Using Multiple Ways to Investigate Cognitive Load Theory in the Context of Physics Instruction

Tianlong Zu
Purdue University

Follow this and additional works at: https://docs.lib.purdue.edu/open_access_dissertations

Recommended Citation
https://docs.lib.purdue.edu/open_access_dissertations/1678

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.
USING MULTIPLE WAYS TO INVESTIGATE COGNITIVE LOAD THEORY IN THE CONTEXT OF PHYSICS INSTRUCTION

by

Tianlong Zu

A Dissertation

Submitted to the Faculty of Purdue University
In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

Department of Physics and Astronomy
West Lafayette, Indiana
December 2017
THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. N. Sanjay Rebello, Chair
Department of Physics and Astronomy
Department of Curriculum and Instruction

Dr. Lynn A. Bryan
Department of Curriculum and Instruction
Department of Physics and Astronomy

Dr. Mark P. Haugan
Department of Physics and Astronomy

Dr. Andrew S. Hirsch
Department of Physics and Astronomy

Dr. Lester C. Loschky
Department of Psychological Sciences, Kansas State University

Approved by:

Dr. John P. Finley
Head of the Graduate Program
This dissertation is dedicated to my family
ACKNOWLEDGMENTS

It is a blessing to go on a journey to pursue a PhD in physics. Upon reflecting my own one, I have found it mostly an exciting and fortunate one.

There is nothing more grateful than having a great mentor in graduate school. I was fortunate to have Dr. N. Sanjay Rebello as my supervisor. Dr. N Sanjay Rebello is very well respected as a physics education research professor and for his excellent teaching in physics classrooms. My supervisor has inspired me with his unmatched insights in education and cognitive science in countless times. And he has influenced and encouraged me with his immense knowledge and extraordinary professionalism. He examined my research work with great care and patience. I could not have imagined of having a better mentor for my PhD study.

Besides my supervisor, I would like to express my gratitude to the rest of my committee members: Dr. Lester. C. Loschky, Dr. Lynn Bryan, Dr. Mark Haugan, and Dr. Andy Hirsh for their insightful comments and encouragement. Their prudent guidance has significantly strengthened my research work from various perspectives.

In addition to these great mentors, I also want to extend my thankfulness to my friends for their company, help, and peer support in numerous times: Erfan Saydanzad, Balram Kaderiya, Javad Robatjazi, Elise Agra, Xian Wu, Bahar Modir, Nandana Nelith, Tugba Yuksel, Yuri Piedrahita Uruena.

Since I arrived at the United States, I have met many great people, colleagues, and friends in my life. I would also like to thank them for their unmatched kindness, help and encouragement: Dr. Carina Rebello, Kim Coy (our beloved secretary from Department of Physics of KSU), Sandy Formica (our beloved secretary from Department of Physics and Astronomy of Purdue University), John Hutson, and Zach Throneberg.

Last but not the least, I am grateful to have a family to support me all the time. Those who love me the most and who I love the most will always be my harbor.
TABLE OF CONTENTS

LIST OF TABLES .................................................................................................................... viii
LIST OF FIGURES ................................................................................................................... ix
ABSTRACT ................................................................................................................................ xi

1. INTRODUCTION ...................................................................................................................... 1
   1.1 Measurement of Cognitive Load .......................................................................................... 2
   1.2 Peer Interaction and Cognitive Load .................................................................................... 5
   1.3 Visual-cueing, Multimedia Video-solutions and Cognitive Load ........................................ 6

2. BACKGROUND ........................................................................................................................ 8
   2.1 Nature of Cognitive Load ..................................................................................................... 8
   2.2 Cognitive Load Theory and the Cognitive Theory of Multimedia Learning ....................... 12
   2.3 Criticisms of Cognitive Load Theory ................................................................................ 15

3. MEASUREMENT OF COGNITIVE LOAD ........................................................................ 17
   3.1 Analytical Approaches ........................................................................................................ 17
   3.2 Empirical Approaches ........................................................................................................ 18
      3.2.1 Subjective measurements ............................................................................................ 18
      3.2.2 Objective Measurements .......................................................................................... 19
   3.3 Measuring the Three Types of Cognitive Load .................................................................. 22
   3.4 Current Limitations ............................................................................................................. 25

4. STUDY ONE: USING EYE MOVEMENTS TO INVESTIGATE COGNITIVE LOAD THEORY .... 28
   4.1 Introduction ........................................................................................................................ 28
   4.2 Methodology ....................................................................................................................... 36
      4.2.1 Manipulation of the Three Kinds of Cognitive Load .................................................. 36
      4.2.2 Procedure .................................................................................................................... 38
      4.2.3 Controlling the Effect of Luminance .......................................................................... 40
   4.3 Results ................................................................................................................................ 41
      4.3.1 Pre-survey, Transfer Test and O-span Task ................................................................ 41
      4.3.2 Non-eyetracking Based Metrics ................................................................................... 41
      4.3.3 Eyetracking Based Metrics ......................................................................................... 44
4.4 Examining the Assumptions of Sweller’s (2010) CLT Model ........................................................................ 52
4.5 Discussion and Implications ....................................................................................................................... 57

5. STUDY TWO: THE EFFECTS OF PEER INTERACTION ON CONCEPTUAL TEST PERFORMANCE AND COGNITIVE LOAD ........................................................................................................... 61
5.1 Introduction .................................................................................................................................................. 61
5.2 Methodology ............................................................................................................................................... 65
  5.2.1 Survey for Measuring Cognitive Load .................................................................................................. 65
  5.2.2 Participants and Materials .................................................................................................................. 70
  5.2.3 Procedures ........................................................................................................................................... 71
5.3 Results ...................................................................................................................................................... 72
  5.3.1 Performance on the DIRECT Test ..................................................................................................... 72
  5.3.2 Cognitive Load ................................................................................................................................... 74
5.4 Discussion and Implications ....................................................................................................................... 81

6. STUDY THREE: EFFECT OF VIDEO SOLUTIONS AND VISUAL CUES ON CONCEPTUAL TASK PERFORMANCE AND COGNITIVE LOAD .......................................................... 84
6.1 Introduction ............................................................................................................................................... 84
6.2 Methodology ............................................................................................................................................... 89
  6.2.1 Materials and Data Collection ........................................................................................................... 89
  6.2.2 Visual Cueing and Video Solutions ................................................................................................... 91
  6.2.3 Near and Far Transfer Tasks ............................................................................................................. 93
6.3 Results ...................................................................................................................................................... 94
  6.3.1 Performance ....................................................................................................................................... 94
  6.3.2 Visual Attention ................................................................................................................................. 96
  6.3.3 Cognitive Load ................................................................................................................................ 100
6.4 Discussion and Implications ....................................................................................................................... 103

7. GENERAL DISCUSSION .......................................................................................................................... 107
7.1 Summary .................................................................................................................................................. 107
7.2 Limitations and Future Directions ........................................................................................................... 109
REFERENCES ............................................................................................................................................... 112
APPENDIX A1. MATERIALS USED IN EXPERIMENT ONE .............................................................................. 129
APPENDIX A2. PRE-SURVEY QUESTIONS FROM EXPERIMENT ONE .......................................................... 131
APPENDIX A3. QUESTIONS FROM TRANSFER TEST IN EXPERIMENT ONE ............ 132
APPENDIX A4. SNAPSHOTS OF MENTAL EFFORT AND DIFFICULTY RATING QUESTION PAGES USED IN EXPERIMENT ONE.............................................................. 133
APPENDIX B. SNAPSHOT OF THE SURVEY FOR MEASURING COGNITIVE LOADS USED IN EXPERIMENT TWO ........................................................................................................... 134
APPENDIX C. CONCEPTUAL PHYSICS TASKS USED IN EXPERIMENT THREE........ 135
LIST OF TABLES

Table 1: Results for the Three Metrics of Cognitive Load from (DeLeeuw & Mayer, 2008)...... 42
Table 2: Replicated Results for the Three Metrics Used in (DeLeeuw & Mayer, 2008) .......... 42
Table 3: Correlations b/w Dependent Measures for All Students of (DeLeeuw & Mayer, 2008)................. 43
Table 4: Replicated Correlations b/w Dependent Measures of All Students......................... 43
Table 5: Correlation b/w Three Measurements and WM capacity........................................ 44
Table 6: Independent t Test for FSL ........................................................................................ 46
Table 7: Hierarchical Regression Analysis Assessing Prediction of IVs By MFD...................... 47
Table 8: Hierarchical Regression Analysis for Prediction of IVs by MFD for Animation AoI... 47
Table 9: Hierarchical Regression Analysis Assessing Prediction of IVs by MSL..................... 48
Table 10: Hierarchical Regression Analysis Assessing Prediction of IVs by NBC..................... 48
Table 11: Hierarchical Regression Analysis Assessing Prediction of IVs by MSD................... 49
Table 12: Hierarchical Regression Analysis Assessing Prediction of IVs by MSPV ................. 49
Table 13: Hierarchical Regression Analysis Assessing Prediction of IVs by RPSC .................. 50
Table 14: Hierarchical Regression Analysis for Prediction of IVs by RPSC for Animation AoI 50
Table 15: Hierarchical Regression Analysis Assessing Prediction of IVs by TBTA................... 51
Table 16: Correlations b/w Dependent Measures for Both Groups........................................... 52
Table 17: Results for the Multilevel Model Analysis of Comparing S-model and A-model...... 55
Table 18: Results for the Multilevel Model Analysis of Comparing SS3-model and S4 model.. 56
Table 19: Summary of the Sensitivity of the Three Measurements to Each Load ..................... 57
Table 20: Summary of the Sensitivity of the Eye Movements Based Metrics to Each Load....... 58
Table 21: Means (S.D) of Eye Movement Parameters for Both Conditions & p-values......... 101
LIST OF FIGURES

Figure 1: Stages of Cognitive Load Theory (Figure Is Adapted from Chen et al. (2016)) .......... 9
Figure 2: Illustration of Different Types of Load (Figure Is from Chen et al. (2016)) ............. 11
Figure 3: Stages of Cognitive Load Theory (Incorporated Results from Sweller, (2010)) ......... 12
Figure 4: Screenshot of Multimedia Lesson ............................................................................. 37
Figure 5: Interview Procedure .................................................................................................. 38
Figure 6: AoIs in Non-Redundant Condition (Left), and Redundant Condition (Right) .......... 45
Figure 7: Procedure of the Experiment ...................................................................................... 71
Figure 8: Performance of Students on the Pretests and Posttests ............................................ 73
Figure 9: Cognitive Load of Students from Individual Pretest and Posttest ......................... 75
Figure 10: Cognitive Load of Students from Pretest Individual and Group ............................ 76
Figure 11: Cognitive Load of Students from Posttest Individual and Group .......................... 78
Figure 12 Total cognitive Load Perceived by Students at All Phases .................................... 79
Figure 13: Total Cognitive Energy Devoted by Students at All Phases .................................. 80
Figure 14: Procedure in Each Problem Set ................................................................................. 90
Figure 15 Solution-Relevant (RED) and Irrelevant (BLUE) AoIs on Two Example Tasks ...... 92
Figure 16: Performance of Both Conditions for the Initial, Near Transfer & Far Transfer Tasks 94
Figure 17: Performance of Initially Incorrect and Correct Participants for Near Transfer Task.. 95
Figure 18: Performance for Far Transfer Tasks ......................................................................... 96
Figure 19: PT/PA over Novice-like Aol & PT/PA over Expert-like Aol for Initial Task .......... 97
Figure 20: PT/PA of Relevant Aol for Initial (Incorrect) & Near Transfer (Correct) Task ....... 98
Figure 21: PT/PA of Relevant Aol for Initial (Incorrect) & Far Transfer (Correct) Task ......... 99
Figure 22: Mean Fixation Duration according to Initial Correctness and Assigned Condition . 102
Figure 23: Pupil Diameter based on Condition and Initial Correctness of Students .............. 103
Figure 24: Five Diagrams Used in Pre-survey ......................................................................... 131
Figure 25: Mental Effort Rating Snapshot ................................................................................. 133
Figure 26: Difficulty Rating Snapshot ...................................................................................... 133
Figure 27: Snapshot of the Survey Used in Study Two ............................................................... 134
Figure 28: Initial Tasks for Speed (Left) and Energy (Right) Tasks, One for Each ............. 135
Figure 29: Training Tasks for Speed (Left) and Energy (Right) Tasks, Four for Each .......... 136
Figure 30: Near Transfer Tasks for Speed (Left) and Energy (Right) Tasks, One for Each...... 136
Figure 31: Far Transfer Tasks for Speed (Left) and Energy (Right) Tasks, One for Each ........ 137
ABSTRACT

Cognitive load theory (CLT) (Sweller 1988, 1998, 2010) provides us a guiding framework for designing instructional materials. CLT differentiates three subtypes of cognitive load: intrinsic, extraneous, and germane cognitive load. The three cognitive loads are theorized based on the number of simultaneously processed elements in working memory. Intrinsic cognitive load depends upon the number of interacting elements in the instructional material that are related to the learning objective. Extraneous cognitive load is the mental resources allocated to processing unnecessary information which does not contribute to learning as caused by non-optimal instructional procedure. It is determined by the number of interacting elements which are not related to learning goal. Both intrinsic and extraneous load vary according to prior knowledge of learners. Germane cognitive load is indirectly related to interacting elements. It represents the cognitive resources deployed for processing intrinsic load, chunking information and constructing and automating schema. Germane cognitive load is related to level of motivation of the learner. Given this triarchic model of cognitive load and their different roles in learning activities, different learning outcomes can be expected depending upon the characteristics of the educational materials, learner characteristics, and instructional setting.

In three experiments, we investigated cognitive load theory following different approaches. Given the triarchic nature of cognitive load construct, it is critical to find non-intrusive ways to measure cognitive load. In study one, we replicated and extended a previous
landmark study to investigate the use of eye movements related metrics to measure the three kinds of cognitive load independently. We also collected working memory capacity of students using a cognitive operation-span task. Two of the three types of cognitive load (intrinsic and extraneous) were directly manipulated, and the third type of cognitive load (germane) was indirectly ascertained. We found that different eye-movement based parameters were most sensitive to different types of cognitive load. These results indicate that it is possible to monitor the three kinds of cognitive load separately using eye movement parameters. We also compared the up-to-date cognitive load theory model with an alternative model using a multi-level model analysis and we found that Sweller’s (2010) up-to-date model is supported by our data.

In educational settings, active learning based methodologies such as peer instruction have been shown to be effective in facilitating students’ conceptual understanding. In study two, we discussed the effect of peer interaction on conceptual test performance of students from a cognitive load perspective. Based on the literature, a self-reported cognitive load survey was developed to measure each type of cognitive load. We found that a certain level of prior knowledge is necessary for peer interaction to work and that peer interaction is effective mainly through significantly decreasing the intrinsic load experienced by students, even though it may increase the extraneous load.

In study three, we compared the effect of guided instruction in the form of worked examples using narrated-animated video solutions and semi-guided instruction using visual cues on students’ performance, shift of visual attention during transfer, and extraneous cognitive load during learning. We found that multimedia video solutions can be more effective in promoting transfer performance of learners than visual cues. We also found evidence that guided instruction
in the form of multimedia video solutions can decrease extraneous cognitive load of students
during learning, more so than semi-guided instruction using visual cues.
1. INTRODUCTION

Cognitive load theory (CLT) as a framework provides us guidelines for designing instructional materials (Sweller, 1988, 1994, 1998, 2010). It is suggested that optimal learning requires the instructional material to impose cognitive load that is neither too low nor too high (Sweller et al., 1998). It has been frequently shown that traditional instructional materials which impose too high a cognitive load do not facilitate learning so finding ways to reduce cognitive load has been the focus of studies for several decades (Paas, 1992; Paas et al., 2003; Sweller, 1988).

Cognitive load theory differentiates three kinds of load: intrinsic, extraneous and germane cognitive load. Intrinsic cognitive load (intrinsic load for short hereafter) reflects the number of interacting elements in instructional material that need to be carried in working memory simultaneously. Intrinsic load depends on the learning goal and knowledge level of learners (Sweller, 2010). Extraneous cognitive load (extraneous load for short hereafter) indicates the mental resources allocated to processing unnecessary information which does not contribute to learning. It can be reduced by optimizing instructional procedures. Germane cognitive load (germane load for short hereafter) represents the cognitive resources deployed for processing intrinsic load, chunking information and constructing and automating schemas. It depends on the motivation of learners and can be improved by instructional design. From the definition of the three kinds of cognitive load; intrinsic load is based on what needs to be learned, which is neutral if it is not too high for students to handle; germane load facilitates this learning process and usually high intrinsic load entails high germane load as well; and extraneous load impedes learning since it is a detrimental deployment of cognitive resources. Any instructional material that aims to promote better and efficient learning should avoid making students process
unnecessary information.

1.1 Measurement of Cognitive Load

For the past decades, several methods have been employed to measure cognitive load. Cognitive load can be measured using different analytical and empirical methods (Linton et al., 1989; Sweller 1988; Xie & Salvendy, 2000). Examples of the empirical approach are self-reported questionnaires (Paas, 1992) and physiological measures. Examples of physiological measures include heart rate variability (Paas & van Merrienboer, 1994), neuroimaging of brain activities (Whelan, 2007), EEG measure of brain activities (Berka et al., 2007; Fitzgibbon et al., 2004; Gevins et al., 1997), skin conductivity (Nourbakhsh et al., 2012), eye movements (Van Orden et al., 2001) and task performance related methods such as dual task techniques which are based on performance on a secondary task (Chandler & Sweller, 1996; Sweller, 1988). Overall, these results can tell us how different metrics respond to tasks that impose different levels of overall cognitive load (Gerjets et al., 2009). But, these metrics have not been used to measure the three different kinds of cognitive load. One of the reasons is that no clear manipulation of the three kinds of cognitive load were included in their experimental designs. It makes one wonder whether these different methods of measuring cognitive load are tapping the same construct or whether some of them are more sensitive to only a subset of the types of cognitive load. It has been found that although several different metrics were able to differentiate tasks of different overall levels of cognitive load, some methods of measurement did not yield results consistent with other methods of measurement. For instance, as EEG alpha band-power decreases indicating a decreasing cognitive load, pupil dilation did not have a significant correlation with EEG alpha bad-power decrease, although each of the two measurements were shown to be responsive to cognitive load (Scharinger et al., 2015) and other similar evidence (Kamp et al.,
This tells us that the complexity of cognitive activity and different metrics might tap different aspects of cognitive load. In a study by DeLeeuw and Mayer (2008), low prior knowledge students were asked to learn from a narrated animated video how an electric motor works, while the three cognitive loads were manipulated independently. They showed that each of the three metrics used in their study were sensitive to each of the three cognitive load types. This work provides evidence to support Sweller’s triarchic construct of cognitive load and suggests that different measures tap different subtypes of cognitive load. DeLeeuw and Mayer (2008) provides us evidence supporting the triarchic nature of cognitive load by using multiple metrics to measure cognitive load. Other studies did include multiple measures of cognitive load. But none of them tested Sweller's theory using multiple measures, each of which was independently shown to be most sensitive to one of the 3 types of load. Despite this important contribution to the literature, the above study had some important limitations. First, all participants in this study were low prior knowledge learners, so this study could not show the effect of prior knowledge on cognitive load. Second, there was an issue with the manipulation of germane load in the study. The authors claimed that participants with higher transfer performance had a higher germane load. This assumption at first makes sense, but upon further consideration, we think the authors failed to consider another potential confounding variable, fluid intelligence, which may be an important factor influencing the germane load. Fluid intelligence is the ability to use logic to analyze novel problems such that reasoning efficiency varies between subjects (Cattell, 1963). This is similar to the generative processing of making connections with what one knows with what needs to be learnt (Mayer, 2010). Thus, fluid intelligence may further affect the level of germane cognitive load. Since fluid intelligence has been shown to be related to reasoning efficiency and performance (Fry & Hale, 1996), Namely,
learners with higher fluid intelligence would be expected to show greater learning form a lesson, though such learning was used to measure germane load. Fluid intelligence was also shown to be significantly correlated with human working memory capacity, which is an integrated part of cognitive load theory (Sweller, 2010) and working memory capacity has been shown to be correlated with mental effort rating and difficulty ratings in different research contexts (Rudner et al., 2012; Schmader & Johns, 2003). Third, two of the measures used by DeLeeuw and Mayer (2008) -- response time to a secondary task and mental effort rating -- are intrusive in nature since learners were interrupted eight times during the multimedia lesson for collecting these data using extrinsic tasks. Finally, two of the measures used -- difficulty rating of the learning experience and the effort rating -- are both self-reported, which brings up the issue of questionable reliability of the study. Specifically, one would prefer to be able to measure the cognitive load based on some measure not requiring introspection by the learners, or at least triangulating the measurement of cognitive load with objective measures. Any future study trying to replicate the study by DeLeeuw and Mayer (2008) should take these matters into consideration.

In study one, we replicated and extended this study in three ways: First, we included a pre-survey which was designed to poll students’ understanding regarding the material that they would be learning. Second, we included a cognitive Operation Span task (O-span task) which was designed to measure working memory capacity (WMC) (Unsworth et al., 2005). Since WMC has been shown to be positively correlated with fluid intelligence (Turner & Engle, 1989), it could potentially moderate the relationship between germane load and extraneous load. Third, we employed eye movements to investigate CLT in addition to the three measurements used in their original study. An eye-tracker can provide non-intrusive measurement, which means data
could be collected in real time without interrupting the learner’s cognitive processes. Eye-based measures such as pupil dilation (Hess & Polt, 1964; Van Gerven et al., 2004), mean fixation duration (Chen et al., 2016; Holmqvist et al., 2011), saccade peak velocity (Di Stasi et al., 2010) and mean saccadic length (Chen et al., 2016; Holmqvist et al., 2011) have been shown to be good indicators of cognitive load.

1.2 Peer Interaction and Cognitive Load

Contrary to traditional lecture based instruction, student-centered instruction has been shown to be very effective in promoting students’ conceptual understanding (e.g. Hake, 1998; Halloun & Hestenes, 1985; McDermott, 1984). Learning environments that encourage dialogue between peers can help develop critical thinking skills and improve conceptual understanding (Anderson et al., 1996, 2001; deCorte, 1996; Matthews, 1996; Reiter, 1994). One of these popular approaches, peer instruction, has been widely adopted in college classrooms (Laws, 1997; Mazur, 2013). Peer instruction allows students to share, discuss ideas with each other and it encourages articulation of one’s ideas to others, often deepening one’s understanding of the learning material. However, many research studies have investigated peer instruction based on behavioral patterns changes caused by interacting with peers (Rogoff, 1998). For example, children, when working with peers, are more active discussing the logic and articulating and examining their generated thoughts (Rogoff, 1998). Nicol and Boyle (2003) found that learning can be affected by the type of dialog and the discussion sequences that were adopted in classrooms. This is not surprising, as learning processes have a direct relation to cognitive load. As per Mayer (2010), intrinsic load is due to essential processing which is about initial selection and lower order organization of information, extraneous load is about processing unnecessary information, and germane load is equal to generative processing, which is about higher order of
organizing and integrating information with knowledge from long term memory. Following this line of thought, cognitive load theory allows us to examine the effect of peer interaction through a converging cognitive load concept. As conjectured by Singh (2005) the fact that students improved their performance after working in pairs might have been caused by a reduction in cognitive load. In study two, we examined the effect of peer interaction on students’ performance as well as on the three kinds of cognitive load. Results of this will shed light on the mechanisms of peer interaction on the cognition.

