.56
Item 10	8.15 (1.62)	.88	.08	.82	.67	.01	-.23	.82
e9 to e10		.37[b]						

[a] Factor loadings fixed to 1 (unit loading identification; Kline, 2011). To test these loadings for statistical significance, additional models were run in which these loadings were freed ,while loadings on subsequent items were fixed to 1; all factor loadings significant at $p < .001$; [b] correlation not factor loading, correlation sig at $p < .01$
IL = Intrinsic Load; EL = Extrinsic Load; GL = Germane Load

Table 2

Correlations between Subscales, McDonald's Omega (ω), and Average Variance Extracted (AVE) Values

	Correlations				
	IL	EL	GL	ω	AVE
Nervous Model					
IL	-	.04	.02	.94	.83
EL		-	–.27*	.93	.82
GL			-	.92	.76
Locomotor Anatomy					
IL	-	.11	.00	.93	.80
EL		-	–.29*	.92	.79
GL			-	.89	.68

*p < .05; IL = Intrinsic Load; EL = Extrinsic Load; GL = Germane Load

Supplement to

"Psychometric Properties of a Cognitive Load Measure When Assessing the Load Associated with a Course"

Missing Data

Missing data was minimal across the cognitive load items. No item was missing more than 3.1% of responses and most items (17 out of 20) were missing no responses or one response only. One participant completed the locomotor anatomy section of the survey, but not the nervous anatomy section (which was presented on a separate page). This participant aside, Little's test revealed the missing datapoints to be missing completely at random, $\chi^2(113) = 121.88$, $p = .268$. To maximize sample size, expectation-maximization was used to generate estimates for missing datapoints (excluding the participant who did not complete the nervous anatomy section of the questionnaire). Missing values were estimated following outlier Windsorization/deletion (see below).

Data Screening

Univariate outliers were assessed using the outlier labelling rule with a 2.2 multiplier (Hoaglin & Iglewicz, 1987). Across the dataset, 16 univariate-outlying datapoints were detected and Windsorized. Mahalanobis distances were used to identify multivariate outliers (using an α of .001). Two multivariate outliers were detected and deleted.

CFA was carried out using maximum likelihood estimation. Maxmimum likelihood estimation is inappropriate when data are severely non-normal. Fabrigar et al. (1999) recommend |skew| < 2 and |kurtosis| < 7. For the locomotor anatomy data, the skew and kurtosis of items was below these recommended thresholds (skew range = –1.32 to 0.56; kurtosis range = –0.80 to 2.00). This was also the case for the skew and kurtosis of the nervous anatomy items (skew range = –1.29 to 0.50; kurtosis range = –1.05 to 1.86).

Follow-Up EFA on Locomotor Anatomy Data

Following Fabrigar et al. (1999), several methods were used in tandem to determine the number of factors to extract: the eigenvalue-greater-than-one rule (also called the Kaiser criterion), inspection of the scree plot, and parallel analysis. Parallel analysis involves comparing the eigenvalues obtained from the sample data to eigenvalues generated from either random datasets or permutations of a raw sample dataset (O'Connor, 2000). O'Connor (2000) recommends comparing the sample eigenvalues to the parallel eigenvalues which correspond the desired percentile (e.g., 99th) of the distribution of random/permuted-raw eigenvalues. A factor is retained if the observed eigenvalue for that factor is greater than the corresponding n^{th} percentile parallel eigenvalue. This process was carried out using

O'Connor's "rawpar" script for SPSS (https://people.ok.ubc.ca/brioconn/nfactors/nfactors.html).

Using principal components extraction, three eigenvalues greater than one were identified (which together explained 82.49% of the variance in scores; see Table S1). The sample eigenvalues are plotted in Figure S2 along with the 99th percentile parallel eigenvalues. As can be seen, there is a precipitous drop going from the 3rd to 4th eigenvalue. Thus, the scree plot and the parallel analysis would also support a three-factor solution. Given the follow-up EFA results, the three-factor solution was retained, and it was concluded that Items 9 and 10 are likely not measuring an additional factor.

