Zavgorodniiaia, Albina; Duran, Rodrigo; Hellas, Arto; Seppala, Otto; Sorva, Juha

Measuring the cognitive load of learning to program

Published in:
UKICER 2020 - Proceedings of the 2020 Conference on United Kingdom and Ireland Computing Education Research

DOI:
10.1145/3416465.3416468

Published: 03/09/2020

Please cite the original version:

This material is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of the repository collections is not permitted, except that material may be duplicated by you for your research use or educational purposes in electronic or print form. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone who is not an authorised user.
Measuring the Cognitive Load of Learning to Program: A Replication Study

Albina Zavgorodniaia, Rodrigo Duran, Arto Hellas, Otto Seppälä, and Juha Sorva
{albina.zavgorodniaia,rodrigo.duran,arto.hellas,otto.seppala,juha.sorva}@aalto.fi
Department of Computer Science, Aalto University, Finland

ABSTRACT
Cognitive load (CL) on a learner’s working memory has emerged as an influential concept in computing education and beyond. CL is commonly divided into at least two components, intrinsic load (IL) and extraneous load (EL). The two are cut from the same cloth: they both stem from contextual factors. One of the major open questions in CLT is how to measure cognitive load; work is also needed on adapting generic cognitive-load measures to the programming domain. We investigate these questions by replicating two studies. We will state our specific research questions in Section 3 after providing some background for them.

Cognitive load theory (CLT) [63, 65] has established itself as one of the leading theoretical frameworks to support instructional design.

2 RELATED WORK

2.1 Basic Concepts of Cognitive Load Theory
Cognitive load theory rests on the widely accepted model of human cognitive architecture whose central components are working memory and long-term memory [63]. Working memory (WM) is extremely limited: it can hold only a handful of elements at a time, for a short time. This is a bottleneck for learning, as all new information must be first consciously processed in WM. On the other hand, long-term memory is virtually unlimited in capacity and stores domain-specific knowledge organized in schemas. Once learned, even a complex schema can be treated as a single element in WM, which is why people can deal with complex situations in familiar domains. For CLT, learning essentially means schema construction.

Cognitive load (CL) refers to the demands imposed by a learning situation—i.e., materials and activities—on a learner’s working memory; this is the intensity of cognitive activity required for a specific learning goal during a narrow time frame [27]. To estimate that intensity, CLT uses the concept of element interactivity, which refers to how many elements the situation requires the learner to simultaneously and consciously hold in working memory; this is the intensity of cognitive activity required for a specific learning goal during a narrow time frame [27]. To estimate that intensity, CLT uses the concept of element interactivity, which refers to how many elements the situation requires the learner to simultaneously and consciously hold in WM. Too much element interactivity means cognitive overload and unsuccessful learning.

CLT identifies two types of load: intrinsic load (IL) and extraneous load (EL). The two are cut from the same cloth: they both stem from contextual factors.
from element interactivity and load WM similarly. The difference is analytic: IL is the load that is unavoidable when aiming for a specific learning objective with a certain level of prior knowledge, and EL is the avoidable load that results from suboptimal materials and activities. EL and IL add up to overall CL. [25, 62, 63]

2.2 Germane Load, Old CLT, and New CLT

The term “germane load” (GL) is associated with the idea that learning does not happen without WM activity; it refers to this positive aspect of cognitive load. However, the more precise definition and role of GL in CLT has shifted during the past decades. The ripples of that change are still evident in the current literature, which draws on various versions of CLT.

The 1998 formulation of the theory [65], which we will call “Old CLT,” defined GL as an independent, comparable source of load alongside IL and EL. In this view, GL is a third, additive component representing load that makes learning happen and should be maximized; IL and EL should be low enough for GL to also “fit in” working memory. Some scholars stressed that GL is not just any useful load but arises from those additional, effortful aspects of learning that go beyond task performance, such as conscious reflection and self-explanation [25, 57]. When interventions based on Old CLT target improvements to motivation and deep engagement with learning materials, they are often phrased as increasing GL.