1.3 Visual-cueing, Multimedia Video-solutions and Cognitive Load

Prior research has shown that, when considering a physics conceptual task containing diagrams consisting of solution-relevant and irrelevant areas, incorrect solvers attend to the solution-irrelevant areas, while correct solvers tend to attend to the solution-relevant areas (Madsen et al., 2013). Visual-cueing is a technique of highlighting the solution-relevant area as a way to redirect attention of students and it has been found to be effective in promoting learning (Rouinfar et al., 2014a, 2014b). But its effect on transfer problem performance is mixed (Rouinfar et al., 2014a, 2014b). Looking at this through the lens of cognitive load theory, visual cueing might be imposing higher extraneous cognitive load due to the ambiguity of the information provided by the visual cues. From the perspective of the modality principle (Mayer, 2014), it might be because information is provided to students only through the visual channel, rather than through both the visual and auditory channels. So, we can hypothesize that the effects of visual cueing can be can be improved upon if we also provide information through the auditory channel.

Worked examples have been shown to be an effective way to reduce extraneous cognitive load (Atkinson et al., 2000; Sweller et al., 1998). The effectiveness of worked examples can be
attributed to the process of facilitating students’ attention to the relevant concepts and information. Thus, learners can avoid processing a lot of unnecessary information and spend energy focusing on relevant principles and a correct solution path. Multimedia-based worked examples have been shown to be effective in promoting exam performance (Mestre et al., 2015). This type of worked example provides information in both modalities (visual and auditory). We anticipate that they will outperform visual cues in promoting learning and transfer. In study three, we investigated the effect of worked examples in the form of narrated animated video solutions and visual cueing on students’ performance and shifts of visual attention on transfer tasks as well as their cognitive load during learning.

This dissertation is organized as follows. Chapter 2 will review the relevant theoretical underpinnings of all our studies such as the basics of cognitive load theory, a brief history of the development of cognitive load theory, and the relationship between cognitive load theory and the cognitive theory of multimedia learning (Mayer, 2014). Chapter 3 presents a review of measurement of cognitive load. Chapter 4 describes and presents the results of our Study One. Chapter 5 presents and discusses design and results of Study Two. Chapter 6 presents and discusses design and results of Study Three. Chapter 7 is a general discussion of all studies and future research directions.
2. BACKGROUND

Cognitive load theory provides us a lens for examining educational interventions (Paas & van Merrienboer, 1994; Sweller, 1988). It suggests that optimal learning requires instruction to avoid overloading the working memory of learners (Sweller, 1988). It is important to know the nature of cognitive load and the history of cognitive load theory and existing efforts devoted to measuring cognitive load.

2.1 Nature of Cognitive Load

A short history of how CLT evolved over time will be discussed (see Figure 1). In the early version of CLT, attention was placed upon schema acquisition and its relation to different types of cognitive processes associated with different problem solving approaches. Schema acquisition is the goal of learning and the core of problem solving. Learners with well-developed schemas are better problem solvers. So, facilitating learners to build schemas is an important step toward improving problem solving performance. However, it is important to ensure that the working memory limit is not exceeded during that process of building schema. The first stage of CLT attributes mental activities not associated with schema building as extraneous load which should be reduced by instructional design (see Stage 1 in Figure 1). Extraneous load alone accounts for the total load experienced by the learner (Sweller, 1988). Later, it was realized that the inherent content of material is irreducible and thus intrinsic load was identified as an independent source of cognitive load (see Stage 2 in Figure 1). Intrinsic load was specified by element interactivity – i.e. the number of interacting elements or pieces of information that the learner must attend to. In Stage 2, the total load includes intrinsic and extraneous load (Sweller, 1994). Lastly, germane load was identified as another source of cognitive load, with the understanding that germane can be increased to promote learning (see Stage 3 in Figure 1).
the three independent cognitive loads emerged, an assumption was made that the three cognitive loads were additive in nature. Based on this assumption, the sum of all three was referred to as the total cognitive load (Sweller et al., 1998). It is difficult, if not impossible, to test this assumption because there might be a complicated interaction between the three kinds of loads.

The following discussion is based on the three stages identified in Figure 1.

![Figure 1: Stages of Cognitive Load Theory (Figure Is Adapted from Chen et al. (2016))](image)

Based on cognitive load theory (CLT), schema construction and automation are critical goals of instruction (Paas et al., 2003). A limited working memory and a relatively unlimited long term memory plays an important role in CLT (Sweller, 1998). Paas and van Merrienboer defined cognitive load as the load imposed on human cognitive system when a learner is doing a given task (Paas & van Merrienboer, 1994). It is a major predictor of students’ learning performance. Many different factors may affect cognitive load. Task related factors include the inherent complexity of task, the representation of the task, time limitation and the pacing of the task. Learner related factors include prior knowledge level, age, and spatial ability (Paas et al., 2003b). This model also differentiates mental load from cognitive load, the former is about the
cognitive capacity demands anticipated by the learner and are thus a good prior estimate of
cognitive load. Mental effort indicates the actual cognitive resources allocated by the learner to
complete the task. Task performance in terms of final achievement of the learning goals can also
be used as a proxy for cognitive load. However, it has also been suggested that mental load,
mental effort and performance may tell us different information about the process of learning
during instruction (Paas et al., 2003b).

In Stage 3 of CLT history, the new framework differentiates three types of cognitive load. Intrinsic
cognitive load reflects the complexity of any learning material which is directly related
to the learning goal. It is affected by prior knowledge of a learner. Extraneous load does not help
learning. It depends upon the way information is presented to learners, such as using video,
audio, text or other formats as well as how the information is organized both spatially and
temporally. These features are under the control of instructional designers. For example, if
information is presented repeatedly, the learner will have to process redundant information
which will unnecessarily use some cognitive resources. Germene cognitive load is associated
with the schema building process and can be manipulated by instructional designers as well. The
best design entails moderate intrinsic cognitive load, as low extraneous cognitive load as
possible, and adequate working memory capacity left over for germane load. The three loads are
assumed to be additive, consistent with a triarchic construct of cognitive load. The measurement
of cognitive load, will be discussed in next chapter.

A much more careful illustration was made by Xie and Salvendy (2000) (Figure 2). They
identified and distinguished several different characteristics of cognitive load. Among them are
instantaneous load, peak load, accumulated load, average load, and overall load. Instantaneous
load fluctuates as time goes on reflecting the dynamics of mental activity. Peak load is the
maximum value of instantaneous load. Accumulated load is the total load imposed on the learner which can be calculated as an integral of instantaneous load over time. Average load is the accumulated load divided by the total time spent on the task. And overall load is the total load experienced during the task.

Figure 2: Illustration of Different Types of Load (Figure Is from Chen et al. (2016))

In model represented in Stage 3 of Figure 1, intrinsic load was based on the idea of element interactivity, while extraneous load and germane load were analyzed differently from case to case (Beckmann, 2010; Schnotz & Kurschner, 2007). As pointed out by Sweller (2010), it is necessary to seek a unified foundation for the three kinds of cognitive load to avoid any misunderstanding. In Stage 4, Sweller (2010) proposed a model in which intrinsic load, extraneous load and germane load were all formulated under the same theoretical idea of element interactivity (Figure 3). As per Sweller (2010), interacting elements related to the learning goal reflected the intrinsic load, and interacting elements not related to the learning goal constitute to the extraneous cognitive load. In other words, cognitive resources devoted to reducing the
number of interacting elements but not contributing to learning, pertain to the extraneous load, while cognitive resources devoted to processing intrinsic load were categorized as germane load. According to this formalism, the total cognitive load is the sum of intrinsic load and extraneous load. This is based the observation that when extraneous load goes down, germane load goes up, but the summation of them does not necessarily remain constant. Stage 4 assumes that learners have high motivation such that they use all available cognitive i.e. working memory resources to process intrinsic load and extraneous load. The working memory resources used to process the intrinsic load constitutes germane load, while the remaining working memory resources are used to process extraneous load. All our experiments conducted and reported in this dissertation were based on this latest (Stage 4) formulation of CLT.

Figure 3: Stages of Cognitive Load Theory (Incorporated Results from Sweller, (2010))

2.2 Cognitive Load Theory and the Cognitive Theory of Multimedia Learning

The cognitive theory of multimedia learning (CTML) uses empirical methods to explore
how instructional material can be designed to facilitate learning. Several empirical design principles were identified and demonstrated by various studies (Mayer, 2014; Mayer & Moreno, 1998, 2002). Ten principles were identified, based on cognitive load theory, which suggested how each type of load can be manipulated to facilitate learning (Mayer, 2014; Mayer & Moreno, 2003).

A. Three principles for managing intrinsic load

1. **Segmenting principle**: Presentation can be broken into multiple bit-size segments. Learners need to attend to only one segment at a point in time. In each segment, they have adequate time to process new information.

2. **Pre-training principle**: Students may receive prior instruction related to information of the main topic. Getting familiar with the concepts or terminology can help decrease the demand of mental resources.

3. **Modality principle**: If verbal information is presented in the form of narration, the demand on the visual channel will be reduced, so more space in visual working memory will be available for processing other visual elements.

B. Five principles for reducing extraneous load

4. **Coherence principle**: Learning material should be concise and coherent such that students will not need to process unnecessary information and focus more on the essential concepts.

5. **Signaling principle**: By cueing or labelling essential information we can help learners select and organize the material.

6. **Redundancy principle**: Information should be presented in the visual format as well as in narration. But identical information should not be provided in both formats so
mental resources will not be allocated to process the same information through two channels.

7. **Spatial contiguity principle**: Pictures and corresponding text should be placed as close in space as possible to make it easier to make connections.

8. **Temporal contiguity principle**: Relevant information should be presented to learners simultaneously instead of consecutively such that connections can be made more easily by the learner.

C. Two principles for fostering germane load

9. **Personalization principle**: If the words used in a multimedia lesson are in a conversational style rather than a formal style, learners will be more willing to engage with the material.

10. **Voice principle**: If the narration of a multimedia instruction is spoken in a friendly human voice rather than using a machine voice, the learners will be more willing to engage in the learning process.

The above mapping between the three types of cognitive load and CTML provides researchers a systematic way of manipulating intrinsic, extraneous, and germane cognitive load to optimize learning. For example, in Study One of this dissertation, we manipulated extraneous cognitive load using the redundancy principle such that learners receiving redundant information will be treated as having high extraneous load. In Study Three, we designed multimedia worked examples following closely the modality, signaling, temporal and spatial contiguity principles, such that multimedia worked examples imposed lower extraneous cognitive load than visual cues.
2.3 Criticisms of Cognitive Load Theory

Cognitive load theory has been successful in explaining why certain ways of instruction are more effective in promoting learning than other ways of instruction. However, there also have been some critiques of the conceptual and methodological issues related to CLT. Gerjets et al. (2009) helped clarify an important issue of CLT -- whether CLT can be treated as a scientific theory. Given the struggles in finding empirical indicators for the three basic types of cognitive load, whether they can be captured by any empirical metrics was brought into question. As argued by Karl Popper (1959, 1963), all assumptions of a scientific theory should be falsifiable. Based on that criterion, to falsify CLT would have to meet two requirements: First, measurements should be able to differentiate between intrinsic load, extraneous load and germane load; Second, these measurements should be made independent of measurements of learning performance. Many existing methods fail to satisfy these two prerequisites and researchers have expressed their skepticism about the possible existence of such measures (Schnotz & Kürschner, 2007). However, this issue can be avoided by taking a structuralist view of theory (Sneed, 1979), according to which certain fundamental assumptions of a theory can be treated as empirically non-testable. Rather, a theory is justifiable in terms of the extent to which its fundamental assumptions can provide a basis for a larger network of theories which allows researchers to develop testable predictions. Specifically, with regard to CLT, taking this structuralist view, we focus on testing the predictions made by CLT, while taking the assumptions of the triarchic model of CLT as a given.

De Jong, an important researcher in the field of CLT, has raised some insightful questions regarding the theory in its current state as well as future direction (De Jong, 2010). De Jong challenges the definition of cognitive load especially the notion that we can make distinctions
between the three types of cognitive load. He argues that since ontologically intrinsic load is based on the number of “objects” and germane load refers to generative “processes”, load in general exists only when “objects” are being “processed”. What is the real distinction between extraneous load and germane load since they are defined based on their function of whether they help or hinder learning? He also raised issues about how we can measure the three types of load due to a lack of clarity in their definition. Issues regarding several different methods were discussed in his critique.

Mayer (2010) addressed some of the issues raised by De Jong (2010) in his critique. Regarding the criticism based on the definitions of the three types of load, Mayer argued in favor of defining the three kinds of load in terms of underlying cognitive processes. Namely, intrinsic load is due to essential processing which is about initial selection and lower order of organization of information, extraneous load is about processing unnecessary information and germane load is equal to generative processing, which involves higher order organizing and integrating information with prior knowledge from long term memory. This conceptualization was partially validated in the study by DeLeeuw and Mayer (2008), which demonstrated that three different measures of cognitive load were sensitive to manipulations of the three kinds of cognitive load, supporting the triarchic construct of cognitive load. As mentioned earlier, Sweller updated CLT based on the idea of element interactivity in an effort to address the concerns and questions raised by others, such as De Jong. This new formalism (Stage 4 in Figure 3) has been introduced earlier in this chapter. All experiments conducted in this dissertation are based on this new formalism of CLT.
3. MEASUREMENT OF COGNITIVE LOAD

As stated in the previous chapter, cognitive load is a good predictor of learning outcomes, so how to measure cognitive load becomes a significant issue that has attracted a lot of attention. Based on Paas and van Merrienboer (1994a), cognitive load can be measured by mental load, mental effort and performance. Efforts to measure cognitive load can be categorized into two groups, one follows an analytical approach which estimates mental load based on expert analysis of learning material and computational modelling. The other one follows an empirical approach which can be further classified into subjective and objective measures. Subjective measures depend on collecting self-reported ratings of mental effort, task difficulty, and others. Objective measures include measuring response time to a secondary task and use of physiological signals to infer the cognitive load. In the following sections, all these different approaches and methods will be discussed and attention will be placed on eye-tracking based methods since they are the focus of this current study.

3.1 Analytical Approaches

There has been a long history of using computational method to analyze the cognitive processes (Langley & Neches, 1981; Larkin et al., 1980; Sweller, 1988). For example, John Sweller used an analytical approach to determine cognitive load (Sweller, 1988). A computational model was constructed to simulate the problem-solving behavior of novices particularly the means-ends analysis approach. The aim was to investigate the cognitive load imposed by conventional problem solving through a means-ends analysis compared to a non-specific goal approach. Results showed that a higher cognitive load could be imposed by the means-ends problem solving method than a non-specific goal approach. In addition to differences in cognitive load, the consequences of learning using both approaches are different.
A non-specific goal approach usually leads to schema construction, while means-ends analysis, which is a goal specific approach, leads to problem solution rather than acquiring of schemata.

3.2 Empirical Approaches

3.2.1 Subjective measurements

Use of subjective ratings to measure cognitive load started with NASA’s task load index or NASA-TLX which was originally designed to obtain a subjective estimate of workload (Hart & Staveland, 1988). The NASA-TLX is a multidimensional measurement tool including measurement of (1) mental demands, (2) physical demands, (3) temporal demands, (4) performance, (5) effort, (6) frustration level on a scale of 0 to 100. An overall load can be measured through a combination of all six dimensions. For CLT studies, researchers have modified the NASA-TLX to include only three dimensions, (1) task demands, (2) effort, and (3) navigational demands with each as a measurement of intrinsic load, extraneous load and germane load respectively (Gerjets et al., 2004, 2006).

Fred Paas was the first to demonstrate the feasibility of using a rating scale by participants to indicate subjective overall load in the context of cognitive load theory (Paas, 1992). Subjective rating presumes that people can reflect and report the mental effort that they exerted during a task. These subjective measures are usually in the form of a questionnaire consisting of multiple levels by which people can indicate the mental burden experienced by them. Most of these types of rating scales will ask people to report on multiple dimensions, such as mental effort, fatigue, and frustration. The ratings across different dimensions are normally highly correlated (Nygren, 1991). A unidimensional measure which asks for rating on only one of the dimensions has been shown to be quite reliable in measuring total cognitive load (Paas & van Merriendoer, 1994b). It has also been demonstrated to be able to differentiate small
differences in cognitive load (Gimino, 2002; Paas et al., 1994).

The subjective cognitive load rating questionnaire developed by Paas (1992) has been widely used to measure cognitive load (Paas et al., 1994, 2003a, 2003b, 2008; Van Gog et al., 2006, 2008). The typical question asked in these questionnaires is: “Please rate the amount of mental effort invested in the task”. Participants need to select an option ranging from “extremely low mental effort” to “extremely high mental effort” (Paas, 1992). These ratings can be collected in the middle of a task (Kalyuga et al., 2001) or immediately after the completion of a task (Paas, 1992; Van Gog et al., 2006, 2008). Although subjective ratings questionnaires seem to work well, the main disadvantage is that they cannot measure the fluctuation of instantaneous cognitive load, because they cannot be used too frequently during a task. Presumably the subjective ratings measure the accumulated load if collected right after a task, or the overall load if conducted after a series of tasks (Antonenko et al., 2010).

3.2.2 Objective Measurements

*Response time to a secondary task:* One objective measure of cognitive load that has also been used a lot is response time to a secondary task (RT). It is based on the idea that when engaged in a task, learners can only use the available space of working memory to respond to a secondary task which usually is irrelevant to the main task. Apparently, the less the available space, longer the response time (Brunken et al., 2003; DeLeeuw & Mayer, 2008; Rolfe, 1971). For example, in DeLeeuw and Mayer (2008), students were asked to press a button as soon as they noticed the screen background color started to change from red to pink as they were learning from multimedia materials.

*Physiological measurements:* Physiological measures of cognitive load include heart rate variability (Paas & van Merrienboer, 1994), variation of skin conductivity (Nourbakhsh et al.,
2012), and hormone levels (Wilson & Eggemeier, 2002). These methods that can discriminate levels of workload have not been widely adopted in the Educational Psychology literature (but have been used for much longer in the Cognitive Psychology literature).

We focus here on eye tracking based indexes that can measure cognitive load in a continuous manner and discuss them in detail. These physiological measures have been explored to a larger extent than the other physiological measures listed above. They are the focus of this dissertation.

Better assessment of instructional design can be achieved if instantaneous cognitive load can be measured and monitored (Antonenko et al., 2010). Oculomotor activity has long been studied in psychology, linguistics, and human computer interaction. Much of this research is based on the so-called “eye-mind assumption” which says that the information people are looking at is concurrently being processed in their mind (Just & Carpenter, 1980). This also provides justification for why eye movements can tell us information about the concurrent cognitive load of students. Thus, the cognitive activity of learners can be inferred by their visual attention, as measured by their eye movements (EMs) and pupillometry (i.e. measurement of pupil size), which can tell us many things about the cognitive load. It is suggested by many authors that eye movements and pupillometry based measures, due to their nonintrusive nature, can be used in human-computer interfaces that designed for instructional purposes (Chen et al., 2011; Gerjets et al., 2014; Gevins & Smith, 2003; Smith et al., 2001).

Pupil dilation has long been investigated and shown to be correlated with level of task difficulty. Hess and Polt (1964) have showed that dilation of pupil size increased with increasing difficulty level of problems. Van Gerven et al. (2004) have shown similar results that pupil size is positively correlated with imposed workload but it is also moderated by aging. Marshall
presented a metric called the Index of Cognitive Activity (ICA), which is based on changes in pupil dilation (Marshall, 2002). The ICA is a measure of the frequency of discontinuities in the continuously recorded pupil size signal. There are two major sources for the discontinuities: one is due to the ambient lighting and the other is due to cognitive processing. It was demonstrated that ICA could differentiate between tasks of different levels of complexity in a math context and was found not to depend on lighting condition.

Mean eye fixation duration is well-known to be a good indicator of cognitive load (for reviews, see Nuthmann & Henderson, 2012; Rayner, 1998). The duration of a fixation means how long one looks at one specific location. Researchers have compared it with pupil dilation. Meghanathan et al. (2015) conducted a visual search task, during which mean fixation duration and pupil dilation were recorded. In their experiment, three, four or five targets were randomly mixed among a set of distractors with both displayed in various orientations. The numbers of targets manipulated the imposed load. Participants were asked to locate all the targets and asked if one of the targets has changed orientation in a subsequent display. They found that fixation duration was correlated with levels of load both within and beyond working memory limit of participants. However, pupil size only correlated with levels of load when it exceeded the limits of working memory. In other words, mean fixation duration was sensitive to memory load and processing load while pupil size was only sensitive to processing load. Here memory load is related to the amount of information needing to be carried in short term memory and processing load is related to processing of information.

A study by Chen et al. (2011) investigated a series of measures based on eye movements and pupil dilation during a video watching task. The task required participants to watch a basketball game clip and identify defenders from attackers and later recall their positions. The
complexity of the task was manipulated and the low, medium and high levels of complexity were validated by a self-reported survey. The results showed that pupil size, saccadic length, blink rate, fixation duration and fixation rate all differentiated the levels of complexity. Especially, saccadic length decreased as required workload increased and fixation duration increased with required workload. All measures were correlated with each other except with the mean and standard deviation of pupil size. No explanation was given in this study regarding why eye movement based measurements were not correlated with pupil dilation data.

Another eye movement based measure, saccadic peak velocity, has been shown to differentiate between different levels of mental workload. Di Stasi et al. (2010) investigated whether saccadic peak velocity can differentiate three different levels of workload in a driving simulation task. The levels of complexity indicating workload were validated by a subjectively rated survey, driving performance, and secondary task performance. They found that saccadic peak velocity could separate low from medium/high workload conditions but not between medium and high workload conditions. Since from the literature, we also knew that the saccadic peak velocity is also strongly correlated with the saccade length and the saccade duration based on main sequence of eye movements (Bahill et al., 1975). It is not surprising to see saccade length can also measure cognitive load in CLT studies. We found that mean saccade length and saccade duration were both sensitive the same type of cognitive load in our Study One.

### 3.3 Measuring the Three Types of Cognitive Load

Since the identification of the three kinds of cognitive load (Sweller et al., 1998; Sweller, 2010), assessing instructional design and predicting performance of learning has been getting more sophisticated. There have been studies investigating possible ways to measure either one of the three kinds of cognitive load or all three of them three together.
Researchers have shown that subjective ratings methods can differentiate tasks when only one type of cognitive load is manipulated. The subjective ratings surveys usually included only one item asking for example the perceived difficulty of the materials (Kalyuga et al., 1998) or mental effort devoted to the learning task (Ayres, 2006). Kalyuga et al. (1998) showed that the perceived difficulty of the materials by the learners could be used to measure the extraneous load in a context where a split-attention effect was investigated. The split-attention effect was used to manipulate the extraneous load in their tasks. Ayres (2006) demonstrated that by appropriately designing a task it is possible to use subjective ratings to measure the intrinsic cognitive load. Ayres recruited 8th and 9th graders to work through an arithmetic bracket-expansion task. He argued that the intrinsic complexity is not evenly distributed among the four operations of the task employed and predicted that a higher error rate would occur with operations of high numbers of interacting elements such as the second operation has higher numbers of interacting elements than the first operation in the formula: - (5 - 6x). He also assumed that there would be minimal extraneous load imposed on problem-solvers if there was no instructional intervention provided. Further, he also assumed that germane load would be minimal and held constant throughout the task since the task was an assessment rather than facilitating building of new schema. Based on these assumptions, Ayres argued that any variation in subjective rating could be attributed to differences of intrinsic load. He found that operations with higher numbers of interacting elements are rated with greater difficulty. Further, he found a statistically significant correlation between subjective ratings and error rates. So, unlike previous researchers who had used a subjective rating to measure overall cognitive load, Ayres demonstrated the possibility of employing subjective rating to measure intrinsic cognitive load. Provided the extraneous load
and germane cognitive load were kept constant, any change in subjective rating would be due to the difference in intrinsic cognitive load.