Calculation of McDonald's Omega (ω) and AVE values

ω values were generated using the Hayes and Coutts' (2020) OMEGA macro (version 0.2) for SPSS (available from http://afhayes.com/spss-sas-and-r-macros-and-code.html). Specifically, the EFA-ML method of calculating ω was utilized. AVE values were constructed by summing the squared factor loadings for each subscale and then dividing by the number of items making up that subscale (Hair et al., 2014, p. 619).

References

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*, 272–299. http://dx.doi.org/10.1037/1082-989X.4.3.272

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2014). *Multivariate data analysis* (7th ed.). Harlow, UK: Pearson.

Hayes, A. F., & Coutts, J. J. (2020). Use omega rather than Cronbach's alpha for estimating reliability. But…. *Communication Methods and Measures, 14*(1), 1-24. https://doi.org/10.1080/19312458.2020.1718629

Hoaglin, D. C., & Iglewicz, B. (1987). Fine tuning some resistant rules for outlier labelling. *Journal of American Statistical Association, 82*, 1147–1149. http://dx.doi.org/10.2307/2289392

O'connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instruments, & Computers*, *32*, 396–402. https://doi.org/10.3758/BF03200807

Table S1

Observed and Parallel Eigenvalues Based on the Locomotor Anatomy Data

Component	Observed Eigenvalues	% of Variance Explained	Cumulative % of Variance	99th Percentile Parallel Eigenvalues
1	3.60	36.01	36.01	1.78
2	2.69	26.89	62.90	1.54
3	1.96	19.59	82.49	1.36
4	0.62	6.09	88.57	1.23
5	0.29	2.86	91.44	1.13
6	0.25	2.54	93.98	1.03
7	0.20	1.95	95.93	0.93
8	0.16	1.62	97.55	0.83
9	0.13	1.25	98.80	0.76
10	0.12	1.20	100.00	0.66

