Cognitive Load Theory in Computing Education Research: A Review

RODRIGO DURAN, Federal Institute of Mato Grosso do Sul, Brazil
ALBINA ZAVGORODNIAIA and JUHA SORVA, Department of Computer Science, Aalto University, Finland

40

One of the most commonly cited theories in computing education research is cognitive load theory (CLT), which explains how learning is affected by the bottleneck of human working memory and how teaching may work around that limitation. The theory has evolved over a number of decades, addressing shortcomings in earlier versions; other issues remain and are being debated by the CLT community. We conduct a systematic mapping review of how CLT has been used across a number of leading computing education research (CER) forums since 2010. We find that the most common reason to cite CLT is to mention it briefly as a design influence; authors predominantly cite old versions of the theory; hypotheses phrased in terms of cognitive load components are rare; and only a small selection of cognitive load measures have been applied, sparsely. Overall, the theory's evolution and recent themes in CLT appear to have had limited impact on CER so far. We recommend that studies in CER explain which version of the theory they use and why; clearly distinguish between load components (e.g., intrinsic and extraneous load); phrase hypotheses in terms of load components *a priori*; look further into validating different measures of cognitive load; accompany cognitive load measures with complementary constructs, such as motivation; and explore themes such as collaborative CLT and individual differences in working-memory capacity.

CCS Concepts: • **Social and professional topics → Computing education;**

Additional Key Words and Phrases: Cognitive load theory, computing education, literature review

ACM Reference format:
Rodrigo Duran, Albina Zavgorodniaia, and Juha Sorva. 2022. Cognitive Load Theory in Computing Education Research: A Review. *ACM Trans. Comput. Educ.* 22, 4, Article 40 (September 2022), 27 pages.
https://doi.org/10.1145/3483843

1 INTRODUCTION

Cognitive load theory (**CLT**) [100, 102, 103] is a theoretical framework that informs instructional design by explaining how learning is affected by the limitations of the human cognitive

This work was partially funded by the Conselho Nacional de Desenvolvimento Científico e Tecnológico Grant no. 201365/2015-4.
Authors' addresses: R. Duran, Rodovia MS–473, km 23, s/n°, Fazenda Santa Bárbara, CEP 79750-000 | Nova Andradina, MS; email: rodrigo.duran@ifms.edu.br; A. Zavgorodniaia and J. Sorva, Department of Computer Science, Konemiehentie 2, Espoo, Finland; emails: {albina.zavgorodniaia, juha.sorva}@aalto.fi.

1946-6226/2022/09-ART40 $15.00
https://doi.org/10.1145/3483843

architecture. Fundamental to CLT is the relationship between working memory, long-term memory, and their distinct characteristics; *cognitive load* refers to the demands that an activity places on working memory.

Cognitive load research has generated many pedagogical recommendations [103], which are often discussed in terms of "effects." CLT originated in the 1980s from studies of problem-solving that demonstrated the superiority of initially studying examples rather than attempting to solve problems [99]—this is the *worked-example effect*. Worked examples, which present a solution for the learner to study, remain the quintessential CLT-based learning activity. Other cognitive load effects include, for example, the *split-attention effect*, which recommends that pieces of related information are presented in integrated format so that the learner's working memory is not burdened by having to connect the information; the *redundancy effect*, which recommends that the same information is not presented in multiple self-contained formats; the *isolated-elements effect*, which recommends that the pieces of very complex information are first learned in isolation before integrating them; and the *self-explanation effect*, which recommends prompting students to generate explanations (to themselves) of how they reason about concepts and procedures.

So-called "compound effects" influence the other effects [103]. These compound effects include the *element interactivity effect*, according to which the effects only apply where sufficiently complex information is involved. The related *expertise reversal effect* explains why the effects vanish or even reverse for learners with prior knowledge of the topic; for instance, high-prior-knowledge learners may benefit more from problem-solving practice than from worked examples.

CLT is widely cited as an influence on instructional design as well as being the focus of many research studies. This influence extends to **computing education research** (**CER**): in a 2014 review of theory use in CER, Malmi et al. [55] noted that CLT is among the theories most frequently cited. In CER, CLT has influenced example-based learning [58, 93], course design [8, 97, 107], the design of learning materials and visualizations [6, 65, 68, 116], practice tasks such as Parsons problems [20, 28], and complexity analysis of program code [18], among other things.

However, like many other influential theories with a history, CLT is neither unitary nor unchanging. There are two main variants currently in use. One is championed by CLT's creators as the simplest and most current version of the theory; another builds on the concept of germane cognitive load that was introduced to CLT in 1998 but has since fallen out of favor with CLT's creators [100, 103]. Different authors in CER and elsewhere continue to cite and use different versions, sometimes deliberately and sometimes apparently by accident or default. At the same time, the limitations of CLT and criticisms of CLR continue to be debated. Some of these criticisms were highlighted in CER by Nelson and Ko [71], who argued that computing education researchers have typically failed to acknowledge the theory's limitations.

Inspired by Nelson and Ko's [71] commentary, our own informal observations of CER articles, and ongoing disagreements in the CLT community (see, e.g., [49]), we set out to review how different CLT variants have been used in CER during the past decade and how recent themes in CLT are reflected in CER.

We conduct a systematic mapping review of CLT in four journals and six conferences that publish CER. This produces a picture of how CLT is cited in CER, which aspects of articles CLT is influencing, which versions of the theory have been used, and which instruments have been used to measure cognitive load. As we conducted our review, we noted some issues in how CER has applied CLT and reported on it; we comment on these concerns and make recommendations about how authors might avoid them in the future. A secondary contribution of this article is to bring CLT's evolution and recent developments to the attention of the CER community.

Section 2 below introduces CLT's main concepts, identifies the theory's two main variants, and points at a number of recent themes and open debates in CLT.[1] Section 3 details the research questions and methods for our mapping review. Section 4 presents the main results of the review; we then comment on those results in Section 5 and accompany them with further observations and recommendations. Section 6 briefly wraps up the article.

2 BACKGROUND: CLT

2.1 Basic Concepts

CLT views learning as the construction of knowledge in *long-term memory*, which is virtually unlimited in capacity. For that construction, information must first be processed in *working memory*, which is very limited in both capacity and duration. Although people's working-memory capacity varies somewhat and can be improved through training [46, 87], it remains the case that any person's working memory can only hold a small number of elements at a time, for a short time. Overcoming this bottleneck through instructional design for effective and efficient learning is CLT's central concern [100].[2]

2.1.1 Elements and Schemas. The key CLT concept of *element interactivity* refers to how many elements a situation requires the learner to process simultaneously in working memory. This does not simply mean the total number of elements to be considered during a lesson: element interactivity reflects the degree that the elements are interconnected in the sense that they must be compared, contrasted, integrated, or otherwise consciously processed together. *Cognitive load* is the intensity of cognitive activity required for a specific learning goal during a narrow time frame [38]; it arises directly from element interactivity. Too much element interactivity means cognitive overload and unsuccessful learning.

Long-term memory stores domain-specific knowledge organized in *schemas*. Once learned, even a complex schema can be retrieved and treated as a single element in working memory, which is why people can deal with complex situations in familiar domains. Schemas also explain the expertise reversal effect [35]: learners with prior knowledge may treat complex information as a single chunk, which reduces element interactivity and may drastically influence which instructional methods are efficient. Cognitive load is thus not purely a feature of the learning activity but also depends crucially on the learner.

2.1.2 Intrinsic and Extraneous Load. There are two types of cognitive load whose existence is unanimously accepted by cognitive load theorists: *intrinsic load* and *extraneous load*. This is an analytical difference. Intrinsic load "needs to be there" according to a CLT-aware observer, whereas it is "safe" and often helpful to reduce extraneous load.

Sweller [100] defines intrinsic cognitive load as "the natural complexity of information that must be understood and material that must be learned, unencumbered by instructional issues such as how the information should be presented." That is, intrinsic load is the component of cognitive load that is unavoidable when aiming for a particular learning objective with a certain level of prior knowledge. Prior knowledge reduces intrinsic load as it enables learners to treat larger amounts of knowledge as a a single element, or even to deal with very familiar information automatically, without burdening working memory. This feature of the human cognitive architecture explains CLT's isolated-elements effect (see above).

[1] N.B. Some passages of text in Section 2 have been replicated from an earlier conference publication [115].
[2] Cognitive load theory is closely related to Mayer's Cognitive Theory of Multimedia Learning [61] and shares many assumptions, concepts, and recommendations with it. We will not debate the two theories' relationship further here, but note briefly that some of the work in CER that discusses "cognitive load" cites Mayer's theory rather than CLT.

Extraneous load, in contrast, involves cognitive processing that is not necessary for learning, at least not necessary *per se*. It results from instructional materials and activities [100]. "Unnecessary content in learning materials, content that is not easy to access while solving a problem but needs to be kept in mind, disturbing background noise, redundant instructions that need to be scanned for new content, instruction that introduces unnecessarily many new topics at once, and verbose lists of examples of factors causing extraneous cognitive load which appear at the beginnings of sentences that fail to start by explaining what it is that they list, are examples of factors causing extraneous cognitive load" [96].

Various cognitive load effects can be explained in terms of extraneous load. For example, materials that require the learner to integrate multiple sources of information cause a split-attention effect, since dealing with the sources in working memory constitutes an extraneous load. Similarly, materials that repeat the same information in multiple formats cause a redundancy effect, as extraneous processing is needed to compare the information feeds.

Consider, too, the worked-example effect. Learners who solve a problem and learners who study an example solution both need the intrinsic load of contemplating each problem state and how to proceed from there. The difference is that the problem-solving learner needs further to consider a wide range of potential "moves" across the problem space, so problem-solving involves a greater extraneous load than studying examples does [100]. Nevertheless, worked examples also bring *some* extraneous load, as the examples themselves are an instructional artifact distinct from the solution strategies that are being learned.

Where intrinsic load is high enough due to complex enough content and low enough prior knowledge, extraneous load should be reduced. On the other hand, where cognitive overload is not a concern, there is no need to reduce extraneous load. In some cases, it is justifiable to increase load (e.g., to introduce interesting but potentially distracting details for motivational purposes [94]).

2.2 The Evolution of Cognitive Load Theory

As there is no established nomenclature for the variants of cognitive load theory, we will use our own. Henceforth, *New CLT* refers to the variant that is currently promoted by Sweller and others as being the most up-to-date. *Old CLT* is the variant that was similarly promoted between roughly 1998 and 2010, which continues to be preferred by some scholars. (Pre-1998 versions of the theory are not relevant for present purposes.[3])

2.2.1 Old CLT. As discussed above, some cognitive load arises from the intrinsic complexity of the learning goal, and another source is the extraneous complexity of instruction. According to Old CLT [102], there is a third source of cognitive load alongside those two: *germane load* arises from engagement with the learning activity. Germane load goes beyond simply coping with intrinsic load; it is needed in order to construct schemas and make learning happen. Together with intrinsic and extraneous load, germane load adds up to overall cognitive load, as shown in Figure 1.

Germane load is "good" cognitive load: the more the better. If there's too much intrinsic and/or extraneous load, there's no "room" for germane load and no learning will occur. On the other hand, even if intrinsic and extraneous loads are low enough, the learning activity may fail to engage the learner to process the information sufficiently; in terms of Old CLT, such an activity has insufficient germane load. When interventions based on Old CLT target improvements to deep engagement with learning materials, they are often said to increase germane load. The same applies to cognitive load effects such as the self-explanation effect, which is interpreted in Old CLT as

[3]In some respects, what we call New CLT is actually a back-to-basics return to a simpler pre-1998 theory; for a history, see [103].

