Measuring Learners' Cognitive Load when Engaged with an Algorithm Visualization Tool

Razieh Fathi
rfathi@smith.edu
Department of Computer Science
Smith College
Northampton, MA 01060 USA

James D. Teresco
jteresco@siena.edu
Department of Computer Science
Siena College
Loudonville, NY 12211 USA

Kenneth Regan
regan@buffalo.edu
Department of Computer Science
University at Buffalo
Buffalo, NY 14260 USA

Abstract

We present results of a preliminary study that applies cognitive load theory (CLT) to investigate how students with different amounts of prior experience learn algorithms. We test the following assertions from the CLT framework: The high CL of algorithm learning comes from intrinsic CL – meaning the complexity of the information being processed. There is also high germane CL – that induced by the instructional intervention – in tasks designed to assess the learned knowledge. Lowering either of these two CLs results in measurable learning gains. Lowering the complexity of incremental steps is the key determinant of success. We investigated the extent to which students' previous knowledge and experience influence the process of learning algorithms. This also involved testing whether an algorithm visualization tool (Map-based Educational Tools for Algorithm Learning, METAL) improves the understanding of graph algorithms. Our study adapted an existing survey instrument developed by Klepsch, et al., to algorithmic thinking tasks and used it as a tool to measure CL components. We explored and measured three types of CL for breadth-first and depth-first graph traversal algorithms, and among three groups of participants, non-Computer Science students, beginning CS students, and more advanced CS students. Results include: (i) Among different types of CL, germane load was the most substantial type for all groups. Students with more background in CS showed lower levels of all types of CL. (ii) The three groups showed similar relative effects of intrinsic, germane, and extraneous CL. We discuss future research and limitations of the study.

Keywords: algorithm visualization, student engagement, cognitive load, student learning

1. INTRODUCTION

As computer science has opened wide to students of diverse backgrounds and different levels of prior experience, the educational community needs progressively better understanding of how to optimize learning across this spectrum. Algorithm visualization (AV) tools have been shown to be effective (Hansen et al. 2002) but they fit a wider picture. We gain insight by conducting an experiment that employs Cognitive Load Theory (CLT) (Paas et al. 2003a; Sweller et al. 2011) to see how students at various levels learn when guided by an AV tool.

The experiment involves university students at three levels of experience with computing: non-computer science majors, those early in the CS major, and those at a more advanced stage of the major. We first describe CLT and how it applies in this context. Then we present results that accord with expectations from previous work in CLT and draw further conclusions.

2. COGNITIVE LOAD THEORY

CLT aims to structure the analysis of what is commonly called the "learning curve." How can we define and measure the effort required to learn concepts? We want to measure the efficiency of a learning process as the proportion that drives acquired knowledge and skills, versus the part expended on incidentals of the learning process. This can inform the design and ordering of instructional materials, and also evaluate the efficacy of automated learning tools. Much of the effort goes into memorizing and retention, which mean both the memory immediately needed to function and the memory of how concepts and procedures are ordered so they can be efficiently recovered from notes.

According to CLT, cognitive workload is the level of measurable mental effort put forth by an individual in response to one or more cognitive tasks (Van Gog & Paas 2008). In other words, cognitive load can be defined as the ratio between the workload that directly leads to the acquisition of knowledge and skills, and the workload that is expended on incidentals of the learning process. In general, there are three categories of cognitive workload (Sweller et al. 2019), reflected also in (Klepsch et al. 2017; Klepsch & Seufert 2020):

- Intrinsic cognitive load (ICL)
- Extraneous cognitive load (ECL)
- Germane cognitive load (GCL)

Intrinsic cognitive load refers to the complexity of the information being processed. It also relates to the concept of element interactivity. Interactive elements have to be processed simultaneously in working memory for learning to begin. Consequently, learning new material with a high number of interacting elements will impose a high cognitive workload. The other two kinds of loads are not considered inherent to learning the material, but as imposed by the design of instructional units for that material. When the load imposed by the design is ineffective or detrimental for learning, it is called extraneous cognitive load; when it is effective for learning it is referred to as germane cognitive load (Sweller et al. 2019).

The overall key to improving learning for novices is reducing the undesirable parts of the cognitive load to allow maximum memory usage for learning (Morrison et al. 2016). One of the original assumptions of CLT is that the three basic types of load (ICL, ECL, and GCL) are additive (Paas et al. 2003); thus, if the ECL is using the capacity of working memory, little can be devoted to the GCL. Because working memory is considered to be a fixed size (Miller 1956), it falls upon the instructional designer to minimize the ECL, design appropriately for the ICL, and emphasize the GCL. To accomplish this, one must be able to measure the specific load components for any pedagogical intervention.