It has also been shown that the subjective ratings methods can be used to measure more than one kinds of cognitive load. These subjective ratings had multiple items each targeting one type of cognitive load. For example, Windell et al. (2006) found that when a comparison between NASA-TLX and subjective difficulty rating was conducted, with clear manipulation of extraneous load and intrinsic load, the NASA-TLX “mental demands” subscale as well as difficulty rating can differentiate tasks imposing different levels of intrinsic load. Further, they found that the NASA-TLX weighted workload score was sensitive to both intrinsic load and extraneous load. A multidimensional modified version of NASA-TLX task was designed to measure the intrinsic, the extraneous, and the germane load (Gerjets et al., 2004, 2006). The intrinsic load was manipulated by the complexity of the materials used in (Gerjets et al., 2004, 2006). The germane load was manipulated by including instructional explanations or self-explanation prompts in (Gerjets et al., 2006). Cierniak et al. (2009), on the other hand, developed a three-item rating survey. In their survey, difficulty level of content was used to measure intrinsic load, difficulty level of the interacting with material was used to measure extraneous load, level of concentration was used to measure germane load. Their results indicated that an increase in extraneous load was associated with a decrease in germane load, but the measurement of intrinsic load was less successful.

The aforementioned subjective ratings methods used only one item to measure each type of the cognitive load. Subjective ratings surveys using multiple items to measure each type of the cognitive load were also developed. For example, Leppink et al. (2013) developed a ten-item subjective rating survey to measure the three types of loads in a statistics context. Each type of
the cognitive load was either measured by three (intrinsic load & extraneous load) or four items (germane load). The results provided evidence supporting the reliability and validity of their survey.

There were also studies that used more than one instrument to investigate the measurement of the three kinds of cognitive load. For example, DeLeeuw and Mayer (2008) conducted a study to answer the question of whether cognitive load is a unitary or triarchic construct and whether the two subjective ratings surveys and one objective response time to a secondary task measure were tapping different load constructs. In their work, the intrinsic cognitive load was manipulated by changing the complexity of the sentences in the learning materials, which served as a within-subject factor; the extraneous cognitive load was manipulated by providing redundant materials, which was a between-subject factor; and the germane cognitive load was measured by the transfer test performance of the participants, which was based on the assumption that the transfer performance was a measure of the mental resources devoted to learning during the instruction phase. They found that response time to a secondary task was most sensitive to the manipulation of the extraneous cognitive load, the subjective mental effort rating was most sensitive to the manipulation of the intrinsic cognitive load, and the overall difficulty rating was most sensitive to the manipulation of the germane cognitive load. A triarchic construct of cognitive load was supported by this work and the three measures were found to be sensitive to manipulations of different kinds of cognitive load.

3.4 Current Limitations

In general, different cognitive load measures are responsive to manipulations of workload depending on the context of the task. However, most of the studies were manipulating the total imposed workload and did not specify which type of load was being manipulated. It is not clear
if the measures employed were tapping the total cognitive load or different sub-types of cognitive load.

Interestingly, several different measures are sensitive to manipulations of different types of cognitive load as demonstrated by DeLeeuw and Mayer (2008). Nevertheless, some limitations of their study should be pointed out. Their objective measure of response time to a secondary task was intrusive in nature and interrupted the learning process, as was the subjective mental effort rating measured during the lesson. The subjective difficulty rating conducted after the lesson was not intrusive but was administered at the end of learning, and thus failed to capture the variations of load during learning. However, with the advantage of computers and multimedia based learning, monitoring the level of cognitive load of learners and providing real-time feedback is possible. Better design of smart human-computer interfaces allows unobtrusive and continuous measuring and monitoring of cognitive load (Gerjets et al., 2014; Gevins & Smith, 2003; Smith et al., 2001). As mentioned above, eye movement based methods and pupil dilation are good candidates to be used in HCI design. It is important to examine the relationship between these measures with cognitive load, specifically, their sensitivity to different sub-types of cognitive load.

In the measurement of cognitive load, we should pay attention to potential interactions between the three kinds of load. For example, in a study by Cierniak et al. (2009), they investigated the split-attention effect in terms of cognitive load theory. One example of split-attention effect is that when diagram presenting the same information is placed at a distant location relative to the equations need to be learned. In their study, participants were randomly assigned to two conditions: one with an integrated format, the other with a spatially separated format. Secondary task performance was employed to measure the overall cognitive load.
Subjective ratings were used to measure intrinsic load, extraneous load, and germane load respectively. The results showed that students had worse posttest performance in the spatially separated format condition, but did not do worse on the secondary task. Students’ subjective ratings on extraneous load was smaller when they were doing with the integrated format than for that in the form of split-attention. In the meanwhile, the ratings for the germane load were lower for the material in the form of split-attention than those in integrated format. This indicates that split-attention effect does not only involve the reduction of extraneous load but also concurrently an increase of germane load. Studies of this kind, which manipulate and measure different kinds of cognitive load have implications for the design of learning materials. The next chapter, describes the first study in this dissertation which uses eye movement measures to investigate the different kinds of cognitive load experienced by learners as they participate in a multimedia lesson.
4. STUDY ONE: USING EYE MOVEMENTS TO INVESTIGATE

COGNITIVE LOAD THEORY

4.1 Introduction

Cognitive load theory (CLT) provides guiding principles for designing effective instructional materials (Sweller, 1988, 1998). According to CLT, learning is best fostered if the cognitive load imposed during instruction does not exceed the working memory capacity of the learners (Mayer, 2005; Mayer & Moreno, 2003; Paas & Sweller, 2014). The reasoning is as follows. First, information to be learned must be processed in working memory before it can be encoded into long-term memory. But, although humans’ long-term memory capacity for learning is unlimited, we have limited working memory capacity (Cowan, 2002). Thus, if cognitive load uses up this limited working memory capacity during instruction, it creates a bottleneck that greatly limits the amount of learning material that can be encoded into long-term memory.

CLT differentiates three distinct kinds of cognitive load: intrinsic load (due to the complexity of the material to be learned), extraneous load (present in the instructional materials, but irrelevant to the learning objectives), and germane load (due to elaborative processing by the student during learning). According to Sweller (2010), the construct of all three kinds of cognitive load can be unified under element interactivity, which is further determined by the number of interacting elements that the learner must attend to. Intrinsic cognitive load (hereafter, intrinsic load) refers to the number of interacting elements that need to be stored and processed simultaneously in working memory to learn a concept. It is conjointly determined by the instructional objective and the concurrent knowledge level of learners. According to Sweller (2010), extraneous cognitive load (hereafter, extraneous load) is also determined by element
interactivity. In the case of extraneous load, however, the element interactivity pertains to those elements processed by the learner that are unrelated to the learning objective. Non-optimal design of instructional materials increases the interactivity of unnecessary elements that need to be processed by the learner and causes high extraneous load. Germane cognitive load (hereafter, germane load) is also indirectly related to element interactivity since it is determined by the cognitive working memory resources devoted to processing intrinsic load. Germane load facilitates learning by reducing number of interacting elements relevant to learning by constructing what you are learning to what you already know. It is theorized that extraneous load and germane load are complementary in a sense that an increase in germane load is accompanied by a decrease in extraneous load.

Given the above multifaceted triarchic nature of cognitive load, it is necessary to find ways to measure the three kinds of cognitive load separately. Many methods of measuring cognitive load have been shown to differentiate tasks imposing different levels of cognitive load, such as subjective ratings (Paas, 1992), measures of reaction time to secondary tasks (Brunken et al., 2003), heart rate variability (Paas & van Merrienboer, 1994), variation of skin conductivity (Nourbakhsh et al., 2012), hormone levels (Wilson & Eggemeier, 1991), and eye-movement based methods (Marshall, 2002; Meghanathan et al., 2015). But efforts to specifically measure CLT’s three types of cognitive load have also been made in the past decade since they were first proposed (Sweller, 1998).

Kalyuga et al. (1998) hypothesized that material which splits the learners’ attention will impose higher extraneous load. They examined this conjecture by asking students how easy or difficult to understand the material was on a Likert scale from 1 (extremely easy) to 7 (extremely
difficult). They found that the self-reported difficulty rating could differentiate these two levels of extraneous load.

Gerjets et al. (2004) compared the effect of molar and modular presentation of solution procedures on performance and cognitive load of students. Solutions in modular form are based on problem categories and their associated solution steps. Solutions in modular form breaks complex solution steps into smaller meaningful solution pieces that can be understood separately. Thus, a modular presentation would impose lower intrinsic load. A three-item self-reported survey developed based on NASA-TLX was administered to measure intrinsic load. They found students had lower ratings for stress, devoted effort, and task demands when they were provided worked solutions in modular form.

Ayres (2006) investigated the possibility of using a subjective mental effort rating to measure intrinsic load in an arithmetic context. The number of interacting elements was manipulated in arithmetic operation tasks. It was also argued that since this task was an assessment task rather than a learning task, extraneous load and germane load would be constant between conditions. It was found that students rated arithmetic operations involving more steps as more effort demanding and made more mistakes.

Windell et al. (2007) investigated the sensitivity of NASA-TLX and subjective difficulty rating to intrinsic load and extraneous load. Intrinsic load was manipulated by adjusting the complexities of instructional material and extraneous load was increased by providing redundant information. They found that the NASA-TLX subscale of mental demands as well as difficulty ratings can differentiate tasks imposing different levels of intrinsic load, and NASA-TLX weighted workload score was shown to be sensitive to both intrinsic load and extraneous load.
Cierniak et al. (2009), on the other hand, developed a three-item rating questionnaire survey. In their survey, difficulty level of content was used to measure intrinsic load, difficulty level of interacting with material was used to measure extraneous load, concentrating level was used to measure germane load. Their results indicated that an increase in extraneous load was associated with a decrease in germane load, but measuring intrinsic load was less successful.

Observing that most rating surveys used included only one item for each type of cognitive load, Leppink et al. (2013) developed a ten-item subjective rating survey to measure the three types of loads in a statistics context. Each type of load was either measured by three (intrinsic load & extraneous load) or four items (germane load). Their results provided evidence supporting the reliability and validity of their survey.

Many studies using self-reported surveys assume that certain questions target certain types of load when one or more types of load were being manipulated. However, it is also important to examine the sensitivity of different measures toward each type of load without presuming their sensitivity. Following this line of thinking, DeLeeuw and Mayer (2008) explicitly manipulated the three types of cognitive load and showed that a subjective mental effort rating was most sensitive to intrinsic load, an objective measure of reaction time to a secondary task was most sensitive to extraneous load, and a subjective difficulty rating was most sensitive to germane load. Their results were consistent with the claim of CLT that cognitive load is not a single construct but rather has a triarchic nature, and that the three types of cognitive load can be measured independently. Nevertheless, the above methods could be intrusive when used in real time monitoring of cognitive load. For example, participants are stopped during learning at the measurement points and perform either a subjective rating task or press a button in response to a secondary task, which themselves may add extraneous load. Finding objective
and non-intrusive ways to measure cognitive load in a continuous manner is necessary not only to test CLT, but could also be advantageous in human-computer interfaces designed to facilitate learning (Chen et al., 2011; Gerjets et al., 2014; Gevins & Smith, 2003; Smith et al., 2001).

One possible set of measures that meet all the above criteria is eye movement metrics. For example, pupil dilation has long been investigated and shown to be correlated with the levels of difficulty of tasks (Ahern & Beatty 1979; Beatty, 1982; Beatty & Lucero-Wagoner, 2000; Chen et al., 2001; Gerven et al., 2004; Granholm et al., 1996; Hess & Polt 1964; Kahneman & Beatty, 1966). A measurement called index of cognitive activity (ICA), which is based on changes in continuously recorded pupil size, was shown to differentiate between levels of mathematical problem complexity and found to be independent of lighting condition (Marshall, 2002, 2007; Marshall et al., 2004).

Mean fixation duration was found to be good indicator of cognitive load as well (Chen et al., 2011). Longer mean fixation duration has been found when a word of low frequency was encountered in reading (Rayner & Duffy, 1986) and is suggested as indicating more extensive processing (Nuthmann & Henderson, 2012; Rayner, 1998). Longer mean fixation duration is also found when information is more difficulty to understand (Carroll et al., 1992). Fixation duration increases as processing demands (task difficulty) increases (Underwood et al., 2004). It was found that, in free viewing, when three, four or five targets were presented with distractors as variation of cognitive load, mean fixation duration increased as workload increased (Meghanathan et al., 2015).

Other eye movement related parameters, such as saccade length, blink rate, and fixation rate were all able to differentiate tasks complexity levels in a video clip watching task (Chen et al., 2011) and saccadic peak velocity, was also shown to differentiate between different levels of
mental workload in a driving simulation task (Di Stasi et al., 2010). It was found that when workload increased, saccadic length decreased and fixation duration increased (Chen et al., 2011). This is consistent with findings from psychological studies (Ringer et al., 2016).

The above eye movement studies showed the feasibility of using eye-movements to measure cognitive load. However, no clear manipulation of the three kinds of load was considered in their designs, such that no conclusion could be made regarding which type of load was measured by these metrics. As mentioned earlier, finding ways to measure the three different kinds of load independently could potentially facilitate improving instructional material design and consequently learning performance. In this work, we investigate the sensitivity of several eye movement parameters to the three kinds of load: intrinsic load, extraneous load, and germane load.

As mentioned earlier, CLT relies on the cognitive structure of working memory (WM). A key component of Baddeley’s seminal WM model is the central executive (Baddeley, 1986), which is similar, if not identical, to executive attention (Baddeley, 1996; Fan et al., 2005). A key function of the central executive is to maintain one’s focus on relevant information by inhibiting irrelevant elements in working memory, which otherwise take up precious WM capacity (Engle et al., 1995). It follows that inhibiting irrelevant distractors in WM requires executive attentional resources. Such irrelevant elements in WM, and the drain on attentional resources needed to inhibit them, can be understood as Sweller’s extraneous load. Executive attentional resources are also essential to the encoding and maintaining of the contents of WM, because both rely on the same fronto-parietal network of brain areas (Awh & Jonides, 1998; LaBar et al., 1999; Linden et al., 2003; Mayer et al., 2007; Zanto et al., 2011). Information maintained and processed in working memory can be attention-demanding or routinized depending on the task and the one’s
prior knowledge and skill level with it. The attentional and WM resources required for storing and processing relevant information in WM can thus be understood as constituting Sweller’s (2010) intrinsic and germane loads.

Thus, it becomes necessary that individual differences in WM capacity (WMC) must be taken into consideration when measuring manipulated cognitive loads. WMC has also been shown to be positively correlated with one’s fluid intelligence (Unsworth et al., 2005). As mentioned earlier, fluid intelligence is the ability to use logic to analyze novel problems such that reasoning efficiency varies between subjects. This is similar to the generative processing of making connections with what one knows with what needs to be learnt (Mayer, 2010). Thus, fluid intelligence may further affect the level of germane cognitive load. Furthermore, according to Sweller’s (2010) latest formalism of CLT, WMC is a summation of extraneous load and germane load provided a high motivation level is presumed, such that WMC is inevitably a factor in determining the three types of cognitive load and should be considered in design and analysis in CLT studies.

Furthermore, relationships between subjective ratings and working memory capacity did not receive a lot of attention in the past. Rudner et al. (2012) showed a negative correlation between WMC and perceived effort in a noisy listening task. In terms of the relationship between WMC and perceived difficulty, women under stereotype threat showed lower WM capacity and higher difficulty ratings toward the working memory task (Schmader & Johns, 2003). If eye-tracking based parameters are found to be sensitive to any type of cognitive load, it is necessary to take differences in WMC of participants into account and confirm whether that correlation remains significant, especially when different physiological measures are related to WMC. We argue that any cognitive load study should take WMC into consideration in the design of
experiments. In this work, we measured WMC of participants by adopting an easy-to-use automatic Operation Span task which can be conducted on a computer (Unsworth et al., 2005). This task is to be introduced in detail shortly.

This study addresses the following research question: 1) To what extent are various eye-tracking based parameters sensitive to the three different kinds of cognitive loads? 2) Are the assumptions underlying Sweller’s (2010) model of CLT valid?

Our methodological approach was to closely replicate the DeLeeuw and Mayer’s (2008) experiment and extend it by 1) tracking learners’ eyes to gather several carefully chosen eye movement metrics that have been shown to measure cognitive load, and 2) measuring learners’ O-span to measure their WMC.

To foreshadow our results, we successfully replicated all of DeLeeuw and Mayer’s (2008) original pattern of results. We then correlated our eye movement measures with each of DeLeeuw and Mayer’s original measures that were argued to represent Sweller’s three types of cognitive load. We found specific eye movement parameters that correlated most strongly with each of Sweller’s three types of cognitive load. Finally, we tested the assumptions of Sweller’s CLT by modeling our results using both DeLeeuw and Mayer’s original measures and our eye movement-based measures of cognitive load together with the O-span measure of WM capacity. These analyses showed that mean fixation duration and mean saccade length and others were most sensitive to extraneous load. Number of transitional saccades between different areas of interest was most sensitive to intrinsic load. Total dwell time spent over a key animation area of interest and mean saccade peak velocity were most sensitive to germane load. The above results remain the same even after controlling for differences in working memory capacity of the
learners. We also found that Sweller’s (2010) model was supported by a multilevel model analysis.

4.2 Methodology

4.2.1 Manipulation of the Three Kinds of Cognitive Load

The participants were N = 51 students (48 females, 3 males) enrolled in an elementary education class in a U.S. Midwestern land grant university. Volunteers were paid $30 each to participate in this 90-minute long study. Most of the participants were sophomores or juniors and fewer than half of the participants had taken a physics class in high school, as indicated by a questionnaire collected at the end of this experiment. The learning material was a multimedia lesson explaining how an electric motors work. Figure 4 shows a screenshot of the learning material. The complete materials can be found in Appendix A1.

We manipulated the three types of cognitive load following the design of DeLeeuw and Mayer (2008). Extraneous load was manipulated as between-subject factor, using the redundancy effect (Mayer, 2005). Redundant effects states that the same information should not be presented in different modalities since it will waste valuable cognitive resources. Participants were randomly assigned to one of two conditions. The non-redundant condition was presented a version of the multimedia lesson with animation and narration. The redundant condition was presented a version with animation and narration as well as concurrent onscreen text that replicated the narration. Since participants needed to process this redundant textual information while listening to the narration, this redundancy increased the extraneous load as evidenced by a lower performance on the transfer post-test (Mayer, 2005).

We administered a pre-survey to gauge the prior knowledge of each of our participants. Based on this survey, all participants in our study were found to have low prior knowledge about
the learning materials. Intrinsic load was manipulated in terms of sentence complexity which determines how many elements need to be processed simultaneously necessary to fully digest the concepts. Since each sentence includes different numbers of key concepts for understanding the materials. Those sentences having high numbers of key concepts which should be understood altogether imposed high intrinsic load. When determining the key concepts, we should take the prior knowledge of the students into consideration. Sentence complexity was a within-subject factor. The multimedia lesson was broken into nine segments. Four of them were identified as having high-complexity and four of them low-complexity. Example low- (as highlight in blue) and high-complexity (as highlighted in red) sentences can be found in figure 4.

![Image of a circuit diagram](image.png)

**Figure 4: Screenshot of Multimedia Lesson.**

Germane load depends on learners’ interaction with the learning material, and consequently their performance on a transfer test was used to designate them as having high or low germane load during learning phase. Transfer tasks included seven conceptual questions presented in a specific order, with a time limit of three minutes per question. If participants had a better than average score on the transfer task, they were designated as having experienced high
germane load, and those with a lower than average score were designated as having experienced low germane load. This is based on the postulation that higher germane load during learning would result in better performance on the transfer test.

4.2.2 Procedure

**Pre-Survey**: The study procedure is shown in Figure 5. A short pre-survey was given to ascertain participants’ prior knowledge about the material. In the pre-survey, five images (from Figure 24) describing specific physics phenomena from a commonly used introductory physics textbook (Hewitt, 2014) were shown to them. One of the images resembled the image used in this study. They rated their knowledge level of each physics phenomenon represented by each image on a Likert scale from 1 (I don’t know anything about this) to 5 (I understand this very well). Written answers reflecting their understanding of the physics related to each image were also collected. The latter was used as confirmation of their Likert scale ratings. The pre-survey was conducted using the online platform Qualtrics. The images were presented in a randomized sequence.

![Figure 5: Interview Procedure](image)

**Multimedia Lesson**: Following the pre-survey was the multimedia lesson. Participants were randomly assigned to the redundant and non-redundant conditions described above. To make our manipulations consistent with DeLeeuw and Mayer (2008), measurements of cognitive load used in that study were also included in this design: (i) Reaction time to a secondary task which asked participants to press the spacebar immediately after noticing the screen background
color starts to change from light gray to purple at the end of each of the nine multimedia lesson segments. (ii) Participants’ *subjective mental effort rating* on a scale from 1 (extremely low mental effort) to 9 (extremely high mental effort) at the end of each of the multimedia lesson segments. Once they rated their current mental effort, the multimedia lesson proceeded to the next segment. (iii) Participants’ *subjective difficulty level rating* of the learning material on a scale from 1 (extremely easy) to 9 (extremely difficult) at the end of the entire lesson. A snapshot of mental effort rating and difficulty rating can be found in Figure 25 and Figure 26. In addition to these measures, we also recorded participants’ eye-movements during the learning phase using an EyeLink 1000 + eye-tracker from SR research. The sampling rate is 1000H with the eye-tracker in tower-mounted mode. Participants were calibrated on a 9-point grid with a maximum error smaller than 1 degree and average error smaller than 0.5 degree.

**Transfer Test:** In the transfer test, participants answered seven questions which assessed their understanding of the material they just learnt. Questions were asked online by using Qualtrics and presented in a pre-determined order. Participants had a maximum of three minutes to answer each question. Their answers were later coded and graded according to an answer rubric based on the correctness of the answer. Questions can be found in Appendix A3.

**Working Memory Span Task:** Individual differences in working memory capacity (WMC) were identified as a potential confounding variable in this cognitive load study. At the end of each interview, we measured each participant’s WMC using the Operation Span task developed by Unsworth et al., (2005). The task used in this study was conducted on a computer. Participants solved simple arithmetic operations such as \(8/4 + 2 =?\) and memorized a letter which was shown immediately after each mathematical operation. In each set, the total number of operations varied from three to seven, hence participants needed to memorize between three to
seven letters in each trial. After each trial, participants needed to report the memorized letters back in the presented order by clicking them from a letter-matrix. In total, there were 75 operations and thus 75 letters. Participants’ data were used only if their performance on operation part satisfied an 85% correctness rate. This guarantees that there is no trade-off between processing the math task and memorizing the letters. A score indicating working memory capacity was automatically generated at the end of the task with a possible maximum score of 75.