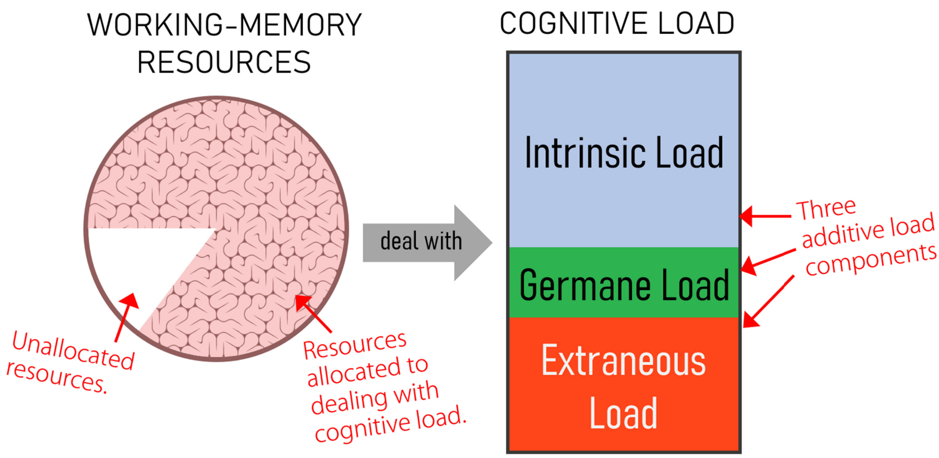

Fig. 1. Old CLT, as interpreted by the present authors. In the abstract scenario depicted, most of a learner's working-memory capacity is devoted to dealing with the cognitive load imposed by a learning activity. (In this scenario, there is no cognitive overload and some resources are left unallocated).

increasing germane load. Some scholars have stressed that germane load is not just any useful load but instead arises from those additional, effortful aspects of learning that go beyond task performance, such as conscious reflection and self-explanation [36, 86].

The precise definition and role of germane load in CLT has shifted during the past decades. Germane load has been at the heart of many critiques of Old CLT, and it is the major difference between Old CLT and New CLT, as described below.

2.2.2 Criticisms of Old CLT. Notable scholars in and around CLT complained that Old CLT suffered from conceptual flaws, failed to explain various empirical results, used a germane-load construct that was unnecessary and unjustified by results, and led to unfalsifiable circular interpretations of findings [16, 23, 36, 86, 100]. These issues were compounded by there being no valid measures of germane load. For example, if a reduction of overall cognitive load was observed in an intervention, researchers might state *post hoc* that the change must have been in extraneous load (if learning was enhanced) or germane load (if learning was hindered), but there was no way of falsifying such interpretations, rendering them nearly worthless [16, 23, 86]. Moreover, numerous studies reported that learning results went up as overall cognitive load went down, which conflicted with Old CLT's notion that a reduction in extraneous load should lead to an equal increase in germane load [100, 103].

Around the same time as the above criticisms surfaced, Moreno [66] pointed at more shortcomings: "CLT is remarkably silent about the relation among load, affect, and motivation. . . . Another omission from CLT is self-regulation." Although Old CLT did have the concept of germane load, which was sometimes linked to pedagogies promoting deeper engagement, Moreno [66] persuasively argued that Old CLT "has not attempted to clearly specify the characteristics of mental processes that promote 'good' versus 'bad' load" and the theory has therefore failed "to unequivocally predict the outcomes of different instructional methods on [load components] and learning."

In part because of such criticisms, CLT was revised in 2010.

2.2.3 New CLT. New CLT [36, 100] is neater and narrower. It defines intrinsic and extraneous load strictly in terms of element interactivity and posits that these two components constitute

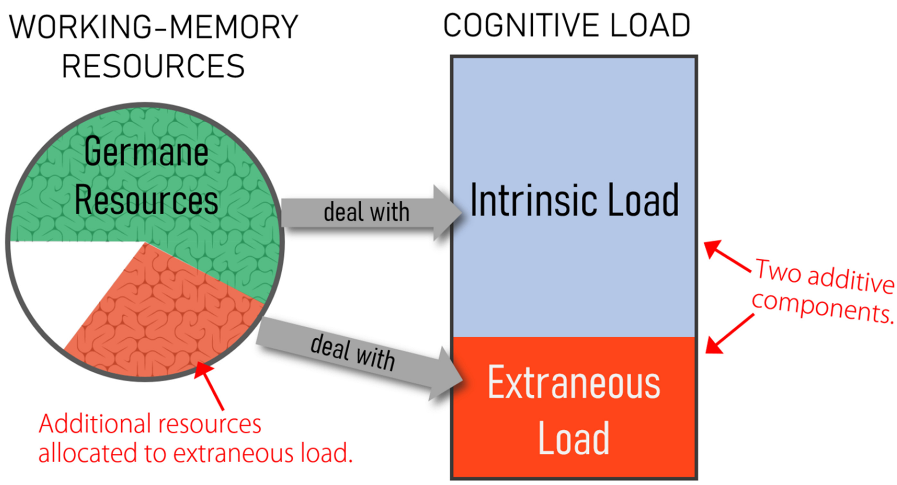

Fig. 2. New CLT, as interpreted by the present authors; cf. Old CLT in Figure 1. In this scenario, extraneous load requires the learner to devote some working-memory resources, but the learner is also able to devote germane resources to intrinsic load, i.e., learning. (This figure, too, depicts a scenario without cognitive overload.)

overall cognitive load. That is, New CLT discards the idea of germane load as a third, additional and additive source of load. Any cognitive load that is not extraneous is intrinsic.

In New CLT, learning results from dealing with intrinsic load. This includes learning from "additional" elaborative processes, which other frameworks might interpret as going beyond task performance. If the learner is intended to, say, compare examples that vary in surface features, or to self-explain, then the activity's learning goals are more ambitious than otherwise, and the intrinsic load is higher [36, 100].

The word "germane" is not gone from New CLT. However, New CLT's concept of "germane load," which is alternatively and perhaps less confusingly also referred to as *germane resources* or *germane processing*, "belongs to a quite different category to intrinsic and extraneous cognitive load. It simply consists of working memory resources used to handle element interactivity associated with intrinsic cognitive load" [100]. It is an entirely different quantity than Old CLT's germane load (see Figure 2).

It is our impression that—compared to the many Old CLT studies that have sought to increase germane load—the concept of germane resources tends to play a reduced role in studies based on New CLT.

One of the criticisms of Old CLT around 2010 was its silence on motivation. New CLT explicitly sidesteps this critique: it is interested in how instructional design affects "ideal" learners that are highly engaged; it *assumes* motivation [36, 100]. Motivation is thus something that is relevant to learning but external to the theory; motivated engagement is not an additional load on a learner's mind, but something that is expected before the theory applies.

2.2.4 Both CLTs are Alive. New CLT is frequently cited as being more up to date, and it bears the cachet of CLT's primary creator, John Sweller. Nevertheless, Old CLT has notable support from some groups of researchers who deliberately opt to use the original formulation of germane load [44, 45, 89, 114]. As things stand, it is contentious which constructs compose cognitive load, how best to interpret cognitive load measurements, and how to phrase hypotheses that involve relationships between load components, overall load, and learning outcomes.

The issue of Old CLT vs New CLT is entangled with a number of other challenges in cognitive load research. We briefly review some of those below.

2.3 Open Issues in Cognitive Load Theory

2.3.1 Measuring Cognitive Load. Researchers have come up with an assortment of methods for approximating cognitive load empirically; for recent reviews and commentaries, see [48, 60, 103, 118]. However, there are arguably no truly well established and confidence-inspiring ways to measure load, and the problem is even worse if one wishes to measure individual load components—a goal that has been referred to as CLT's "holy grail" [3] and "mission impossible" [44]. This means that it is difficult to compare versions of CLT empirically and, more generally, that it is challenging to do research based on CLT.

Subjective measures of cognitive load are common. They are typically based on retrospective self-assessment, in which learners report on their perceived mental effort or other aspects of a preceding learning episode. Paas's [74] unidimensional nine-point scale for mental effort from 1992 is the time-tested instrument that continues to be used, but that simple scale cannot differentiate between load types and suffers from noisy measurements [49]. Various groups of researchers have created or adapted questionnaires for assessing separate load components with a battery of questions (e.g., [44, 45, 50, 51, 89, 112, 113]); this work on instrument development extends to CER as well [69, 115]. Doubts remain over these instruments' construct validity, however, and especially over the questionnaire items that target germane load [34, 49, 51, 60, 115]. On the basis of empirical findings, some researchers have discarded the name "germane load" in favor of "self-perceived learning" [5] or modified an instrument to focus on intrinsic and extraneous load only, as per New CLT [9]; others target germane load in the Old CLT sense [44, 89]. Ultimately, it remains an open question precisely what the existing instruments succeed in measuring.

Objective measures of cognitive load do not depend on self-assessment. They are usually based on physiological data sources, such as eye-tracking [32], pupillometry [64, 106], heart rate [76], or neuronal [1, 109] or electrodermal [73, 90] activity. This avoids the problem that learners' self-reports are often unreliable; another benefit is that many objective methods can be applied during learning as an index to real-time changes in load, which is valuable since working memory operates on a timescale of seconds. On the downside, many objective methods are intrusive and distracting, and evidence of their construct validity, too, is limited. Moreover, although some attempts have been made to differentiate load components using objective methods [3], a breakthrough appears unlikely in the near future.

2.3.2 Motivation, Productive Failure, and CLT's Scope. CLT has been at the heart of many debates between proponents of "explicit (or direct) instruction" and proponents of "constructivism" (see, e.g., [26, 42, 57, 96, 105]). Some (but not all) cognitive load researchers have placed themselves firmly in the former camp, advancing in particular the argument that studying worked examples that teach novices what to do is superior to problem-solving practice, especially when that problem-solving is insufficiently scaffolded.

Lately, such discussions have frequently featured the pedagogical pattern of *productive failure* [39, 92], in which a problem-solving activity that fails by design precedes other activities that teach a solution. On the one hand, productive failure has credible empirical support, but on the other, it ostensibly contradicts the well established worked-example effect. Various scholars—and this includes voices from within cognitive load research—have recently argued that efforts to reconcile CLT with productive failure and other complementary frameworks have been insufficient [38, 49, 53]. Some work toward such a reconciliation is emerging (see, e.g., [12–14, 24, 25, 53, 70, 92, 94]).

Motivational factors are key to various pedagogical strategies that emphasize problem-solving, inquiry, or discovery and that are often perceived to contradict CLT's prescription of worked examples. CLT's assumption of motivated learners is a strength and a weakness. Cordoning off motivation makes 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. Perhaps the answer is that CLT's role is to "micro-manage" [38] instruction at a fine grain, while other perspectives complement CLT by addressing macro levels that involve motivation, self-regulation, and similar concepts. Such possibilities continue to be debated (see, e.g., [79, 85, 101]). Moreover, it has been suggested that cognitive load should be studied as an influence on motivation [22] and that CLT could be revised to encompass tradeoffs between risking high extraneous load and risking low learner engagement [94].

2.3.3 Collaborative Learning. Almost all research on cognitive load has focused on individual learning, which has brought the longstanding criticism that CLT ignores collaborative learning. Recent critiques have suggested, for instance, that the effects of collaboration may explain the tension between CLT and productive failure [49].

Matters have changed somewhat during the last decade, with a number of studies exploring the influence of collaboration on cognitive load effects. In one study, for example, collaborative problem-solving was found to be superior to problem-solving on one's own, but collaboration was counterproductive when studying worked examples [83]. An extension of CLT, *collaborative cognitive load theory*, has been proposed [41, 43] and related research is emerging (see, e.g., [33]). Nevertheless, the interplay between collaborative learning and cognitive load remains understudied, which is another factor that researchers in CER—and teachers—should bear in mind as they consider the theory's recommendations.

2.3.4 Estimating Element Interactivity. While formulating intrinsic and extraneous load in terms of element interactivity for New CLT, Sweller [100] suggested that "these two forms of cognitive load can readily be distinguished using *a priori* analyses of materials prepared for particular learners." *A priori* analysis of load interests practitioners, too, as Mason et al. [59] observed in CER: "there are circumstances in which a decision on learning materials must be made before those materials are developed and deployed, . . . the assessment of the likely imposition of cognitive load is made by the instructors and informs the development of the learning materials."

In many studies of cognitive load, researchers have estimated the load imposed by learning materials, classifying the materials as high or low in element interactivity (see, e.g., [12, 13, 27, 54, 72]). Findings are then interpreted in the light of such estimates. However, the estimates are typically based on relatively informal and sometimes opaque argumentation rather than an explicit method. There is no broad agreement on what counts as high or low complexity, and even the definition of element interactivity as a measurable quantity has been challenged [40, 49]. To our knowledge, there are no established guidelines on how to determine element interactivity for different materials (e.g., computing content), how to compare element interactivity across contexts, or how to deal with prior knowledge in such analyses. The last point is crucial, since prior knowledge greatly reduces element interactivity and cognitive load.

2.3.5 Other Developments. Recent work has examined CLT's relationship with themes such as individual differences in working-memory capacity [11, 84], emotions [30, 79, 94, 114], procedural vs conceptual knowledge [14], human movement [88], the physical environment [15], and the testing effect [27, 40, 49], among others.