Context and Relevant Work

Sweller (Sweller 1988) proposed CLT, which articulated the association between cognitive resources and task demands in creating cognitive load. Key elements defined in (Groth-Marnat & Wright 2016) and (Van Gog & Paas 2008) are *schemata* and *schemas*. The former means cognitive structures representing generic knowledge, i.e., structures that do not contain information about particular entities, instances or events, but rather about their general form. People use schemata to organize current knowledge and provide a framework for future understanding. Examples of schemata include academic rubrics. Schemas, on the other hand, are single information elements that combine to form schemata.

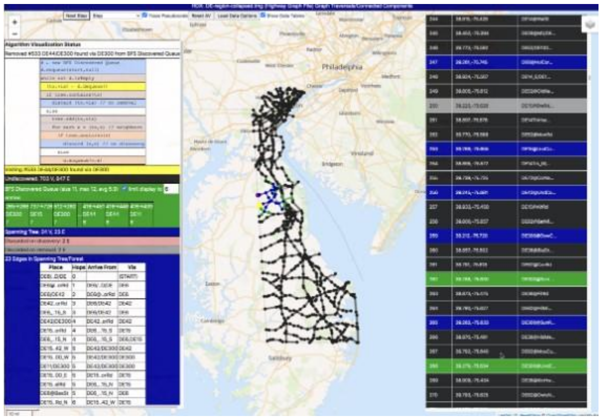
Learning is considered to happen through schema construction, elaboration, and automation. Automation means execution without controlled processing through intensive and consistent practice (Van Gog & Paas 2008). Cognitive load is metered by resources that learners consume while performing tasks. In this model, working memory is a cognitive resource, but is a limited one; only a small fraction of elements can be consciously handled per unit time, especially when they are novel or unfamiliar. However, long-term memory provides the ability to overcome the limitation of working memory, with the help of schemas (Xie et al. 2017).

In the field of educational research, CLT is mainly used to explain the effects of various forms of instructional design (Sweller et al. 2011). According to this theory, ICL is not directly affected by instructional design. It is related to element interactivity in learning materials and learners' prior knowledge. The level of ICL of a specific task is usually treated as depending on the level of element interactivity (Xie et al. 2017). An element can be anything that will be or has been presented, for example a concept or a procedure. Instructional materials with low element interactivity allow single (or several) element(s) to be processed with little or even no reference to other elements, thus resulting in a low ICL; however, high element interactivity materials contain elements that heavily interact with each other and cannot be processed separately, leading to a high ICL. The theory supports the position that GCL is directly beneficial to learning, whereas ECL only is detrimental to learning. In particular, GCL is imposed by cognitive processes of active schema construction, such as clarifying, inferring, and organizing, whereas ECL obstructs schema construction and automation.

The total cognitive load during information processing is the sum of the three kinds of cognitive loads. One important objective of instructional design is to ensure that the total cognitive load is within the learner's cognitive capacity, in order to avoid cognitive overload (Paas et al. 2003b). Techniques used to measure cognitive load include subjective rating scales, dual-task performance, and physiological measures (Antonenko et al. 2010; Paas et al. 2003b; Whelan 2007). Paas (Paas 1992) introduced the mental effort scale, which was a modified version of Bratfisch, Borg, and Dornic's scale (Bratfisch et al. 1972) for measuring perceived task difficulty. Paas's 9-point mental effort scale included one item that asked learners to report how much mental effort they invested when learning the material. Since then, the mental effort or perceived difficulty scale has been widely used in research in the field of learning and instruction because it is easy to administer, is non-invasive, and has good reliability and validity (Paas et al. 2003b)

3. USING METAL AS AN INTERACTIVE ALGORITHM VISUALIZATION

Transfer of learning occurs when people apply information, strategies, and skills they have learned to a new situation or context (Olson 2015). Transferring this learning performance generally requires additional instructional support that allows learners to go below the schema level in the level of hierarchy of learning and to understand the rationale of individual solution steps. One promising avenue to support learners in this type of reasoning is to embed interactive visualizations within an example-based hypermedia environment (Van Merriënboer et al. 2003). This study uses a variant of this idea, based on an interactive AV tool. The AV system provided by the Map-based Educational Tools for Algorithmic Learning (METAL) project (Teresco et al. 2018) has several advantages: scalability, a customizable API, visualizations that show the progress of algorithms overlaid on Leaflet Maps, color-coded tables showing contents of data structures, and example real-world data sets in a variety of sizes. These all enhance student engagement (Teresco et al. 2018). Figures 1 and 2 show a snapshot of METAL's AV system in action for the two algorithms used in our study: breadth-first search within a graph (BFS) and depth-first search within a graph (DFS).