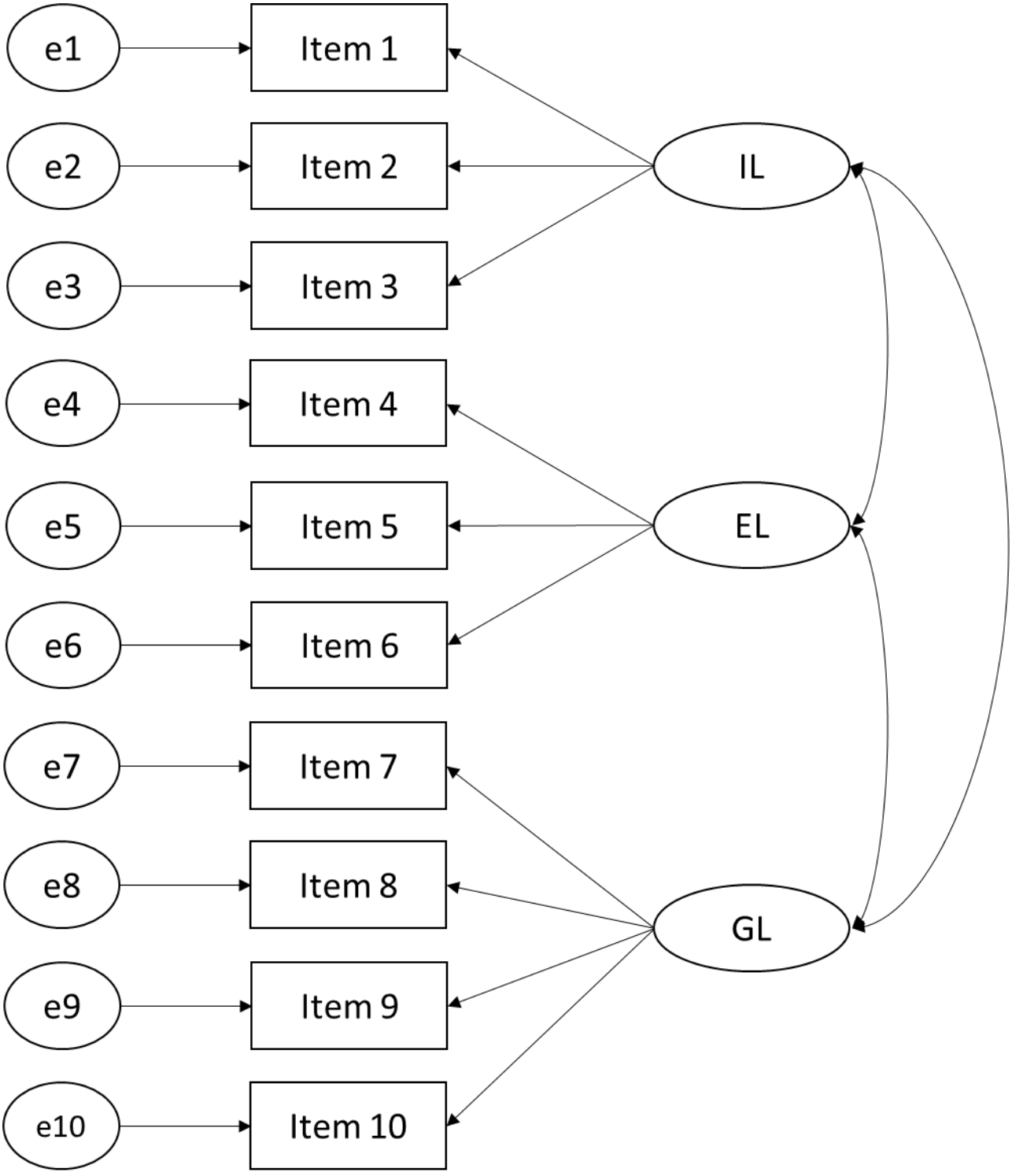

Figure S1. Structure of the models tested as part of the CFA.

Note. IL = Intrinsic Load; EL = Extrinsic Load; GL = Germane Load; e = Error

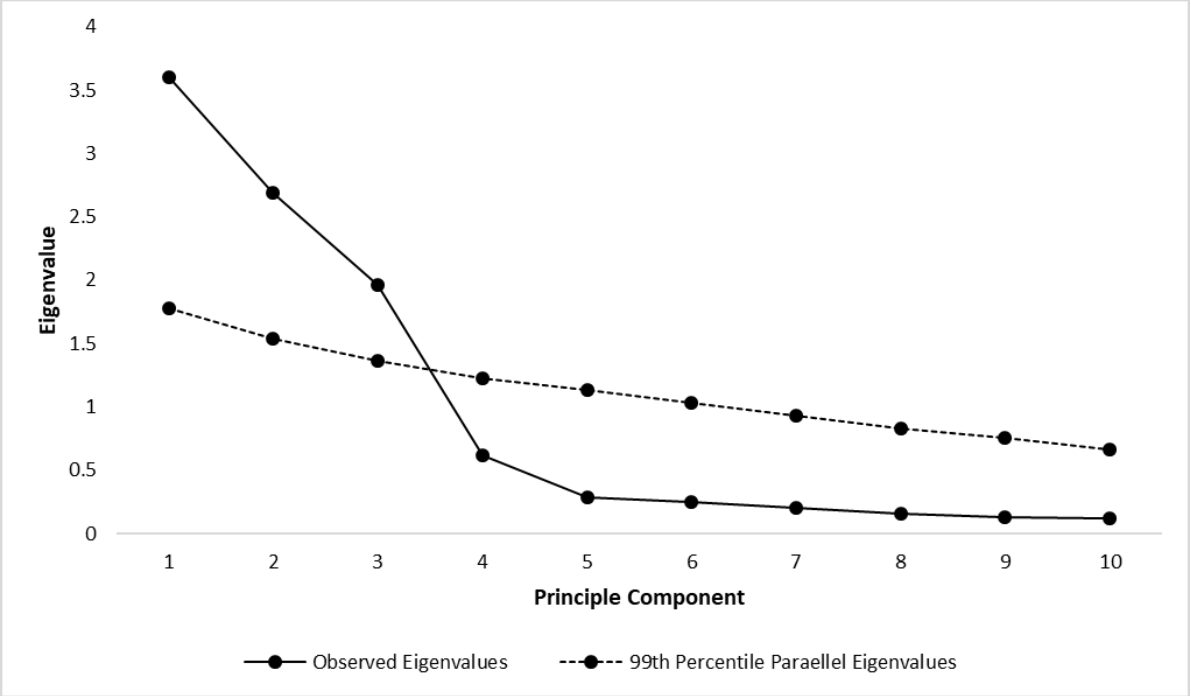

Figure S2. Scree plot with observed and 99th percentile parallel eigenvalues.