Section Summary: There are multiple versions of cognitive load theory, many unanswered questions about methodology, longstanding debates about the theory's fundamental concepts, and numerous new threads of research.

These issues are of interest to researchers in computing education as well as practitioners. In the rest of this article, we will examine how CLT and its developments are reflected in the CER literature.

3 METHOD

We conducted a systematic mapping review of cognitive load theory in the context of computing education research. It builds on the guidelines for systematic reviews presented by Petersen et al. [78].

3.1 Research Questions

We sought answers to the following research questions:

RQ1. *How much does CLT feature in CER?*
- (a) How many publications actually *use* CLT (e.g., design studies or explain findings using CLT)?
- (b) How many publications at least *cite* CLT?

RQ2. *Which versions of CLT are being used or cited in CER?*
- (a) How many publications use Old CLT compared to New CLT?
- (b) How many publications cite Old CLT compared to New CLT?
- (c) Have these proportions changed over the last decade?

RQ3. *How common is it to measure cognitive load in CER, and with which tools?*

3.2 Selection of Publications

3.2.1 Selected Venues. Since our focus was on the CER community, we decided to define the scope of our publication search by venue. We looked for publications published in these journals:

- ACM Transactions on Computing Education (TOCE),
- Computer Science Education (CSE),
- IEEE Transactions on Education (ToE), and
- Computers & Education (C&E).[4]

And these conferences, whose proceedings series were readily available via the ACM Digital Library:

- ACM SIGCSE Technical Symposium (SIGCSE TS),
- ACM International Computing Education Research Conference (ICER),
- Conference on Innovation and Technology in Computer Science Education (ITiCSE),
- Koli Calling International Conference on Computing Education Research (Koli),
- The United Kingdom and Ireland Computing Education Research conference (UKICER), and
- ACM Conference on Human Factors in Computing Systems (CHI).[5]

[4] Although C&E does not focus solely or even mainly on CER, we included it since it is a prominent journal that publishes high-quality CER.
[5] Although CER is only one of many themes covered at the massive CHI conference, we included CHI as we were aware that many studies published there mention cognitive load.

Table 1. Search Terms for Each of the Selected Venues

Venue	Search String
TOCE	"query": {"cognitive load"} "filter": {Publication Date: (01/01/2010 TO 03/31/2021), Published in ACM Transactions on Computing Education, ACM Content: DL}
CSE	[All: "cognitive load"] AND [in Journal: Computer Science Education] AND [Publication Date: (01/01/2010 TO 03/31/2021)]
C&E	"cognitive load" from "Computers & Education" from 2010 to 2021
ToE	("Publication Title": IEEE Transactions on Education) AND ("All Metadata":"cognitive load"), filter added 2010–2021
ICER	"query": {AllField:("cognitive load")} "filter": {Article Type: Research Article, Conference Collections: ICER: International Computing Education Research Workshop, Publication Date: (01/01/2010 TO 03/31/2021), ACM Content: DL}
SIGCSE TS	"query": {AllField:("cognitive load")} "filter": {Conference Collections: SIGCSE: Computer Science Education, Article Type: Research Article, Publication Date: (01/01/2010 TO 03/31/2021), ACM Content: DL}
Koli	"query": {AllField: ("cognitive load")} "filter": {Article Type: Research Article, Conference Collections: Koli Calling: Koli Calling International Conference on Computing Education Research, Publication Date: (01/01/2010 TO 03/31/2021)}
ITiCSE	"query": {AllField: ("cognitive load")} "filter": {Article Type: Research Article, Conference Collections: ITiCSE: Innovation and Technology in Computer Science Education, Publication Date: (01/01/2010 TO 03/31/2021)}
UKICER	"query": {AllField: ("cognitive load")} "filter": {Article Type: Research Article, Conference Collections: UKICER: UK & Ireland Computing Education Research, Publication Date: (01/01/2010 TO 03/31/2021)}
CHI	"query": {AllField: ("cognitive load")} "filter": {Article Type: Research Article, Conference Collections: CHI: Conference on Human Factors in Computing Systems, Publication Date: (01/01/2010 TO 03/31/2021)}

Our selection of venues could certainly be expanded. However, we believe it to be fairly uncontroversial and to cover most major CER forums from the time period of interest.

3.2.2 Search Criteria. In 2010, Sweller [100] published what we consider the seminal New CLT article. Since our primary interest is recent work, one of our research questions concerns the adoption of New CLT, and 2010 is a nice round number, we chose to focus on the time interval between 2010 and the present time, which we delimited on March 31st, 2021.

The publications were available in the ACM Digital Library database, with the exception of three journals: ToE (IEEE Xplore), CSE (Taylor & Francis Online), and C&E (Science Direct). In all the databases, we searched for "cognitive load." Table 1 details the search criteria.

3.2.3 Filtering the Publications. Our initial search yielded 1,253 publications. We filtered these results using the following inclusion criteria:

— Must be a research publication. (This includes both short and long papers from Koli Calling.)
— Must be written in English.
— Must count as computing education research (see below).

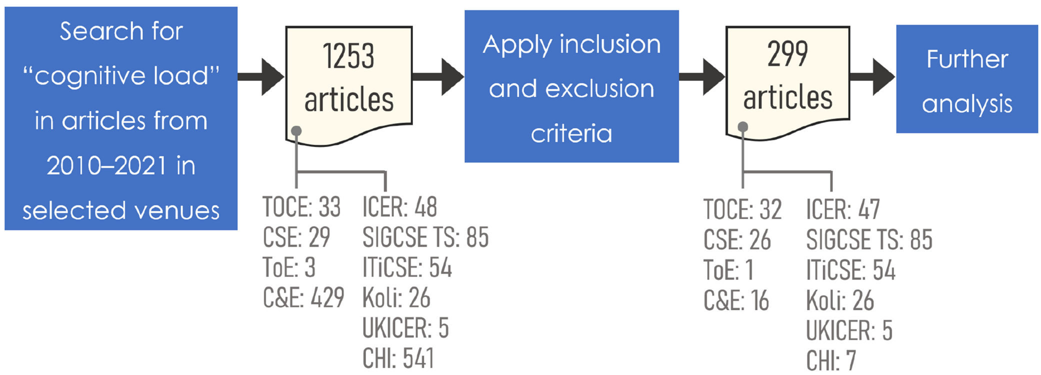

Fig. 3. Selection of publications to analyze, with counts by venue.

And these exclusion criteria:

— Short documents: posters, abstract, or extended abstract only.
— Meta and other non-standard publications: editorials, panels, keynotes, submissions to doctoral consortia.

We construed CER broadly as research related to the teaching and learning of computing. This includes, for example, computational literacy [17] but excludes the general use of computers for learning non-computing topics. The venues' nature meant that this criterion was broadly met, with the exception of the C&E journal and CHI conference, from which two we excluded many publications. This filtering was done by screening the abstracts.

After applying the inclusion and exclusion criteria, we had 299 publications for further analysis. Figure 3 summarizes the filtering process and shows the publication counts for each venue.

3.3 Procedure

Each publication was analyzed by either the first author, the second author, or both. After an initial round of classification, we had a refinement round to solve unclear cases and disagreements.

We inspected each publication for CLT usage, CLT citations, any CLT versions that were cited and/or used, cognitive load measurement, mentions of cognitive load types, and any other CLT-related argumentation. Figure 4 shows an overview of the main questions that we asked during this analysis in order to classify the publications. The figure also lists the classification labels.

To answer the questions and classify the publications, we searched each document for keywords such as "cognitive load theory," "intrinsic," "extraneous," and "germane," among others. The specific criteria for classification are explained in more detail below.

Our dataset with each publication's classification is available online [19].

Note: As illustrated in Figure 4, we distinguished between *using* CLT and *citing* it, and examined each separately. Often, studies that used CLT would also cite it, but some did only one or the other—or neither, if "cognitive load" was mentioned in passing without a citation.

3.4 Classification Criteria

3.4.1 Which CLT Version Was Cited (If Any)? For each of the publications in our dataset, we determined whether it cited cognitive load theory. If the theory was cited, we attempted to determine the version that was being referred to. This information helped us answer RQ1b *(How many*

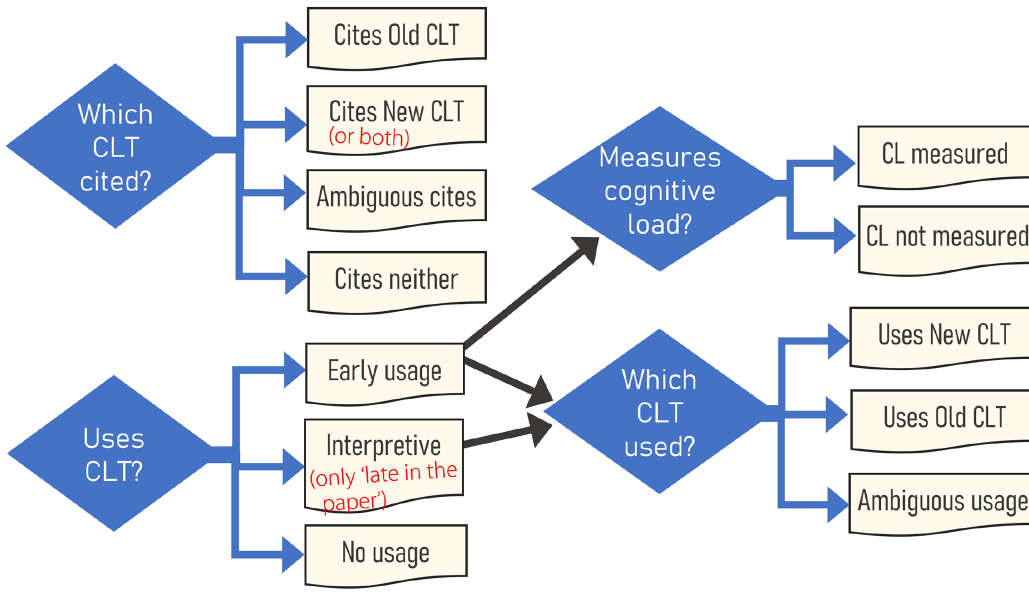

Fig. 4. Questions asked during the analysis of the selected publications, and the corresponding categories.

publications at least cite CLT?), RQ2b *(How many publications cite Old CLT compared to New CLT?)*, and RQ2c *(Have these proportions changed over the last decade?)*.

More specifically, we looked at whether each publication cited at least one "primary" or "secondary" CLT source. We considered "primary" sources to be publications by Sweller and colleagues that discuss CLT or its effects, whereas "secondary" sources are any other publications with clear and obvious connections to cognitive load, such as Mayer's theory of multimedia learning [61].

Publications that did not cite any primary or secondary CLT sources were classified as *Cites neither*. If at least one CLT source was cited, we determined the cited CLT version mostly on the basis of references to primary sources and the publication's description of CLT and cognitive load components. If New CLT or both versions were cited, we labeled the publication as *Cites New CLT*; we used the label *Cites Old CLT* if only that version was cited. In some cases, it was not possible to assert which version of the theory was cited; we labeled these publications as *Ambiguous cites*.

3.4.2 How Was CLT Used? To answer RQ1a *(How many publications actually use CLT?)*, we checked each publication for evidence of CLT use. Each publication was given one of three labels: *Early usage*, *Interpretive*, or *No usage*.

Early usage means that we saw evidence of CLT influencing the research already at an "early" stage; that is, CLT has helped hypothesize about results, inspired an intervention design, influenced data collection (e.g., CLT measurement), or informed theoretical framing. Depth of usage varied greatly within this category, ranging from publications that briefly mention CLT as an inspiration for a specific tool feature (e.g., [95]) to others whose study design is grounded in CLT (e.g., [91]).

Early usage contrasts with the *Interpretive* category, which we used for publications where CLT *only* appeared "late in the paper," as a suggested post hoc explanation for empirical findings. One example is Bennedsen and Schulte's [4] conjecture that cognitive load may have played a role in why students did not benefit from additional features in a debugger.

Finally, the publications that did not match either of the other two categories fell under the *No usage* label.