Algorithm Visualization Status

Removed #533 DE44/DE300 found via DE300 from BFS Discovered Queue

Undiscovered: 703 V, 847 E

Spanning Tree: 24 V, 23 E

23 Edges in Spanning Tree/Forest

Place	Hops	Arrive From	Via
[illegible]	0		(START)
[illegible]	1	[illegible]	[illegible]
DE6/DE42	2	[illegible]	DE6
DE42_orRd	3	DE6/DE42	DE42
[illegible]	3	[illegible]	[illegible]
DE42/DE300	4	DE42_orRd	DE42
[illegible]	4	[illegible]	[illegible]
[illegible]	4	[illegible]	[illegible]
[illegible]	5	DE42/DE300	DE42
[illegible]	5	DE42/DE300	DE300
DE11/DE300	5	DE42/DE300	DE300
[illegible]	5	[illegible]	[illegible]
[illegible]	5	[illegible]	[illegible]
[illegible]	5	[illegible]	DE6
[illegible]	6	[illegible]	[illegible]

Fig. 1. A METAL AV in progress using the BFS algorithm on the Delaware region graph. The violet dot shows the starting vertex. Blue vertices and edges have been found to be part of the spanning tree. Green vertices and edges are candidates have been "discovered" and are in the queue of candidates to be added to the spanning tree in subsequent steps. The yellow edge and vertex just came out of the discovered queue as the next candidate to be added to the spanning tree.

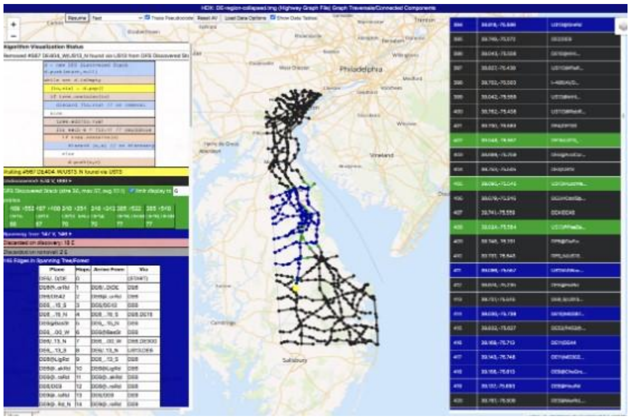


Fig. 2. A METAL AV in progress using the DFS algorithm on the Delaware region graph. Colors here match those described in Figure 1, except here the green discovered vertices and edges are stored in a stack rather than a queue.

4. METHOD

Our cognitive load measurement survey is adapted from Klepsch, et al (Klepsch et al. 2017; Klepsch & Seufert 2020). It includes elements to measure each of the three cognitive load components (English translations).

- For this task, many things needed to be kept in mind simultaneously (ICL).
- This task was very complex (ICL).
- For this task, I had to highly engage myself (GCL).
- For this task, I had to think intensively what things meant (GCL).
- During this task, it was exhausting to find the important information (ECL).
- The design of this task was very inconvenient for learning (ECL).
- During this task, it was difficult to recognize and link the crucial information (ECL).

Participants responded to each item using a Likert score from 0 ("absolutely wrong") to 7 ("absolutely right"). For the study herein, the questions were as follows:

- For BFS/DFS, many things needed to be kept in mind simultaneously (ICL).
- BFS/DFS was very complex (ICL).
- I made an effort, not only to understand several details, but to understand the overall context (GCL).
- My point while dealing with BFS/DFS was to understand everything correctly (GCL).
- The learning task consisted of elements supporting my comprehension of the task (GCL).
- During this task, it was exhausting to find the important information (ECL).
- The design of this task was very inconvenient for learning (ECL).
- During this task, it was difficult to recognize and link the crucial information (ECL).

Study Participants

Participants were undergraduate students, aged 18-24, in a university in the United States. Participants included both Computer Science (CS) majors, and students majoring in other disciplines. The CS majors were furthered divided into those at the CS1 stage of computer science and those at CS2 or higher (by the ACM classification). Thus, our participants are divided into 3 groups of 15 students each: (i) Non-CS majors (denoted hereafter as "NCS"), (ii) CS1 students (denoted as "CS1"), and (iii) CS2 or higher level students (denoted as "CS2+"). The cognitive load survey was administered after students were exposed to learning tasks and interview questions. Surveys were online and took 10 minutes to complete for each task. During data collection, cognitive load surveys were monitored in a Zoom meeting. No invalid surveys were returned. A total of 90 cognitive load surveys, 45 each for BFS and for DFS were collected.