4.2.3 Controlling the Effect of Luminance

Since the initial study by Hess and Polt (1964), pupil size variance has been explored extensively especially in CLT studies (Chen et al., 2011; Van Gerven et al., 2004). However, we know that pupil size can be affected by other factors such as visual field luminance of the source image, which is of core interest to CLT studies (Klingner, 2010; Klingner et al., 2001; Kun, 2012). Because luminance imposed at different locations of the image used in this study varies, we used a parameter titled normalized pupil size change (NPSC) from a baseline to indicate the percentage change of the pupil size. We collected baseline pupil data for each participant. Different locations from the images of the multimedia lessons were Fourier transformed into white noise while maintaining the same overall luminance. For the non-redundant condition, three representative locations from the multimedia lesson display were chosen (from the animation); and six representative locations were chosen for the redundant conditions (three from the animation part and three from the text part). Areas of interest (AoS) were created surrounding each representing location. Baseline data were collected right before the start of the multimedia lesson. Participants were asked to fixate on a small red cross at each representative location for 10 seconds. Baseline pupil size was determined by the minimum stable magnitude of
pupil size at that location. NPSC was calculated as the difference between actual pupil size recorded and the baseline pupil size recorded at a location within the same AoI. This was done to ensure that NPSC could be attributed to the imposed cognitive load variance, by controlling for the variance due to luminance.

4.3 Results

4.3.1 Pre-survey, Transfer Test and O-span Task

The mean (± S.D.) pre-survey participants’ rating of their understanding of the learning material was 1.29±0.67 on a scale of 1 through 5, which indicates that they all had low prior knowledge about electric motors. This result was consistent with participants’ written answers on the pre-survey. The mean (± S.D.) transfer test score was 7.14±3.20. A mean-split rendered 28 participants (below the mean) in the low-germane load group and 18 participants (above the mean) in the high-germane load group. The O-span task scores were normally distributed with a mean (± S.D.) of 32.72±15.90, with half participants scoring above the mean and half below the mean.

4.3.2 Non-eyetracking Based Metrics

Response Time to Secondary Task: A 2 x 2 mixed ANOVA was conducted to test the effect of the within-subject factor (sentence complexity) and the between-subject factor (redundancy) on response time to the secondary task. We found that there is a significant main effect of redundancy on response time to the secondary task, F(1, 44) = 4.985, p = .031, η² = .102. The redundant condition (high extraneous load) took a longer time to press the spacebar than those from the non-redundant (low extraneous load) condition. No significant main effect of sentence complexity was found. An independent samples t-test to compare the response time of low- and high-germane load groups, found no significant difference, t(44) = -0.400, ns. (Table
2). Internal-reliability of reaction time to the secondary task was measured by Cronbach’s alpha: 

\[ \alpha = .864 \] for low complexity videos and \[ \alpha = .814 \] for high complexity videos.

<table>
<thead>
<tr>
<th>Measures of cognitive load</th>
<th>Redundant</th>
<th>Non-redundant</th>
<th>High</th>
<th>Low</th>
<th>High (n = 45)</th>
<th>Low (n = 51)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Response time</td>
<td>2918(872)</td>
<td>2520(797)</td>
<td>2677(869)</td>
<td>2769(974)</td>
<td>2569(848)</td>
<td>2859(847)</td>
</tr>
<tr>
<td>Effort rating</td>
<td>5.67(1.59)</td>
<td>4.99(1.67)</td>
<td>5.47(1.66)</td>
<td>5.21(1.68)</td>
<td>5.58(1.58)</td>
<td>5.13(1.71)</td>
</tr>
<tr>
<td>Difficulty Rating</td>
<td>5.33(1.83)</td>
<td>5.21(1.74)</td>
<td>4.82(1.92)</td>
<td>5.67(1.56)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. Highlighted boxes indicate significant differences (\( p < .05 \))

<table>
<thead>
<tr>
<th>Measures of cognitive load</th>
<th>Redundant</th>
<th>Non-redundant</th>
<th>High</th>
<th>Low</th>
<th>High (n = 18)</th>
<th>Low (n = 28)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Response time</td>
<td>1245(512)</td>
<td>968(305)</td>
<td>1113(512)</td>
<td>1101(446)</td>
<td>1139(515)</td>
<td>1086(393)</td>
</tr>
<tr>
<td>Effort rating</td>
<td>5.45(1.67)</td>
<td>4.41(1.51)</td>
<td>5.17(1.47)</td>
<td>4.70(1.48)</td>
<td>4.88(1.71)</td>
<td>4.97(1.27)</td>
</tr>
<tr>
<td>Difficulty Rating</td>
<td>5.39(1.50)</td>
<td>5.22(1.28)</td>
<td>4.78(1.59)</td>
<td>5.64(1.13)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. Highlighted boxes indicate significant differences (\( p < .05 \))

*Mental Effort Rating:* A 2 x 2 mixed ANOVA was conducted to test the effect of within-subject factor (sentence complexity) and between-subject factor (redundancy) on subjective mental effort rating. We found that there is a significant main effect of redundancy on mental effort rating, \( F(1, 44) = 6.652, p = .013, \eta^2_p = .102 \). The redundant condition reported higher
mental effort is than the non-redundant condition. We also found a significant main effect of sentence complexity, $F(1, 44) = 24.9, p < .001, \eta^2_p = .361$, with participants rating their mental effort higher with high-complexity sentences than lower-complexity sentences. An independent samples $t$-test to compare the means of subjective mental effort rating of low- and high-germane load groups found no significant difference, $t(44) = .223, ns.$ (Table 2). Internal reliability of subjective mental effort rating is measured by Cronbach’s alpha: $\alpha = .892$ for low complexity videos and $\alpha = .927$ for high complexity videos.

Table 3: Correlations b/w Dependent Measures for All Students of (DeLeeuw & Mayer, 2008)

<table>
<thead>
<tr>
<th>Measure</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Response time measure</td>
<td>--</td>
<td>.12</td>
<td>.13</td>
<td>-.30**</td>
</tr>
<tr>
<td>2 Mental effort rating</td>
<td>--</td>
<td>--</td>
<td>.33**</td>
<td>.11</td>
</tr>
<tr>
<td>3 Difficulty rating</td>
<td>--</td>
<td>--</td>
<td>-.22*</td>
<td>--</td>
</tr>
<tr>
<td>4 Posttest score</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
</tbody>
</table>

Note: $N = 46$. * $p < .05$. *** $p < .001$.

**Difficulty Rating:** Self-reported difficulty ratings were collected at the end of the learning phase. An independent samples $t$-test to test the effect of redundancy and transfer performance

Table 4 Replicated Correlations b/w Dependent Measures of All Students

<table>
<thead>
<tr>
<th>Measure</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Response time measure</td>
<td>--</td>
<td>.03</td>
<td>-.05</td>
<td>-.12</td>
</tr>
<tr>
<td>2 Mental effort rating</td>
<td>--</td>
<td>--</td>
<td>.51***</td>
<td>-.11</td>
</tr>
<tr>
<td>3 Difficulty rating</td>
<td>--</td>
<td>--</td>
<td>-.35*</td>
<td>--</td>
</tr>
<tr>
<td>4 Posttest score</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
</tbody>
</table>

Note: $N = 46$. * $p < .05$. *** $p < .001$.

found no significant differences between difficulty ratings of the redundant and non-redundant conditions, $t(44) = .423, ns.$ The high-germane load condition rated the material as significantly less difficult than the low-germane load condition, $t(44) = 2.157, p = .037$. Our results (Table 2) are consistent with those of DeLeeuw and Mayer (2008) (Table 1).
Correlation Between Different Metrics: We determined the correlation between response time to secondary task, difficulty rating, mental effort rating and post-test score. Our results (Table 4) are consistent with that from DeLeeuw and Mayer (2008) (Table 3).

Correlation Between Three Measurements and WM capacity: Linear regression analyses showed no significant correlation between response time and working memory span of participants. But there were significant negative correlations between subjective mental effort rating and difficulty rating such that participants of higher WM span reported lower mental effort as well as lower difficulty (Table 5). This result suggests that WM capacity can potentially be a confounding variable in explaining our results. To see the effect of cognitive load manipulation alone on any eye-tracking based parameters, variation in these parameters that is due to working memory capacity should be factored out. So, in the hierarchical multiple regression analysis used in this study, WM capacity is treated as a controlled variable. Adding this to the model is used to confirm that the variability in the measured eye-movement parameters is not due to variability of WMC of the participants.

<table>
<thead>
<tr>
<th>DVs</th>
<th>Correlation with O-span score</th>
<th>β</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difficulty rating</td>
<td>-.157</td>
<td>.025*</td>
<td></td>
</tr>
<tr>
<td>Mental effort rating</td>
<td>-.134</td>
<td>.010*</td>
<td></td>
</tr>
<tr>
<td>Response time</td>
<td>-.022</td>
<td>.671</td>
<td></td>
</tr>
</tbody>
</table>

4.3.3 Eyetracking Based Metrics

To address the research questions, several eye movement based parameters were examined against different types of load.

Percentage Dwell Time: In the multimedia learning material, depending on the condition (redundant, non-redundant), different areas of interest (AoIs) were defined and used as units of
analysis: the redundant condition contained two AoIs: animation and text; the non-redundant condition contained only the animation AoI (Figure 6). The percentage of dwell time spent in an AoI was calculated by the total fixation time spent over a certain AoI divided by the total fixation time for the trial. This is a measure of the visual attention to that AoI.

Figure 6: AoIs in Non-Redundant Condition (Left), and Redundant Condition (Right)

First, we investigated whether adding text to the animation and narration shifted participants’ attention. An independent samples t-test ($t(366) = 29.021, p < .001$) found that the percentage dwell time spent in the animation AoI in the non-redundant (low extraneous load) condition (86%) was significantly higher than in the redundant (high extraneous load) condition (38%). This means adding text shifted participants’ attention from the animation to the text.

Furthermore, in the redundant condition, we found that the high-germane load group had a significantly higher ($t(182) = -5.752, p < .001$) percentage dwell time (47%) in the animation AoI than the low-germane load group (31%). This result is consistent with the redundancy effect (Mayer, 2005). Such that the percentage dwell time spent looking at the animation AoI is sensitive to germane load in the redundant condition. Dwell time being most sensitive to germane load is consistent with the notion that creating schemata takes time and is a cumulative result. It is consistent with the interpretation that schema acquisition and automation do not happen in a short amount of time. Rather knowledge building is a slow-paced process happening
over a relatively long time period. Therefore, the amount of time that a learner spends dwelling on, or attending to, relevant information is sensitive to germane load.

*First Saccade Latency (FSL):* Text and animation compete for visual attentional resources (Wickens, 2002). Participants must choose which one to attend to first: text or animation. If they choose to attend to the animation first, will attending animation AoI be faster or slower than when there is no text at all? With the onset of information, the first saccade latency to that information is the reaction time for participants’ eye movements. It is calculated as the difference between the time when they make their first saccade to that area of interest and the onset time of that information.

<table>
<thead>
<tr>
<th></th>
<th>Redundant (first saccade to animation) &amp; Redundant (first saccade to text)</th>
<th>Redundant (first saccade to animation) &amp; Non-redundant (first saccade to animation)</th>
<th>Redundant (first saccade to text) &amp; Non-redundant (first saccade to animation)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Means(SD)</td>
<td>289(100)</td>
<td>289(100)</td>
<td>247(126)</td>
</tr>
<tr>
<td>T test</td>
<td>t (182) = 1.452</td>
<td>t (124) = 2.296</td>
<td>t (268) = 6.553</td>
</tr>
<tr>
<td>P test</td>
<td>p = .148</td>
<td>p = .041</td>
<td>p &lt; .001</td>
</tr>
</tbody>
</table>

A first saccade latency analysis indicated that, in the redundant condition, participants first attended to the text in 164/184 cases. In only 20/184 cases did the participants attend to the animation first. Furthermore, an independent samples t-test did not find a significant difference between first saccade latency to the animation AoI in the redundant condition compared to the first saccade latency to the text AoI in same condition. We found a significant difference between first saccade latency to the text AoI in the redundant condition and first saccade latency to the animation AoI in the non-redundant condition, such that participants attended to text significantly faster in the redundant condition than in the non-redundant condition. This result
(Table 6) seems to support the notion that processing of linguistic information is more automatic than processing of diagrammatic information. In a sense, it is similar as the Stroop effect (Stroop, 1935).

**Table 7: Hierarchical Regression Analysis Assessing Prediction of IVs By MFD**

<table>
<thead>
<tr>
<th>Variables</th>
<th>Model 1</th>
<th></th>
<th>Model 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>β</td>
<td>p</td>
<td>β</td>
<td>p</td>
</tr>
<tr>
<td>Intrinsic load</td>
<td>.049</td>
<td>.268</td>
<td>.049</td>
<td>.262</td>
</tr>
<tr>
<td>Extraneous load</td>
<td>-.547***</td>
<td>.000</td>
<td>-.549***</td>
<td>.000</td>
</tr>
<tr>
<td>Germane load</td>
<td>.054</td>
<td>.219</td>
<td>.099*</td>
<td>.031</td>
</tr>
<tr>
<td>O-SPAN</td>
<td></td>
<td></td>
<td>-.143**</td>
<td>.002</td>
</tr>
</tbody>
</table>

Note. N = 46. R² = .30*** for model 1; ΔR² = .02** for model 2; * p < .05. ** p < .01. *** p < .001

_Mean Fixation Duration (MFD):_ Fixation duration records how long participants look at a particular location. A multiple regression analysis was conducted to examine the correlation between independent variables (IVs) and MFD. In the first model, IVs included intrinsic load, extraneous load and germane load. For each load, we had two levels: low and high. We included the O-span score as an IV in the second model to control the effect of working memory capacity.

**Table 8: Hierarchical Regression Analysis for Prediction of IVs by MFD for Animation AoI**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Model 1</th>
<th></th>
<th>Model 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>β</td>
<td>p</td>
<td>β</td>
<td>p</td>
</tr>
<tr>
<td>Intrinsic load</td>
<td>.025</td>
<td>.605</td>
<td>.025</td>
<td>.604</td>
</tr>
<tr>
<td>Extraneous load</td>
<td>-.350***</td>
<td>.000</td>
<td>-.352***</td>
<td>.000</td>
</tr>
<tr>
<td>Germane load</td>
<td>.058</td>
<td>.242</td>
<td>.087</td>
<td>.093</td>
</tr>
<tr>
<td>O-SPAN</td>
<td></td>
<td></td>
<td>-.093</td>
<td>.072</td>
</tr>
</tbody>
</table>

Note. N = 46. R² = .123*** for model 1; ΔR² = .008 is ns for model 2; * p < .05. ** p < .01. *** p < .001

We found that extraneous load can uniquely predict MFD in the first model, and that extraneous load still can uniquely predict MFD after the effect of working memory is included in the second model (Table 7).

We did a similar analysis for MFD in the animation AoI alone. The result was similar to the results over the entire viewing area (Table 8). It seems that shorter mean fixation durations
are related to higher extraneous load. This is consistent with the redundancy effect that high extraneous load is caused by the participants’ need to simultaneously process the animation and text in the redundant condition. According to studies of eye movements in reading, processing text is usually a fast process unless unfamiliar words are being read (Rayner, 1977).

Table 9: Hierarchical Regression Analysis Assessing Prediction of IVs by MSL

<table>
<thead>
<tr>
<th>Variable</th>
<th>Model 1</th>
<th></th>
<th>Model 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>β</td>
<td>p</td>
<td>β</td>
<td>p</td>
</tr>
<tr>
<td>Intrinsic load</td>
<td>-.004</td>
<td>.924</td>
<td>-.004</td>
<td>.923</td>
</tr>
<tr>
<td>Extraneous load</td>
<td>.578***</td>
<td>.000</td>
<td>.576***</td>
<td>.000</td>
</tr>
<tr>
<td>Germane load</td>
<td>.003</td>
<td>.935</td>
<td>.051</td>
<td>.252</td>
</tr>
<tr>
<td>O-SPAN</td>
<td></td>
<td></td>
<td>-.151**</td>
<td>.001</td>
</tr>
</tbody>
</table>

Note: N = 46. $R^2 = .34$*** for model 1; $\Delta R^2 = .02$** for model 2; ** p < .01. *** p < .001.

Mean Saccade Length (MSL): Mean saccade length records how far on average participants’ shift their attention from one location to another. The same multiple regression analysis as before was completed for MSL.

Table 10: Hierarchical Regression Analysis Assessing Prediction of IVs by NBC

<table>
<thead>
<tr>
<th>Variable</th>
<th>Model 1</th>
<th></th>
<th>Model 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>β</td>
<td>p</td>
<td>β</td>
<td>p</td>
</tr>
<tr>
<td>Intrinsic load</td>
<td>.000</td>
<td>.998</td>
<td>.000</td>
<td>.998</td>
</tr>
<tr>
<td>Extraneous load</td>
<td>.158**</td>
<td>.003</td>
<td>.159**</td>
<td>.002</td>
</tr>
<tr>
<td>Germane load</td>
<td>-.07</td>
<td>.175</td>
<td>-.102</td>
<td>.061</td>
</tr>
<tr>
<td>O-SPAN</td>
<td></td>
<td></td>
<td>.101</td>
<td>.065</td>
</tr>
</tbody>
</table>

Note. N = 46. $R^2 = .3$* for model 1; $\Delta R^2$ is ns for model 2, * p < .05. *** p < .001

We found that extraneous load can uniquely predict MSL in the first model. It is worth noting that extraneous load still can uniquely predict MSL after the effect of WMC is accounted for in the second model. Again, higher MSL is related to high extraneous load, according to the redundancy principle. If participants attend predominantly to text, they should have high extraneous load as well as longer saccade (Table 9).
**Normalized Blink Count (NBC):** It is the number of blinks made per unit time by each participant. We conducted a hierarchical multiple regression analysis for NBC, using the same two models as before. We found that extraneous load can uniquely predict NBC even after the effect of working memory capacity is accounted for (Table 10).

**Mean Saccade Duration (MSD):** Saccade duration is the time spent making a saccade. The same analysis as before was completed to investigate the predictability of IVs on MSD. We found that extraneous load is most sensitive to MSD, in that longer MSD is related to higher extraneous load. This is not surprising, since attending to text requires longer saccades, thus longer saccadic duration (Bahill et al., 1975) (Table 11).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Model 1</th>
<th></th>
<th></th>
<th>Model 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>β</td>
<td>p</td>
<td>β</td>
<td>p</td>
<td></td>
</tr>
<tr>
<td>Intrinsic load</td>
<td>-.007</td>
<td>.896</td>
<td>-.007</td>
<td>.896</td>
<td></td>
</tr>
<tr>
<td>Extraneous load</td>
<td>.165**</td>
<td>.001</td>
<td>.164**</td>
<td>.002</td>
<td></td>
</tr>
<tr>
<td>Germaine load</td>
<td>.101</td>
<td>.052</td>
<td>.128*</td>
<td>.019</td>
<td></td>
</tr>
<tr>
<td>O-SPAN</td>
<td></td>
<td></td>
<td>-.086</td>
<td>.112</td>
<td></td>
</tr>
</tbody>
</table>

Note. N = 46. $R^2 = .04**$ for model 1; $\Delta R^2$ is ns for model 2; * p < .05. ** P < .01

<table>
<thead>
<tr>
<th>Variable</th>
<th>Model 1</th>
<th></th>
<th></th>
<th>Model 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>β</td>
<td>p</td>
<td>β</td>
<td>p</td>
<td></td>
</tr>
<tr>
<td>Intrinsic load</td>
<td>-.035</td>
<td>.496</td>
<td>-.035</td>
<td>.49</td>
<td></td>
</tr>
<tr>
<td>Extraneous load</td>
<td>.096</td>
<td>.06</td>
<td>.093</td>
<td>.065</td>
<td></td>
</tr>
<tr>
<td>Germaine load</td>
<td>.214***</td>
<td>.000</td>
<td>.269***</td>
<td>.000</td>
<td></td>
</tr>
<tr>
<td>O-SPAN</td>
<td></td>
<td></td>
<td>-.176**</td>
<td>.001</td>
<td></td>
</tr>
</tbody>
</table>

Note. N = 46. $R^2 = .06***$ for model 1; $\Delta R^2 = .03**$ for model 2; * p < .05. *** p < .001

**Mean Saccade Peak Velocity (MSPV):** When participants shift their attention from one fixation to the next one, they make a saccade with non-uniform velocity. Saccade peak velocity has been shown to be positively correlated with saccade length in a visually guided saccade task.
(Bahill et al., 1975). We conducted a hierarchical regression analysis for mean saccade peak velocity, using the same two models as before. We found that germane load can uniquely predict MSPV even after the effect of working memory is controlled for. Although MSL is shown to be most sensitive to extraneous load, here MSPV is shown to be most sensitive to germane load. We do not have an explanation for this result (Table 12).

**Ratio of Pupil Size Change (RPSC):** This is defined as \( \frac{PS - PS_{baseline}}{PS_{baseline}} \) where \( PS_{baseline} \) is the pupil size when no information is presented to participants. Change of pupil size during learning can be said to be affected by cognitive load. We conducted a hierarchical regression analysis for RPSC, using the same two models as before. We found that germane load and extraneous load can both uniquely predict RPSC even after the effect of working memory capacity is controlled for. Larger RPSC is related to higher extraneous load and lower germane load. This seems to suggest that increase in extraneous load is associated with a decrease in germane load or vice versa (Table 13).

| Table 13: Hierarchical Regression Analysis Assessing Prediction of IVs by RPSC |
|-----------------------------------------|--------|--------|--------|--------|
| Variable                  | \( \beta \) | \( p \) | \( \beta \) | \( p \) |
| Intrinsic load             | .005   | .929   | .005   | .929   |
| Extraneous load            | .13*   | .013   | .129*  | .013   |
| Germane load               | -.131* | .012   | -.118* | .032   |
| O-SPAN                     |        |        | -.042  | .442   |

Note. \( N = 46 \). \( R^2 = .03 \) ** for model 1; \( \Delta R^2 \) is ns for model 2; * \( p < .05 \). ** \( p < .01 \)

A similar analysis was also completed for RPSC for the animation AoI only. Here RPSC is only sensitive to extraneous load even after the effect of working memory is controlled for. These results (Table 14) seem to indicate that pupil size over the animation AoI is affected by the existence of extraneous information even though the redundant text information is not directly
being looked at in that moment. It suggests a potential interaction between the foveal processing of animation and peripheral processing of text information. This requires further investigation.

Table 14: Hierarchical Regression Analysis for Prediction of IVs by RPSC for Animation AoI

<table>
<thead>
<tr>
<th>Variable</th>
<th>Model 1</th>
<th>Model 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>β</td>
<td>p</td>
</tr>
<tr>
<td>Intrinsic load</td>
<td>.029</td>
<td>.562</td>
</tr>
<tr>
<td>Extraneous load</td>
<td>.270***</td>
<td>.000</td>
</tr>
<tr>
<td>Gerane load</td>
<td>-.09</td>
<td>.077</td>
</tr>
<tr>
<td>O-SPAN</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. N = 46. $R^2 = .078$*** for model 1; $\Delta R^2$ is ns for model 2; *** p < .01

Transition between Text and Animation AoIs (TBTA): This parameter records how many saccades are made between the text and animation AoIs. This analysis is only done for the redundant condition, because the non-redundant condition does not have a text AoI. Because the analysis is done only for the redundant (high extraneous load) condition, it cannot be completed with respect to extraneous load. We conducted a hierarchical regression analysis for TBTA, using the same two models as before, except without extraneous load as an IV (Table 15). We found that only intrinsic load can uniquely predict TBTA even after the effect of WMC is factored out.