Old CLT was found to be problematic. Leading figures in CLT complained that it suffered from conceptual flaws, failed to explain various empirical results, used a GL construct that was unnecessary and unjustified by results, and led to unfalsifiable circular interpretations of findings [25, 57, 62]. Some who developed measurement instruments for Old CLT [33] eventually gave up on the GL part of the theory [5, 9, 34]. In 2010, CLT was reformulated.

“New CLT” [25, 62] is nearer and narrower. The major change is the removal of GL as a third, additive component of CL. All learning— even that resulting from “additional” reflection—results from dealing with IL. New CLT tends to view motivation as something that is relevant to learning but external to the theory; motivated engagement is not an additional load on a learner’s mind, but a part of dealing with IL. In New CLT, the term “germane load” has been (perhaps confusingly) repurposed to mean something conceptually quite distinct from IL, EL, and Old CLT’s GL. It refers to the actual use of WM resources by a learner to process IL, which happens during learning assuming there is sufficient motivation. Overall, GL tends to play a reduced role in studies that are based on New CLT.

The interpretation of the GL construct is important (to the present work and otherwise) not least because GL is central to debates around empirical measurement of CL, as discussed below.

2.3 Measuring Cognitive Load

Methods of CL measurement are commonly grouped into objective and subjective ones. We will briefly comment on objective measures but focus on those subjective measures that relate directly to our work. For broader and deeper reviews, see, e.g., [31, 63, 69].

2.3.1 Objective measures of CL. So-called objective CL measures do not require learners to reflect on and evaluate their mental effort or learning. Many of these measures are based on physiological data sources, such as such as eye-tracking [21], pupillometry [16, 45], heart rate [54], and neuronal [3, 67] and electrodermal [52, 58] activity. Objective measures also include dual-task experiments where some learners are required to perform a secondary task during learning so that cognitive load may be estimated from the differences in resulting performance (e.g. [6]).

Objective methods do not depend on learners’ self-assessment abilities. Another benefit is that many of them can be applied “online”—i.e., during learning—as an index to real-time changes in load. On the downside, many objective methods are intrusive and distracting. Moreover, most objective measures cannot differentiate between the different types of load, which lofty goal has been suggested as CLT’s “holy grail” [4] and “mission impossible” [29]. Differential measurement of loads would enable more detailed predictions and add incisiveness to interpretations, and it might be useful to practitioners as well as researchers. Although some attempts have been made to differentiate loads using objective methods [4], most attempts to date rely on subjective ratings instead.

2.3.2 Subjective measures of CL. Subjective CL measures ask learners to report on their perceived mental effort or other aspects of a learning episode that are expected to indicate cognitive load. Paas’s [53] unidimensional nine-point scale for mental effort from 1992 is the time-tested instrument that continues to be used. However, this simple scale cannot differentiate between load types and suffers from noisy measurements [32].

Several groups of researchers have created, adopted, and adapted various questionnaires for subjectively assessing types of CL separately. We will comment on a few.

In 2013, Leppink et al. [33] developed a CL self-assessment questionnaire (quoted in our web appendix [1] for easy reference), which has been since adapted and used by the same group and various others. It asks learners to indicate the level of their agreement with ten claims such as “The activity covered formulas that I perceived as very complex” on an eleven-point scale; IL, EL, and GL are targeted by three, three, and four items, respectively. This initial study took place in a statistics class, but Leppink et al. suggested that the instrument could be readily adapted to other complex knowledge domains by replacing words such as “statistics” and “formula” appropriately. Through factor analysis, Leppink et al. [33] uncovered a three-factor model that fit the data, with factors matching the IL-, EL-, and GL-targeting items, respectively. However, the study failed to support the hypothesis that high scores on the GL-targeting items correlated with better learning, and similar results from a follow-on study [34] led the authors to suggest that the third factor does not represent germane load and to throw their support behind New CLT. In later work, some of the same authors have discarded the name “germane load” in favor of “self-perceived learning” [5] or modified the instrument to focus on IL and EL only [9].