Design of the Study

The study involves two families of algorithms: breadth-first search (BFS) and depth-first search (DFS). Both BFS and DFS are accessible at some level to both an advanced CS major and a non-major. The procedure of the study is divided into two blocks. One for BFS and one for DFS. After completing informed consent, the process below was used first for the BFS algorithm, then repeated for the DFS algorithm.

(1) The participant was asked several interview questions relevant to the algorithm (CS1 and CS2+ only, as NCS participants are assumed to be unfamiliar with the algorithms).

(2) The participant watched the METAL AV tutorial video for the algorithm.

(3) The participant uses the interactive METAL AV for the algorithm.

(4) The participant was asked several knowledge questions with different levels of complexity regarding the interaction with METAL and knowledge related to the algorithm.

ISSN: 2473-4901
v8 n5776

(5) The participant completes the cognitive load survey.

5. RESULTS AND DISCUSSIONS

The experimental design gives several axes for comparisons. First, we compare the cognitive load experienced in the ICL, GCL and and ECL categories, which are reflected by three groups of questions: Q1-Q2, Q3-Q5, and Q6-Q8. Then we compare the three groups of non-majors, CS1, and CS2, repeating for BFS (Figures 3-5) and DFS (Figures 6-8). Finally we compare experiences with BFS vs. DFS holding other factors the same (Figures 9-11).

In these comparisons, we try to refute a null hypothesis expressing that there is no difference between groups, i.e., that the inputs from the groups come from one underlying distribution. The basic Analysis of Variance (ANOVA) test presumes that this distribution is normal. For one student, the responses are drawn from the 0 to 7 Likert scale. We avoid the controversial presumption that each individual student's responses can be treated as drawn from a normal distribution centered somewhere on the Likert scale. Instead, for each of the eight questions, and within each group of 15 students, we average the responses to the question and input the mean as one data item. By appeal to the Central Limit Theorem, these means are representative of a normal distribution. As also stated in the footnoted excerpt from (Willett nd), we are shy of the conventional 30 for such appeal, but we compensate by having three groups of 15 and by the observation that our individual Likert responses do not have extreme polarization—that is, they do not have two modes on the 0-7 scale where some students strongly agree and others strongly disagree. We still find significance even with just 8 data points per item in the ANOVA, while our results are conservative compared to the alternative procedure of treating individual responses as normally distributed. We plot sums out of 105 rather than averages out of 7 for each group; this makes no difference to the ANOVA.

The first null hypothesis (NH) we try to refute is that there is no difference in the students' experience of cognitive load between the items identified as ICL, GCL, and ECL. Under NH, we would be supposing that the sampled questions are indistinguishable from randomly drawn responses about stages of cognitive load. As per the original design in Klepsch, et al., there is sufficient homogeneity in what each of Q1,...,Q8 addresses—this also reflects the intent to divide the learning exercise into "steps" that are reasonably uniform.

BFS Algorithm Studies

Participants in all groups first worked with the BFS algorithm and completed the surveys.

Non-CS Majors. Recall that Q1 and Q2 are intended to measure ICL; Q3-Q5 measure GCL and Q6-Q8 measure ECL.

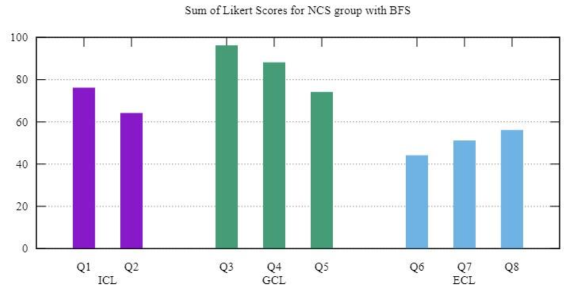
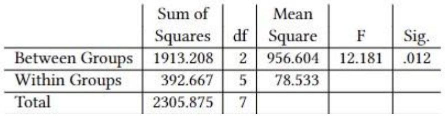


Fig. 3. Sum of Likert scores by question for the BFS algorithm study's 15 NCS participants.

In Figure 3, we see the GCL is highest, followed by ICL, with ECL the lowest. To test if there is a significant difference among these three, we use an ANOVA analysis to obtain a probability (p-value). This was computed as 0.012, less than a significance threshold of 0.05, indicating a significant difference. Details of this analysis are shown in Table 1.



	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1913.208	2	956.604	12.181	.012
Within Groups	392.667	5	78.533		
Total	2305.875	7			

Table 1. Results of the ANOVA analysis for the BFS algorithm study's NCS participants. "Groups" here are the three types of cognitive load measured.