Table 15: Hierarchical Regression Analysis Assessing Prediction of IVs by TBTA

<table>
<thead>
<tr>
<th>Variables</th>
<th>Model 1</th>
<th>Model 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>β</td>
<td>p</td>
</tr>
<tr>
<td>Intrinsic load</td>
<td>.351**</td>
<td>.000</td>
</tr>
<tr>
<td>Germane load</td>
<td>-.061</td>
<td>.381</td>
</tr>
<tr>
<td>O-SPAN</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. N = 23. $R^2 = .127$*** for model 1; $\Delta R^2 = .009$ is ns for model 2; *** p < .001

Correlation between eye-tracking parameters: If different parameters target the same underlying construct, they should be significantly correlated with each other. A correlation analysis was conducted with the effect of WMC controlled for (Table 16). We found that most of
the parameters are significantly correlated with each other. This is because most of them are sensitive to extraneous cognitive load.

Table 16: Correlations b/w Dependent Measures for Both Groups

<table>
<thead>
<tr>
<th>Measure</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Mean fixation duration</td>
<td>-</td>
<td>-.44***</td>
<td>-.37***</td>
<td>-.90***</td>
<td>-.21***</td>
<td>-.16**</td>
<td>-.08**</td>
</tr>
<tr>
<td>2. Mean saccade amplitude</td>
<td>-</td>
<td>.15**</td>
<td>.45***</td>
<td>.36***</td>
<td>.49***</td>
<td>.05</td>
<td></td>
</tr>
<tr>
<td>3. Normalized blink count</td>
<td>-</td>
<td>.51***</td>
<td>.08</td>
<td>.14*</td>
<td>.39***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Normalized fixation count</td>
<td>-</td>
<td>.21***</td>
<td>.18**</td>
<td>.24***</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Mean saccade duration</td>
<td>-</td>
<td>.71***</td>
<td>-.004</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Mean saccade peak velocity</td>
<td>-</td>
<td>-</td>
<td>-.03</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Ratio of pupil size change</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: N = 46. * p < .05. ** p < .01. *** p < .001.

4.4 Examining the Assumptions of Sweller’s (2010) CLT Model

Cognitive load theory has undergone several significant improvements since it was first proposed (Sweller, 1988). In the introduction, we briefly reviewed the history of CLT from its genesis to its current form. This endeavor will continue as more researchers join this enterprise and further explore CLT. In the latest model of CLT (Sweller, 2010),

"… germane cognitive load is purely a function of the working memory resources devoted to the interacting elements that determine intrinsic cognitive load. The more working memory resources that must be devoted to extraneous cognitive load, the fewer will be available to deal with intrinsic cognitive load, reducing learning. This formulation assumes that motivation is high and all available working memory resources are being devoted to dealing with intrinsic and extraneous cognitive load"

In this latest formalism, Sweller (2010) assumes high motivation of learners, such that working memory capacity is completely used up to attend to the extraneous load together with the germane load (which attends to the intrinsic load).
However, students’ motivation is not always high, as evidenced by many PER studies (Docktor & Mestre, 2014). So an alternative model is proposed here. Given the results from our experiments, we should not assume that extraneous load and germane load will use up available space of working memory under all circumstances. Especially when intrinsic load and extraneous load are both relatively low, we should not assume that germane load will be higher than the intrinsic load resulting in overlearning. Given this, we suggest that intrinsic load should be used as a criterion for determining the level of germane load that students need to devote to learning. So, it will be more reasonable to adopt the following assumption: mental energy devoted (germane load) to processing intrinsic load is equal to the required load (intrinsic load) as determined by the element interactivity related to learning objective if this required load is smaller than the available space in working memory when part of which has already been used as extraneous load. When required mental energy to process the intrinsic load is too high to be accommodated by the working memory capacity, then germane load will be determined as whatever working memory remains after allocating working memory resources to attend to the extraneous load. Mathematically:

$$\text{Germane load} = \min (\text{Intrinsic load}, WMC - \text{Extraneous load})$$

We also assume that germane load cannot be directly manipulated, rather, it is a function of independent manipulations of intrinsic load and extraneous load as described by the equation above.

The Study One allows us to discuss these new assumptions of Sweller’s (2010) model because of two features of this study. First, we can use not only subjective ratings methods or performance of a secondary task, but also eye movements related metrics to measure the three kinds of cognitive load. Second, we have measured the working memory capacity of our
participants by using the O-span task (Unsworth et al., 2005), which allows us to test the model described by the mathematical equation above.

In general, we used a multilevel model analysis to compare Sweller’s (2010) model and the newly proposed alternative model. The levels of cognitive load cannot be directly measured without using any proxy metrics. Our Study One provides us multiple metrics that are sensitive to different types of cognitive load. In order to fully test each model, we need to have a complete set of proxy metrics. A complete set implies we should have one proxy metric for the intrinsic load, one proxy metric for the extraneous load, one proxy metric for the germane load and a metric for working memory capacity. Furthermore, all proxies selected should be converted to Z-scores to make them comparable since different metrics have different units.

As mentioned above, the germane load is predicted rather than manipulated, so predictions of the germane load can be made using the mathematical expressions of the two models are as follows. We use ‘S model’ to refer to Sweller’s (2010) model and ‘A model’ to refer to the alternative model. So, mathematically:

**S model:** \[ Z_{Germ} = Z_{WM} - Z_{E} \]

**A model:** \[ Z_{Germ} = \min (Z_{Intr}, Z_{WM} - Z_{E}) \]

We examined the comparison based only on the data from the redundant condition since we only have complete sets of proxies for that condition. This was because one of the proxies for intrinsic load (Transitions between text AoI and animation AoI) would only be relevant in the redundant condition, as the non-redundant condition did not have a text AoI.

The proxy for the germane load of the redundant condition is the percentage dwell time over the animation AoI. We selected 6 different sets of proxies for the intrinsic load and extraneous load as can be seen in Table 17. Two predicted germane load values were calculated...
following the mathematical expressions for the S model and the A model and later compared to
the measured proxy the percentage dwell time over the animation AoI using a multilevel model
analysis. Akaike Information Criterion (AIC) values were adopted to compare which model was
a better fit for the data. The criterion is that the smaller the AIC value, the better the fit to the
corresponding model. The results can be found in Table 17 and Sweller’s (2010) model was
assessed as a better fit for the data.

<p>| Table 17: Results for the Multilevel Model Analysis of Comparing S-model and A-model |
|---------------------------------|---------------------------------|----------------|---------|----------------|</p>
<table>
<thead>
<tr>
<th>Intrinsic load proxy</th>
<th>Extraneous load proxy</th>
<th>In favor of</th>
<th>(Y/N)</th>
<th>AIC value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Mental effort rating</td>
<td>Mean saccade length</td>
<td>S model</td>
<td>Y</td>
<td>434.5090</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A model</td>
<td>N</td>
<td>434.6768</td>
</tr>
<tr>
<td>2 Mental effort rating</td>
<td>Response time to a secondary task</td>
<td>S model</td>
<td>N</td>
<td>441.6155</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A model</td>
<td>Y</td>
<td>439.2373</td>
</tr>
<tr>
<td>3 Mental effort rating</td>
<td>Mean fixation duration</td>
<td>S model</td>
<td>Y</td>
<td>423.3338</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A model</td>
<td>N</td>
<td>430.4264</td>
</tr>
<tr>
<td>4 Transitions between AoIs</td>
<td>Mean saccade length</td>
<td>S model</td>
<td>Y</td>
<td>423.3338</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A model</td>
<td>N</td>
<td>439.0050</td>
</tr>
<tr>
<td>5 Transitions between AoIs</td>
<td>Response time to a secondary task</td>
<td>S model</td>
<td>N</td>
<td>441.6155</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A model</td>
<td>Y</td>
<td>440.9331</td>
</tr>
<tr>
<td>6 Transitions between AoIs</td>
<td>Mean fixation duration</td>
<td>S model</td>
<td>Y</td>
<td>434.5090</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A model</td>
<td>N</td>
<td>439.2617</td>
</tr>
</tbody>
</table>

As reviewed in the development history of CLT in Chapter 2. Sweller updated the
definitions of the three kinds of cognitive load based on the idea of element interactivity and
redefined the germane load as all available mental resources left in the working memory as in the
Stage 4 of Figure 3. This new formalism differs from the previous Stage 3 in two ways: 1) the
intrinsic load is not in the working memory any more, 2) there is no available space left in
working memory capacity in the Stage 4 model. Sweller updated this model based on several
results from literature (Sweller, 2010) and presumed a high motivation of all learners.
With the different proxies for the three kinds of cognitive load, and the individual data of working memory capacity as measured by an O-span task (Unsworth et al., 2005), we compared the Stage 3 and the Stage 4 model of CLT by using a multilevel model analysis.

Table 18: Results for the Multilevel Model Analysis of Comparing SS3-model and S4 model

<table>
<thead>
<tr>
<th>Intrinsic load proxy</th>
<th>Extraneous load proxy</th>
<th>In favor of</th>
<th>(Y/N)</th>
<th>AIC value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Mental effort rating</td>
<td>Mean saccade length</td>
<td>SS3 model</td>
<td>N</td>
<td>441.4105</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S4 model</td>
<td>Y</td>
<td>434.5090</td>
</tr>
<tr>
<td>2 Mental effort rating</td>
<td>Response time to a secondary task</td>
<td>SS3 model</td>
<td>Y</td>
<td>441.2516</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S4 model</td>
<td>N</td>
<td>441.6155</td>
</tr>
<tr>
<td>3 Mental effort rating</td>
<td>Mean fixation duration</td>
<td>SS3 model</td>
<td>N</td>
<td>438.8781</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S4 model</td>
<td>Y</td>
<td>423.3338</td>
</tr>
<tr>
<td>4 Transitions between AoIs</td>
<td>Mean saccade length</td>
<td>SS3 model</td>
<td>N</td>
<td>440.7435</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S4 model</td>
<td>Y</td>
<td>434.5090</td>
</tr>
<tr>
<td>5 Transitions between AoIs</td>
<td>Response time to a secondary task</td>
<td>SS3 model</td>
<td>N</td>
<td>442.5058</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S4 model</td>
<td>Y</td>
<td>423.3338</td>
</tr>
<tr>
<td>6 Transitions between AoIs</td>
<td>Mean fixation duration</td>
<td>SS3 model</td>
<td>N</td>
<td>437.6957</td>
</tr>
<tr>
<td></td>
<td></td>
<td>S4 model</td>
<td>Y</td>
<td>423.3338</td>
</tr>
</tbody>
</table>

Again, we only have a complete set of proxies for the redundant condition. We converted all the proxies to Z-scores to make them comparable. The Stage 3 model of CLT allows available space in working memory capacity. Since there is no way to determine how large this free space was in our experiment. We assume all students had high motivation such that all remaining mental resources were used as the germane load. We call this model the Simplified Stage 3 model (SS3). We compared this simplified model with the Stage 4 model of CLT (S4) with both under the same assumption of high motivation. The mathematical expressions of the two models are as follows:

SS3 model: \[ Z_{\text{Germane load}} = Z_{\text{WMC}} - Z_{\text{Extraneous load}} - Z_{\text{Intrinsic load}} \]

S4 model: \[ Z_{\text{Germane load}} = Z_{\text{WMC}} - Z_{\text{Extraneous load}} \]
The proxy for the germane load of the redundant condition is the percentage dwell time over the animation AoI. We selected 6 different sets of proxies for the intrinsic load and extraneous load as can be seen in Table 18. Two predicted germane load values were calculated following the mathematical expressions for the SS3 and S4 models and later compared to the measured proxy the percentage dwell time over the animation AoI using a multilevel model analysis. The results can be found in Table 18 and the S4 model (Sweller’s (2010) model) was assessed as a better fit for the data.

### 4.5 Discussion and Implications

Our results showed findings consistent with DeLeeuw and Mayer (2008) in terms of non-eye-tracking based measurements, namely subjective mental effort rating was most sensitive to intrinsic load; response time to a secondary task was most sensitive to extraneous load; subjective difficulty rating was most sensitive to germane load. (see Table 19).

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subjective mental effort rating</td>
<td>√</td>
</tr>
<tr>
<td>Response time for a secondary task</td>
<td>√</td>
</tr>
<tr>
<td>Subjective difficulty rating</td>
<td>*</td>
</tr>
</tbody>
</table>

√ indicates sensitivity to each type of the three loads, * indicates a negative correlation.

To address the sensitivity of several eye movement based metrics to different types of cognitive load manipulations, we used a hierarchical multiple regression analysis to investigate the sensitivity of different eye movement based metrics to each load type. Due to large individual differences in working memory capacity (WMC), this confounding variable should be controlled for. Our analysis showed the results are robust even after controlling for the effect of WMC.
We found that *mean fixation duration* (longer for the non-redundant than for the redundant condition), *mean saccadic length* (longer for the redundant than for the non-redundant condition), *normalized blink counts* (greater for the redundant than for the non-redundant condition), and *mean saccade duration* (longer for redundant than for non-redundant condition) were most sensitive to *extraneous* load.

<table>
<thead>
<tr>
<th>Eye movement metrics</th>
<th>Sensitivity to…</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Intrinsic load</td>
</tr>
<tr>
<td>Mean fixation duration</td>
<td>√</td>
</tr>
<tr>
<td>Mean fixation duration for animation only</td>
<td>√</td>
</tr>
<tr>
<td>Mean saccade length</td>
<td>√</td>
</tr>
<tr>
<td>Normalized blink count</td>
<td>√</td>
</tr>
<tr>
<td>Mean saccade duration</td>
<td>√</td>
</tr>
<tr>
<td>Mean saccade peak velocity</td>
<td>√</td>
</tr>
<tr>
<td>Ratio of pupil size change</td>
<td>√</td>
</tr>
<tr>
<td>Ratio of pupil size change for animation only</td>
<td>√</td>
</tr>
<tr>
<td>Transitions between the text and animation AoIs</td>
<td>√</td>
</tr>
<tr>
<td>Dwell time over the animation AoI</td>
<td></td>
</tr>
</tbody>
</table>

√ indicates sensitivity to each type of the three loads, ∗ indicates a negative correlation.

*Mean saccade peak velocity* (greater for high-transfer than for low-transfer participants), and *percentage dwell time over animation AoI* (greater for high-transfer than for low-transfer participants) were most sensitive to *germane* load. It appears that high-transfer participants did not need to attend to the text to learn the concepts. Rather, they could attend to the animation and listen to the narration to learn from the multimedia lesson.
Transitions between the text and animation AoIs were most sensitive to intrinsic load, with more transitions between the text and animation AoIs for the high-complexity videos than for the low-complexity videos.

Mean ratio of pupil size change was most sensitive to both extraneous and germane load with a higher ratio of pupil size change for redundant and low-transfer condition than for non-redundant and high-transfer condition. (see Table 20).

We also found that in the redundant (high extraneous load) condition, adding concurrent onscreen text shifted participants’ attention from the animation to the text as evidenced by participants spending a significant amount of time looking at the text. Probably processing text is more automatic than processing animation when it comes to learning. This is further supported by the first saccade latency analysis which showed that participants attended to the text significantly faster than attending to the animation when the text was not presented.

Results in this study confirmed that eye-tracking based metrics can be used to measure different types of cognitive load. These results add evidence to literature that a triarchic construct of cognitive load is supported.

The results of Study One allows us to have more than one set of proxies to measure the three kinds of cognitive load and a measure of working memory capacity of students. Having all of these measures allows us to theoretically compare Sweller’s (2010) model with an alternative model which is based on realistic thinking and Sweller’s (2010) model with a previous CLT model, namely, the Stage 3 model as shown in Figure 3. The method used for the two comparisons was a multilevel model analysis, it revealed that the latest Sweller’s (2010) model was a better fit of the data we collected in the Study One.
As we have discussed in this chapter, different eye movements related metrics can be used to measure the three types of cognitive load. A learning system in the form of human-computer interface will greatly benefit from these findings (Chen et al., 2011; Gerjets et al., 2014; Gevins & Smith, 2003; Smith et al., 2001). To remind ourselves, intrinsic load is related to the learning objective, such that it should be neither too high, nor too low. Extraneous load is the energy used for processing information that hinders learning, such that it should be as low as possible. We want students to have as high a germane load as possible in order to achieve a maximum learning outcome. Therefore, we can incorporate the functionality of eye-tracking into learning systems in the form of a human-computer interface, such that eye movements can be collected and analyzed in real-time as learners are using the system. The system itself could decide whether some help should be provided to the learners with the monitoring of the three kinds of cognitive load. For example, when intrinsic load of students was high as detected by a specific parameter, the system can provide the same information in smaller chunks instead of presenting them at one time. This affords us the opportunity to provide differentiated i.e. personalized instruction that is adaptive in real time for students’ based knowledge of their working memory capacity and the cognitive load demands that they experience.
5. STUDY TWO: THE EFFECTS OF PEER INTERACTION ON CONCEPTUAL TEST PERFORMANCE AND COGNITIVE LOAD

5.1 Introduction

Traditional science classrooms are predominately lecture based. It means that students sit passively in the classroom and teachers do most of the talking. A vast body of physics and science education research has converged on the conclusion that lecture-based instructional approaches are not the most conducive to facilitating learning. On the other hand, it has been shown that classes using interactive-engagement methods are superior to the traditional lecture format as evidenced by improvements conceptual understanding as measured by research-validated instruments (Hake, 1998).

Peer instruction is one of the interactive-engagement methods that has been widely adopted in college physics classrooms (Mazur, 2001, 2013) because of its low bar for adoption compared to some other interactive engagement approaches, it can be adapted into the lecture format. Advantages of peer instruction include engaging students in class, allowing students to actively discuss the topics taught, providing opportunities for students to articulate their thought process, receiving feedback from peers immediately and allowing for instructors to know where students are in terms of their understanding and adjust the instruction accordingly just in time. Peer collaboration has also used in many other different contexts, e.g. facilitating problem solving skills in context-rich problems (Heller & Anderson, 1992) and instruction in a workshop style classroom (Laws, 1997).

Singh (2005) investigated the impact of peer interaction on conceptual test performance. CSEM (Conceptual Survey of Electricity and Magnetism) was used to test students’ knowledge about topics taught in an E&M class (Maloney et al., 2001). The design consists of three
conditions. The first condition took the CSEM test individually and then worked on the same test in pairs shortly thereafter; the second condition worked in pairs first followed by working on the same test individually shortly thereafter; the control condition worked on the same test twice individually with a two-week delay for the second individual test. It was shown that students’ final performance on the conceptual test in the first two conditions outperformed those from the control condition. In explaining the results, Singh (2005) made a conjecture that when working in groups the cognitive load will be shared between partners thus it would not be as high as if they work individually. According to cognitive load theory (CLT), low cognitive load can lead to better learning/testing performance (Sweller, 1998). However, cognitive load is a complex, multifaceted concept. As we know, CLT differentiates three kinds of cognitive load: intrinsic cognitive load, extraneous cognitive load and germane cognitive load. The three loads have different ramifications for learning. As discussed earlier, intrinsic load is directly related the learning goal, extraneous load is related to irrelevant information, and germane load is related to learning the relevant information. To complement the conjecture made by Singh regarding the reduction of total cognitive load, impact of peer interaction on the three kinds of load should also be discussed. Specifically, we hypothesize that germane load will be increased by peer interaction, since students exchanged more ideas with each other, reached a deeper level of understanding of the material and better extent of justification of arguments when teachers were absent (Hogan et al., 2000). Meanwhile, extraneous load might be increased as well, since when students converse with each other, the discussion might not always be focusing on relevant topics. Others may perceive as distracting information, ideas that are offered by some students. We may hypothesize that intrinsic load might be reduced, since different members may have different levels of knowledge, higher level knowledge peers would offer ideas that are
scientifically correct and help lower peers learn, such that intrinsic load for lower learners and therefore on average will be reduced.

On the other hand, the type of instruction affects students of different knowledge level differently. For example, the expertise reversal effect proposes that novices learn better from guided instruction and higher knowledge level learners are better facilitated with less guided instruction such as pure problem solving (Kalyuga et al., 2003). Reisslein et al. (2006) investigated the effect of problem-example pairs and example-problem pairs. In the problem-example pair condition, participants need to solve a problem followed by a worked example which contains solution to a similar problem; in the example-problem pair condition, students need to learn from a worked example followed by solving a similar problem. The results showed that expert participants benefited more from learning from problem-example pairs while novices benefited more from learning from example-problem pairs which indicates experts benefit more from problem solving and novices learn better with direct instruction. It is important to investigate how knowledge level of students moderate the effect of peer interaction on their test performance. Based on CLT, knowledge level will affect the cognitive load of students. For example, Newton’s second law, $\vec{F} = m\vec{a}$, for a beginning learner, has four interacting elements. But it can be only one interacting element for experts since they are familiar with this principle and have chunked it as one piece of information. The ability to recognize relevant information from irrelevant information also depends on level of knowledge. Such that that effect of peer interaction on cognitive load experienced by students might be moderated by level of knowledge as well.

In this work, we will answer the following research questions:

1. How does peer interaction affect performance of students on a conceptual test?
2. How does peer interaction affect cognitive loads experienced by students on a conceptual test?

3. How does level of knowledge of students moderate the above two effects?

Below we introduce the test used in this study and how the various cognitive loads were measured.

We propose two hypotheses for the effect of knowledge level on the cognitive load of students.

1) We hypothesize that as knowledge level increases, the intrinsic load and the extraneous load will decrease. This is because high knowledge level students have the necessary schema to understand the materials such that the intrinsic load is low, and they are also able to differentiate irrelevant information from relevant information (Sweller, 2010).

2) According to the Stage 4 model of Sweller (2010), when the extraneous load decreases, germane load increases. So, as a result of a decrease in the extraneous load, we hypothesize that as students’ knowledge level increases, their germane load will increase.

We propose hypotheses for the effect of peer interaction on the cognitive load of students, based on the results from literature.

1) We hypothesize that the intrinsic load will decrease when students are working in groups. Since students may possess different levels of prior knowledge. Discussing each other will help each other learn, such that their intrinsic load will decrease.

2) We hypothesize that the extraneous load will increase. According to literature, when students worked in a group, the discussion was not always focusing on the task at hand. There were often some off-task behaviors and disengagement from the main tasks (Keil et al., 2015)
and disruptions from one or more group members (Pawlak et al., 2015). Students may have perceived this as distracting information or behavior.

3) We hypothesize that the germane load will increase. According to literature, working in groups facilitated students to exchange more ideas with each other (Beatty, 2005), causing them to reach a deeper understanding of the learning materials and provided better justifications of their arguments (Hogan et al., 2000). They also showed evidence of co-constructing knowledge (Singh, 2005). For all these reasons, we hypothesize that the germane load of students will increase.