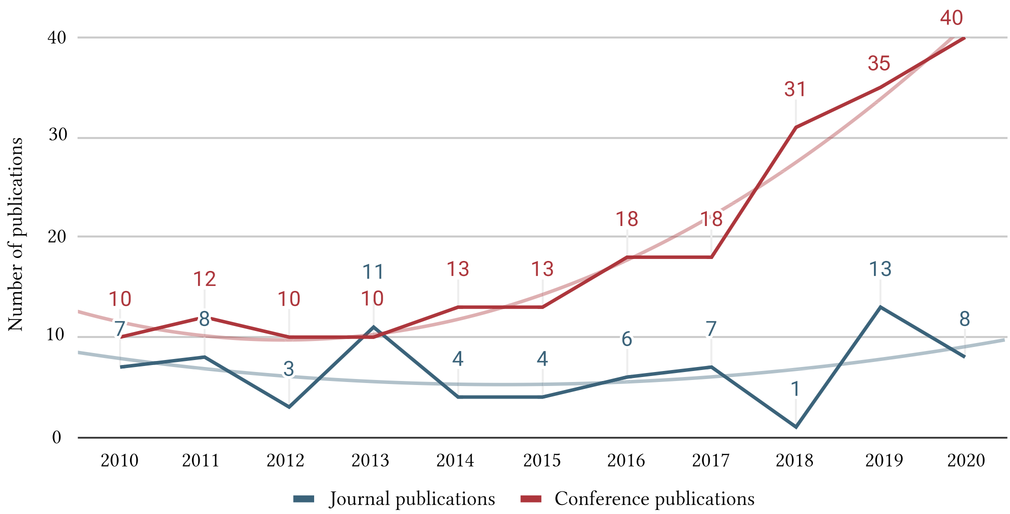

Fig. 5. Total number of publications containing the phrase "cognitive load" and published at the selected venues between 2010 and 2020. The faint red line is the polynomial trend for conferences and the faint blue line for journals.

These categories are meant to describe which aspects of research CLT has influenced. They are not a commentary on the publications' quality, which we did not assess.

3.4.3 Which CLT Version Was Used? This analysis only involved those documents that had been classified as *Early usage* or *Interpretive*; each of those publications received an additional label of *Uses New CLT*, *Uses Old CLT*, or *Ambiguous usage*. We decided which CLT version was used by examining the authors' descriptions of cognitive load components and their references to specific instruments for measurement. This step provided answers to RQ2a *(How many publications use Old CLT compared to New CLT?)* and RQ2c *(Have these proportions changed over the last decade?)*.

Cited and used CLT versions did not always go hand in hand. In particular, even if a publication cited a given CLT version, it was not necessarily clear which version was actually used in the study. In some rare cases, a publication used CLT and described that usage in enough detail that we could assess the version despite the publication not citing CLT.

3.4.4 Was Cognitive Load Measured? In RQ3, we asked: *How common is it to measure cognitive load in CER, and with which tools?* Any studies that measured cognitive load ended up in the *Early usage* category, so we went through those publications, recording which studies applied which instrument(s) for measurement.

3.4.5 Additional Analysis. We recorded whether the publications mentioned any specific cognitive load components (extraneous, intrinsic, germane), and if they did, which ones. During analysis, we also made notes on how CLT was discussed in the documents that we studied; we will comment on some of these observations in Section 5.

4 RESULTS

4.1 RQ1: How Much Does CLT Feature in CER?

Since 2010, the number of publications that feature the phrase "cognitive load" has grown from 17 articles in 2010 to 48 articles in 2020. Figure 5 shows the publication counts for each full year in our time interval. Visual inspection suggests a positive trend that is especially strong from

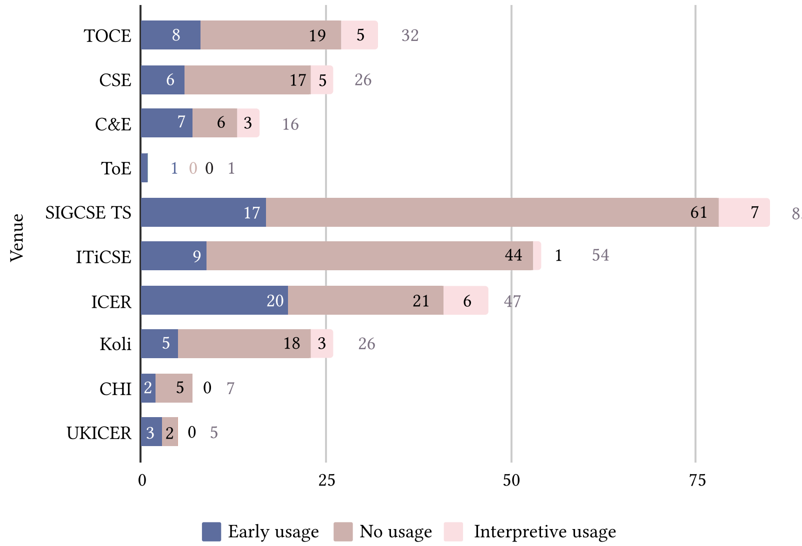

Fig. 6. How many articles *use* CLT?

2014 onward for conferences. The recent numbers are boosted especially by the large conferences SIGCSE TS and ITiCSE.

4.1.1 RQ1a How Many Publications Actually Use CLT? Of the 299 analyzed, 78 publications (26%) were labeled *Early usage*, 28 (9%) *Interpretive*, and 193 (65%) *No usage*. That is, the majority of the publications mentioning "cognitive load" did not actually use the theory in the senses described in Section 3.4 above. Usage varied somewhat by venue, as shown in Figure 6.

4.1.2 RQ1b How Many Publications at Least Cite CLT? Many of the reviewed publications (126; 42%) mention "cognitive load" without any reference to either version of the theory or other obviously related research. This is shown in Figure 7, where *Cites neither* has a high frequency at most venues. Among the publications that did cite CLT sources, 105 (35%) clearly referred to one or both versions of CLT. In the remaining 68 (23%), it appeared that they were citing CLT or related literature, but we could not tell which version (*Ambiguous cites* in Figure 4).

If we focus only on those publications that used CLT (either *Early usage* or *Interpretive*), we find that 73 (69%) cited one or both versions of CLT, 14 (13%) were ambiguous about which version they cited, and 19 (18%) did not cite CLT despite making use of the theory.

4.2 RQ2: Which Versions of CLT are Being Used or Cited in CER?

Our results indicate that the CER community uses and cites both versions of CLT, with Old CLT prevalent.

4.2.1 RQ2a: How Many Publications Use Old CLT Compared to New CLT? Studies that used CLT (*Early usage* or *Interpretive*) were often ambiguous in which CLT they used: 59% of these studies described their usage of the theory in insufficient detail to suggest a particular version of the theory.

Uses Old CLT (36% of all publications using CLT) was far more common than *Uses New CLT* (5%). Figure 8 shows these results by venue and suggests that there are differences between venues. In

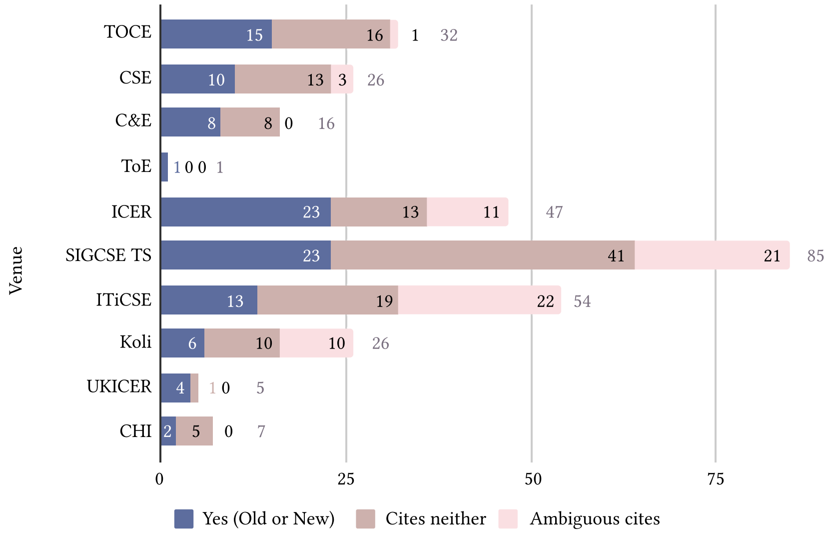

Fig. 7. Citations to CLT in publications featuring "cognitive load," by venue.

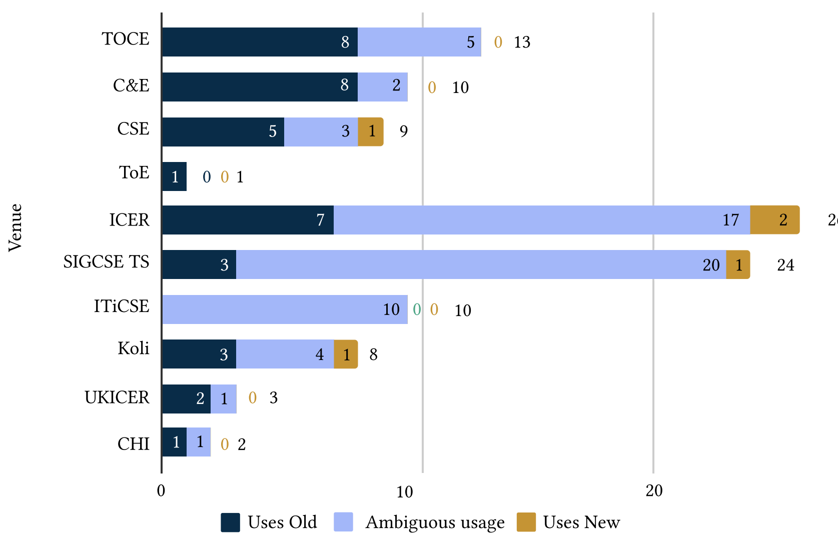

Fig. 8. Usage of Old CLT and New CLT, by venue.

general, journal articles fell into the *Ambiguous usage* category less often (30%) than conference papers (72%) did.

4.2.2 *RQ2b: How Many Publications Cite Old CLT Compared to New CLT?* For 105 publications, we could identify a version of CLT that was cited. Most, 76, cited Old CLT, with 29 papers citing New CLT. The distribution for each venue is shown in Figure 9.

We did not compute statistics on the most frequently cited papers, but can point to Sweller et al. [102], Chandler and Sweller [10], and Sweller [99] as examples of commonly cited references for Old CLT, with New CLT commonly cited via Sweller [100].

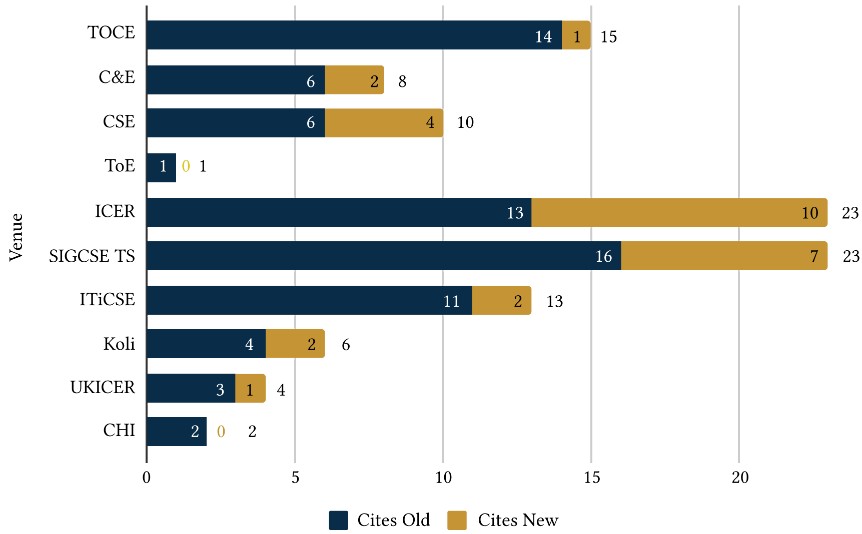

Fig. 9. Citations of Old CLT and New CLT, by venue.

