Heinonen, Ava; Hellas, Arto

Exploring the instructional efficiency of representation and engagement in online learning materials

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

DOI:
10.1145/3416465.3416470

Published: 03/09/2020

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

Please cite the original version:
Heinonen, A., & Hellas, A. (2020). Exploring the instructional efficiency of representation and engagement in online learning materials. In *UKICER 2020 - Proceedings of the 2020 Conference on United Kingdom and Ireland Computing Education Research* (pp. 38-44). ACM. https://doi.org/10.1145/3416465.3416470

Exploring the Instructional Efficiency of Representation and Engagement in Online Learning Materials

Ava Heinonen
ava.heinonen@aalto.fi
Aalto University
Espoo, Finland

Arto Hellas
arto.hellas@aalto.fi
Aalto University
Espoo, Finland

ABSTRACT

Using two 2 x 3 in-situ experiments in an introductory programming course, we study the effect of representation and engagement on the instructional efficiency of learning materials. In the experiments, we controlled for used representation and the level of engagement and accounted for prior experience and prior cognitive effort. We observe that *analogical representations* with little engagement are more beneficial for those already familiar with the topic. No significant effect from engagement or prior experience was observed when students studied using *traditional representations.* Low cognitive effort before studying was related to studying being less cognitively demanding, regardless of the condition. No single way of presenting information seems to work better or worse universally for all participants.

CCS CONCEPTS

• Social and professional topics → Computer science education; **• Applied computing → Interactive learning environments**; **E-learning**.

KEYWORDS

instructional materials, instructional efficiency, cognitive load, prior experience, learning programming, engagement taxonomy, multimedia learning materials

ACM Reference Format:
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), September 3–4, 2020, Glasgow, United Kingdom.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3416465.3416470

1 INTRODUCTION

A variety of pedagogical practices that can increase students' retention and learning in introductory programming courses exist [19, 30, 41]. These include crafting learning materials that are relevant to the particular audience, such as contextualizing the topic using e.g. media as the main theme [12], as well as using practices that pace students' learning and increase classroom collaboration [30]. One particular form of support is the instructional materials. The way how materials are structured and presented can improve learning outcomes [21, 38], and the type of engagement constructed into the learning materials, e.g. reading vs answering questions, also has an effect on learning outcomes [26, 35, 42].

In the last decades, there has been an increase in the number of blended and online opportunities for learning programming. Currently, there are several initiatives both in the UK and around the world that aim to bring computing to the masses [2, 6, 10]. With the increasing number of participants and consequently a broader variety of learners' backgrounds, it is crucial to re-assess how instructional materials should be constructed. In this work, we are interested in the role of representation and engagement in introductory programming materials.

Our work focuses on the instructional efficiency of learning materials in an introductory programming course. Instructional efficiency is measured using a procedure proposed by Tuovinen and Paas [37]. In our experiments, students were shown instructional materials with two different representations (*analogical* vs *traditional*) and three different engagement levels (plain text, static visualizations and text, interactive slideshow with embedded text and questions). Our goal is to *quantify the combined effect of representation and engagement on the instructional efficiency of instructional materials used in learning programming.*

The closest matches to our work are the works on engagement taxonomy, e.g. [14, 25], observing that increased engagement leads to better learning, the works on notional machines [9], which study pedagogical devices for teaching programming, and the work on cognitive load on e-learning, e.g. [40], suggesting ways for constructing instructional materials. Our work takes steps towards combining these threads of research.

This article is structured as follows. Next, we outline the theoretical frameworks on which this work builds, including relevant theories from learning and cognition and designing instructional materials. Section 3 outlines the methodology of this research, including the research questions and details of the context in which this study took place. Section 4 outlines the results of our experiments, which are further discussed in Section 5. Finally, Section 6 summarizes our findings and outlines directions for future research.

2 BACKGROUND

2.1 Learning and cognition

The most widely used framework for representing human cognitive architecture divides memory into three sections: sensory memory, working memory, and long-term memory [15, 22, 27]. Learning

UKICER '20, September 3–4, 2020, Glasgow, United Kingdom

ACM ISBN 978-1-4503-8849-8/20/09…$15.00
https://doi.org/10.1145/3416465.3416470

occurs when information is processed in working memory and stored in the long term memory. While long-term memory is almost unlimited in capacity, the working memory can hold only a few pieces of information at a time [22]. Learning complex knowledge involves processing information in working memory, chunking that information into information structures, i.e. *schemas* [36], which are then stored into the long-term memory. A schema can be later retrieved from long-term memory, and although it may consist of complex structured information, it will take only "a single slot" from working memory.

In essence, schemas govern the way we acquire new knowledge [20]. If no existing schema for a problem exists, we apply generic problem-solving strategies to find appropriate solutions, creating a new schema. If a relevant schema exists, we recall it to our working memory and either remodel the existing schema to incorporate any new relevant information, or if the new information does not fit within our existing schema, reject it [20, 29].