Drawing on Old CLT, Klepsch et al. set out to develop a CL component questionnaire that would be even easier to transfer to different domains than Leppink’s. Klepsch et al. also phrased (especially) the GL-targeting items differently, seeking to probe whether students had attempted to understand the lesson completely and holistically (as shown in our web appendix [1]). Much like Leppink et al. [33, 34], Klepsch et al. found that a three-factor model fit their data but the GL-targeting items did not correlate with learning.
A somewhat different questionnaire was developed by Yeung et al. [68], who identified a four-factor model to predict cognitive load, with factors representing learners’ perceptions of difficulty, incompetence, negative affect, and investment of effort.

A recent report by Jiang and Kalyuga [24] sought evidence for New CLT in the language-learning domain. The authors fitted a two-factor model against data collected using a questionnaire similar to that of Leppink et al. [33] (see [1]). They hypothesized that since New CLT’s GL represents the allocation of working-memory resources on IL, the IL-targeting and GL-targeting questionnaire items might capture the same construct from different perspectives.

A confirmatory factor analysis bore this out: a two-factor model fit vanic skin response and eye blinks. Duran et al. [14] have proposed from two earlier publications. We ask:

RQ1 Morrison et al. [49] report good discriminant validity and internal reliability for a programming-domain adaptation of Leppink’s [33] CL questionnaire, which is based on the three-component Old CLT. Do these findings replicate with a different cohort and a different programming tutorial?

RQ2 In a language-learning setting, Jiang and Kalyuga [24] used a questionnaire similar to Leppink’s [33] and found support for New CLT’s two-component model of CL, with GL redundant. Do the findings replicate in the programming domain?

Exploring the robustness of those earlier findings is an incremental step towards solving the broader issues in CL measurement in CER and beyond. Answering these questions will not—and we are not attempting here to—show which version of CLT is “correct.” The reader will note that our questions do not directly address construct validity (i.e., whether the instruments actually measure what they are meant to), which is a topic that we will return to in Discussion.

4 METHODS

4.1 Procedure

Participants were recruited and paid using Amazon MTurk¹, a platform for crowd-sourcing work that requires human intelligence. Although already established as a powerful resource for research, crowd-sourcing platforms suffer from variable quality of data [7, 23]. To alleviate this issue, we only accepted participants with 500 approved tasks on MTurk and a task-approval rate of at least 98%.

Participation consisted of three stages: (1) a demographic survey, (2) an instructional video on programming, and (3) a post-instructional survey. The demographic survey asked about the participants’ previous programming experience among various other background questions and was designed to avoid hinting at the subsequent programming-related instruction. The instructional video is described below. The post-instructional survey contained a cognitive-load questionnaire (see below) and a post-test on programming that is outside the scope of this article.

4.1.1 Instructional Video

The 24-minute instructional video was a beginner-level lesson to reading Python code, in English. The video introduced the concepts of variable, expression, and value and taught the participants to reason about short fragments of imperative code consisting of assignment statements and print commands. The video covered several short example programs, explaining the code constructs and tracing each programs’ behavior in detail. The video was configured so that once the participant clicked Play, it could not be paused, rewound, or forwarded.

Because we collected this data in the context of an experiment on presentation modalities in instructional video (not reported here), the participants were randomly given one of three variants of the same video. The variants differed in whether the program examples were accompanied by text, audio, or both. As we observed no significant differences in CL scores between these groups, we are combining the groups for the analysis presented herein.

4.1.2 Cognitive-Load Questionnaire

To measure cognitive load, we used the instrument previously adapted for programming by Morrison et al. [49] from Leppink et al.’s [33] original. Shown in Figure 1, Items 1 to 3 target IL; items 4 to 6 EL, and items 7 to 10 GL. Below, we will refer to the items as I1–I3, E1–E3, and G1–G4, respectively. A minor difference between our instrument and Morrison’s is that we used a ten-point scale while they used an eleven-point one.

1https://www.mturk.com/
participants who stated that they had no or little previous programming experience. The rest were excluded for a variety of reasons, such as log data suggesting they had “cheated” during the experiment, and missing responses.