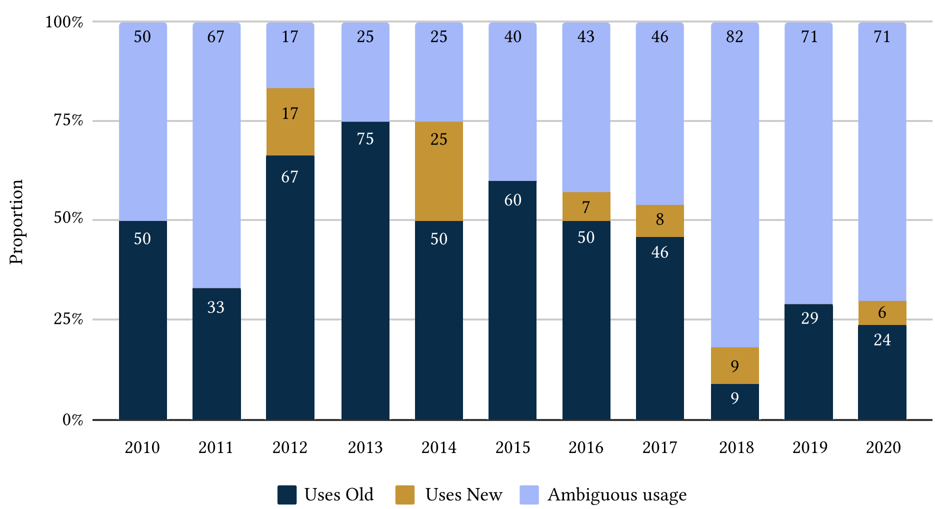

Fig. 10. Usage of CLT versions between 2010 and 2020, aggregated across venues.

4.2.3 RQ2c: Have These Proportions Changed Over the Last Decade? In terms of *usage* (whether *Early* or *Interpretive*), New CLT has remained quite rare throughout the past decade. Between 2010 and 2020, it was used by at most one publication per year (in 2012, 2014, 2016–2018, and 2020).

Old CLT's usage has varied over the years, with peaks in 2013 (six publications) and 2017 (seven). Most recently, in 2019 and 2020, Old CLT was used in four publications each.

Ambiguous usage, on the other hand, has been increasing over the evaluated time frame, from 2 publications in 2015 to 6 in 2016–2017, 9 in 2018, 10 in 2019, and 12 in 2020.

Visual inspection of Figure 10 suggests little has changed between the two CLT versions, but from 2018 onward, a higher proportion of studies used ambiguous versions of the theory.

In addition to trends in CLT usage, we looked for possible trends in citations. Among articles that cite either Old CLT or New CLT, citing (only) Old CLT happened 76% of time between 2010

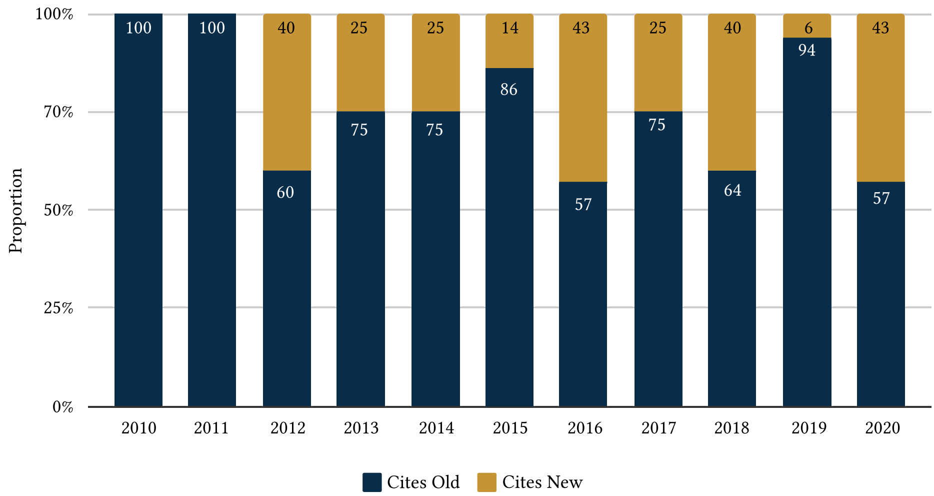

Fig. 11. Citing of CLT versions between 2010 and 2020, aggregated across venues.

and 2020. Figure 11 shows the proportions for each year. Despite the exception in 2019, there might be a minor emerging trend toward citing New CLT, with its proportion reaching or exceeding 40% in multiple recent years.

4.3 RQ3: How Common is it to Measure Cognitive Load in CER, and with Which Tools?

Out of 78 *Early usage* studies, 22 measured cognitive load empirically. The most common instruments used in these publications are Morrison's [69] adaptation of Leppink's questionnaire (11 publications) and Paas's [75] rating scale (seven publications). Other instruments were used less frequently. These include those by Leppink et al. [50], Kalyuga et al. [37], and Ayres [2], as well as the NASA-TLX [29] and PSA [52]. A few studies (e.g., [28, 31]) used more than one of these instruments, such as a combination of Paas's instrument with NASA-TLX or Leppink's questionnaire.

4.4 Mentions of Cognitive Load Components

As noted above, we additionally looked at how often the publications mentioned specific cognitive load components. At least one component was mentioned in 69 publications out of our entire dataset of 299; this subset largely overlaps with studies that not only cited but actually used CLT. The most frequently mentioned component was extraneous load (65), followed by intrinsic load (48) and germane load (40).

5 DISCUSSION

5.1 CLT Features Often in CER

Cognitive load theory is increasingly influencing CER. Since 2010, the number of annual publications that at least mention "cognitive load" has nearly tripled. In quite a few articles, "cognitive load" comes up vaguely and casually, and the authors may not have meant the term in the CLT sense. Nevertheless, many authors did refer to the theory.

Often, CLT was briefly mentioned either as a design influence or as a possible explanation for findings. Few articles engaged with the theory in a deeper sense. Below, we provide some observations on these different kinds of CLT usage.

5.1.1 Justifying Design Decisions by Mentioning CLT. Many articles in CER cite CLT as a design influence for a pedagogical decision or a learner-facing technology: for instance, block-based programming [77, 82], Parsons problems [119], worked examples [62, 108, 110], audio explanations of code [21], and syntax exercises [98] have been justified on the grounds of cognitive load. Some of these mentions are not concerned with CLT as a whole but more with a particular CLT-related recommendation, such as the worked-example effect or the modality effect. In other articles, the intention is more generally to keep complexity in check (e.g., complexity due to syntax).

There is nothing inherently wrong with mentioning CLT in such a way. We note, however, that such mentions were often not accompanied by an attempt to measure load, to systematically analyze the learning situation from a CLT perspective, or to debate the relevant CLT effects critically.

Anecdotally, we observe that it is fairly common in CER to justify a design choice by mentioning "cognitive load" and citing a single paper that supports a somewhat similar design in a single other context, without citing CLT itself. It seems rare in CER to discuss assumptions, boundary conditions, and mixed findings concerning the various cognitive load effects. This is a potential pitfall, since practically all of CLT's recommendations come with notable provisos; even the venerable worked-example effect is being considered in new light (see Section 2.3 above). Moreover, complications in transferring cognitive load effects to the computing domain have been documented [67, 116, 117], and cross-contextual transfer is a perennial challenge in general.

Recommendations: Go ahead and cite cognitive load effects as design influences, but do look into the limitations of the relevant effects. Consider measuring how your design influences cognitive load. Don't cite CLT merely to name drop a theory of learning where you want a simple design. Avoid "cognitive load" as a generic phrase for any effortful mental work.

5.1.2 Mentioning CLT While Interpreting Findings. Researchers in CER are looking to CLT for explanations of empirical results. Some articles briefly mentioned CLT in discussion only: after reporting a study, the authors brought up cognitive load in order to explain learning outcomes or to speculate about possible causes for observed phenomena (e.g., [4, 63, 80, 81]). Many such interpretations appeared post hoc (i.e., cognitive load was not an element of the study design).

Even where CLT is not a major component of a study, speculation about a result in terms of CLT can be a perfectly reasonable thing to do and may lead to valuable further research. We note, however, that we did not spot concrete cases where such speculations were tested in a later publication. Furthermore, without measurements of cognitive load itself, researchers must be careful not to overstate what is actually known about participants' cognitive load. As we noted in Section 2.2.2, Old CLT especially has been criticized for the unfalsifiability trap, where any outcome can be conveniently explained post hoc, and similar concerns have been raised in CER as well [71].

Recommendations: When interpreting empirical findings, go ahead and use CLT to generate speculations about what happened. Then conduct studies to test those speculations. State *a priori* hypotheses in terms of overall cognitive load and/or load components. Don't use (Old or indeed any) CLT as a *deus ex machina.*

5.1.3 Deeper Uses of CLT are Uncommon. Articles that engage more deeply with cognitive load theory remain fairly rare in CER. They certainly do exist, however, and have explored a variety of research questions in terms of CLT. We list a few examples here to illustrate:

— Shin [91] and Tasir and Pin [104] sought to reduce extraneous load in learning materials by addressing cognitive load effects. They measured load using Paas's instrument.

— Morrison et al. [69] adapted and validated a cognitive load measuring instrument for programming.
— Yeh et al. [111] prompted germane processing in working memory through multimedia learning materials. Their work is fairly unusual in CER in that it explicitly brings up prior knowledge as a moderating variable for available cognitive resources and thus cognitive load.
— Ericson et al. [20] compared the effects of two forms of practice—Parsons problems and program writing—on learning outcomes, efficiency, and cognitive load. They formulated hypotheses in terms of CLT and measured cognitive load.
— Margulieux and Catrambone [56] measured cognitive load as part of a study on subgoal labels in programming education.

5.2 Variants of CLT and Load Components

5.2.1 New CLT Has Not Really Caught On. In CER articles, Old CLT continues to feature more prominently than New CLT, despite more than a decade having passed since New CLT was formulated. We suspect that there are multiple factors behind this.

In some cases, citing Old CLT is clearly appropriate because the authors mean to point at the beginnings of the theory and the seminal studies.

In some articles that do not engage deeply with CLT, it matters little which version of the theory is brought up (or if the version is ambiguous). All versions of CLT share many core assumptions and concepts, such as limited working memory and the consequent need to avoid cognitive overload while learning. Moreover, many of the specific pedagogical recommendations (e.g., worked examples) can, arguably, be justified in terms of either Old CLT or New CLT.

It seems clear, however, that part of the reason for Old CLT's prevalence is simply that authors in CER are not aware of developments in the theory. This also suggests that many authors citing CLT may be unaware of the conceptual and practical challenges in applying CLT reliably, to which New CLT is a partial response, many of which continue to be debated. Similarly, the last decade's research on the caveats and boundary conditions of specific cognitive load effects may have gone unnoticed by many researchers in CER.

Anecdotally, we also observed that a few articles cited New CLT but nevertheless proceeded to treat cognitive load as three additive components as per Old CLT, without addressing this apparent contradiction. Moreover, the term "germane load" was sometimes used in a way that left unclear whether it was meant in the Old CLT or New CLT sense, which muddled the authors' intended meaning.

As we mentioned in Background, Old CLT continues to have support among the general literature on CLT. Some authors in CER, too, might have deliberately opted for Old CLT despite being aware of New CLT. We saw little or nothing to suggest such a preference, however, although a few studies that used or adapted an existing Old CLT instrument did mention the theory's evolution since the original instrument's introduction.

Recommendations: Decide which version of CLT you use to design your study and why. Report clearly on that decision. If you use New CLT for CER, consider which of the open issues affect your work. If you use Old CLT, additionally consider whether your work is affected by the criticisms that led to New CLT.

5.2.2 Vague Discussion of Load Components. CER articles frequently bring up cognitive load as a generic concept, without discussing specific load components. Often this is fine, of course, such as when briefly mentioning the theory. In other cases, we felt that the authors' argumentation might have been sharper if they had distinguished between types of load. For example, many articles stated that there was a need to reduce cognitive load without reference to intrinsic or extraneous load. While reading them, we often got the impression that a particular learning situation was *assumed* to be complex and that too much *extraneous* load was assumed to be present, but these matters were not discussed by the authors.

In our dataset, cognitive load was generally framed as a sort of antagonist, with the researchers looking to reduce or eliminate it. This framing resonates with CLT's goal of avoiding cognitive overload. On the other hand, it is not *always* appropriate to make every effort to reduce intrinsic or extraneous load. Those efforts can be valid when load is too high, which depends not only on specific learning objectives but also on learners' prior knowledge. We observed that CER articles seldom discuss cognitive load as something to be optimized rather than simply reduced (Shin [91] is an exception). The expertise reversal effect, which implies that there is great variation in which teaching approaches are effective for which students, appears to have attracted little comment in CER, perhaps due to a focus on complete beginners.

In Old CLT, germane load is something that teachers want to increase as much as possible, since it is the beneficial, additional load associated with schema formation and learning. In New CLT, on the other hand, learning results from allocating resources to intrinsic load. This issue, which is a further source of confusion in interpreting authors' arguments, received little attention in the CER articles that we reviewed.