We follow this by using Tukey's HSD test (Tukey 1949) post hoc to indicate which groups in the sample differ. Tukey's test uses the honest significant difference, a number that represents the distance between groups, to compare every mean with every other mean. The results of this test indicate a significant difference between GCL and ECL. For an example of this, we quote Willett (Willett nd) from the Simulation Canada website: "[It] is common to see ordinal data analyzed using parametric tests, such as the t-test or an ANOVA. Sometimes this is appropriate and sometimes it is not. So when can parametric tests, which are generally more sensitive and more powerful, be used? Only when the ordinal data meets all of the assumptions of the parametric test. These are: 1. The sampling distribution (not necessarily the data itself) is normally distributed. This will be true if 1. Sample size (n) is greater than 30; or 2. n<30 and the data appears to be normally distributed on inspection.”

CS1 students. Figure 4 shows that the three questions that measure GCL have the highest individual cognitive load for participants, and account more than half the total cognitive load.

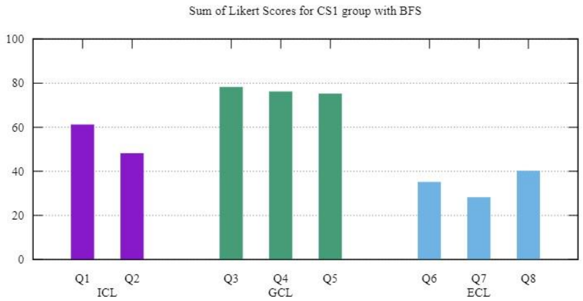
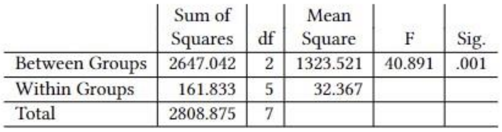


Fig. 4. Sum of Likert scores by question for the BFS algorithm study’s 15 CS1 participants.

As expected, ECL is the lowest. The BFS algorithm is new to this group of participants, but there is a difference in learning BFS between CS1 and NCS. CS1 students have some familiarity with the queue data structure at the core of the BFS algorithm, so we would expect this group to have some better schema creation, resulting in lower GCL, based on this previous experience that NCS participants did not have. This is observed in the GCL questions. ICL and ECL are also lower for the CS1 group compared to the NCS group.

For this group, ANOVA analysis (Table 2) gives a p-value of 0.001, again below the significance threshold of 0.05, indicating significant differences among the three types of cognitive load. The post hoc Tukey test also indicates significant differences between each pair of cognitive load types.



	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2647.042	2	1323.521	40.891	.001
Within Groups	161.833	5	32.367		
Total	2808.875	7			

Table 2. Results of the ANOVA analysis for the BFS algorithm study’s CS1 participants. “Groups” here are the three types of cognitive load measured.

CS2+ Students. Similarly to the NCS and CS1 groups, Figure 5 shows that the GCL is the largest component of cognitive load among the CS2+ group. CS2+ participants were familiar with the BFS algorithm but had not seen it recently (based on their pre-test interview) and had studied relevant data structures. The BFS AV for these students was more of a refresher. There was no significant difference between CS1 and CS2+ participants’ cognitive loads.

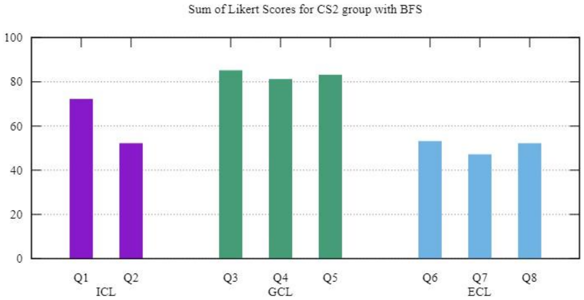
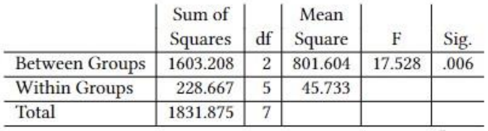


Fig. 5. Sum of Likert scores by question for the BFS algorithm study’s 15 CS2+ participants.

The ANOVA analysis for CS2+ (Table 3) shows a significant overall difference among three cognitive loads (p-value 0.006). The post hoc Tukey test indicates a significant difference between ICL and GCL, and between ECL and GCL, but not between ICL and ECL.



	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1603.208	2	801.604	17.528	.006
Within Groups	228.667	5	45.733		
Total	1831.875	7			

Table 3. Results of the ANOVA analysis for the BFS algorithm study’s CS2+ participants. “Groups” here are the three types of cognitive load measured.