Processing too many pieces of information at the same time overloads the working memory capacity, which leads to a situation where information cannot be processed effectively, compromising learning [29]. The burden that processing of information causes to the working memory is referred to as *cognitive load.* Cognitive load can be caused by properties of information itself (intrinsic cognitive load), properties of instructional material (extraneous cognitive load), or the burden to the working memory caused by processing information (germane cognitive load) [24].

2.2 Instructional materials

Ideally, instructional materials should aim to decrease extraneous and intrinsic cognitive load and promote germane cognitive load. Multiple ways of how the design of instructional material affects cognitive load have been identified [39]. For example, intrinsic cognitive load can be managed using a low-to-high-fidelity strategy, increasing the number of interacting elements over time by increasing the details in the material [40]. This strategy aids novices with no existing schema on the topic as low-fidelity material with less interacting elements causes less intrinsic cognitive load.

In the context of computer science education research and program visualizations, one particular theory of interest to us is the engagement taxonomy. It argues that learning from visualizations is improved when the visualization engages students in an active learning activity [26]. In principle, higher levels of engagement between a student and a visualization lead to higher levels of understanding and consequently better learning [26]. The taxonomy details six levels of engagement between a student and a visualization, ranging from no viewing, where there is no visualization to engage with, to presenting, where self-constructed visualizations are presented to others [25]. In our work, we consider the first three levels of the engagement taxonomy, which are: (1) No viewing (no visualization), (2) viewing (visualization is passively viewed), and (3) responding (student is prompted to interact with the visualization) [26].

2.3 Notional machines

Teaching and learning programming is heavily intertwined with learning the inner workings of a computer – or an abstraction of it – to form an understanding of how a program is executed. In computer science education research, such abstractions are often referred to as *notional machines* [7, 8, 33], originally framed as "*the idealised model of the computer implied by the constructs of the programming language*" [8].

More recent research refers to the notional machine as a pedagogical device used for supporting the understanding of some aspect of a program or programming, often used in the context of learning e.g. a programming concept [9]. Such notional machines include e.g. tools that generate visualizations from code such as Jeliot [18], UUhistle [34] and Online Python Tutor [11] as well as manually drawn visualizations present e.g. in textbooks. In our work, we use two types of hand-drawn visualizations, where one type links the learned concepts to real-world phenomena, and the other type follows a more traditional textbook-like format with arrows and boxes.

2.4 Evaluating instructional efficiency

Instructional efficiency can be evaluated in multiple ways. On a macro level, one can use course pass rates [32] to measure the instructional efficiency of teaching approaches [41]. A step further, leading to micro-level assessment, is the use of pre- and post-testing to assess learning gain (difference in knowledge between pre- and post-testing) of instructional activity. Neither of these approaches, however, account for the mental effort invested during learning.

Several approaches for measuring the mental effort and cognitive load exist. These range from subjective rating scales [13, 23, 27, 28] to dual-task methodology, where the performance in a secondary task is used as a proxy for cognitive load [3, 31], and to using physiological measures such as skin conductivity or heart rate variance for assessing mental effort [1, 27].

The cognitive load caused by a learning task and the learning outcomes can be used to calculate instructional efficiency scores [39]. They can be used to measure the quality of learning outcomes – acquisition of efficient cognitive schemata is indicated by combinations of high performance and low mental effort [39]. In our work, for assessing instructional efficiency, we use the instructional efficiency formula by Tuovinen and Paas [37] which is as follows.

$$Efficiency = \frac{P - E_l - E_t}{\sqrt{3}}$$

In the formula, $Efficiency$ stands for instructional efficiency, P is normalized test performance, E_l is normalized learning effort, and E_l is normalized test effort.

3 METHODOLOGY

3.1 Research questions

The research questions of our work are as follows:

- **RQ1** How do representation and engagement level affect the instructional efficiency of learning materials?
- **RQ2** How does prior experience on the topic under study affect the instructional efficiency of the learning material?
- **RQ3** How does cognitive effort prior to studying affect the instructional efficiency of the learning material?

In this work, the term *representation* refers to two representation types (see Fig. 1): *Analogical representations* anchor the presentation to real-world artefacts and concepts such as presenting a list as a (shopping) list on a notepaper. *Traditional representations* follow a format typically seen in programming text-books and program visualization tools, such as presenting a list as a horizontally aligned set of boxes. Traditional representations are also the main format in the context in which the study was conducted. Conversely, the term *engagement level* refers to the Engagement taxonomy (discussed in subsection 2.2). Our focus is on the first three levels of the engagement taxonomy; no viewing, viewing, and responding. The no viewing level corresponds to plain text (i.e. no visualizations), viewing corresponds to plain text and static visualizations, and responding refers to interactive slideshow with embedded text, visualizations, and questions.