Recommendations: Consider whether your arguments are about specific load components rather than overall cognitive load. When reporting an attempt to reduce cognitive load, specify the type(s) of load. Try to design teaching that optimizes cognitive load for learners with different characteristics.

5.3 Estimating Cognitive Load

5.3.1 Empirical Measurement. Measuring cognitive load is rare in CER. We found an average of about two articles per year that had sought to measure either overall load or one or more load components. Even among the subset of studies that actually used CLT (rather than merely mentioning it), this is clearly a minority. At least in principle, many more CER studies might have benefited measuring cognitive load than actually did measure it; the lack of measurement is explicitly recognized as a limitation by some authors (e.g., [7, 47]).

Subjective self-assessment instruments were by far the most common.[6] Of those, CER has largely relied on the Paas single-question instrument [74] and Morrison and colleagues' [69] programming-domain adaptation of Leppink and colleagues' [50] 10-item questionnaire. This is unsurprising, as the Paas and Leppink instruments are also popular outside of CER.

As we discussed in Section 2.3.1 above, trustworthy measurement of cognitive load and its components is an open problem. The Paas scale has been criticized as unreliable and does not measure load components; there are concerns over the construct validity of the Morrison/Leppink instrument; many other instruments exist, but they are not without issues either.

CER is not alone with these problems. Leppink recently lamented the status quo: "researchers continue to use their own definitions and measurement tools without mention of alternative views, as if no one ever questioned for instance the role of germane load or any of the issues with the use of self-report measurements" [49].

[6] We might have seen some more use of eye-tracking or pupillometry as a cognitive load measure had our review included certain other (low-volume) venues. However, we do not believe that would have changed the big picture.

Recommendations: Design studies that measure cognitive load so that CLT-based explanations for findings may be evaluated. Do not assume that because an instrument for measuring cognitive load has been published and used, it works. Study the caveats of your chosen instruments; consider alternative explanations for the results you get. Contribute to the quest for trustworthy measurements.

5.3.2 A Priori Analysis. Another open issue in CLT is how to estimate, *a priori*, the element interactivity inherent in a learning situation for a particular type of learners (see Section 2.3.4 above). In our review, we found very few papers that clearly attempted something in this vein.

One of those few is the article by Mason et al. [59], which sets out to "flip" the assessment of cognitive load by having instructors examine the interacting elements present in a programming environment. Another is the work of Duran et al. [18], who proposed (but did not empirically evaluate) an approach for analyzing the complexity of computer programs in terms of the hierarchical schemas required to understand the programs.

Recommendations: Continue looking for ways to estimate element interactivity in programming and other subdomains of computing. Evaluate the approaches proposed so far.

5.4 Other Themes in Cognitive Load Research

Many novel threads of research have influenced cognitive load theory during the past decade, and more are emerging. Some of these themes were highlighted in Section 2.3 above; they include the tension between CLT and productive failure, CLT's relationships with emotion and motivation, collaborative cognitive load theory, and individual differences in working-memory capacity, among others. Some of these developments are very new, but others have existed for a decade or more. Overall, our review uncovered barely any evidence of these threads of research influencing CER substantially.

Recommendations: Consider CLT only one piece in instructional design. Complement cognitive load measurements by measuring other relevant constructs (e.g., motivation). Explore topical themes from the CLT literature—such as productive failure, collaborative CLT, and individual differences—in computing contexts.

6 CONCLUSION

For more than three decades, cognitive load theory has made valuable contributions to research on instructional design. The theory has been questioned, expanded, and reformulated. That evolution is still happening: CLT is not the finished article now, nor should we expect it to be "complete" at any future time. All this is a testament to the theory's ongoing relevance.

In this article, we have looked at CLT's presence in computing education research. During the past decade, citing CLT in CER has become increasingly common. Many authors mention the theory in passing; others have delved deeper. It is clear that CLT is a source of inspiration for researchers in our field.

However, many articles appear to rely on old versions of the theory, often without acknowledging alternatives. Attempts to measure cognitive load or estimate it analytically are rare. Design recommendations derived from CLT are frequently cited without any discussion of their boundary conditions or other limitations. Overall, we find that CER has not quite kept up to speed with recent developments in cognitive load research.

Many of the issues facing computing education researchers are shared by the CLT community at large. For instance, measuring cognitive load empirically is a vexing problem, as trustworthy instruments continue to elude us; future work in this area may be decisive to the theory's enduring impact.

Recommendations: Treat CLT as a living, evolving theory: it has much to commend it and holds the promise of more, but it also comes with many open issues and limitations. Examine CLT—its concepts, tools, and recommendations—in earnest, with a critical eye. Keep up to date with research on CLT.

ACKNOWLEDGMENTS

We thank the anonymous reviewers, whose comments helped us improve this article.

REFERENCES

[1] Pavlo Antonenko, Fred Paas, Roland Grabner, and Tamara van Gog. 2010. Using electroencephalography to measure cognitive load. *Educational Psychology Review* 22, 4 (2010), 425–438. https://doi.org/10.1007/s10648-010-9130-y

[2] Paul Ayres. 2006. Using subjective measures to detect variations of intrinsic cognitive load within problems. *Learning and Instruction* 16, 5 (2006), 389–400. https://doi.org/10.1016/j.learninstruc.2006.09.001

[3] Paul Ayres. 2017. Subjective measures of cognitive load: What can they reliably measure? In *Cognitive Load Measurement and Application*, Robert Z. Zheng (Ed.). Routledge, 9–28. https://doi.org/10.4324/9781315296258-2

[4] Jens Bennedsen and Carsten Schulte. 2010. BlueJ visual debugger for learning the execution of object-oriented programs? *ACM Transactions on Computing Education* 10, 2 (June 2010), Article 8, 22 pages. https://doi.org/10.1145/1789934.1789938

[5] Esther M. Bergman, Anique B. H. de Bruin, Marc A. T. M. Vorstenbosch, Jan G. M. Kooloos, Ghita C. W. M. Puts, Jimmie Leppink, Albert J. J. A. Scherpbier, and Cees P. M. van der Vleuten. 2015. Effects of learning content in context on knowledge acquisition and recall: A pretest-posttest control group design. *BMC Medical Education* 15 (2015), 133. https://doi.org/10.1186/s12909-015-0416-0

[6] Leon Bernard, Sagar Raina, Blair Taylor, and Siddharth Kaza. 2021. Minimizing cognitive load in cyber learning materials—An eye tracking study. In *Symposium on Eye Tracking Research and Applications (ETRA'21)*. ACM, 1–6. DOI: 10.1145/3448018.3458617

[7] Dennis Bouvier, Ellie Lovellette, John Matta, Bedour Alshaigy, Brett A. Becker, Michelle Craig, Jana Jackova, Robert McCartney, Kate Sanders, and Mark Zarb. 2016. Novice programmers and the problem description effect. In *Proceedings of the 2016 ITiCSE Working Group Reports (ITiCSE'16)*. Association for Computing Machinery, New York, NY, 103–118. https://doi.org/10.1145/3024906.3024912

[8] Michael E. Caspersen. 2007. *Educating Novices in the Skills of Programming*. Ph.D. Dissertation. Department of Computer Science, University of Aarhus.

[9] Raquel Cerdan, Carmen Candel, and Jimmie Leppink. 2018. Cognitive load and learning in the study of multiple documents. *Frontiers in Education* 3, 59 (2018), 1–7. DOI: 10.3389/feduc.2018.00059

[10] Paul Chandler and John Sweller. 1991. Cognitive load theory and the format of instruction. *Cognition and Instruction* 8, 4 (1991), 293–332. DOI: 10.1207/s1532690xci0804_2

[11] Ouhao Chen, Juan C. Castro-Alonso, Fred Paas, and John Sweller. 2017. Extending cognitive load theory to incorporate working memory resource depletion: Evidence from the spacing effect. *Educational Psychology Review* 30, 2 (Sept. 2017), 483–501. https://doi.org/10.1007/s10648-017-9426-2

[12] Ouhao Chen and Slava Kalyuga. 2019. Exploring factors influencing the effectiveness of explicit instruction first and problem-solving first approaches. *European Journal of Psychology of Education* 35, 3 (2019), 607–624. https://doi.org/10.1007/s10212-019-00445-5

[13] Ouhao Chen, Endah Retnowati, and Slava Kalyuga. 2020. Element interactivity as a factor influencing the effectiveness of worked example–problem solving and problem solving–worked example sequences. *British Journal of Educational Psychology* 90 (2020), 210–223. DOI: https://doi.org/10.1111/bjep.12317

[14] Ouhao Chen, Geoff Woolcott, and Slava Kalyuga. 2021. Comparing alternative sequences of examples and problem-solving tasks: The case of conceptual knowledge. *The Educational and Developmental Psychologist* (2021), 1–13. https://doi.org/10.1080/20590776.2021.1915098

[15] Hwan-Hee Choi, Jeroen J. G. Van Merriënboer, and Fred Paas. 2014. Effects of the physical environment on cognitive load and learning: Towards a new model of cognitive load. *Educational Psychology Review* 26 (2014), 225–244. DOI: https://doi.org/10.1007/s10648-014-9262-6

[16] Ton De Jong. 2010. Cognitive load theory, educational research, and instructional design: Some food for thought. *Instructional Science* 38, 2 (2010), 105–134. DOI: https://doi.org/10.1007/s11251-009-9110-0

[17] Andrea A. DiSessa. 2001. *Changing Minds: Computers, Learning, and Literacy*. MIT Press.

[18] Rodrigo Duran, Juha Sorva, and Sofia Leite. 2018. Towards an analysis of program complexity from a cognitive perspective. In *Proceedings of the 2018 ACM Conference on International Computing Education Research (ICER'18)*. Association for Computing Machinery, New York, NY, 21–30. https://doi.org/10.1145/3230977.3230986

[19] Rodrigo Duran, Albina Zavgorodniaia, and Juha Sorva. 2021. *Cognitive Load Theory Review Dataset*. Retrieved April 16th 2022 from https://osf.io/wbct6/?view_only=25d16ee808374c32be971d4a55834e33.

[20] Barbara J. Ericson, Lauren E. Margulieux, and Jochen Rick. 2017. Solving parsons problems versus fixing and writing code. In *Proceedings of the 17th Koli Calling International Conference on Computing Education Research (Koli Calling'17)*. Association for Computing Machinery, New York, NY, 20–29. https://doi.org/10.1145/3141880.3141895

[21] Barbara J. Ericson, Kantwon Rogers, Miranda Parker, Briana Morrison, and Mark Guzdial. 2016. Identifying design principles for CS teacher ebooks through design-based research. In *Proceedings of the 2016 ACM Conference on International Computing Education Research (ICER'16)*. Association for Computing Machinery, New York, NY, 191–200. https://doi.org/10.1145/2960310.2960335

[22] David F. Feldon, Gregory Callan, Stephanie Juth, and Soojeong Jeong. 2019. Cognitive load as motivational cost. *Educational Psychology Review* 31, 2 (2019), 319–337. https://doi.org/10.1007/s10648-019-09464-6

[23] Peter Gerjets, Katharina Scheiter, and Gabriele Cierniak. 2009. The scientific value of cognitive load theory: A research agenda based on the structuralist view of theories. *Educational Psychology Review* 21, 1 (2009), 43–54. https://doi.org/10.1007/s10648-008-9096-1

[24] Inga Glogger-Frey, Corinna Fleischer, Lisa Grüny, Julian Kappich, and Alexander Renkl. 2015. Inventing a solution and studying a worked solution prepare differently for learning from direct instruction. *Learning and Instruction* 39 (2015), 72–87. https://doi.org/10.1016/j.learninstruc.2015.05.001

[25] 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 (2016), 26–35. https://doi.org/10.1016/j.learninstruc.2016.11.002

[26] Mark Guzdial. 2015. What's the best way to teach computer science to beginners? *Communications of the ACM* 58, 2 (2015), 12–13. 10.1145/2714488.