DFS Algorithm Studies

Participants in all groups next worked with the DFS algorithm and completed the surveys

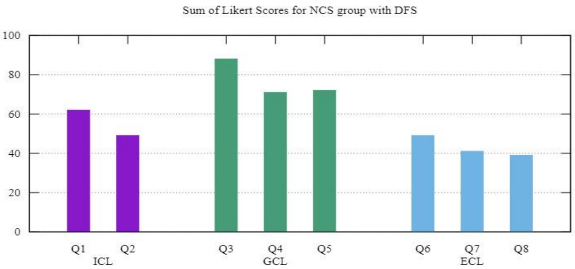


Fig. 6. Sum of Likert scores by question for the DFS algorithm study’s 15 NCS participants.

Non-CS Majors. Results for the NCS group’s interaction with DFS in Figure 6 show that all types of cognitive load are lower than for the same group with BFS, indicating a lower level of

ISSN: 2473-4901
v8 n5776

difficulty (unsurprising, since they had done BFS first). As was the case with BFS for this group, GCL is higher than the other loads for DFS.

NCS participants had a significant difference among the three types of load, as indicated by the p-value 0.009 obtained from the ANOVA analysis (Table 4). The post hoc test shows there is a significant difference only between ECL and GCL.

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1764.375	2	882.188	13.677	.009
Within Groups	322.500	5	64.500		
Total	2086.875	7			

Table 4. Results of the ANOVA analysis for the DFS algorithm study's NCS participants. "Groups" here are the three types of cognitive load measured.

CS1 Students. With DFS, the CS1 group (Figure 7) GCL is higher than ICL and ECL.

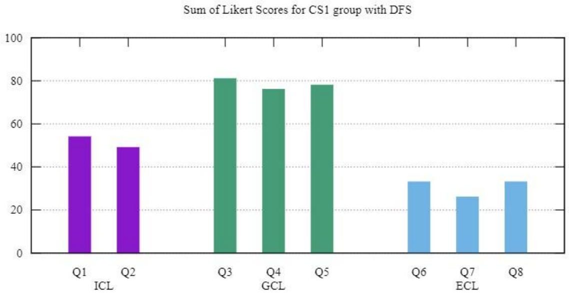


Fig. 7. Sum of Likert scores by question for the DFS algorithm study's 15 CS1 participants.

These results are very similar to this group's results with BFS in all three types of cognitive load, and indicate that the DFS and BFS algorithm for CS1 participants were about the same level of difficulty.

The ANOVA analysis here (Table 5) gives a p-value 0.001, demonstrating significant differences among the three types of cognitive load. The post hoc Tukey test shows there is a significant difference between each pair of ICL, ECL, and GCL.

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	3421.667	2	1710.833	147.911	<.001
Within Groups	57.833	5	11.567		
Total	3479.500	7			

Table 5. Results of the ANOVA analysis for the DFS algorithm study's CS1 participants. "Groups" here are the three types of cognitive load measured.

CS2+ Students. Again for DFS with the CS2+ group, Figure 8 shows that GCL is highest.

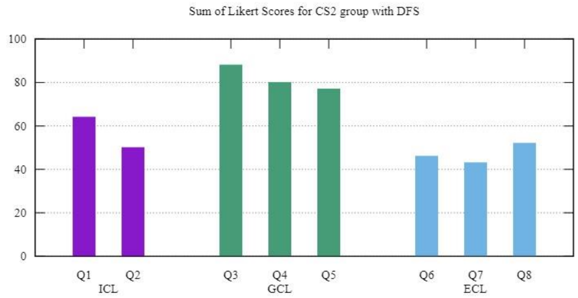


Fig. 8. Sum of Likert scores by question for the DFS algorithm study's 15 CS2+ participants.

ANOVA analysis (Table 6) gives a significant difference among the types of load (p-value 0.003). The post hoc Tukey test shows there is a significant difference between GCL and ECL, and between GCL and ICL, but not between ICL and ECL. CS2+ participants have the most familiarity with the algorithm and have little need for schema creation to store new knowledge, meaning less difference between ICL and ECL.

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1883.333	2	941.667	23.005	.003
Within Groups	204.667	5	40.933		
Total	2088.000	7			

Table 6. Results of the ANOVA analysis for the DFS algorithm study's CS2+ participants. "Groups" here are the three types of cognitive load measured.

Comparisons Between BFS and DFS Figures 9 (for the NCS group), 10 (for the CS1 group), and 11 (for the CS2+ group), show side-by-side comparisons of the results for BFS and DFS surveys presented earlier in this section. For NCS participants, we see that the cognitive loads are smaller for DFS than for BFS. For the CS1 and CS2+ groups, we observe a reduction in ICL and ECL for DFS compared to BFS.