3.2 Study context

The study was conducted in two instances of a 14-week introductory Java programming course, offered in Finnish. The difference between the course instances was the delivery format; one was offered fully online, while the other had voluntary lectures and labs. Both instances use the same materials: an online workbook that contains text sections detailing programming concepts blended with questionnaires and programming exercises. When students create an account to the course system, they are asked for permission to use their data for research. Data only from students who gave permission for research was used for this study. In this article, we have merged the responses from the two instances and refer to them as one course.

3.3 Experiments and participants

Two experiments were conducted on the 11th week of the course. The part teaches internal implementations of ArrayList and HashMap and working with multidimensional data using arrays. Experimental instructional materials were embedded into the two chapters, referred later to as the "Hash table experiment" and the "Multidimensional arrays experiment". Students in the course had previously learned to work with Java's HashMap class and with one-dimensional arrays.

The study followed a 2 x 3 design with representation type and engagement level as the factors. A total of 123 unique students were included in the study. 121 students completed the hash table experiment and 114 completed the multidimensional arrays experiment. Students were randomly assigned to one of the six treatment groups (controlled by a learning management system), resulting in group sizes shown in Table 1.

The materials used for the experiment consisted of a pre-test questionnaire, treatment group -specific learning materials, and a post-test questionnaire. Students received course points for answering the questionnaires.

The pre-test questionnaire contained three multiple-choice questions about the experiment topic, a self-evaluation prior knowledge question about the topic, the NASA Task Load Index questionnaire (NASA-TLX) [13] asking students to rate their mental and physical effort during the last hour, and the Paas mental effort scale [28] used for rating their mental effort answering the pre-test questionnaires.

Once the students had studied the instructional materials, they were given a post-test questionnaire. Students were asked to rate

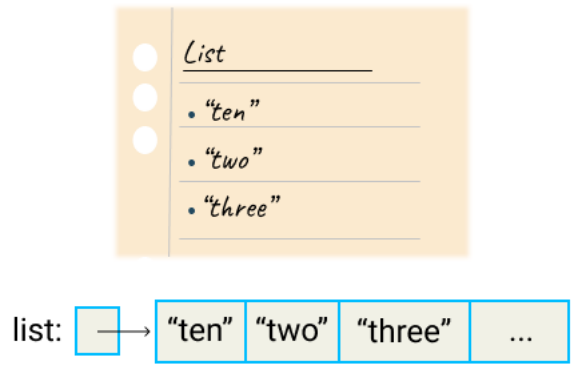

Figure 1: Example showing a list using analogical representation style (on top) and traditional representation style (at the bottom).

Table 1: Treatment groups and their sizes. Ht refers to Hash table experiment, and Ma Multidimensional arrays experiment. Total column has the total number of participants in each group.

Group no.	Representation	Engagement	Ht	Ma	Total
1	Analogical	No viewing	16	15	31
2	Analogical	Viewing	15	12	27
3	Analogical	Responding	24	23	47
4	Traditional	No viewing	30	21	51
5	Traditional	Viewing	16	21	37
6	Traditional	Responding	20	22	42

the mental effort of studying the material using the Paas mental effort scale, answer three multiple-choice questions that assessed students understanding of the topic, and rate the mental effort of answering the multiple-choice questionnaire again using the Paas mental effort scale.

3.4 Analysis

Instructional efficiency was measured using the approach proposed by Tuovinen and Paas [37] (discussed in subsection 2.4). To measure the learning gain for each treatment group, the average difference between the pre- and post-test scores was calculated.

Statistical significance of the effects was tested using a linear model. The dependent variable was instructional efficiency, and the explanatory variables were treatment group, experiment, prior knowledge, pre-test score, pre-test cognitive load, NASA-TLX answers, and interaction of treatment group and experiment. Interactions among the other explanatory variables were examined, but they were not statistically significant. Moreover, the homogeneity of the variance and the normality of the distribution of the residual term was checked using scatterplots and normal probability plots.

Unless otherwise noted, group comparisons are conducted using ANOVA. When reporting results, up to 4 digits of the F-test statistic are presented, and the p-value is rounded to three decimals. When discussing statistical significance, we use $p < 0.05$ as a boundary for statistical significance. When reporting p-values, we do not correct for multiple comparisons, but instead outline all conducted tests for transparency.

4 RESULTS

4.1 Overview

Table 2 describes the mean prior knowledge, pre-test score, pre-test cognitive load, study cognitive load, post-test score, test cognitive load and learning gain for combined data and for both experiments.

Table 2: Mean values of measured variables in the combined data set, Hash table (Ht) experiment, and Multidimensional arrays (Ma) experiment.