[27] Jose Hanham, Wayne Leahy, and John Sweller. 2017. Cognitive load theory, element interactivity, and the testing and reverse testing effects. *Applied Cognitive Psychology* 31, 3 (2017), 265–280. https://doi.org/10.1002/acp.3324

[28] Kyle James Harms, Jason Chen, and Caitlin L. Kelleher. 2016. Distractors in Parsons problems decrease learning efficiency for young novice programmers. In *Proceedings of the 2016 ACM Conference on International Computing Education Research (ICER'16)*. Association for Computing Machinery, New York, NY, 241–250. https://doi.org/10.1145/2960310.2960314

[29] Sandra G. Hart. 2006. NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 50. Sage publications Sage CA: Los Angeles, CA, 904–908. 10.1177/154193120605000909

[30] Benjamin S. Hawthorne, Dianne A. Vella-Brodrick, and John Hattie. 2019. Well-being as a cognitive load reducing agent: A review of the literature. *Frontiers in Education* 4, (Oct. 2019), 1–11. https://doi.org/10.3389/feduc.2019.00121

[31] Ava Heinonen and Arto Hellas. 2020. Exploring the instructional efficiency of representation and engagement in online learning materials. In *United Kingdom & Ireland Computing Education Research Conference. (UKICER'20)*. Association for Computing Machinery, New York, NY, 38–44. https://doi.org/10.1145/3416465.3416470

[32] Kenneth Holmqvist, Marcus Nyström, Richard Andersson, Richard Dewhurst, Jarodzka Halszka, and Joost van de Weijer. 2011. *Eye Tracking: A Comprehensive Guide to Methods and Measures*. Oxford University Press, Oxford New York.

[33] Jeroen Janssen and Paul A. Kirschner. 2020. Applying collaborative cognitive load theory to computer-supported collaborative learning: Towards a research agenda. *Educational Technology Research and Development* 68 (2020), 783–805. https://doi.org/10.1007/s11423-019-09729-5

[34] Dayu Jiang and Slava Kalyuga. 2020. Confirmatory factor analysis of cognitive load ratings supports a two-factor model. *Tutorials in Quantitative Methods for Psychology* 16 (2020), 216–225. http://www.tqmp.org/RegularArticles/vol16-3/p216/p216.pdf.

[35] Slava Kalyuga. 2007. Expertise reversal effect and its implications for learner-tailored instruction. *Educational Psychology Review* 19, 4 (2007), 509–539. https://doi.org/10.1007/s10648-007-9054-3

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

[37] Slava Kalyuga, Paul Chandler, and John Sweller. 1999. Managing split-attention and redundancy in multimedia instruction. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition* 13, 4 (1999), 351–371. https://doi.org/10.1002/(SICI)1099-0720(199908)

[38] Slava Kalyuga and Anne-Marie Singh. 2016. Rethinking the boundaries of cognitive load theory in complex learning. *Educational Psychology Review* 28 (2016), 831–852. https://doi.org/10.1007/s10648-015-9352-0

[39] Manu Kapur. 2008. Productive failure. *Cognition and Instruction* 26, 3 (2008), 379–424. 10.1080/07370000802212669

[40] Jeffrey D. Karpicke and William R. Aue. 2015. The testing effect is alive and well with complex materials. *Educational Psychology Review* 27, 2 (2015), 317–326. https://doi.org/10.1007/s10648-015-9309-3

[41] Femke Kirschner, Fred Paas, and Paul A. Kirschner. 2009. A cognitive load approach to collaborative learning: United brains for complex tasks. *Educational Psychology Review* 21 (2009), 31–42. https://doi.org/10.1007/s10648-008-9095-2

[42] Paul A. Kirschner, John Sweller, and Richard E. Clark. 2006. Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist* 41, 2 (2006), 75–86. 10.1207/s15326985ep4102_1

[43] Paul A. Kirschner, John Sweller, Femke Kirschner, and Jimmy Zambrano R. 2018. From cognitive load theory to collaborative cognitive load theory. *International Journal of Computer-Supported Collaborative Learning* 13, 2 (2018), 213–233. https://doi.org/10.1007/s11412-018-9277-y

[44] Melina Klepsch, Florian Schmitz, and Tina Seufert. 2017. Development and validation of two instruments measuring intrinsic, extraneous, and germane cognitive load. *Frontiers in Psychology* 8 (2017), 1997. https://doi.org/10.3389/fpsyg.2017.01997

[45] Melina Klepsch and Tina Seufert. 2020. Understanding instructional design effects by differentiated measurement of intrinsic, extraneous, and germane cognitive load. *Instructional Science* 48, 1 (2020), 45–77. https://doi.org/10.1007/s11251-020-09502-9

[46] Torkel Klingberg. 2010. Training and plasticity of working memory. *Trends in Cognitive Sciences* 14, 7 (July 2010), 317–324. https://doi.org/10.1016/j.tics.2010.05.002

[47] Edwin M. Knorr. 2020. Worked examples, cognitive load, and exam assessments in a senior database course. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education (SIGCSE'20)*. Association for Computing Machinery, New York, NY, 612–618. https://doi.org/10.1145/3328778.3366915

[48] Andreas Korbach, Roland Brünken, and Babette Park. 2018. Differentiating different types of cognitive load: A comparison of different measures. *Educational Psychology Review* 30 (2018), 503–529. https://doi.org/10.1007/s10648-017-9404-8

[49] Jimmie Leppink. 2020. Revisiting cognitive load theory: Second thoughts and unaddressed questions. *Scientia Medica* 30 (2020), 1–8. https://doi.org/10.15448/1980-6108.2020.1.36918

[50] Jimmie Leppink, Fred Paas, Cees P. M. van der Vleuten, Tamara van Gog, and Jeroen J. G. van Merriënboer. 2013. Development of an instrument for measuring different types of cognitive load. *Behavior Research Methods* 45 (2013), 1058–72. https://doi.org/10.3758/s13428-013-0334-1

[51] Jimmie Leppink, Fred Paas, Tamara van Gog, Cees P. M. van der Vleuten, and Jeroen J. G. van Merriënboer. 2014. Effects of pairs of problems and examples on task performance and different types of cognitive load. *Learning and Instruction* 30 (2014), 32–42. https://doi.org/10.1016/j.learninstruc.2013.12.001

[52] Shu-Sheng Liaw. 2008. Investigating students' perceived satisfaction, behavioral intention, and effectiveness of e-learning: A case study of the Blackboard system. *Computers & Education* 51, 2 (Sept. 2008), 864–873. https://doi.org/10.1016/j.compedu.2007.09.005

[53] Vicki Likourezos and Slava Kalyuga. 2017. Instruction-first and problem-solving-first approaches: Alternative pathways to learning complex tasks. *Instructional Science* 45, 2 (2017), 195–219. https://doi.org/10.1007/s11251-016-9399-4

[54] Jia Lu, Slava Kalyuga, and John Sweller. 2020. Altering element interactivity and variability in example-practice sequences to enhance learning to write Chinese characters. *Applied Cognitive Psychology* 34, 4 (2020), 837–843. https://doi.org/10.1002/acp.3668

[55] Lauri Malmi, Judy Sheard, Simon, Roman Bednarik, Juha Helminen, Päivi Kinnunen, Ari Korhonen, Niko Myller, Juha Sorva, and Ahmad Taherkhani. 2014. Theoretical underpinnings of computing education research: What is the evidence? In *Proceedings of the 10th Annual Conference on International Computing Education Research (ICER'14)*. Association for Computing Machinery, New York, NY, 27–34. https://doi.org/10.1145/2632320.2632358

[56] Lauren Margulieux and Richard Catrambone. 2017. Using learners' self-explanations of subgoals to guide initial problem solving in app inventor. In *Proceedings of the 2017 ACM Conference on International Computing Education Research (ICER'17)*. Association for Computing Machinery, New York, NY, 21–29. https://doi.org/10.1145/3105726.3106168

[57] Lauren E. Margulieux, Paul Denny, Kathryn Cunningham, Michael Deutsch, and Benjamin R. Shapiro. 2021. When wrong is right: The instructional power of multiple conceptions. In *Proceedings of the 17th ACM Conference on International Computing Education Research (ICER'21)*. ACM, 184–196. 10.1145/3446871.3469750

[58] Lauren E. Margulieux, Briana B. Morrison, and Adrienne Decker. 2020. Reducing withdrawal and failure rates in introductory programming with subgoal labeled worked examples. *International Journal of STEM Education* 7 (2020), 1–16. https://doi.org/10.1186/s40594-020-00222-7

[59] Raina Mason, Simon, Graham Cooper, and Barry Wilks. 2016. Flipping the assessment of cognitive load: Why and how. In *Proceedings of the 2016 ACM Conference on International Computing Education Research (ICER'16)*. Association for Computing Machinery, New York, NY, 43–52. https://doi.org/10.1145/2960310.2960321

[60] Myrto F. Mavilidi and Lijia Zhong. 2019. Exploring the development and research focus of cognitive load theory, as described by its founders: Interviewing John Sweller, Fred Paas, and Jeroen van Merriënboer. *Educational Psychology Review* 31 (2019), 499–508. https://doi.org/10.1007/s10648-019-09463-7

[61] Richard E. Mayer. 2020. *Multimedia Learning* (3rd ed.). Cambridge University Press.

[62] Craig S. Miller and Amber Settle. 2011. When practice doesn't make perfect: Effects of task goals on learning computing concepts. *ACM Transactions on Computing Education* 11, 4 (Nov. 2011), Article 22, 16 pages. https://doi.org/10.1145/2048931.2048933

[63] Craig S. Miller and Amber Settle. 2019. Learning to get literal: Investigating reference-point difficulties in novice programming. *ACM Transactions on Computing Education* 19, 3 (May 2019), Article 28, 17 pages. https://doi.org/10.1145/3313291

[64] Ritayan Mitra, Karen S. McNeal, and Howard D. Bondell. 2016. Pupillary response to complex interdependent tasks: A cognitive-load theory perspective. *Behavior Research Methods* 49, 5 (2016), 1905–1919. https://doi.org/10.3758/s13428-016-0833-y

[65] Jan Moons and Carlos De Backer. 2013. The design and pilot evaluation of an interactive learning environment for introductory programming influenced by cognitive load theory and constructivism. *Computers & Education* 60, 1 (2013), 368–384. https://doi.org/10.1016/j.compedu.2012.08.009

[66] Roxana Moreno. 2010. Cognitive load theory: More food for thought. *Instructional Science* 38, 2 (2010), 135–141. https://doi.org/10.1007/s11251-009-9122-9

[67] Briana B. Morrison. 2016. *Replicating Experiments from Educational Psychology to Develop Insights into Computing Education: Cognitive Load as a Significant Problem in Learning Programming*. Ph.D. Dissertation. Georgia Institute of Technology.

[68] Briana B. Morrison. 2017. Dual modality code explanations for novices: Unexpected results. In *Proceedings of the 2017 ACM Conference on International Computing Education Research (ICERv'17)*. Association for Computing Machinery, New York, NY, 226–235. https://doi.org/10.1145/3105726.3106191

[69] Briana B. Morrison, Brian Dorn, and Mark Guzdial. 2014. Measuring cognitive load in introductory CS: Adaptation of an instrument. In *Proceedings of the 10th Annual Conference on International Computing Education Research (ICER'14)*. Association for Computing Machinery, New York, NY, 131–138. https://doi.org/10.1145/2632320.2632348

[70] Valentina Nachtigall, Katja Serova, and Nikol Rummel. 2020. When failure fails to be productive: Probing the effectiveness of productive failure for learning beyond STEM domains. *Instructional Science* 48, 6 (2020), 651–697. https://doi.org/10.1007/s11251-020-09525-2

[71] Greg L. Nelson and Amy J. Ko. 2018. On use of theory in computing education research. In *Proceedings of the 2018 ACM Conference on International Computing Education Research (ICER'18)*. Association for Computing Machinery, New York, NY, 31–39. https://doi.org/10.1145/3230977.3230992

[72] Bing Hiong Ngu, Huy P. Phan, Kian Sam Hong, and Hasbee Usop. 2016. Reducing intrinsic cognitive load in percentage change problems: The equation approach. *Learning and Individual Differences* 51 (2016), 81–90. https://doi.org/10.1016/j.lindif.2016.08.029

[73] Nargess Nourbakhsh, Yang Wang, Fang Chen, and Rafael A. Calvo. 2012. Using galvanic skin response for cognitive load measurement in arithmetic and reading tasks. In *Proceedings of the 24th Australian Computer-Human Interaction Conference (OzCHI'12)*. Association for Computing Machinery, New York, NY, 420–423. https://doi.org/10.1145/2414536.2414602

[74] Fred Paas. 1992. Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology* 84, 4 (1992), 429–434. https://doi.org/10.1037/0022-0663.84.4.429

[75] Fred Paas, Juhani E. Tuovinen, Huib Tabbers, and Pascal W. M. van Gerven. 2003. Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist* 38, 1 (2003), 63–71.