5.2 Methodology

5.2.1 Survey for Measuring Cognitive Load

A subjective rating questionnaire was developed to measure the three kinds of cognitive load (intrinsic load, extraneous load and germane load) experienced by students in this study. The survey was created based on existing surveys from literature. A modified version of the NASA-TLX questionnaire (Hart & Staveland, 1988) was used to measure the three kinds of cognitive load by Gerjets, et al. (2004). Three items chosen were: effort (How hard did you have to work in your attempt to understand the content of the learning environment?), stress (How insecure, discouraged, irritated, stressed, and annoyed did you feel during the learning task?), and task demands (How much mental and physical activity was required, e.g., thinking, deciding, calculating, remembering, looking, searching etc.?). Each of the three cognitive load items was rated on a scale ranging from 0 (low cognitive load) to 100 (high cognitive load). The results showed item effort is sensitive to intrinsic load which was the only one being manipulated in their paper. According to Sweller (2010), higher intrinsic load requires higher germane load as well. Due to the limitation of only intrinsic load being manipulated by Gerjets et al. (2004), it is
not completely convincing that effort is only sensitive to intrinsic load and not to germane load also.

Windell et al. (2007) investigated the sensitivity of NASA-TLX and subjective difficulty rating to intrinsic load and extraneous load. Intrinsic load was manipulated by adjusting the complexities of instructional material and extraneous load was increased by providing redundant information. They found that NASA-TLX subscale mental demands as well as difficulty rating can differentiate tasks imposing different levels of intrinsic load, and NASA-TLX weighted workload score was shown to be sensitive to both intrinsic load and extraneous load.

Research has shown that the difficulty that learners perceived from the learning materials was higher when information in a split-source format than an integrated format (Kalyuga et al., 1998). These results have been interpreted to suggest that this perceived difficulty could be attributed to the extraneous load, rather than the intrinsic load.

Ayres (2006) investigated the possibility of using a subjective mental effort rating to measure intrinsic load in an arithmetic operation context. Number of interacting elements was manipulated in arithmetic operation tasks. It was also argued that since no instruction was involved and only knowledge was assessed rather than being taught, extraneous load and germane load were kept minimal and constant. It was found that students rated arithmetic operations involving more steps as more effort demanding and made more mistakes.

Researchers have also argued that the mental effort reported by learners is a measure of how hard they concentrated during a task, and therefore is an indicator of their germane load (Salomon, 1984). This is because the greater the mental effort that learners devote to a task, the greater the level of engagement and consequently the greater the learning outcome.

Cierniak et al., (2009) used a secondary task as a measure of overall cognitive load, and
three subjective questions as measures of the three cognitive loads. There were subtle
differences between the wordings of the three questions to gauge the three different loads. The
intrinsic load is based on the amount that the learner needs to learn, so the corresponding
question was “How difficult was the learning content for you?” The extraneous load, as argued
by Kalyuga et al. (1998) is based on how the content is presented i.e. the learning material. So,
the corresponding question was “How difficult was it for you to learn with the material?” Note
the subtle difference between this question and the question for intrinsic load. While the intrinsic
load question focuses on the difficulty of the content, the extraneous load question focuses on the
difficulty presented to the learner by the characteristics of the material. Finally, the germane load
refers to the mental effort or level of concentration devoted by the learner to the task at hand,
therefore the corresponding question, following Salomon (1984) asked “How much did you
concentrate during learning?”

The results showed that reaction time to a secondary task is not different between the two
conditions that differ in extraneous load, such that the total load is the same for the two groups.
Some caveats were mentioned regarding this result. Since it has been previously shown that
secondary task performance is related to extraneous load, an increase in extraneous load is
associated with a decrease in germane load. Among the three cognitive load measurements, the
intrinsic load measure was the least successful, while extraneous load and germane load
measurements were successful. Item questions targeting intrinsic load and extraneous load might
have been framed as too similar to each other, such that students treated them as not different
from each other.

Given the lack of using multiple items measuring each one of the three types of cognitive
load, Leppink et al. (2013) developed a ten-item subjectively rated survey measuring the three
cognitive loads separately. This survey is used to assess the three types of cognitive load in any activities (lecture, class, discussion session, skills training or study session) that students have just completed. Students rate their agreement with each item on a scale from 0 (not at all the case) to (10) (completely the case). The ten items are: 1) The topic/topics covered in the activity was/were very complex; 2) The activity covered formulas that I perceived as very complex; 3) The activity covered concepts and definitions that I perceived as very complex; 4) The instructions and/or explanations during the activity were very unclear; 5) The instructions and/or explanations were, in terms of learning, very ineffective; 6) The instructions and/or explanations were full of unclear language; 7) The activity really enhanced my understanding of the topic(s) covered; 8) The activity really enhanced my knowledge and understanding of statistics; 9) The activity really enhanced my understanding of the formulas covered; and 10) the activity really enhanced my understanding of concepts and definitions. The results of the study (Leppink et al., 2013) showed that 1, 2, 3 measure intrinsic load, 4, 5, 6 measure extraneous load, and 7, 8, 9, 10 measure germane load.

In our study, we adapted the survey by Leppink et al. (2013) to develop an 8-item survey measuring students’ cognitive load during taking a conceptual test. Complexity level of concepts, definitions, equations, formulas are closely tied to the concept of intrinsic load which relies on number of interacting elements (Leppink et al., 2013; Sweller, 2010). Confusing and distracting levels are related to extraneous load since they are the irrelevant information. As mentioned earlier, level of concentration is related to the concept of germane cognitive load as per (Salomon, 1984). Due to the distinction made between extraneous load and germane load such that extraneous load is the mental resources devoted to processing unnecessary information, and germane load is mental resources devoted to processing relevant information (Sweller, 2010;
Leppink et al., 2013). We developed a survey with items 1 through 3 measuring intrinsic load, items 4 through 6 measuring extraneous load, and items 7 and 8 measuring germane load. Each item asked students to rate their agreement with the statement on a scale from 1 (not at all the case) to 9 (completely the case). The wording of the items was a bit different from those used by Leppink et al. (2013), because our survey was designed to measure the three type of cognitive load on a test, while Leppink et al. (2013) survey was designed to measure cognitive load during learning.

The items on our survey were:

1. The topics covered on the physics test were very complex.
2. The physics test covered formulas that I perceived to be very complex.
3. The test covered concepts and definitions that I perceived as very complex.
4. The questions on the physics test had confusing language that was not clear to me.
5. It was very hard to identify what information is relevant to answering the questions on the physics test.
6. There was a lot of distracting information in the question statements on the physics test.
7. I concentrated a lot as I answered the questions on the physics test.
8. I devoted a lot of mental effort in finding and applying the relevant concepts needed to answer the questions on the physics test.

The purpose of this survey is for students to subjectively report their levels of the three types of cognitive load after taking a conceptual test. We used the average rating of all items targeting the same cognitive load to ascertain the level of that type of cognitive load. For example, the average values of items 1, 2, 3 were used to ascertain the level of intrinsic load of
students; the average value of 4, 5, 6 was used to ascertain the level of extraneous load; and average value of 7, 8 was used as level of germane load of students.

A snapshot of the survey can be found in the Appendix (Figure 27), a unique feature about this survey is that this is an integrated format. The first line allows students to rate the statement after working on a test individually and the second line allows students to rate the same statements after working in groups. The advantage of this integrated format is that students can use their responses to each item in the first administration of the survey (i.e. after the individual test) as a reference point for their rating of that item in the second administration of the survey (i.e. after the group test). This increased the test-retest reliability of the survey. Anecdotal conversations with several students right after the study confirmed they indeed used their ratings on the first administration (i.e. after the individual test) of the survey as a reference points for their rating on the second administration (i.e. after the group test) of the survey.

### 5.2.2 Participants and Materials

71 preservice elementary teachers who were concurrently taking a physics class for elementary education majors at a large U.S. Midwestern land grant university participated in this study as part of their lab session. Less than half of the students had previous experience of physics as evidenced by a survey collected in the end of the interview. The conceptual survey used in this study is DIRECT (Determining and Interpreting Resistive Electric Circuit Concepts Test) a well-known conceptual test which was developed by Engelhardt and Beichner (2004). DIRECT is a diagnostic test which helps instructors identify students’ misconceptions of electric circuits.
5.2.3 Procedures

This experiment had two phases (Figure 7), the first phase of the study was completed in the 1st week of the semester during the pre-test at the beginning of a unit on electric circuits and the second phase in the 7th week of the same semester, during the post-test at the end of a unit on electric circuits. Each phase essentially consisted of the same activities. Students were typically in group of 3-4 sitting around the same table. In the beginning of each phase, students were given about 30-40 minutes to complete the DIRECT test, recording their responses to each question on a Scantron. When each student finished the DIRECT test, they handed back the Scantron, and were given the cognitive load survey. They completed the first part of the cognitive load survey in about 5-10 minutes in which they responded to the eight items described earlier based on their experience of taking DIRECT individually. After completing the cognitive load survey, they put it face down on the table.

![Figure 7: Procedure of the Experiment](image)

After all students in the group had finished both the DIRECT test individually and the cognitive load survey, they were asked to work as a group on the DIRECT test again. They were each given a new Scantron on which to record their answers to each question on the DIRECT test after discussing them with their group mates. When any group finished the test, they would complete the second part of the cognitive load survey individually. The second part had the same eight items as before, but this time they were asked to rate each item based on their experience taking the DIRECT test as a group.
5.3 Results

5.3.1 Performance on the DIRECT Test

*Question 1: Does students’ knowledge increase from pretest to posttest?*

Since this is a within-subject design, all measures were repeated measures. A paired sample t test was conducted to compare the mean score of students from individual level of pretest and mean score of students from individual level of posttest. We found a significant improvement from pre-test to post-test, $t(65) = -6.684, p < .001$. The mean individual score on the pre-test was 6.85 (S.D. = 2.37), it increased to 9.65 (S.D. = 2.88) on the individually taken post-test (Figure 8).

*Question 2: Does students’ performance improve from individual to group on the pre-test?*

A paired sample t test was conducted to compare the mean individual score of students on the pre-test and mean group score of students. We did not find a significant main effect of peer interaction on students’ performance on DIRECT test, $t(65) = -1.562, p = .123$. The mean individual score from pre-test was 6.85 (S.D. = 2.37) and the mean group score on the pre-test was 7.41 (S.D. = 2.18). This indicates that students’ knowledge did not significantly increase their performance from working individually to working in a group on the pre-test (Figure 8).

*Question 3: Does students’ performance improve from individual to group on the post-test?*

A paired sample t test was conducted to compare the mean individual score of students on the post-test and with the mean group score of students on the post-test. We found a significant main effect of peer interaction on students’ performance on DIRECT test, $t(65) = -4.724, p < .001$. The mean individual score on the post-test was 9.65 (S.D. = 2.88) and the mean
group score on the post-test was 11.36 (S.D. = 2.44). These results indicate that students’ knowledge significantly increased from working individually to working in group on the post-test (Figure 8).

*Question 4: How does knowledge level moderate the effect of peer instruction on students’ performance?*

On the pre-test, we found that 25 students had a lower score after working as a group than they had after working individually, 34 students had a higher score after working as a group than they had after working individually, and 12 students had a same score after working as a group as they had after working individually.

![Figure 8: Performance of Students on the Pretests and Posttests](image)

On the post-test, we found that 13 students had a lower score after working as a group than they had after working individually, 48 students had a higher score after working in group than they had after working individually, and seven students had a same score after working as a group than they had after working individually.
Fisher’s exact test was conducted to investigate this gain on the DIRECT test. We found a main effect of knowledge level on gain direction, \( p = .0182 \). Knowledge level is moderating the effect of peer interaction on students’ performance on a conceptual test such that when students had low prior knowledge (i.e. on the pre-test), peer interaction almost had no effect on their test performance, but when students had a higher prior knowledge (i.e. on the post-test), peer interaction clearly had a positive effect on test performance.

5.3.2 Cognitive Load

Question 5: Is this cognitive load survey a valid measure of intrinsic load, extraneous load and germane load?

This survey clearly has face validity since all eight items on the survey were clearly in line with the definitions of the three kinds of load based on Sweller (2010), and they were all adapted based on previous literature which confirmed the validity of each of them. However, we still need to provide evidence to confirm that different items targeting the same cognitive load construct converge together. A confirmatory factor analysis was conducted on the individual survey data collected on the pre-test and confirmed that a three-component model is supported. To be more specific, items 1, 2, 3, are more in line with component I, items 4, 5, 6 are more in line with component II, and items 7, 8 are more in line with component III. After examination, we believe component I, II, and III are intrinsic load, extraneous load and germane load respectively.

Question 6: Does students’ intrinsic load, extraneous load and germane load change from individual pretest to individual posttest?

A paired sample t test was conducted to compare the mean intrinsic load of students on the individual pre-test with their mean intrinsic load on the individual post-test. We found a
significant drop in intrinsic load from individual pre-test to individual post-test, \( t(65) = 3.229, p = .002 \). The mean intrinsic load on the individual pre-test was 5.30 (S.D. = 1.81) and the mean intrinsic load on the individual post-test was 4.63 (S.D. = 1.50). This drop in intrinsic cognitive load is consistent with increased knowledge on the post-test compared to the pre-test. Students were more knowledgeable about the concepts on the post-test compared to the pre-test (as evidenced by the statistically significant improvement on individual DIRECT score), and therefore they experienced a lower intrinsic cognitive load on the post-test compared to the pre-test.

![Cognitive Load of Students](image)

**Figure 9: Cognitive Load of Students from Individual Pretest and Posttest**

A paired sample t test was conducted to compare the mean extraneous load of students on the individual pre-test with the mean extraneous load of students on the individual post-test. We found a significant drop in extraneous load from individual pre-test to individual post-test, \( t(65) = 2.467, p = .016 \). The mean extraneous load on the individual pre-test is 4.31 (S.D. = 1.69) and the mean extraneous load on the individual post-test is 3.72 (S.D. = 1.58).
A paired sample t test was conducted to compare the mean germane load of students on the individual pre-test with the mean germane load of students on the individual post-test. We found a significant drop in germane load as a result of knowledge increase, \( t(65) = 10.025, p < .001 \). The mean germane load on the individual pre-test is 5.61 (S.D. = 1.91) and the mean germane load on the individual post-test is 4.67 (S.D. = 1.35) (Figure 9).

**Question 7:** Does students’ intrinsic load, extraneous load and germane load change from individual to group on the pre-test?

![Cognitive Load of Students](image)

**Figure 10: Cognitive Load of Students from Pretest Individual and Group**

A paired sample t test was conducted to compare the mean intrinsic load of students from individual pre-test with the mean intrinsic load of students on the group pre-test. We did not find a significant main effect of peer interaction on students’ intrinsic load imposed by DIRECT test, \( t(65) = .284, p = .777 \). The mean intrinsic load on the individual pre-test in week 1 was 5.30 (S.D. = 1.81) and the mean intrinsic load on the group pre-test was 5.21 (S.D. = 1.86). Thus, students’ intrinsic load did not significantly change from working individually to working in group on the pre-test.
A paired sample t test was conducted to compare the mean extraneous load of students from individual pre-test with the mean extraneous load of students on the group pre-test. We did not find a significant main effect of peer interaction on students’ extraneous load imposed by DIRECT test, \( t(65) = .088, p = .930 \). The mean extraneous load from individual pre-test is 4.31 (S.D. = 1.69) and the mean extraneous load from group pre-test is 4.28 (S.D. = 1.81). This result indicates that students’ extraneous load did not significantly change from working individually to working in group on the pre-test.

A paired sample t test was conducted to compare the mean germane load of students from the individual pre-test with the mean germane load of students from group pre-test. We did not find a significant main effect of peer interaction on students’ germane load imposed by the DIRECT test, \( t(65) = -.516, p = .607 \). The mean germane load on the individual pre-test is 5.61 (S.D. = 1.91) and the mean germane load on the group pre-test is 5.77 (S.D. = 1.78). Thus, students’ germane load did not significantly change from working individually to working in group on the pre-test (Figure 10).

**Question 8: Does students’ intrinsic load, extraneous load and germane load change from individual to group on the post-test?**

A paired sample t test was conducted to compare the mean intrinsic load of students from the individual post-test with the mean intrinsic load of students on the group post-test. We found a significant main effect of peer interaction on students’ intrinsic load imposed by the DIRECT test, \( t(65) = 8.122, p < .001 \). The mean intrinsic load from the individual post-test is 4.63 (S.D. = 1.50) and the mean intrinsic load from the group post-test is 3.45 (S.D. = 1.41). Thus, students’ intrinsic load significantly dropped from working individually to working in group on the post-test.
A paired sample t test was conducted to compare the mean extraneous load on the individual post-test with the mean extraneous load on the group post-test. We found a significant main effect of peer interaction on students’ extraneous load imposed by DIRECT test, $t(65) = -7.432, p < .001$. The mean extraneous load from individual post-test is 3.72 (S.D. = 1.58) and the mean extraneous load from group post-test is 4.52 (S.D. = 1.57).

![Cognitive Load of Students](image)

A paired sample t test was conducted to compare the mean germane load of students from individual post-test with the mean germane load on the group post-test. We did not find a significant main effect of peer interaction on students’ germane load imposed by DIRECT test, $t(65) = -1.265, p = .210$. The mean individual germane load on week 7 was 4.67 (S.D. = 1.35) and the mean group germane load on the post-test was 4.86 (S.D. = 1.58). Thus, students’ germane load did not significantly change from working individually to working in group on the post-test (Figure 11).

**Question 9: Does students’ total load change?**

According to Sweller (2010), total load is sum of intrinsic load and extraneous load, since they are both imposed by learning material. Here, we operationalized the total load as the summation of average intrinsic load and average of extraneous load.
A paired sample t test was conducted to compare the total cognitive load of students from individual pre-test with the total cognitive load of students from individual post-test. We found a significant drop in total cognitive load as a result of knowledge increase from pre-test to post-test, \( t(65) = 3.113, p = .003 \). The total cognitive load from the individual pre-test is 9.61 (S.D. = 3.22) and the total cognitive load from individual post-test is 8.35 (S.D. = 2.95).

We did not find a significant difference between total cognitive load of students from individual pre-test and total cognitive load of students on the group pre-test as evidenced by a paired sample t test, \( t(65) = .202, p = .84 \). The total cognitive load from individual pre-test was 9.61 (S.D. = 3.22) and the total cognitive load from group pre-test was 9.50 (S.D. = 3.39).

![Total cognitive load perceived by Students](image)

**Figure 12 Total cognitive Load Perceived by Students at All Phases**

**Question 10:** Does students’ total energy devoted to the DIRECT test change?

We found a significant difference between total cognitive load of students from individual post-test and total cognitive load of students from group level post-tests as evidenced by a paired sample t test, \( t(65) = 2.06, p = .043 \). The total cognitive load from individual post-test...
is 8.35 (S.D. = 2.95) and the total cognitive load from group post-test is 7.96 (S.D. = 2.82). Thus, the total load decreased significantly from individual to group on the post-test (Figure 12).

According to Sweller (2010), total devoted energy is sum of extraneous load and germane load, since they are the mental resources devoted to processing interacting elements. We operationalize total devoted energy as the summation of average extraneous load and average of germane load.

A paired sample t test was conducted to compare the total devoted energy of students from individual pre-test with the total devoted energy of students from individual post-test. We found a significant drop in total devoted energy as a result of knowledge increase, $t(65) = 4.24, p < .001$. The total devoted energy from individual pre-test is 9.93 (S.D. = 2.76) and the total devoted energy from individual post-test is 8.39 (S.D. = 2.64).

We did not find a significant difference between total devoted energy of students from individual pre-test and total devoted energy of students from group pre-test as evidenced by a
paired sample t test, \( t(65) = -.295, p = .769 \). The total devoted energy from individual pre-test is 9.93 (S.D. = 2.76) and the total devoted energy from group pre-test is 10.06 (S.D. = 2.71).

We found a significant difference between total devoted energy of students from individual post-test and total devoted energy of students from group post-test as evidenced by a paired sample t test, \( t(65) = -4.822, p < .001 \). The total devoted energy from individual post-test was 8.39 (S.D. = 2.64) and the total devoted energy from group post-test was 9.38 (S.D. = 2.82). Thus, the devoted energy increased significantly from individual to group on the post-test.

5.4 Discussion and Implications

Our results indicate that knowledge level of students increased significantly from pre-test to post-test as evidenced by their performance in the conceptual DIRECT test when they took it individually. However, students’ performance did not improve a lot from working individually to working in groups on the pre-test. On the other hand, their performance did improve significantly from working individually to working in a group on the post-tests. This result seems to support the idea that peer interaction will influence students’ performance on a conceptual test only when they have certain level of prior knowledge. This provides an instructional insight. For example, before instruction, peer interaction might not be as effective. After instruction, allowing students to interact with each other can significantly improve their performance.

We tried to understand why this could happen through the lens of cognitive load theory. We used a subjective survey to measure self-reported cognitive loads of students. We found that when students had low prior knowledge, as on the pre-test, their intrinsic load, extraneous load, and germane load levels were not affected by peer interaction at all. No wonder there is no improvement in test performance. However, when students had a certain level of knowledge, such as on the post-test, intrinsic load, extraneous load and germane load of students decreased
from the individual pre-test to the post-test. This is consistent with predictions of cognitive load theory, for higher knowledge level students, intrinsic load, extraneous load and germane load will all be lower for the same test. What is exciting is that on the post-test, when working in a group, intrinsic load was further decreased from individual level. As result of that, performance of students increased significantly to a higher level. This is true although extraneous load increased significantly from individual to group on the post-test and germane load remained steady from individual to group on the post-test. These results seem to suggest that peer interaction allows students to contribute their own knowledge to discussion which decreased the intrinsic load of the whole group. Meanwhile, the contribution of others’ ideas into the discussion increased their extraneous load. It is surprising that germane load was not affected by peer interaction. Independently, we also found that total load did not change much from individual to group on the pre-tests, but total load decreased significantly on the post-test compared to the pre-test. This decrease was primarily due to reduction in intrinsic load. The total devoted energy did not change much from individual to group on the pre-test. But it increased significantly from individual to group on the post-tests, which was primarily caused by increment in extraneous load.

The results of our Study Two seem to suggest that peer interaction is most effective when applied after students have acquired a certain level of knowledge. However, we have to admit that the result is only based on one group of students and one particular conceptual physics test. Further investigation is required before we can generalize these results to other groups of students or other assessments.

The cognitive load survey we developed, although only partially validated, can be very useful in classroom instruction or any learning system in the form of human computer interface.
It is possible for the instructor to administer the cognitive load survey in classroom to assess the cognitive load imposed by classroom assessments, providing the instructor information to make adjustments in the assessments used or the material covered in the class.
6. STUDY THREE: EFFECT OF VIDEO SOLUTIONS AND VISUAL CUES ON CONCEPTUAL TASK PERFORMANCE AND COGNITIVE LOAD

6.1 Introduction

Sweller and colleagues have proposed the cognitive load theory (CLT) which suggests that all instructional methods should take the limited working memory capacity of humans into account (Sweller, 1988, 1994). The design of instruction should avoid overloading of working memory to facilitate optimal learning outcomes. Cognitive studies have shown that human working memory has limited capacity to process information simultaneously and it is known to hold only seven plus or minus two pieces of information at a time (Miller, 1956). More recently, research has shown that the working memory capacity can be as low as four pieces when processing, rather than just retaining information (Cowan, 2001).