Of the 175 participants, 78 self-identified as men and 97 as women, with no-one picking the other options. Most (148) were from the US, fourteen were from India, and the remainder from various countries. English was 158 participants’ native language; the rest spoke a mix of languages. Seven of the participants were between 18–24 years of age, 62 were between 25–34, 49 between 35–44, 37 between 45–54, and 20 were at least 55 years old. 108 of the participants held a bachelor’s degree, 21 a master’s, one a doctoral degree, and 45 were school graduates. 118 participants reported having some background in the humanities, 41 in natural and technical sciences, and 16 reported no education beyond primary school. 123 had no programming experience at all; 52 reported “a little.”

4.3 Statistical Analyses
Morrison et al. [49] found a three-factor interpretation for their data, with items targeting IL, EL, and GL forming the three factors, respectively. To replicate that analysis, we grouped the questionnaire items so that I1–I3, E1–E3, and G1–G4 formed the factors. We computed Cronbach’s alpha to assess the internal reliability of each of these factors separately.

On the other hand, Jiang and Kalyuga [24] found a two-factor interpretation with GL-targeting and IL-targeting items aligned. To replicate that analysis, we grouped the questionnaire items so that E1–E3 again formed one factor and the other seven items (I1–I3 and G1–G4) together formed another.

To investigate whether the three-factor or two-factor model would best fit our data, we conducted two confirmatory factor analyses (CFA) using the two hypothetical models above. Given that our data are ordinal and violate normality assumptions, the DWLS (diagonally weighted least squares) estimator would usually be appropriate. However, Li [35] suggests that with a small sample size ($N < 200$), the ML (maximum likelihood) estimator is more trustworthy.

Following Jiang and Kalyuga [24], we reversed the GL-targeting items’ scores for the CFA and item correlations. That is, a score of 10 is treated as a 1 and other scores are similarly converted.

We used the package lavaan in R (version 4.0).

5 RESULTS
5.1 Reliability
Cronbach’s alpha for all the items was 0.87, compared to an acceptance threshold of 0.70. The alpha for the IL-targeting items alone was 0.96; for the EL-targeting items it was 0.89; and for the GL-targeting items it was 0.97. The IL-targeting and GL-targeting questions together had an alpha of 0.83. The squared multiple correlations ($R^2$) of each item varied between 0.61 and 0.95, thus being above the 0.25 threshold of item reliability.

These results suggest that the instrument and each of the proposed factors is internally reliable.

5.2 Factor Analyses
Table 1 shows the fit statistics for the three- and two-factor models.

![Table 1: Fit indicators for the hypothesized factor models.](image)

The Comparative Fit Index (CFI) measures whether the model fits the data better than a more restricted baseline model; the Tucker-Lewis Index (TLI) is a more conservative estimate of fit. CFI and TLI values above 0.95 are considered a good fit [22]. As Table 1 shows, both CFI and TLI can be considered good for the three-factor model, whereas the two-factor model fits poorly.

The Akaike Information Criterion (AIC) and the related Bayesian Information Criterion (BIC) estimate information loss in the model. In general, lower values of AIC and BIC are better fit estimates. Table 1 shows a lower AIC and BIC for the three-factor model, whereas the two-factor model fits poorly.

RMSEA (The Root Mean Square Error of Approximation) summarizes how closely the model reproduces data patterns. Models with RMSEA values of around 0.07 can be considered a good fit, with lower values preferred. RMSEA confirms the fit of the three-factor model but not the two-factor one.

The Standardized Root Mean Square Residual (SRMR) is defined as the standardized difference between observed and predicted correlations. SRMR values under 0.08 are considered a good fit [22]; again, only the three-factor model meets this criterion.

![Figure 1: The CL self-assessment (based on [33, 49]).](image)
Table 2: Factor loadings and descriptive statistics for each questionnaire item.