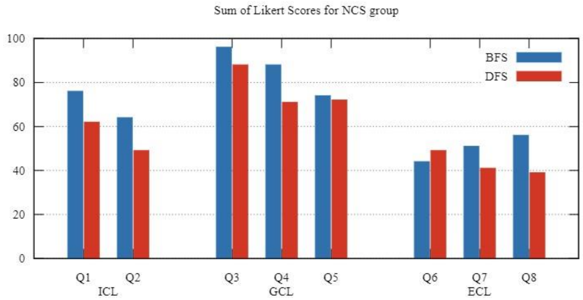


Fig. 9. Side-by-side comparison of the sum of Likert scores by question for the BFS algorithm study and DFS algorithm study's 15 NCS participants.

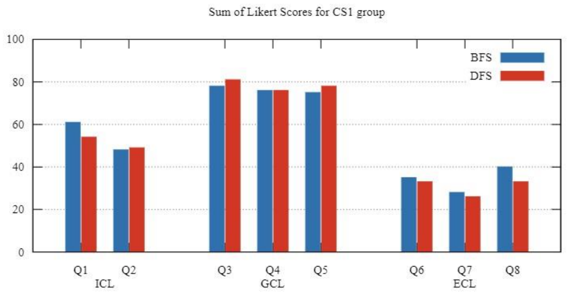


Fig. 10. Side-by-side comparison of the sum of Likert scores by question for the BFS algorithm study and DFS algorithm study's 15 CS1 participants.

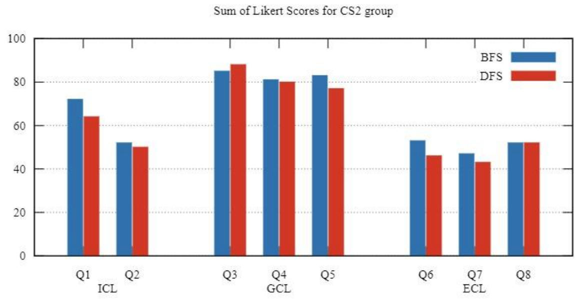


Fig. 11. Side-by-side comparison of the sum of Likert scores by question for the BFS algorithm study and DFS algorithm study's 15 CS2+ participants.

Discussion

In summary, our BFS and DFS studies produced the following results:

- Students with more CS background (CS1 and CS2+) showed lower levels of all three types of cognitive load.
- Among types of cognitive load, GCL was the most substantial for all groups.
- There was less difference between the CS1 and CS2+ groups in regard to cognitive load compared to the NCS group.

Comparing the differences between the BFS and DFS algorithms, we obtained these results:

- There was not significant difference between the two algorithms for the groups that have a background in CS (CS1 and CS2+).
- There was a higher level of cognitive load for BFS than DFS for the NCS group.
- GCL was the highest type of cognitive load for both algorithms.

According to CLT literature, we can reduce ICL in two ways:

1. The segmenting principle (Mayer and Moreno 2010). The goal of this principle is to reduce element interactivity by presenting information step by step. This process helps learners without prior knowledge to organize the incoming information. The METAL AV user interface presents the algorithm in a step-by-step manner, possibly explaining the low ICL in spite of the highly intrinsic nature of algorithmic learning.
2. The pre-training principle (Mayer & Pilegard 2005). According to this principle, ICL is reduced by providing the learner with information about the content before starting with the learning material. Increasing the learner's prior knowledge supports the integration of new information. In the design of our study, students first watched a video of the algorithms to gain some familiarity with the topic. That might be another factor that helped to reduce ICL.

6. LIMITATIONS

Due to pandemic protocols, the algorithm learning experiment was conducted fully online through use of Zoom web conferencing. Participants were monitored during all the steps of study. If we could repeat the experiment in person, it is unclear if we would obtain similar results. Also, the motivation of our subjects to participate in the study is another key factor to consider. The background and previous knowledge of the population under study can affect the results. The study's population size is also a limiting factor. As discussed below, this experiment has the potential to be repeated with a larger sample size. The fact that DFS was learned before BFS is potentially a confounding factor. We are hoping to repeat the experiment with random order and then see the differences.

7. CONCLUSIONS AND FUTURE WORK

We hypothesized that the high level of cognitive load related to algorithm learning comes from ICL. We found that METAL has a positive impact on the learning process over all participants by helping to reduce ICL relative to GCL. We found a significant difference between ICL and GCL within each of the CS1 and CS2+ groups. This indicates that METAL was most effective for those already with some CS background. This runs counter to intuition that visual tools may have greatest impact for neophytes and argues their aptness within core CS curricula.