Measurement (max)	Combined	Ht	Ma
Prior knowledge (3)	1.39	1.96	0.78
Pre-test score (3)	1.14	0.93	1.35
Pre-test cognitive load (9)	2.98	3.22	2.73
Study cognitive load (9)	3.30	3.63	2.96
Post-test score (3)	2.46	2.70	2.19
Post-test cognitive load (9)	2.99	2.77	3.23
Learning gain (3)	1.32	1.77	0.84

4.2 Engagement and representation

When combining data from both experiments, there were no statistically significant differences between the treatment groups in mean instructional efficiency ($F(5,220)=0.6406$, $p=0.669$) or learning gain ($F(5,220)=0.6483$, $p=0.663$). Table 3 describes the mean instructional efficiency, study cognitive load, test cognitive load and post-test scores by treatment group for the two experiments.

The differences in mean instructional efficiency between the treatment groups in the hash table experiment were not statistically significant ($F(5,112)=2.166$, $p=0.064$), and there were no statistically significant differences between the mean post-test scores ($F(5,115)=1.637$, $p=0.156$) or learning gain ($F(5,115)=1.249$, $p=0.291$) between the treatment groups.

In the multidimensional arrays experiment the differences in mean instructional efficiency ($F(5,105)=0.8264$, $p=0.510$) or in mean learning gain ($F(5,108)=0.5379$, $p=0.747$) between the treatment groups were also not statistically significant.

4.3 Prior knowledge

Three measures of prior knowledge were used: self-reported prior knowledge, pre-test score, and experiment identifier. In the pre-test questionnaire, participants rated their knowledge on the upcoming topic and answered a multiple-choice questionnaire about the topic, providing prior knowledge and pre-test score respectively. The participants' backgrounds coming into the experiments were also different - they had studied an introductory chapter on hash tables previously on the course, which gave all participants some

Table 3: Instructional efficiency, study cognitive load, test cognitive load and test result means by treatment group. Group labels are A=Analogical, T=Traditional, N = No viewing, V = Viewing, and R Responding.

Hash table experiment				
Group	Inst. eff.	Study CL	Test CL	Post-test score
A-N	0.75	2.94	2.19	2.88
A-V	0.64	3.20	2.53	3.00
A-R	0.03	3.38	3.04	2.55
T-N	-0.13	4.00	3.03	2.61
T-V	-0.01	3.94	2.94	2.70
T-R	0.12	4.00	2.55	2.70
Multidimensional arrays experiment				
Group	Inst. eff.	Study CL	Test CL	Post-test score
A-N	-0.27	3.07	3.27	2.13
A-V	-0.74	3.42	3.75	1.92
A-R	0.05	3.09	2.91	2.40
T-N	-0.15	2.71	3.29	2.31
T-V	-0.17	2.76	3.29	2.31
T-R	-0.10	2.91	3.14	2.22

prior knowledge on the topic, while the they had not studied multidimensional arrays on the course yet, making it a new topic to majority of the participants.

In the hash table experiment, from the 121 participants, 116 evaluated their prior knowledge as high, 5 evaluated their prior knowledge as mediocre, and no participant considered that they had no prior knowledge. In the multidimensional arrays experiment from the 114 participants, 18 evaluated their prior knowledge as high, 53 evaluated their prior knowledge as mediocre, and 43 considered that they had no prior knowledge. Self-evaluated prior knowledge had no significant effect on the instructional efficiency.

When combining data from both experiments, there were statistically significant differences in mean instructional efficiency between participants with different pre-test scores ($F(3,220)=9.48$, $p=0.002$). Similarly, statistically significant differences were observed in the hash table experiment ($F(1,112)=5.00$, $p=0.003$). Post-hoc comparisons using Tukey HSD indicated that the mean instructional efficiency was significantly different for participants with pre-test score of 0 ($M=-0.27$ $SD=1.53$ in combined data, $M=0.07$ $SD=1.53$ in hash table experiment) than for participants with pre-test score of 3 ($M=0.25$ $SD=1.23$ in combined data, $M=0.80$ $SD=1.18$ in hash table experiment). In the multidimensional arrays experiment the differences were not statistically significant ($F(1,105)=3.47$, $p=0.065$). The pre-test score did not have any significant interaction effect with the treatment group in any data-set. Figure 2 shows instructional efficiency by pre-test score for both experiments.

The interaction between the treatment group and the experiment ($F(5,210)=3.89$, $p=0.002$) is shown in Figure 3. Post-hoc comparison using Tukey HSD indicates significant differences in mean instructional efficiency of treatment groups 1 and 2 between the two experiments.