CLT differentiates three distinct kinds of cognitive load. According to (Sweller, 2010), the construct of all three kinds of cognitive load can be unified under the concept of element interactivity. Element interactivity is determined by the number of interacting elements that the learner attends to. Low element interactivity allows learners to process concept elements separately. On the contrary, high element interactivity necessitates reference to other elements in learning certain concepts (Sweller, 2010). Intrinsic cognitive load refers to the natural complexity of instructional material that is related to the learning objective. Intrinsic load is determined by the number of interacting elements that need to be carried simultaneously in working memory for learning a concept. It can be affected by prior knowledge level of learners. Learners with higher prior knowledge chunk information, and therefore may experience fewer interacting elements in the learning materials or task than learners with lower prior knowledge. Extraneous cognitive load refers to mental energy devoted to processing information that is not
relevant to learning objective. Extraneous load, according to Sweller (2010), can also be formalized through element interactivity. For example, non-optimal design of instructional materials forces learners to process information that does not pertain to learning goal, thus this non-optimal design causes high element interactivity and high extraneous load. Extraneous load can be decreased by optimized design of the learning procedure. Different from intrinsic or extraneous load, germaine cognitive load is indirectly related to element interactivity since it is determined by the cognitive resources of the working memory devoted to processing intrinsic load. Germaine load facilitates learning by reducing number of interacting elements relevant to learning by constructing and automating schema. It is theorized that extraneous load and germaine load are complementary in a sense that an increase in germaine load is accompanied by a decrease in extraneous load (Sweller, 2010; Cierniak et al., 2009). This is necessarily true when extraneous load or intrinsic load are sufficiently high such that there are no unused working memory resources available to accommodate an increase in extraneous load.

The most well-known way to decrease extraneous cognitive load is through worked examples (Atkinson et al., 2000; Sweller et al., 1998). Worked examples have been shown to be effective across different domains such as geometry (Paas et al., 1994; Tarmizi & Sweller, 1988), statistics (Paas, 1992; Quilici & Mayer, 1996), algebra (Carroll, 1994; Cooper & Sweller, 1987; Sweller & Cooper, 1985) and physics (Van Gog et al., 2006; Ward & Sweller, 1990). However, the worked examples from most of these studies have been written examples, which means information is only provided visually. Worked examples which provide information in two modalities (visual and auditory) have been used to improve exam performance (Mestre et al., 2015). It has also been shown that participants who learned from animated-narrated worked-out solutions significantly outperformed participants in the control condition that did not receive any
instructional support in a post-learning transfer task (Schroeder et al., 2015). These results are consistent with what CLT predicts: worked examples are effective, especially for novice learners since working through the problem decreases their extraneous load by providing explicit guidance to help novice learners distinguish the relevant interacting elements in the problem and solution from the irrelevant ones, thus reducing the possibility of processing information that does not help learning.

The ability to transfer what is learned from instruction in one situation to a novel situation is an important goal of education. There have been many studies on promoting transfer (for review, see Mestre, 2006). Prior research (Madsen et al., 2012, 2013; Rouinfar et al., 2014a, 2014b) investigated the use of visual cues during training to promote transfer. In their study, participants were randomly assigned to four conditions depending on whether visual cues or outcome feedback were provided during training. They found that providing visual cues led to improved transfer performance. They also found participants in the cueing conditions spent less time attending to the solution-relevant area in the diagram in the transfer tasks, thereby indicating that receiving visual cues facilitated shift of attention of participants to reach automaticity of information processing. Although the performance aided by visual cues is promising, there is reason to believe that it can be improved upon. One reason is that visual cues provide information only via the visual channel, which is also used to convey solution-relevant information. Thus, the presentation of visual cues may overload the visual channel. Participants may experience a higher cognitive load in trying to understand the visual cues. A model of the working memory suggests that our working memory has two partially independent processors, one deals with visual spatial input and the other processes auditory information (Baddeley, 1986). If information can be provided through visual presentation as well as narration, it may be
best use the dual-channel functionality of working memory. This is also supported by Mayer’s
(2001) modality principle which states that people learn better if information is presented
through dual channels i.e. in the auditory as well as visual modes rather than only the visual
mode (Mayer, 2001). Considering all the above arguments, we hypothesize that worked
examples in form of narrated animated solutions (video solutions) will have better effect in
promoting transfer than visual cues only. Video solutions present information in both auditory
and visual channel. The nature of the video solution will be introduced in next section.

We argue that video solutions will not only improve transfer performance but also
impose a lower cognitive load on learners, compared to visual cues. Very few multimedia based
worked example studies have reported the cognitive load experienced by participants. In this
study, we will investigate the cognitive load imposed by two different treatments. Traditionally,
cognitive load can be measured in different ways, using analytical and empirical methods
(Linton et al., 1989; Sweller, 1988; Xie & Salvendy, 2000). Examples of empirical methods are
self-rated reports (Paas, 1992), physiological techniques including heart rate variability (Paas &
van Merrienboer, 1994), brain activity (Whelan, 2007), skin conductivity (Nourbakhsh et al.,
2012), eye blinking, and pupil dilation (Hess & Polt, 1964; Van Gerven et al., 2004). Examples
of task performance related methods for measuring cognitive load include secondary task
performance (Chandler & Sweller, 1996) and response time to a secondary task (Marcus et al.,
1996).

In this study, we use eye movement based metrics to measure the cognitive load imposed
by learning material during training. For example, pupil dilation has long been investigated and
shown to be correlated with the levels of difficulty of tasks (Ahern & Beatty, 1979; Beatty, 1982;
Beatty & Lucero-Wagoner, 2000; Chen et al., 2001; Gerven et al., 2004; Granholm et al., 1996;
Hess & Polt, 1964; Kahneman & Beatty, 1966). A measurement called index of cognitive activity (ICA), which is based on changes in continuously recorded pupil size, was shown to differentiate between levels of mathematical problem complexity and found to be independent of lighting condition (Marshall, 2002, 2007; Marshall et al., 2004). Pupil size variation is a more promising continuous measurement of cognitive load compared to many other techniques such as subjective ratings (Rosch & Vogel-Walcutt, 2013). As found in our Study One, change of pupil size from baseline was shown to be positively correlated with extraneous load and negatively correlated with germane load.

Mean fixation duration was found to be good indicator of cognitive load as well (Chen et al., 2011). Longer mean fixation duration has been found when a low frequency i.e. uncommonly used word was encountered in reading (Rayner & Duffy, 1986) and is suggested more extensive processing (Rayner, 1998). Also, longer mean fixation duration is found when information is more difficulty to understand (Carroll et al., 1992). Also, fixation duration increases as processing demands (task difficulty) increases (Underwood et al., 2004). It was found that, in free viewing, when three, four or five targets were presented with distractors to manipulate the cognitive load, mean fixation duration increased as workload increased (Meghanathan et al., 2015). In our Study One, we found that mean fixation duration is most sensitive to extraneous load, specifically higher mean fixation duration is related to lower extraneous load.

Other eye movement related parameters, such as saccadic length, blink rate and fixation rate all could differentiate tasks complexity levels in a video clip watching task (Chen et al., 2011) and saccadic peak velocity, was also shown to differentiate between different levels of mental workload in a driving simulation task (Di Stasi et al., 2010). It was found that when workload increased, saccadic length decreased and fixation duration increased as found in (Chen
et al., 2011). It is worthy pointing out in our study one, we found that saccade length, saccade
duration, saccade average velocity as well as saccade peak velocity, are all positively correlated
with extraneous load.

Based on this background, in this study, we address the following interrelated research
questions:

1. **Performance** Question: How does performance on transfer tasks compare when
   participants are trained using guided (video solutions) vs. semi-guided (visual cues)
   instruction?

2. **Attention** Question: How does participants’ visual attention on transfer tasks compare
   after receiving guided vs. semi-guided instruction?

3. **Cognitive Load** Question: How does the cognitive load experienced by participants
   during training compare when participants are trained with guided vs. semi-guided
   instructions?

We will discuss participants’ performance, visual attention during transfer, and cognitive
load during training to address the above research questions.

6.2 **Methodology**

6.2.1 **Materials and Data Collection**

N=26 participants enrolled in an algebra-based physics class at a large U.S. Midwestern
land grant university received extra credit equal to 1% of their course grade as incentive for
participation in our study. Informed consent was obtained from all individual participants
included in the study. The materials were conceptual physics tasks covering the topics of energy
and speed similar as those in previous studies (Madsen et al., 2013; Rouinfar et al., 2014a,
2014b).
The problems were designed purposefully to have diagrams that included solution-relevant areas and solution-irrelevant areas. The solution-irrelevant areas typically contain the most salient features which tend to attract participants’ attention. They are also designed based on the participants’ misconceptions reported in literature (McDermott et al., 1987).

![Figure 14: Procedure in Each Problem Set](image)

Participants were randomly assigned to two conditions. Each participant solved two sets of tasks in each condition. Each set contained an initial task, a training session, a near transfer task, and a far transfer task as shown in Figure 14. (Tasks used can be found in Appendix C, Figure 28, Figure 29, Figure 30, Figure 31). The two sets of tasks were presented in counterbalanced order and the training tasks within each set were randomized to eliminate order effect. The initial, near transfer, and far transfer tasks were identical across the two conditions but the training session differed between the conditions. During the training session, participants in the guided instruction condition (video condition), watched two consecutive multimedia video solutions. Participants in the semi-guided instruction condition (cue condition) solved four isomorphic training tasks with visual cues overlaid on the tasks. The tasks used for both training sessions were isomorphic to the initial task and to each other. The training session was designed to ensure equal time-on-task across conditions such that the two videos had the same viewing time as four cued problems. We did not have a third (control) condition receiving neither solutions nor cues because our previous research has already shown that cued condition outperformed the control condition (Madsen et al., 2013; Rouinfar et al., 2014a, 2014b).
Participants were interviewed individually. Each interview lasted around 30 minutes. For
the validation of the scientific correctness and pedagogical appropriateness of the videos, three
physics professors and two graduate TAs who were either teaching or had recently taught an
algebra-based physics class were asked to inspect the video solutions. All the professors and the
TAs deemed the video solutions to be both scientifically correct and pedagogically appropriate
for this audience. Our goal was to determine the impact of the videos and cues on participants’
performance on the transfer tasks as well as the visual attention and cognitive load experienced
by participants. Participants’ responses to the initial tasks, training tasks (in the cue condition)
and transfer tasks were coded as correct if both their answer and reasoning were scientifically
correct. Alternatively, participants’ responses were coded as incorrect if either the answer or the
reasoning was scientifically incorrect. Eye movements were recorded during the whole
experiment with EyeLink 1000+ Eye tracker from SR research. All procedures performed in
studies involving human participants were in accordance with the Institutional Review Board
(IRB) of the University and ethical standards of the institutional and/or national research
committee and with the 1964 Helsinki declaration and its later amendments or comparable
ethical standards.

6.2.2 Visual Cueing and Video Solutions

As mentioned earlier, the diagrams of tasks used contain two distinct areas: solution-
relevant and irrelevant area of interest (AoI) (Figure 15). The visual cues, designed according to
our previous research (Rouinfar et al., 2014a, 2014b; Chi et al., 1981), used color to highlight the
solution-relevant area to redirect participants’ attention to that area.

The multimedia video solutions used in this study are in the same style as the animated
narrated solutions used by Mestre et al. (2015). In our video solution, we started with
underscoring the common mistakes made by participants, based on research on students’
difficulties documented in the literature (McDermott et al., 1987). This was followed by the
introduction of the related physics principles, e. g. principle of energy conservation. Then the
solution explained, step by step, how the principle was applied to solve an example task.

Figure 15 Solution-Relevant (RED) and Irrelevant (BLUE) AoIs on Two Example Tasks

The design of the video solution employed several principles of Mayer’s (2001) cognitive
theory of multimedia learning. Mayer’s (2001) modality principle states that better learning can
be achieved if verbal information is presented orally as narration instead of visually as text. The
redundancy principle states that to facilitate learning, oral narration should not be duplicated by
written text. This is because, when verbal information is presented as text, it competes for visual
resources needed to process other visual information in the problem, such as the picture. Many
research studies have provided evidence to supporting this idea by showing lower cognitive
effort during instruction and better testing performance (Kalyuga et al., 1999; Kalyuga et al.,
2000; Tindall-Ford et al., 1997). In keeping with these principles verbal instructions in the video
solutions were provided through narration only via the auditory channel, and the visual channel
was utilized to provide pictorial information. In keeping with Mayer’s (2001) signaling principle,
whenever pictorial information, such as critical words from the problem statement or a diagram
were mentioned in the narration, this information was highlighted in the visualization. The 
narration and visual information was synchronized, based on Mayer’s (2001) spatial and 
temporal contiguity principles as well as modality principle. Spatial contiguity principle states 
that people learn better if related words and pictures are shown closer together rather than farther 
apart. Temporal contiguity principle states that people will learn better if corresponding words 
and pictures are presented simultaneously rather than successively.

In this study, since both groups of participants were from the same physics course and 
had roughly the same prior knowledge we can assume that they experienced the same intrinsic 
load. Learners of the same level of prior knowledge are said to have the same level of intrinsic 
load facing materials conveying the same learning goal (Sweller, 2010). Extraneous load was 
deliberately manipulated by condition. Such that our worked-examples video solutions would 
impose lower extraneous load than visual cues. Germaine load was not manipulated, but can be 
considered complementary to extraneous load as per prior research (e.g. Cierniak et al., 2009; 
Sweller, 2010).

6.2.3 Near and Far Transfer Tasks

Transfer tasks are different from the training examples in certain ways. Successfully 
solving the transfer task is a measure of transfer of learning from training examples. Barnett and 
Ceci (2002) identified nine different dimensions along which distance of transfer can be 
determined. In summary, these dimensions relate to the ways in which the learning context 
differs from the transfer context. In our study, we manipulated two aspects of the tasks: the 
representation (graphical, pictorial) and context (e.g. roller coaster vs. projectile). We 
operationally defined a task as near transfer if it can be solved with the same underlying principle 
but in a different context, using the same representation. We operationally defined a task as far
transfer if it could be solved with the same underlying principle but was presented in a different context using a different representation. For example, a distance-time graph is a different representation than a motion diagram of successive snapshots in time. A ball dropping from a roof is a different context than a ball rolling along a track.

6.3 Results

We present our results in three sections. We discuss problem solving performance on the initial and transfer tasks, followed by visual attention results on these tasks, and the cognitive load imposed by learning materials on participants during the training session.

6.3.1 Performance

The overall correctness rate, aggregated over both problem sets, of participants’ responses to the initial, near transfer, and far transfer problems for both conditions are shown in Figure 16. The Pearson Chi square test showed no significant differences between the two groups on the initial task, $\chi^2 (1, 52) = 0.361, p = .382$. Thus, participants in both conditions can be considered as equivalent with respect to their prior knowledge of the relevant concepts. For the near transfer task, we found a significant difference between the two conditions, $\chi^2 (1, 52) =$
6.933, \( p = .009 \), with the video condition outperforming the cue condition. We found no significant difference between the two conditions on the far transfer task, \( \chi^2 (1, 52) = 1.926, p = .133 \).

To gain deeper insight, we investigated how participants with different initial task performance responded to the two (video and cue) training conditions. Specifically, we compared participants who solved initial tasks correctly and those who did not. Results for near transfer task are shown in Figure 17. No significant main effect of treatment was found for participants who were initially incorrect, \( \chi^2 (1, 36) = 3.009, p = .128 \). There is a marginally significant main effect of treatment for correctness rate of participants who were initially correct during near transfer task, \( \chi^2(1,16) = 4.747, p = .062 \). Note all the initially correct participants in the video condition answered the near transfer tasks correctly, which indicates a ceiling effect.

The performance results for far transfer task are shown in Figure 18. A Chi square analysis was conducted to investigate the effect of treatment on far transfer performance for both initially correct and incorrect problem solvers. The two treatments did not render significant difference for initially incorrect participants, \( \chi^2(1,36) = .000, p = 1.000 \). We find that the video
condition outperforms the cue condition for initially correct participants, $\chi^2 (1,16) = 6.349, p = .041$.

![Correctness Rate for Far Transfer](image)

**Figure 18: Performance for Far Transfer Tasks**

Taken together, these results seem to suggest that participants from the video condition outperformed those from cue condition on both near and far transfer tasks, and this was particularly true for participants who answered the initial task correctly.

### 6.3.2 Visual Attention

Visual attention is typically characterized using a metric called percent time per percent area or PT/PA. It is the percentage of time spent viewing an AoI divided by the percentage of area that AoI occupies (Rouinfar et al., 2014b). This metric is also referred to as the domain relative ratio (Fletcher-Watson et al., 2008).

In this study, we compared the video and cue conditions with respect to their visual attention over expert AoIs for all students and we also investigated this by looking at both initially correct and incorrect participants as they proceeded from the initial tasks to far transfer tasks. A one-way ANOVA was conducted to compare the initial correctness on domain relative ratio on novice-like AoI, we did not find a significant main effect of initial correctness on domain relative ratio on novice-like AoI, $F (46, 1) = 1.780, p = .189$. A one-way ANOVA was
conducted to compare the initial correctness on domain relative ratio on expert-like AoI, and we found a significant main effect of initial correctness on domain relative ratio on expert-like AoI, \( F(46, 1) = 4.188, p = .046 \). Thus, students who solved the initial task correctly had a significantly higher domain relative ratio over expert-like AoI (Figure 19). These results are different than those from Rouinfar et al. (2014b) where initially incorrect problem solvers had a significant higher domain relative ratio over novice-like AoI, and no significant difference between domain relative ratio over expert-like AoI and domain relative ratio over novice-like AoI was found for initially correct learners.

![Figure 19: PT/PA over Novice-like AoI & PT/PA over Expert-like AoI for Initial Task](image)

Clearly the above results demonstrated that spending time over solution-relevant AoI is important for answering a question correctly. It is important to see those who showed learning after training shifted their visual attention. Especially we want to know which treatment did better in facilitating this process, video solution or visual cueing. Participants who demonstrated learning were those who were initially incorrect, but could solve near transfer or far transfer tasks correctly. We want to investigate how PT/PA over expert-like AoI changes from initial task to near transfer tasks and far transfer tasks for this sub-group of learners.
A repeated measure (between-factor (cue vs. video) condition x within-factor (initial vs. transfer) task) ANOVA was conducted to compare the effect of condition on PT/PA over relevant AoI from initial task to near transfer task. No main effect of task was found, $F(22, 1) = .101, p = .754$. No significant interaction between task and condition was found, $F(22, 1) = .772, p = .389$. No main effect of condition was found, $F(22, 1) = .132, p = .720$. (Figure 20). An ANCOVA was conducted for the same data, we did not find a significant main effect of condition on PT/PA over expert-like AoI for near transfer task when PT/PA over expert-like AoI for initial task was adjusted, $F(21, 1) = .030, p = .864$. (Figure 20).

A repeated measure (between-factor (cue vs. video) condition x within-factor (initial vs. transfer) task) was conducted to compare the effect of condition on PT/PA over relevant AoI from initial task to far transfer task. Since Mauchly’s assumption of sphericity is violated ($p < .001$), a Greenhouse-Geisser correction was applied to the degrees of freedom for the simple main effects as well as interactions. An alpha value of 0.025 was used to determine the level of significance. We found a main effect of task on PT/PA over relevant AoI, $F(1, 13) = 31.643, p < .001$. partial eta squared = .709. We also found a significant interaction between task and
condition, $F(1, 13) = 10.111$, $p = .007$. partial eta squared $= .437$. No main effect of condition was found, $F(1, 13) = 2.455$, $p = .141$. partial eta squared $= .159$. (Figure 21).

The above results are consistent with Rouinfar et al. (2014b) in that they support the idea of automaticity. In that study participants (Rouinfar et al., 2014b) who received training with cues had a lower PT/PA in the relevant AoI than those who did not receive cues. That means visual cueing is a good training technique to foster students to solve transfer problems more automatically. Here our results show that participants who received training through video solutions show more automaticity in solving not only near transfer tasks but also far transfer tasks, than those who received training through visual cues. The improvement in automaticity can be attributed to the different functionality of video solution and visual cues. Visual cueing highlights relevant information implicitly, leveraging bottom-up attentional processes. On the other hand, video solution explaining the concepts underlying the problem solution, leverage top-down executive processes. Based on these results of automaticity, we may also hypothesize that participants in the video solutions condition experienced less extraneous cognitive load. Sweller (1988) showed that using means-ends analysis would impose very high extraneous load. Unlike
the visual cues condition, the video solutions provided explicit information, and the learner did not have to engage in means-ends analysis to solve the problem. In the next section, we will test the hypothesis above.

6.3.3 Cognitive Load

We investigated students’ extraneous cognitive load by collecting eye movements of students during the training session for the two conditions. Specifically, we examined the effect of condition on pupil diameter, mean fixation duration, saccade peak velocity, mean saccade duration, mean saccade length, and saccade average velocity.

A one-way ANOVA to compare the mean fixation duration (MFD) between video and cue conditions found a significant main effect of condition on mean MFD was found, $F(1, 142) = 33.685$, $p < .001$, such that the video condition had a significantly higher MFD than the cue condition. A one-way ANOVA to compare the mean pupil size (PS) between video and cue conditions found a significant main effect of condition on pupil diameter, $F(1, 142) = 7.918$, $p = .006$, such that the video condition had a significantly higher pupil diameter than the cue condition. A one-way ANOVA to compare the saccade peak velocity (SPV) between video and cue conditions found no significant main effect of condition on SPV, $F(1, 142) = .162$, $p = .688$. A one-way ANOVA to compare the mean saccade duration (MSD) between video and cue conditions also found no significant main effect of condition on MSD, $F(1, 142) = .013$, $p = .910$. A one-way ANOVA to compare the mean saccade length (MSL) between video and cue conditions found a significant main effect of condition on MSL, $F(1, 142) = 23.526$, $p < .001$, such that the video condition had a significantly shorter MSL than the cue condition. Finally, a one-way ANOVA to compare the saccade average velocity (SAV) between video and cue
conditions found a significant main effect of condition on SAV, $F(1, 142) = 44.871$, $p < .001$, such that the video condition had a significantly smaller SAV than the cue condition. (Table 21).

Table 21: Means (S.D) of Eye Movement Parameters for Both Conditions & $p$-values

<table>
<thead>
<tr>
<th></th>
<th>Video</th>
<th>Cue</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean fixation duration (ms)</td>
<td>307.83 (94.63)</td>
<td>245.24 (33.64)</td>
<td>&lt;.001*</td>
</tr>
<tr>
<td>Pupil diameter (arbitrary unit)</td>
<td>1988.37 (348.14)</td>
<td>2340.52 (830.18)</td>
<td>.006*</td>
</tr>
<tr>
<td>Saccade peak velocity (degree/s)</td>
<td>303.06 (108.66)</td>
<td>308.60 (56.56)</td>
<td>.688</td>
</tr>
<tr>
<td>Mean saccade duration (ms)</td>
<td>82.98 (62.89)</td>
<td>83.95 (39.49)</td>
<td>.910</td>
</tr>
<tr>
<td>Mean saccade length (degree)</td>
<td>5.04 (1.32)</td>
<td>6.15 (1.29)</td>
<td>&lt;.001*</td>
</tr>
<tr>
<td>Saccade average velocity (degree/s)</td>
<td>101.22 (15.29)</td>
<td>121.66 (17.81)</td>
<td>&lt;.001*</td>
</tr>
</tbody>
</table>

Out results from study one indicate that a smaller mean fixation duration, larger pupil diameter change, longer mean saccade length, longer mean saccade duration and longer saccade average velocity are all correlated with a higher extraneous load. Consistent with what we found in Study one, we found evidence supporting a smaller extraneous cognitive load during training in the video condition compared to the cue condition as indicated by larger mean fixation duration, smaller pupil diameter, smaller saccade length, and a smaller saccade average velocity induced by video solutions. This could be attributed to the lack of guidance in the cue condition that made it more difficult, than in the video condition, for participants to differentiate between relevant and irrelevant information.