<table>
<thead>
<tr>
<th>Latent Factor</th>
<th>Item</th>
<th>Mean</th>
<th>Std.Err.</th>
<th>Loading</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intrinsic</td>
<td>I1</td>
<td>4.03</td>
<td>0.122</td>
<td>0.887</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>“Intrinsic”</td>
<td>I2</td>
<td>4.07</td>
<td>0.124</td>
<td>0.973</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>“Intrinsic”</td>
<td>I3</td>
<td>4.09</td>
<td>0.124</td>
<td>0.982</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>“Extraneous”</td>
<td>E1</td>
<td>1.76</td>
<td>0.090</td>
<td>0.804</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>“Extraneous”</td>
<td>E2</td>
<td>1.71</td>
<td>0.082</td>
<td>0.896</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>“Extraneous”</td>
<td>E3</td>
<td>1.71</td>
<td>0.086</td>
<td>0.866</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>“Germane”</td>
<td>G1</td>
<td>6.65*</td>
<td>0.144</td>
<td>0.907</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>“Germane”</td>
<td>G2</td>
<td>6.59*</td>
<td>0.135</td>
<td>0.967</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>“Germane”</td>
<td>G3</td>
<td>6.82*</td>
<td>0.137</td>
<td>0.986</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>“Germane”</td>
<td>G4</td>
<td>6.75*</td>
<td>0.134</td>
<td>0.966</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>

Overall, the three-factor model fits our data considerably better than the two-factor model. An ANOVA test confirmed that the three-factor model is a better fit: $\chi^2(2) = 1175.3, p < 0.001$. Table 2 presents the latent factors and factor loadings of the three-factor model. The table also shows the items’ means and standard errors.

There are significant, positive correlations between the factors targeting IL and EL (0.405, $p < 0.001$), and between the factors targeting EL and (reversed) GL (0.447, $p < 0.001$), respectively. We found no significant correlation between the IL- and GL-targeting factors (0.053, $p = 0.49$). Figure 2 illustrates the relationships between all the questionnaire items.

Both Figure 2 and Table 2 suggest a floor effect for E1–E3. Note that GL-targeting scores are not reversed.

6 DISCUSSION AND LIMITATIONS

6.1 RQ1: Measuring CL in Programming

In RQ1, we asked whether Morrison and colleagues’ [49] results concerning their CL questionnaire for programming replicate with a different programming tutorial and a different cohort. We found that the questions targeting each CL type aligned with the other questions targeting that type, which supports the claim that the questionnaire has internal reliability. Moreover, we confirm strong support for the questionnaire’s discriminant validity—that is, the IL-targeting, EL-targeting, and GL-targeting items appear to be measuring three distinct constructs.

By extension, our findings also strengthen the support for the internal reliability and discriminant validity of Leppink’s [33] instrument from which Morrison et al. [49] derived theirs.

6.2 RQ2: The Two-Factor Model for New CLT

In RQ2, we asked whether the two-factor model that fit Jiang and Kalyuga’s [24] language-learning data well replicates in our programming data. We found a poor fit for the two-factor model that combined IL- and GL-targeting items into a single factor, the other factor consisting of the EL-targeting items. This contrasts with the excellent fit of the three-factor model to our data, with the GL-targeting items in a separate factor. Although the IL- and GL-targeting items aligned in Jiang and Kalyuga’s context, they did not in ours.

We feel that the theoretical arguments for New CLT are compelling and that the theory is consistent and parsimonious. However, GL-targeting self-assessment questions do not appear to be redundant with IL across contexts, so their value as empirical evidence for New CLT may be limited. It seems likely that context-dependent and instrument-dependent factors confound the relationship of the IL- and GL-targeting items.

6.3 Construct Validity in CL Measurement

Despite the popularity of self-assessments as a way to measure CL types, their construct validity remains largely unproven [32]. There are question marks over whether learners are able to differentiate IL and EL [44], and the self-assessment of GL is more problematic still. The thread of research to which we add ours has not produced a convincing case for the construct validity of GL-targeting self-assessments. Our study does not resolve this issue either.

What research has shown is that several instruments’ GL-targeting items do measure a third construct that is distinct from EL and IL and that this construct replicates in the programming domain. At face value, the third construct (or third constructs; there are differences between instruments [1]) appears to be related to phenomena such as effective learning, motivation, engagement, reflection, efficiency, intensity of cognition, and/or satisfaction with teaching methods.

Many authors (e.g. [29, 33, 34]) have noted that learners can interpret CL assessments differently from what was intended; subtle changes to wording may be significant. This is a concern in our study as well, and a possible partial explanation for the discrepancy between our findings and those of Jiang and Kalyuga [24], whose questionnaire was similar but not identical to ours [1]. There may also be differences between second-language learning and programming learning—or learners’ perceptions of those domains—that contribute to that discrepancy.