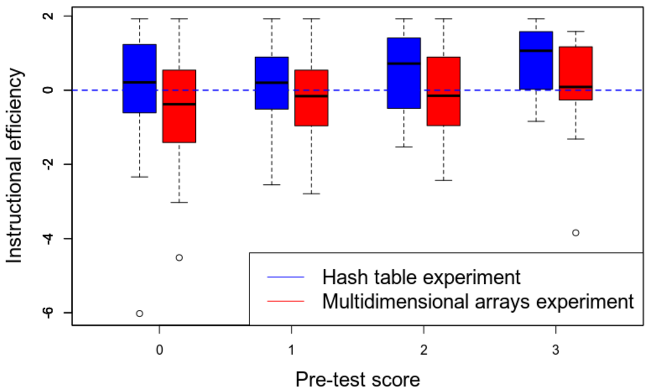

Figure 2: Instructional efficiency by pre-test score for both experiments

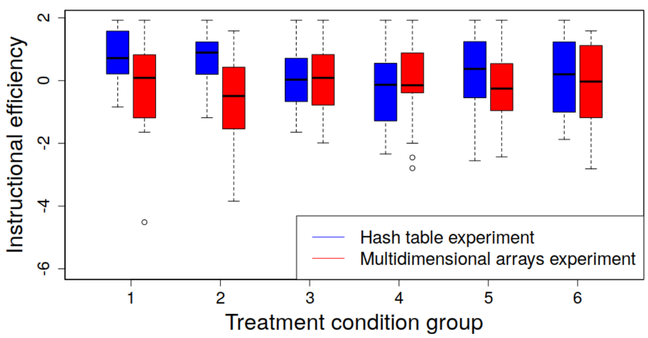

Figure 3: Instructional efficiency by treatment group for both experiments. Treatment groups 1-3 correspond to analogical representations with No viewing, Viewing, and Responding. Treatment groups 4-6 correspond to traditional representations with No viewing, Viewing, and Responding.

4.4 Prior cognitive load

Participants' cognitive load prior to the test was measured using two approaches. First, participants reported on their effort invested during the last hour using the NASA-TLX questionnaire. After answering the pre-test questions participants were also asked to rate the mental effort of answering the questions.

There were no statistically significant differences in mean instructional efficiency between participants who reported different levels of prior cognitive load in the NASA-TLX questionnaire.

There were statistically significant differences in mean instructional efficiency between participants who reported different levels of cognitive load related to responding to the pre-test questions ($F(8,210)=84.010$, $p < 0.001$). Post-hoc comparisons using Tukey HSD confirm that there are significant differences in mean instructional efficiency between participants who reported low cognitive load answering the pre-test questions and participants who reported high cognitive load. No significant interaction effect was observed between the cognitive load related to responding to the pre-test questions and the treatment group. Figure 4 shows the relationship of instructional efficiency and pre-test cognitive load.

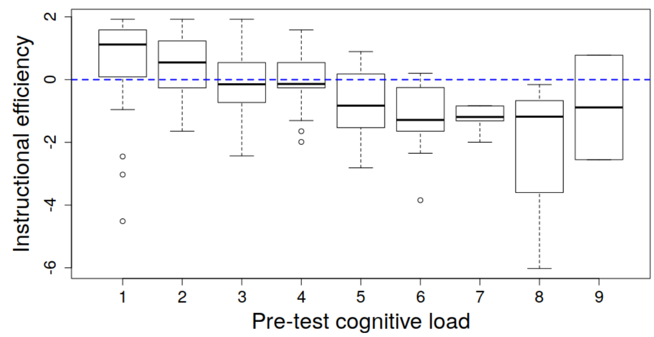

Figure 4: Instructional efficiency by pretest cognitive load score (combined)

When considering the hash table experiment and the multidimensional arrays experiment separately, there are also statistically significant differences in mean instructional efficiency between participants who reported different levels of cognitive load answering the pre-test questions ($F(8,112)=63.56, p<0.001$ in the hash table arrays experiment and $F(8,105)=34.16, p < 0.001$ in the multidimensional arrays experiment). Again post-hoc comparisons using Tukey HSD confirm that there are significant differences in mean instructional efficiency between participants who reported low cognitive load answering the pre-test questions and participants who reported high cognitive load.

5 DISCUSSION

5.1 Engagement and representation

Concerning the compound effect of engagement and representation, we expected that higher engagement would lead to greater learning efficiency, evidenced by lower test cognitive load and higher test results for both representation types. We also expected analogical representations to lead to lower study cognitive load for all engagement levels, leading to higher instructional efficiency.

Our expectations were largely proven wrong. We did not identify statistically significant differences between the treatment groups when looking at the combined data. This indicates that although certain treatment groups performed better for certain subcategories of participants, no single way of presenting information seems to work better or worse universally for all participants.

5.2 Prior knowledge and instructional efficiency

Concerning the effect of prior knowledge on instructional efficiency, we expected that analogical representations would outperform traditional representations for those with low or no prior knowledge.