Further, we investigated the how the cognitive load experienced by students during training in each condition was affected by whether or not they solved the initial task correctly. A two-way ANOVA (condition x initial correctness) was conducted to investigate the effect of condition and initial correctness on mean fixation duration of students during training. A
significant main effect of condition on mean fixation duration was found, \( F (1, 140) = 47.714, p < .001, \eta^2 = .254 \). A simple main effect reveals that both initially correct and incorrect learners had a larger mean fixation duration when they were trained with video solutions. We found a significant main effect of initial correctness on mean fixation duration, \( F (1, 140) = 9.438, p = .003, \eta^2 = .063 \). We also found a significant interaction, \( F (1, 140) = 13.843, p < .001, \eta^2 = .09 \) (Figure 22).

![Figure 22: Mean Fixation Duration according to Initial Correctness and Assigned Condition](image)

A two-way ANOVA (condition x initial correctness) was conducted to investigate the effect of condition and initial correctness on pupil size of students during training. A marginal significant main effect of condition on pupil size was found, \( F (1, 140) = 3.874, p = .051, \eta^2 = .027 \). A simple main effect revealed that initially incorrect learners who were trained with visual cues had a significantly larger pupil diameter. We did not find main effect of initial correctness on pupil size, \( F (1, 140) = .014, p = .905 \). We did not find a significant interaction either, \( F (1, 140) = 2.069, p = .153, \eta^2 = .015 \) (Figure 23).

It seems that effect of video solution on extraneous load of students during training does not depend on whether or not they solved initial task correctly. This suggests even students who
solved initial task correctly, in other words higher prior knowledge students, would still benefit more from learning from multimedia video solutions than from the visual cues.

![Extraneous load during training](image)

Figure 23: Pupil Diameter based on Condition and Initial Correctness of Students

### 6.4 Discussion and Implications

We investigated the effects of providing visual cues and multimedia formatted worked examples (video solutions) on participants’ transfer task performance, shift of visual attention from initial to transfer sessions, as well as cognitive load imposed on learners during training. The overall results showed that video condition outperformed the cue condition on the transfer tasks, with a significant difference in performance between conditions on the near transfer task, albeit not on the far transfer task. This result provides evidence supporting effectiveness of the video solution over visual cues. This is consistent with the fact that the video solution provides explicit guidance for solving the problem. The video solution includes the pertinent conceptual knowledge as well as the procedural knowledge to solve the problem. Further, consistent with the modality principle (Mayer, 2001), it provides information through two modalities: visual and auditory. The visual cue, on the other hand, provides only implicit guidance for visual attention, through only the visual modality. It does not provide any explicit conceptual or procedural guidance.
Our results also show that initial task performance moderates the performance on the transfer tasks. Participants who solved the initial task incorrectly, the two treatments (video & cue) did not lead to significant difference in performance on either the near or far transfer tasks. For participants who solved the initially task correctly, the two treatments did not show significant difference in performance on the near transfer task, but on the far transfer task, the video condition significantly outperformed cue condition. The fact that the video condition outperformed the cue condition can be attributed to two factors: shifts in visual attention and reduction of extraneous cognitive load. We discuss each of these factors separately below.

Shifts in visual attention can provide useful insights into the cognitive processes of students. For example, on the initial task, as measured by domain relative ratio, students spend more time looking at expert-like AoI when they could solve initial tasks successfully. This supported the idea of spending more time processing solution-relevant information is important, and is consistent with previous research in visual attention in physics problem solving (Madsen et al, 2013). Further, previous research has also showed that learners who redirected their attention due to visual cueing showed better training as well as transfer performance than those who did not (Rouinfar et al., 2014b). In our study, we found the similar pattern in favor of using video solution over visual cueing. We also found that use of video solutions facilitate more automatic successful problem solving as evidenced by a smaller domain relative ratio shown by those who demonstrated learning from video condition than cue condition. We think this is because of top down guidance prevented participants from using means-ends analysis since the video solutions provided them with explicit guidance regarding underlying physics conceptual and procedural information for solving the problem. On the contrary, visual cues may not rule out the possibility that participants used means-ends analysis since visual cues highlighted
relevant information only as implicit, not explicit guidance. This indicates that top down process facilitated by video solutions, may help more than bottom up process facilitated by visual cues, in fostering automaticity.

Other evidence supporting the success of participants who received video solutions over those that received visual cues pertains to the reduction in extraneous cognitive load in the visual cueing condition. Extraneous cognitive load during the training sessions was measured using several eye-movement based metrics, namely, mean pupil diameter, mean fixation duration, mean saccade length, saccade peak velocity, saccade average velocity, mean saccade duration. Taken together, these results, based on the conclusions of Study One in which we established that these eye movement parameters are sensitive to extraneous load, converge on the conclusion that multimedia video solutions imposed lower extraneous load on the learners during training session. We also found that among participants who solved the initially task incorrectly, the video condition showed a significantly smaller mean pupil size indicating a smaller extraneous cognitive load than the cue condition. However, we did not find a significant difference between conditions for participants who solved the initial task correctly. These results can partly be interpreted in light of the expertise reversal effect.

The expertise reversal effect posits that lower prior knowledge learners may experience lower cognitive load from instruction that offers more guidance rather than from instruction that offers less guidance. This is because guidance enables lower prior knowledge learners to distinguish between the relevant and irrelevant interacting elements, thereby reducing the extraneous cognitive load. If we assume that initial task performance is a proxy for prior knowledge in that participants who solved the initial task incorrectly can be categorized as having a lower prior knowledge than participants who solved the initial task correctly, our results
seem to indicate that among the lower prior knowledge participants (i.e. participants who solved the initial transfer task incorrectly), the video condition experienced lower cognitive load than the cue condition.

We discuss the implications for instruction based on the results of the Study Two. From (Rouinfar et al., 2014), we know that providing students visual cues can promote their problem-solving performance. Visual cueing investigated by (Rouinfar et al., 2014) clearly can be incorporated in the human-computer interface of a learning system. Not providing direct instruction immediately to students who have a certain level of knowledge might be beneficial as per expertise reversal effect (Kalyuga, 2007). Since students of high knowledge know the material, providing direct (guided) instruction may impose greater extraneous load than unguided instruction. However, for students of low knowledge level, providing visual cues might not be enough adequate due to the high extraneous cognitive load imposed by the visual cues. Our results indicate that providing students worked examples in the form of multimedia instruction can further decrease students’ extraneous load. If we can also combine this with the results of the Study One, which allows us to measure the three kinds of cognitive load, we can make the human computer interface of learning systems more intelligent in determining when students are experiencing high cognitive load and what the kinds of assistance that the system should provide. This merits further research.
7. GENERAL DISCUSSION

7.1 Summary

The current state of cognitive load theory (Sweller, 2010) posits a triarchic model of cognitive load with three independent types of cognitive load. Intrinsic load pertains to the learning goal itself. It is determined by the number of interacting elements i.e. number of pieces of information that the learner must simultaneously attend to, directly related to the learning goal. Typically, given the level of prior knowledge of a learner and the learning goal established in the learning materials, the intrinsic load cannot be altered.

Extraneous load pertains to the features of the learning material that distract from the learning goal. In other words, it is determined by the number of interacting elements that are unrelated to the learning goal. Novice learners who are unable to distinguish between information that is relevant and irrelevant to the learning goal, in the learning material may experience a higher extraneous load. Furthermore, choices made by the instructional developer in the design of the learning material can also influence the extraneous cognitive load.

Finally, the germane cognitive load pertains to the working memory resources allocated by the learner to accomplish the learning goal. In other words, the germane cognitive load is directly related to the intrinsic load. Because the working memory capacity of humans is limited, the greater the extraneous load, the greater will be the working memory resources allocated to attend to irrelevant information, leaving fewer working memory resources available to process the intrinsic load. Therefore, a greater extraneous load often results in a smaller germane load, and also consequently less learning.

We conducted three studies to explore cognitive load theory in different educational contexts. In the first study, we explored the possibility of using eye-tracking technology to
measure the three kinds of cognitive load: intrinsic load, extraneous load, and germane load with a multimedia lesson in a physics context. We found evidence supporting the sensitivity of certain eye-tracking based parameters to each type of the three loads.

In the second experiment, we developed a multiple-item subjective rating survey to investigate the three kinds of cognitive loads when students were taking a paper-and-pencil, multiple choice conceptual test, first individually, and later as a group engaging in peer interaction. We found that when learners with adequate amount of prior knowledge, such that those taking a post-test after instruction, interact with each other, peer interaction can improve students’ performance by decreasing their intrinsic load, although the process of engaging in peer interaction may increase extraneous cognitive load. We found that none of these effects of peer interaction – improvement in performance, decrease in intrinsic load and increase in extraneous load – occurred on the pre-test when the interacting students did not have adequate prior knowledge.

In the third experiment, we compared the effect of narrated-animated worked examples with visual cueing, both used the service of improving near and far transfer on conceptual physics problems. We found that worked examples in the form of video solutions can better improve performance on transfer tasks than visual cues. The reason for the success of video solutions over visual cues are twofold. First, video solutions cause greater automaticity with regard to the shift of visual attention from the solution-irrelevant area to the solution-relevant area on the transfer problems. Second, video solutions also impose lower extraneous cognitive load on initially incorrect solvers, than the visual cues. This is because video solutions provide explicit guidance versus visual cues that provide implicit guidance. The use of explicit guidance
is particularly beneficial to low prior knowledge students and helps reduce their cognitive load by facilitating them to distinguish between relevant and irrelevant information.

7.2 Limitations and Future Directions

Study One: The first limitation for this study is that we used only one set of ways to manipulate the three kinds of cognitive load: intrinsic load was manipulated using sentence complexity, extraneous load was manipulated using the redundancy effect, and germane cognitive load was not manipulated, rather it was ascertained based on transfer test performance. Clearly, there are several other ways to manipulate each of these loads. For instance, as described in the introduction, there is a clear mapping between the three types of loads and various multimedia learning principles. Based on these principles it is possible to design other ways to manipulate the three different kinds of cognitive loads and use these manipulations in an experimental design to further investigate how various eye movement parameters are sensitive to different kinds of cognitive load.

The second limitation for this study is that although we showed several eye-tracking based parameters were most sensitive to the three subtype cognitive loads respectively, these values were either cumulative (such as total dwell time) or averaged over each trial (such as mean fixation duration). The dynamic characteristics that can potentially be captured using eye movements were averaged over time. The main reason for this is that the research design did not allow for real time manipulation of extraneous load, rather participants were placed in either the low or high extraneous load condition. In future, online learning systems that are responsive in real time to learner interaction would require monitoring more than just the mean parameters over a trial. Rather, they would need to be able to use the dynamic value of each eye movement
metric in a continuous stream of data to respond to the learner in real time. We plan to study this issue in the future.

The third limitation is that in this study, our participants were primarily low prior knowledge learners. This limits the application or generalizability of our results to this population. The results of the study would be more complete if we our pool of participants included higher prior knowledge learners as well. We will expand this study to address this issue in the future.

The fourth limitation is that this study is limited to only one physics context which is how electric motor works. To further test the generalizability of the results, we must expand this study to include different contexts relevant to physics and other STEM disciplines.

**Study Two:** In this study, we examined effect of peer interaction on cognitive load of as students completed a circuits conceptual test – DIRECT -- first as a pre-test and later as a post-test. It is important to expand this investigation to include more students of different majors and different tests such as the Force Concept Inventory (Hestenes, 1992) or BEMA (Ding et al. 2006) to increase the potential impact of our results. In our study, we used a subjective survey as way to measure the three kinds of cognitive load. Future studies should include other kinds of measurements, such as eye-movements and EEG.

**Study Three:** The main limitation of this study is clearly the small number of participants. Another important limitation is that we have conducted this study with only two sets of tasks, both of which are in area of introductory mechanics. Similar studies with a wider range of applicability would be needed to study if the results would be more generally applicable. In the discussion, we operationally assumed the prior knowledge level of participants based on performance on initial task. Clearly, performance on initial task is by no means a clearly valid
measure of prior knowledge. To ascertain the participants’ prior knowledge, a more exhaustive pre-test would be necessary. Another very important limitation pertains to the ecological validity of this study. The study was conducted in a laboratory with participants recruited from an introductory physics course. It would be interesting to conduct another study with participants who use visual cues or video solutions to solve problems as part of their usual coursework. It also informs us that some more work is needed for understanding what kinds of scaffolding treatments might be facilitating participants of different levels of prior knowledge. Finally, the study tells us that more research is needed to improve our understanding of cognitive load theory and the associated expertise reversal effect.
REFERENCES


Tassinary, L. G. (2000). The skeletomotor system: Surface electromyography. JT Cacioppo, LG Tassinary, & GG Berntson (Eds.), Handbook of psychophysiology (pp. 163–199).


Kun, A. L., Palinko, O., & Razumenic, I. (2012, October). Exploring the effects of size and luminance of visual targets on the pupillary light reflex. In *Proceedings of the 4th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 183-186). ACM.


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.


APPENDIX A1. MATERIALS USED IN EXPERIMENT ONE

“When the motor is switched on, electrons flow from the negative terminal of the battery through the yellow wire to the end of the yellow wire. Electrons flow from the end of the yellow half of the commutator, then through the wire loop to the red half of the commutator. Electrons flow from the red half of the commutator to the end of the red wire and through the red wire to the positive terminal of the battery. A magnetic field runs from the north end of the magnet on top to the south end of the magnet on the bottom, cutting across the path of the electrical flow in the wire loop. When the downward moving magnetic field intersects the right-to-left moving electron flow on the yellow half of the wire loop, a force is exerted pushing the electrons (and the yellow half of the wire loop) forward toward where you are sitting. When the downward moving magnetic field intersects the left-to-right moving electron flow in the red half of the wire loop, a force is exerted pushing electrons (and the red half of the wire loop) backward away from where you are sitting. Thus, the wire loop begins to rotate. When the wire loop has rotated 180 degrees, the red half of the wire loop has rotated to the bottom and the yellow half of the wire loop has rotated to the top.

The red half of the commutator makes contact with the end of the yellow half of the wire loop and the yellow half of the commutator makes contact with the end of the red half of the wire loop. Electrons flow from the negative terminal through the yellow wire to the yellow half of the commutator. Then electrons flow from the red half of the commutator through the red wire to the positive terminal of the battery. Electrons flow leftward along the red half of the wire loop (on the bottom) and rightward along the yellow half of the wire loop (on the top). A magnetic field runs from the top magnet to the bottom magnet, cutting across the path of the electrical flow in the wire loop. When the downward moving magnetic field intersects the left-to-right moving electron flow on the yellow half of the wire loop (now on top), a force is exerted pushing the
electrons (and the yellow half of the wire loop) backwards away from where you are sitting. When the downward moving magnetic field intersects the right-to-left moving electron flow on the red half of the wire loop (now on the bottom), a force is exerted which pushes electrons (and the red half of the wire loop) forward toward where you are sitting. Thus, the wire loop continues to rotate in the same direction.”
APPENDIX A2. PRE-SURVEY QUESTIONS FROM EXPERIMENT ONE

The question is the same for all five diagrams used in this session

“Is this following figure familiar to you? Please rate on a scale from 1 (I do not know anything about this diagram) to 5 (I understand this diagram very well). What kind of physics phenomenon is it representing? Write down several sentences as your summary.”

Figure 24: Five Diagrams Used in Pre-survey

See saw task        Induction task        Electric motor task

Electric generator task        Free falling task

Figure 24: Five Diagrams Used in Pre-survey
APPENDIX A3. QUESTIONS FROM TRANSFER TEST IN EXPERIMENT ONE

1. What could you do to increase the speed of the electric motor, that is, to make the wire loop rotate more rapidly?

2. What could you do to increase the reliability of the electric motor, that is, to make sure it would not break down?

3. Suppose you switch on an electric motor, but nothing happens. What could have gone wrong?

4. What could you do to reverse the movement of the electric motor, that is, to make the wire loop rotate in the opposite direction?

5. Why does the wire loop move?

6. If there was no momentum, how far would the wire loop rotate when the motor is switched on?

7. What happens if you move the magnets further apart? What happens if you connect a larger battery to the wires? What happens if you connect the negative terminal to the red wire and the positive terminal to the yellow wire?
APPENDIX A4. SNAPSHOTS OF MENTAL EFFORT AND DIFFICULTY RATING
QUESTION PAGES USED IN EXPERIMENT ONE

Figure 25: Mental Effort Rating Snapshot

Figure 26: Difficulty Rating Snapshot
APPENDIX B. SNAPSHOT OF THE SURVEY FOR MEASURING COGNITIVE LOADS USED IN EXPERIMENT TWO

<table>
<thead>
<tr>
<th></th>
<th>The topics covered on the physics test were very complex.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(when I was taking the test individually)</td>
</tr>
<tr>
<td></td>
<td>1 2 3 4 5 6 7 8 9 (completely the case)</td>
</tr>
<tr>
<td></td>
<td>(when I was taking the test in the group)</td>
</tr>
<tr>
<td></td>
<td>1 2 3 4 5 6 7 8 9 (completely the case)</td>
</tr>
<tr>
<td></td>
<td>(not at all the case)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>The questions on the physics test had confusing language that was not clear to me.</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>(when I was taking the test individually)</td>
</tr>
<tr>
<td></td>
<td>1 2 3 4 5 6 7 8 9 (completely the case)</td>
</tr>
<tr>
<td></td>
<td>(when I was taking the test in the group)</td>
</tr>
<tr>
<td></td>
<td>1 2 3 4 5 6 7 8 9 (completely the case)</td>
</tr>
<tr>
<td></td>
<td>(not at all the case)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>I concentrated a lot as I answered the questions on the physics test.</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>(when I was taking the test individually)</td>
</tr>
<tr>
<td></td>
<td>1 2 3 4 5 6 7 8 9 (completely the case)</td>
</tr>
<tr>
<td></td>
<td>(when I was taking the test in the group)</td>
</tr>
<tr>
<td></td>
<td>1 2 3 4 5 6 7 8 9 (completely the case)</td>
</tr>
<tr>
<td></td>
<td>(not at all the case)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>The physics test covered formulas that I perceived to be very complex.</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>(when I was taking the test individually)</td>
</tr>
<tr>
<td></td>
<td>1 2 3 4 5 6 7 8 9 (completely the case)</td>
</tr>
<tr>
<td></td>
<td>(when I was taking the test in the group)</td>
</tr>
<tr>
<td></td>
<td>1 2 3 4 5 6 7 8 9 (completely the case)</td>
</tr>
<tr>
<td></td>
<td>(not at all the case)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>It was very hard to identify what information is relevant to answering the questions on the physics test.</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>(when I was taking the test individually)</td>
</tr>
<tr>
<td></td>
<td>1 2 3 4 5 6 7 8 9 (completely the case)</td>
</tr>
<tr>
<td></td>
<td>(when I was taking the test in the group)</td>
</tr>
<tr>
<td></td>
<td>1 2 3 4 5 6 7 8 9 (completely the case)</td>
</tr>
<tr>
<td></td>
<td>(not at all the case)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>I devoted a lot of mental effort in finding and applying the relevant concepts needed to answer the questions on the physics test.</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>(when I was taking the test individually)</td>
</tr>
<tr>
<td></td>
<td>1 2 3 4 5 6 7 8 9 (completely the case)</td>
</tr>
<tr>
<td></td>
<td>(when I was taking the test in the group)</td>
</tr>
<tr>
<td></td>
<td>1 2 3 4 5 6 7 8 9 (completely the case)</td>
</tr>
<tr>
<td></td>
<td>(not at all the case)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>The test covered concepts and definitions that I perceived as very complex.</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>(when I was taking the test individually)</td>
</tr>
<tr>
<td></td>
<td>1 2 3 4 5 6 7 8 9 (completely the case)</td>
</tr>
<tr>
<td></td>
<td>(when I was taking the test in the group)</td>
</tr>
<tr>
<td></td>
<td>1 2 3 4 5 6 7 8 9 (completely the case)</td>
</tr>
<tr>
<td></td>
<td>(not at all the case)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>There was a lot of distracting information in the question statements on the physics test.</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>(when I was taking the test individually)</td>
</tr>
<tr>
<td></td>
<td>1 2 3 4 5 6 7 8 9 (completely the case)</td>
</tr>
<tr>
<td></td>
<td>(when I was taking the test in the group)</td>
</tr>
<tr>
<td></td>
<td>1 2 3 4 5 6 7 8 9 (completely the case)</td>
</tr>
<tr>
<td></td>
<td>(not at all the case)</td>
</tr>
</tbody>
</table>

For each statement below, circle a number that best describes your agreement on a scale that ranges from 1 (not at all the case i.e. completely DISAGREE) 9 (completely the case i.e. completely AGREE).

You will do this TWICE: 1st time after taking test INDIVIDUALLY 2nd time, after taking test in a GROUP.

Figure 27: Snapshot of the Survey Used in Study Two
APPENDIX C. CONCEPTUAL PHYSICS TASKS USED IN EXPERIMENT THREE

Figure 28: Initial Tasks for Speed (Left) and Energy (Right) Tasks, One for Each

Two balls roll along the paths shown. A snapshot of the position of the balls is taken every second. At what point in time does Ball B have the same speed as Ball A?

How does the final speed of cart A compare to the final speed of cart B, if the mass of the carts is the same and they both start at rest? (Frictional effects can be ignored)

Two balls roll along the paths shown. A snapshot of the position of the balls is taken every second. At what point in time does Ball B have the same speed as Ball A?

How does the final speed of cart A compare to the final speed of cart B, if the mass of the carts is the same and they both start at rest? (Frictional effects can be ignored)
Figure 29: Training Tasks for Speed (Left) and Energy (Right) Tasks, Four for Each

Two balls roll along the paths shown. A snapshot of the position of the balls is taken every second. At what point in time does Ball B have the same speed as Ball A?

Ball A

Ball B

Figure 30: Near Transfer Tasks for Speed (Left) and Energy (Right) Tasks, One for Each

How does the final speed of cart A compare to the final speed of cart B, if the mass of the carts is the same and they both start at rest? (Frictional effects can be ignored)

Initial A Final Initial B Final

Two balls roll along the paths shown. A snapshot of the position of the balls is taken every second. At what point in time does Ball B have the same speed as Ball A?

Ball A

Ball B

Ball A begins riding downward in an elevator at the same time Ball B is dropped from the roof of an adjacent building. A snapshot of the balls is taken every second. At what point in time does Ball B have the same speed as Ball A?

Ball A

Ball B

Two identical balls roll down a hill. How does the final speed of ball A compare to the final speed of ball B if the masses are the same and they both start at rest? (Frictional effects can be ignored)

Initial A Final Initial B Final
The motion of two objects is represented in the graph. When are the two objects moving with the same speed?

Two identical stones A and B are shot from cliff with identical initial speeds. How do the speeds of the stones compare the instant before they hit the ground?

Figure 31: Far Transfer Tasks for Speed (Left) and Energy (Right) Tasks, One for Each