[76] Fred Paas, Jeroen J. G. van Merriënboer, and Jos J. Adam. 1994. Measurement of cognitive load in instructional research. *Perceptual and Motor Skills* 79, 1 (1994), 419–430. https://doi.org/10.2466/pms.1994.79.1.419

[77] Radek Pelánek and Tomáš Effenberger. 2020. Design and analysis of microworlds and puzzles for block-based programming. *Computer Science Education* 0, 0 (2020), 1–39. https://doi.org/10.1080/08993408.2020.1832813

[78] Kai Petersen, Robert Feldt, Shahid Mujtaba, and Michael Mattsson. 2008. Systematic mapping studies in software engineering. In *12th International Conference on Evaluation and Assessment in Software Engineering (EASE'08)*. 1–10. DOI: 10.14236/ewic/EASE2008.8

[79] Jan L. Plass and Slava Kalyuga. 2019. Four ways of considering emotion in cognitive load theory. *Educational Psychology Review* 31, 2 (2019), 339–359. https://doi.org/10.1007/s10648-019-09473-5

[80] James Prather, Raymond Pettit, Kayla McMurry, Alani Peters, John Homer, and Maxine Cohen. 2018. Metacognitive difficulties faced by novice programmers in automated assessment tools. In *Proceedings of the 2018 ACM Conference on International Computing Education Research (ICER'18)*. Association for Computing Machinery, New York, NY, 41–50. https://doi.org/10.1145/3230977.3230981

[81] James Prather, Raymond Pettit, Kayla Holcomb McMurry, Alani Peters, John Homer, Nevan Simone, and Maxine Cohen. 2017. On novices' interaction with compiler error messages: A human factors approach. In *Proceedings of the 2017 ACM Conference on International Computing Education Research (ICER'17)*. Association for Computing Machinery, New York, NY, 74–82. https://doi.org/10.1145/3105726.3106169

[82] Susan Reardon and Brendan Tangney. 2014. Smartphones, studio-based learning, and scaffolding: Helping novices learn to program. *ACM Transactions on Computing Education* 14, 4 (Dec. 2014), Article 23, 15 pages. https://doi.org/10.1145/2677089

[83] Endah Retnowati, Paul Ayres, and John Sweller. 2017. Can collaborative learning improve the effectiveness of worked examples in learning mathematics? *Journal of Educational Psychology* 109, 5 (2017), 666–679. https://psycnet.apa.org/doi/10.1037/edu0000167.

[84] Veerle Ross, Ellen M. M. Jongen, Weixin Wang, Tom Brijs, Kris Brijs, Robert A. C. Ruiter, and Geert Wets. 2014. Investigating the influence of working memory capacity when driving behavior is combined with cognitive load: An LCT study of young novice drivers. *Accident Analysis & Prevention* 62 (Jan. 2014), 377–387. https://doi.org/10.1016/j.aap.2013.06.032

[85] Katharina Scheiter. 2020. Embracing complexity in research on learning from examples and from problem solving. *Applied Cognitive Psychology* 34, 4 (2020), 906–911. https://doi.org/10.1002/acp.3702

[86] Wolfgang Schnotz and Christian Kürschner. 2007. A reconsideration of cognitive load theory. *Educational Psychology Review* 19 (2007), 469–508. https://doi.org/10.1007/s10648-007-9053-4

[87] Matthias Schwaighofer, Frank Fischer, and Markus Bühner. 2015. Does working memory training transfer? A meta-analysis including training conditions as moderators. *Educational Psychologist* 50, 2 (April 2015), 138–166. https://doi.org/10.1080/00461520.2015.1036274

[88] Stoo Sepp, Steven J. Howard, Sharon Tindall-Ford, Shirley Agostinho, and Fred Paas. 2019. Cognitive load theory and human movement: Towards an integrated model of working memory. *Educational Psychology Review* 31 (2019), 293–317. https://doi.org/10.1007/s10648-019-09461-9

[89] Justin L. Sewell, Christy K. Boscardin, John Q. Young, Olle ten Cate, and Patricia S. O'Sullivan. 2016. Measuring cognitive load during procedural skills training with colonoscopy as an exemplar. *Medical Education* 50, 6 (2016), 682–692. https://doi.org/10.1111/medu.12965

[90] Yu Shi, Natalie Ruiz, Ronnie Taib, Eric Choi, and Fang Chen. 2007. *Galvanic Skin Response (GSR) as an Index of Cognitive Load*. Association for Computing Machinery, New York, NY, 2651–2656. https://doi.org/10.1145/1240866.1241057

[91] Shin-Shing Shin. 2016. Evaluation of model driven architecture-based instruction for understanding phase transitions in object-oriented analysis and design. *ACM Transactions on Computing Education* 16, 4 (Aug. 2016), Article 17, 21 pages. https://doi.org/10.1145/2914797

[92] Tanmay Sinha and Manu Kapur. 2021. When problem solving followed by instruction works: Evidence for productive failure. *Review of Educational Research* 91, 5 (2021), 761–798. 10.3102/00346543211019105

[93] Ben Skudder and Andrew Luxton-Reilly. 2014. Worked examples in computer science. In *Proceedings of the 16th Australasian Conference on Computing Education (ACE'14) (CRPIT, Vol. 148)*, Jacqueline Whalley and Daryl D'Souza (Eds.). Australian Computer Society, 59–64.

[94] Alexander Skulmowski and Kate Man Xu. 2021. Understanding cognitive load in digital and online learning: A new perspective on extraneous cognitive load. *Educational Psychology Review* 34 (2022), 171–196. https://doi.org/10.1007/s10648-021-09624-7

[95] Mario Soflano, Thomas M. Connolly, and Thomas Hainey. 2015. An application of adaptive games-based learning based on learning style to teach SQL. *Computers & Education* 86 (2015), 192–211. https://doi.org/10.1016/j.compedu.2015.03.015

[96] Juha Sorva. 2012. *Visual Program Simulation in Introductory Programming Education*. Doctoral dissertation. Department of Computer Science and Engineering, Aalto University. https://aaltodoc.aalto.fi/bitstream/handle/123456789/3534/isbn9789526046266.pdf.

[97] Juha Sorva and Otto Seppälä. 2014. Research-based design of the first weeks of CS1. In *Proceedings of the 14th Koli Calling International Conference on Computing Education Research (Koli Calling'14)*. Association for Computing Machinery, New York, NY, 71–80. https://doi.org/10.1145/2674683.2674690

[98] Shelsey Sullivan, Hillary Swanson, and John Edwards. 2021. *Student Attitudes Toward Syntax Exercises in CS1*. Association for Computing Machinery, New York, NY, 782–788. https://doi.org/10.1145/3408877.3432399

[99] John Sweller. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive Science* 12, 2 (1988), 257–285.

[100] John Sweller. 2010. Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational Psychology Review* 22 (2010), 123–138. https://doi.org/10.1007/s10648-010-9128-5
[101] John Sweller and Fred Paas. 2017. Should self-regulated learning be integrated with cognitive load theory? A commentary. *Learning and Instruction* 51 (2017), 85–89. https://doi.org/10.1016/j.learninstruc.2017.05.005
[102] John Sweller, Jeroen J. G. van Merriënboer, and Fred Paas. 1998. Cognitive architecture and instructional design. *Educational Psychology Review* 10 (1998), 251–296. https://doi.org/10.1023/A:1022193728205
[103] John Sweller, Jeroen J. G. van Merriënboer, and Fred Paas. 2019. Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review* 31, 2 (2019), 261–292. https://doi.org/10.1007/s10648-019-09465-5
[104] Zaidatun Tasir and Ong Chiek Pin. 2012. Trainee teachers' mental effort in learning spreadsheet through self-instructional module based on Cognitive Load Theory. *Computers & Education* 59, 2 (2012), 449–465. https://doi.org/10.1016/j.compedu.2012.01.009
[105] Sigmund Tobias and Thomas M. Duffy. 2009. *Constructivist Instruction: Success or Failure?* Taylor & Francis.
[106] Pascal W. M. van Gerven, Fred Paas, Jeroen J. G. van Merrienboer, and Henk G. Schmidt. 2004. Memory load and the cognitive pupillary response in aging. *Psychophysiology* 41, 2 (March 2004), 167–174. https://doi.org/10.1111/j.1469-8986.2003.00148.x
[107] Jeroen J. G. van Merriënboer. 1990. Strategies for programming instruction in high school: Program completion vs program generation. *Journal of Educational Computing Research* 6, 3 (1990), 265–285.
[108] Camilo Vieira, Alejandra J. Magana, Michael L. Falk, and R. Edwin Garcia. 2017. Writing in-code comments to self-explain in computational science and engineering education. *ACM Transactions on Computing Education* 17, 4 (Aug. 2017), Article 17, 21 pages. https://doi.org/10.1145/3058751
[109] Robert R. Whelan. 2007. Neuroimaging of cognitive load in instructional multimedia. *Educational Research Review* 2, 1 (2007), 1–12. https://doi.org/10.1016/j.edurev.2006.11.001
[110] Joseph B. Wiggins, Fahmid M. Fahid, Andrew Emerson, Madeline Hinckle, Andy Smith, Kristy Elizabeth Boyer, Bradford Mott, Eric Wiebe, and James Lester. 2021. *Exploring Novice Programmers' Hint Requests in an Intelligent Block-Based Coding Environment.* Association for Computing Machinery, New York, NY, 52–58. https://doi.org/10.1145/3408877.3432538
[111] Yu-Fang Yeh, Mei-Chi Chen, Pi-Hsia Hung, and Gwo-Jen Hwang. 2010. Optimal self-explanation prompt design in dynamic multi-representational learning environments. *Computers & Education* 54, 4 (2010), 1089–1100. https://doi.org/10.1016/j.compedu.2009.10.013
[112] Alexander Seeshing Yeung, Cynthia Fong King Lee, Isabel Maria Pena, and Jeni Ryde. 2000. Toward a subjective mental workload measure. In *Paper Presented at the International Congress for School Effectiveness and Improvement.* http://files.eric.ed.gov/fulltext/ED446117.pdf access on 16.04.2022.
[113] John Q. Young, Majnu John, Krima Thakker, Karen Friedman, Rebekah Sugarman, Justin L. Sewell, and Patricia S. O'Sullivan. 2021. Evidence for validity for the Cognitive Load Inventory for Handoffs. *Medical Education* 55, 2 (2021), 222–232. https://doi.org/10.1111/medu.14292
[114] John Q. Young, Krima Thakker, Majnu John, Karen Friedman, Rebekah Sugarman, Jeroen J. G. van Merriënboer, Justin L. Sewell, and Patricia S. O'Sullivan. 2021. Exploring the relationship between emotion and cognitive load types during patient handovers. *Advances in Health Sciences Education* 0123456789 (2021), 1–27. https://doi.org/10.1007/s10459-021-10053-y
[115] Albina Zavgorodniaia, Rodrigo Duran, Arto Hellas, Otto Seppala, and Juha Sorva. 2020. Measuring the cognitive load of learning to program: A replication study. In *United Kingdom & Ireland Computing Education Research Conference (UKICER'20).* Association for Computing Machinery, New York, NY, 3–9. https://doi.org/10.1145/3416465.3416468
[116] Albina Zavgorodniaia, Arto Hellas, Otto Seppälä, and Juha Sorva. 2020. Should explanations of program code use audio, text, or both? A replication study. In *Proceedings of the 20th Koli Calling International Conference on Computing Education Research (Koli Calling'20).* Association for Computing Machinery, New York, NY, Article 5, 10 pages. https://doi.org/10.1145/3428029.3428050
[117] Albina Zavgorodniaia, Artturi Tilanterä, Ari Korhonen, Otto Seppälä, Arto Hellas, and Juha Sorva. 2021. Algorithm visualization and the elusive modality effect. In *Proceedings of the 17th ACM Conference on International Computing Education Research (ICER'21).* Association for Computing Machinery, New York, N, 368–378. https://doi.org/10.1145/3446871.3469747
[118] Robert Z. Zheng (Ed.). 2017. *Cognitive Load Measurement and Application.* Routledge.
[119] Rui Zhi, Min Chi, Tiffany Barnes, and Thomas W. Price. 2019. Evaluating the effectiveness of Parsons problems for block-based programming. In *Proceedings of the 2019 ACM Conference on International Computing Education Research (ICER'19).* Association for Computing Machinery, New York, NY, 51–59. https://doi.org/10.1145/3291279.3339419

Received January 2021; revised September 2021; accepted October 2021