When the data from the two experiments was analyzed together, we did not see any differences between the treatment groups for participants with different levels of prior knowledge but high pre-test scores led to better instructional efficiency overall. We measured significant differences in the instructional efficiency of the no viewing and viewing groups with analogical representation between the two experiments. In the hash table experiment, analogical representations with no viewing outperformed all other groups, while higher engagement levels led to a decrease in instructional efficiency. Overall we measured lower study cognitive load on analogical representations than on traditional representations.

For the multidimensional arrays experiment, where students mostly had no previous knowledge on the topic, we saw lower instructional efficiency in treatment groups 1 and 2 compared to the hash table experiment, but within the experiment no statistically significant differences between the treatment groups were identified. For both representation types, the viewing group performed slightly worse than the no viewing or responding groups, and the responding group performed the best. For the multidimensional arrays experiment, we measured lower study cognitive load on traditional representations than on analogical representations.

We hypothesize that the high performance of analogical representations with no viewing in the hash table test can be explained by the expertise reversal effect [17]. The analogical representations are easily understandable for participants with some prior knowledge and visual or interactive elements may be redundant and hinder learning. Overall, the results between the experiments indicate that participants with higher self-reported prior knowledge achieve higher instructional efficiency with analogical representations in a text format. On the other hand, participants with low self-reported prior knowledge achieve higher instructional efficiency on traditional representations that prompt students to engage with them.

5.3 Prior cognitive load and instructional efficiency

Cognitive load caused by previous tasks can limit the available cognitive resources [4, 5]. Thus we expected that participants reporting high prior cognitive load would benefit from the limited intrinsic cognitive load of analogical representations. Our expectations were proven wrong. Participants' rating of their mental effort expended before the test did not have any effect on the instructional efficiency.

High cognitive load on answering the pre-test questionnaire led, in general, to worse instructional efficiency. There was no difference between the treatment groups which indicates that no type of instructional material was better or worse for participants with high pre-test cognitive load. We hypothesize that this, too, could be related to previous knowledge: it is possible that answering questions is less demanding for those who already know the topic.

5.4 Limitations

The study was conducted in-situ within one course. We do not know how the participants completed the experiments and cannot, for example, make claims about the generalizability of the results, account for factors present in the participants' environment, and we also do not know to what extent the participants were engaged to studying the learning materials. In addition, we do not know if participants in the different groups used the same amount of time for studying.

We used instructional efficiency as a measure of the goodness of the instructional material. The time-tested cognitive load instrument and the instructional efficiency formula by Paas [27] do not, however, differentiate between the types of cognitive load. High study cognitive load could, in some situations, be germane cognitive load, and thus high study cognitive load does not always mean ineffective learning. We also acknowledge that there is research that considers that cognitive load has only two factors, intrinsic load and extraneous load (e.g. [16]).

It is also possible that the representations used in the materials do not match the view of the world of the students, i.e. what the material authors considered as analogical real-world examples may not match the views of the students. We sought to mitigate this concern by testing the materials with our colleagues before the study, but it is possible that some of the differences between the representation types can be explained by differences in the material design. Moreover, as the study was conducted close to the end of the introductory programming course, it is possible that the students were already accustomed to the default representations, which were traditional. That is, it is possible that the analogical representations may have been better in some cases, but that the students may have been used to the traditional representations, and thus have had to spend more effort on studying the analogical representations.

6 CONCLUSION

We studied the compound effect of representation and engagement on the instructional efficiency of an online instructional material in the context of a university-level introductory programming course. To summarize, our research questions and answers are as follows.

RQ1: How do representation and engagement level affect the instructional efficiency of learning materials? **Answer:** When combining data from both experiments, none of the treatments outperformed others in terms of learning gain or instructional efficiency.

RQ2 How does prior experience on the topic under study affect the instructional efficiency of the learning material? **Answer:** In the hash table experiment, where students had previous knowledge on the topic, analogical representations without engagement yielded the highest instructional efficiency. At the same time, when participants had no previous experience on the topic in the multidimensional arrays experiment, responding conditions performed the best. In addition, lower study cognitive load was observed on traditional representations.

RQ3 How does cognitive effort prior to studying affect the instructional efficiency of the learning material? **Answer:** While prior effort, when measured using the NASA-TLX scale, had no impact on learning gain or instructional efficiency, with the exception of prior physical exertion marginally improving instructional efficiency, cognitive load related to answering the pre-test questionnaire contributed significantly (negatively) to instructional efficiency of the learning materials.

Overall, we did not identify any single way of presenting information that would be universally good for all participants. Based on our observations, it is possible that adjusting representation and engagement level for individual students could lead to higher instructional efficiency. As a part of our future work, we are looking into replicating our experiments in another course material as well as studying the differences in the representations in more detail: for example, the traditional representations often also imply something about the underlying memory model, which may be absent in the analogical representations – if we choose one representation type over another, what are the implications in the longer run?