We see this study as a first step that can enable many additional studies. The game plan for this includes the following:

(1) Find a more accurate and reliable measurement of cognitive load, especially as related to learning algorithms.

(2) Expand the use of the cognitive load survey in larger size algorithms classrooms.

(3) Short of being able to scale up the study in its entirety, we can consider partial questionnaires given to larger groups that can provide supplementary information relevant enough to buttress the conclusions.

(4) After the pandemic, it will be possible to employ physiological measurement tools such as eye-trackers to measure the cognitive process of algorithm learning. As a visual tool, METAL is suitable for this as well and will be a key component in this plan.

(5) Replicate the study with the different groups of students and changing the order in which BFS and DFS are introduced.

Points 1–3 raise the following general research question: Can we obtain results measuring cognitive load as accurately as the complex and time-consuming study that was used here with a streamlined study that can more reasonably be scaled to larger groups That could also mean we could do multiple studies or have multiple groups for a study within one larger cohort, gathering much more data in much less time.

8. REFERENCES

Antonenko, P., Paas, F., Grabner, R., & Van Gog, T. (2010). Using electroencephalography to measure cognitive load. Educational psychology review, 22(4), 425-438.

Bratfisch, O. (1972). Perceived Item-Difficulty in Three Tests of Intellectual Performance Capacity.

Groth-Marnat, G. (2009). Handbook of psychological assessment. John Wiley & Sons.

Hansen, S., Narayanan, N. H., & Hegarty, M. (2002). Designing educationally effective algorithm visualizations. Journal of Visual Languages & Computing, 13(3), 291-317.https://doi.org/10.1006/jvlc.2002.0236

Klepsch, M., Schmitz, F., & Seufert, T. (2017). Development and validation of two instruments measuring intrinsic, extraneous, and germane cognitive load. Frontiers in psychology, 8, 1997.https://doi.org/10.3389/fpsyg.2017.01997

Klepsch, M., & Seufert, T. (2020). Understanding instructional design effects by differentiated measurement of intrinsic, extraneous, and germane cognitive load. Instructional Science, 48(1), 45-77.

Mayer, R. E., & Moreno, R. E. (2010). Techniques that reduce extraneous cognitive load and manage intrinsic cognitive load during multimedia learning.

Mayer, R. E. (2005). The Cambridge Handbook of Multimedia Learning: Principles for Managing Essential Processing in Multimedia Learning: Segmenting, Pretraining, and Modality Principles.

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

Morrison, B. B., Margulieux, L. E., Ericson, B., & Guzdial, M. (2016, February). Subgoals help students solve Parsons problems. In Proceedings of the 47th ACM Technical Symposium on Computing Science Education (pp. 42-47).

Olson, M. H. (2015). Introduction to theories of learning. Routledge.

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. Educational psychologist, 38(1), 1-4.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. Cognitive science, 12(2), 257-285. John Sweller, Paul Ayres, and Slava Kalyuga. 2011. Measuring cognitive load. In Cognitive Load Theory. Springer, 71–85.

Sweller, J., van Merriënboer, J. J., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. Educational Psychology Review, 31(2), 261-292.

Teresco, J. D., Fathi, R., Ziarek, L., Bamundo, M., Pengu, A., & Tarbay, C. F. (2018, February). Map-based algorithm visualization with METAL highway data. In Proceedings of the 49th ACM Technical Symposium on Computer Science Education (pp. 550-555).

Tukey, J. W. (1949). Comparing individual means in the analysis of variance. Biometrics, 99-114.

Van Gog, T., & Paas, F. (2008). Instructional efficiency: Revisiting the original construct in educational research. Educational psychologist, 43(1), 16-26.

Van Merriënboer, J. J., Kirschner, P. A., & Kester, L. (2003). Taking the load off a learner's mind: Instructional design for complex learning. Educational psychologist, 38(1), 5-13.

Whelan, R. R. (2007). Neuroimaging of cognitive load in instructional multimedia. Educational Research Review, 2(1), 1-12.

Timothy Willett. n.d.. Analyzing Likert Scale Data: The Rule of N=30. URL. https://www.sim-one.ca/community/tip/analyzing-likert-scale-data-rule-n30.

Xie, H., Wang, F., Hao, Y., Chen, J., An, J., Wang, Y., & Liu, H. (2017). The more total cognitive load is reduced by cues, the better retention and transfer of multimedia learning: A meta-analysis and two meta-regression analyses. PloS one, 12(8), e0183884.