3In the general case, since IL and EL are analytically separated by which aspects of instruction are suboptimal, learners would be required to know instructional design. (And in fact one study did teach learners CLT [29].) Intuition suggests that some forms of EL are much easier for learners to discern from IL than others.
6.4 Motivation and Cognitive Load
Motivation and other learner-dependent factors are another caveat of the present study. We did not assess the participants’ motivation, and our varied cohort of crowd-sourced learners means that motivations, approaches to learning [41], and self-assessment skills may have varied considerably among the participants.

In 2010, Moreno [47] wrote: “CLT is remarkably silent about the relation among load, affect, and motivation.” New CLT sidesteps this critique: it is interested in how instructional design affects “ideal” learners that are highly engaged; it assumes motivation [25, 62]. This is a strength and a weakness. Cording off motivation makes New CLT sharper and more robust but also means that the theory must be complemented by other theories and instruments in order to account for real-world situations where students lack motivation.

The further research that is needed on the relationship between CLT and motivation has already started outside of CER (e.g., [10, 17, 36]).

6.5 Rapid Changes in Cognitive Load
An inherent limitation of our study is that we measured CL only after the participants completed a sizeable learning activity. Working memory operates on a timescale of seconds, so many moment-to-moment changes in load are to be expected during a 24-minute video such as ours or the even longer lectures studied by others. Even though subjective post-instruction ratings might provide a general indication of CL levels during instruction, their validity is nonetheless compromised by the extended time frame [25, 26].

Research in CLT has begun to look into “real-time” measurement of CL types [26, 31]; CER should follow suit. In addition to the rapid changes in cognitive load, future research should consider the (comparatively slow but nevertheless fairly frequent) fluctuations in motivation during episodes of learning [13].

6.6 Summary of Recommendations
Given our results, researchers in CER and practitioners in programming education may use Morrison’s [49] questionnaire with increased confidence in its reliability. That being said, we advise caution in interpreting its GL-targeting scores especially, as their construct validity and theoretical foundation are debatable.

We recommend that researchers in CLT and CER continue to seek empirical evidence for New CLT and instruments to measure CL types. Our study suggests that Jiang and Kalyuga’s [24] findings do not necessarily replicate in other contexts. However, even if IL-targeting and GL-targeting subjective assessment items do not combine into a single factor, that does not imply an additive GL as per Old CLT; it merely implies that a third construct is being measured by the instrument.

We agree with authors such as Klepsch et al. [29, 30] that measuring a third, engagement-related construct is meaningful. That construct may not be a discrete source of cognitive load, however. Even if motivation and engagement are external to CLT—as New CLT suggests—research is needed that combines assessments of motivation (e.g. [55]) with measures of cognitive load.

To improve programming education, instructional designs must be evaluated to see how they motivate students and which combinations of CL-driven designs and other activities best facilitate complex learning. CER, as a field, might learn from current research in educational psychology [10, 17, 18, 27, 36], which explores the interplay of CLT-based tasks—whose goal is schema construction—with other learning tasks whose goals are different, such as activating learners’ intuitive knowledge, promoting deep approaches to learning, or motivating learning by raising learners’ awareness of their knowledge gaps. Productive failure [28] is one example of a framework that might be so used as a complement to CLT (even though there are challenges in reconciling these perspectives [64]).

We join the call [29, 33, 34] for using a combination of qualitative and quantitative methods to study wording effects in CL questionnaires. We also recommend that CER explore the temporal variation in cognitive load and motivation during learning.

Some of the current CLT-based research in CER cites Old CLT and its constructs seemingly unaware of advances in CLT during the last decade. We hope to raise awareness of some of those advances and encourage the development of ever better cognitive-load measures for programming.

REFERENCES
[18] Inga Glogger-Frey, Katharina Gaus, and Alexander Renkl. 2016. Learning from direct instruction: Best prepared by several self-regulated or guided invention activities? Learning and Instruction 51, 26–35.