REFERENCES

[1] Lauri Ahonen, Benjamin Cowley, Jari Torniainen, Antti Ukkonen, Arto Vihavainen, and Kai Puolamäki. 2016. Cognitive collaboration found in cardiac physiology: Study in classroom environment. *PloS one* 11, 7 (2016), e0159178.

[2] Neil CC Brown, Sue Sentance, Tom Crick, and Simon Humphreys. 2014. Restart: The resurgence of computer science in UK schools. *ACM Transactions on Computing Education (TOCE)* 14, 2 (2014), 1–22.

[3] Roland Brünken, Jan L Plass, and Detlev Leutner. 2003. Direct measurement of cognitive load in multimedia learning. *Educational psychologist* 38, 1 (2003), 53–61.

[4] Ouhao Chen, Juan C Castro-Alonso, Fred Paas, and John Sweller. 2018. Extending cognitive load theory to incorporate working memory resource depletion: evidence from the spacing effect. *Educational Psychology Review* 30, 2 (2018), 483–501.

[5] Ouhao Chen and Slava Kalyuga. 2020. Cognitive Load Theory, Spacing Effect, and Working Memory Resources Depletion: Implications for Instructional Design. In *Form, Function, and Style in Instructional Design: Emerging Research and Opportunities*. IGI Global, 1–26.

[6] Tom Crick and Sue Sentance. 2011. Computing at school: stimulating computing education in the UK. In *Proceedings of the 11th Koli Calling International Conference on Computing Education Research*. 122–123.

[7] Benedict Du Boulay. 1986. Some difficulties of learning to program. *Journal of Educational Computing Research* 2, 1 (1986), 57–73.

[8] Benedict du Boulay, Tim O'Shea, and John Monk. 1981. The black box inside the glass box: presenting computing concepts to novices. *International Journal of Man-Machine Studies* 14, 3 (1981), 237–249.

[9] Sally Fincher, Johan Jeuring, Craig S. Miller, Peter Donaldson, Benedict du Boulay, Matthias Hauswirth, Arto Hellas, Felienne Hermans, Colleen Lewis, Andreas Mühling, Janice L. Pearce, and Andrew Petersen. 2020. Capturing and Characterising Notional Machines. In *Proceedings of the 2020 ACM Conference on Innovation and Technology in Computer Science Education* (Trondheim, Norway) *(ITiCSE '20)*. Association for Computing Machinery, New York, NY, USA, 502–503. https://doi.org/10.1145/3341525.3394988

[10] Joanna Goode, Julie Flapan, and Jane Margolis. 2018. Computer science for all. *Diversifying digital learning: Online literacy and educational opportunity* (2018), 45.

[11] Philip J Guo. 2013. Online python tutor: embeddable web-based program visualization for cs education. In *Proceeding of the 44th ACM technical symposium on Computer science education*. 579–584.

[12] Mark Guzdial. 2003. A media computation course for non-majors. In *Proceedings of the 8th annual conference on Innovation and technology in computer science education*. 104–108.

[13] 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.

[14] Tim N Höffler and Detlev Leutner. 2007. Instructional animation versus static pictures: A meta-analysis. *Learning and instruction* 17, 6 (2007), 722–738.

[15] Weidong Huang, Peter Eades, and Seok-Hee Hong. 2009. Measuring effectiveness of graph visualizations: A cognitive load perspective. *Information Visualization* 8, 3 (2009), 139–152.

[16] Dayu Jiang and Slava Kalyuga. 2020. Confirmatory factor analysis of cognitive load ratings supports a two-factor model. 16, 3 (2020), 216–225.

[17] Slava Kalyuga, Paul Chandler, and John Sweller. 2000. Incorporating learner experience into the design of multimedia instruction. *Journal of educational psychology* 92, 1 (2000), 126.

[18] Ronit Ben-Bassat Levy, Mordechai Ben-Ari, and Pekka A Uronen. 2003. The Jeliot 2000 program animation system. *Computers & Education* 40, 1 (2003), 1–15.

[19] Andrew Luxton-Reilly, Ibrahim Albluwi, Brett A Becker, Michail Giannakos, Amruth N Kumar, Linda Ott, James Paterson, Michael James Scott, Judy Sheard, and Claudia Szabo. 2018. Introductory programming: a systematic literature review. In *Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education*. 55–106.

[20] Sandra P Marshall. 2009. *Schemas in Problem Solving*. Cambridge University Press.

[21] Richard E Mayer. 2002. Multimedia learning. In *Psychology of learning and motivation*. Vol. 41. Elsevier, 85–139.

[22] George A Miller. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review* 63, 2 (1956), 81.

[23] Briana B Morrison, Brian Dorn, and Mark Guzdial. 2014. Measuring cognitive load in introductory CS: adaptation of an instrument. In *Proceedings of the tenth annual conference on International computing education research*. ACM, 131–138.

[24] Duygu Mutlu-Bayraktar, Veysel Cosgun, and Tugba Altan. 2019. Cognitive load in multimedia learning environments: A systematic review. *Computers & Education* 141 (2019), 103618.

[25] Niko Myller, Mikko Laakso, and Ari Korhonen. 2007. Analyzing engagement taxonomy in collaborative algorithm visualization. In *ACM SIGCSE Bulletin*, Vol. 39. ACM, 251–255.

[26] Thomas L Naps, Guido Rößling, Vicki Almstrum, Wanda Dann, Rudolf Fleischer, Chris Hundhausen, Ari Korhonen, Lauri Malmi, Myles McNally, Susan Rodger, et al. 2002. Exploring the role of visualization and engagement in computer science education. In *ACM Sigcse Bulletin*, Vol. 35. ACM, 131–152.

[27] Fred Paas, Juhani E Tuovinen, Huib Tabbers, and Pascal WM Van Gerven. 2003. Cognitive load measurement as a means to advance cognitive load theory. *Educational psychologist* 38, 1 (2003), 63–71.

[28] Fred GWC Paas, Jeroen JG Van Merriënboer, and Jos J Adam. 1994. Measurement of cognitive load in instructional research. *Perceptual and motor skills* 79, 1 (1994), 419–430.

[29] Jan L Plass, Roxana Moreno, and Roland Brünken. 2010. *Cognitive load theory*. Cambridge University Press.

[30] Leo Porter and Beth Simon. 2013. Retaining nearly one-third more majors with a trio of instructional best practices in CS1. In *Proceeding of the 44th ACM technical symposium on Computer science education*. 165–170.

[31] Cornelia Schoor, Maria Bannert, and Roland Brünken. 2012. Role of dual task design when measuring cognitive load during multimedia learning. *Educational Technology Research and Development* 60, 5 (2012), 753–768.

[32] Simon, Andrew Luxton-Reilly, Vangel V. Ajanovski, Eric Fouh, Christabel Gonsalvez, Juho Leinonen, Jack Parkinson, Matthew Poole, and Neena Thota. 2019. Pass Rates in Introductory Programming and in Other STEM Disciplines. In *Proceedings of the Working Group Reports on Innovation and Technology in Computer Science Education* (Aberdeen, Scotland Uk) *(ITiCSE-WGR '19)*. Association for Computing Machinery, New York, NY, USA, 53–71. https://doi.org/10.1145/3344429.3372502

[33] Juha Sorva. 2013. Notional machines and introductory programming education. *ACM Trans. Comput. Educ.* 13, 2 (2013), 8–1.

[34] Juha Sorva and Teemu Sirkiä. 2010. UUhistle: a software tool for visual program simulation. In *Proceedings of the 10th Koli Calling International Conference on Computing Education Research*. 49–54.

[35] Juha Sorva and Teemu Sirkiä. 2015. Embedded questions in ebooks on programming: useful for a) summative assessment, b) formative assessment, or c) something else?. In *Proceedings of the 15th Koli Calling Conference on Computing Education Research*. 152–156.

[36] John Sweller, Jeroen JG Van Merrienboer, and Fred GWC Paas. 1998. Cognitive architecture and instructional design. *Educational psychology review* 10, 3 (1998), 251–296.

[37] Juhani E Tuovinen and Fred Paas. 2004. Exploring multidimensional approaches to the efficiency of instructional conditions. *Instructional science* 32, 1-2 (2004), 133–152.

[38] Eunjoon Um, Jan L Plass, Elizabeth O Hayward, Bruce D Homer, et al. 2012. Emotional design in multimedia learning. *Journal of educational psychology* 104, 2 (2012), 485.

[39] Tamara Van Gog and Fred Paas. 2008. Instructional efficiency: Revisiting the original construct in educational research. *Educational Psychologist* 43, 1 (2008), 16–26.

[40] Christa M van Mierlo, Halszka Jarodzka, Femke Kirschner, and Paul A Kirschner. 2012. Cognitive load theory in e-learning. In *Encyclopedia of cyber behavior*. IGI Global, 1178–1211.

[41] Arto Vihavainen, Jonne Airaksinen, and Christopher Watson. 2014. A Systematic Review of Approaches for Teaching Introductory Programming and Their Influence on Success. In *Proceedings of the Tenth Annual Conference on International Computing Education Research* (Glasgow, Scotland, United Kingdom) *(ICER '14)*. Association for Computing Machinery, New York, NY, USA, 19–26. https://doi.org/10.1145/2632320.2632349

[42] Arto Vihavainen, Craig S Miller, and Amber Settle. 2015. Benefits of Self-explanation in Introductory Programming. In *Proceedings of the 46th ACM Technical Symposium on Computer Science Education*. 284–289.