Retrieval-Based Argument Mapping Promotes Learning Transfer

by

Qing Liu

M.A., Simon Fraser University, 2011

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

in the
Educational Technology and Learning Design Program
Faculty of Education

SIMON FRASER UNIVERSITY
Spring 2020

Approval

Name:	**Qing Liu**
Degree:	**Doctor of Philosophy**
Title:	**Retrieval-Based Argument Mapping Promotes Learning Transfer**
Examining Committee:	**Chair:** Robert Williamson Assistant Professor
	John Nesbit Senior Supervisor Professor
	Phil Winne Supervisor Professor
	Kevin O'Neill Internal Examiner Associate Professor
	Matthew McCrudden External Examiner Associate Professor Educational Psychology, Counseling, and Special Education Pennsylvania State University
Date Defended/Approved:	January 30th, 2020

Ethics Statement

The author, whose name appears on the title page of this work, has obtained, for the research described in this work, either:

a. human research ethics approval from the Simon Fraser University Office of Research Ethics

or

b. advance approval of the animal care protocol from the University Animal Care Committee of Simon Fraser University

or has conducted the research

c. as a co-investigator, collaborator, or research assistant in a research project approved in advance.

A copy of the approval letter has been filed with the Theses Office of the University Library at the time of submission of this thesis or project.

The original application for approval and letter of approval are filed with the relevant offices. Inquiries may be directed to those authorities.

Simon Fraser University Library
Burnaby, British Columbia, Canada

Update Spring 2016

Abstract

The purpose of my thesis was to investigate if the effects of retrieval practice are enhanced by having learners recall studied information in the format of an argument map. A sample of 120 university students was randomly divided into three treatment groups: a restudy group, a retrieval practice group, and a retrieval-based dialectical map construction group. After reading a text about wind power, the restudy group reread the text. The retrieval practice group completed two cycles of unstructured retrieval practice of the text. The dialectical map group constructed argument maps in the absence of the text with the aid of a web-based argument visualization tool called the Dialectical Map (DMap). Participants returned within two weeks to complete the outcome tests, including a free recall test, a short-answer test, and an argument essay. The latter two measures required transfer and application of knowledge acquired from the text. The results indicated that retrieval-based argument mapping did not yield superior recall, but it did promote knowledge transfer. Argument mapping as a retrieval activity contributed to greater short-answer and argument essay test achievement relative to restudy and free recall testing. Unexpectedly, participants who engaged in free recall practice after reading the text and those who reread the text performed similarly on all three measures. The interaction effect between need for cognition and study strategy was not statistically detectable. This research is the first to integrate retrieval practice and argument mapping and provides new insight into the phenomenon of test-enhanced learning.

Keywords: test-enhanced learning; retrieval practice; argument visualization; retention; learning transfer; argumentation

To my family

Acknowledgements

Looking back on my PhD journey, plenty of inspiring and heartwarming moments stand out in my mind. Words seem powerless to express my appreciation to those who have uplifted me along the way and made this journey memorable and rewarding.

First of all, I would like to express my deepest gratitude to my senior supervisor Dr. John Nesbit. I could not have reached this milestone without his continuous support and encouragement. When we first met in 2008, John showed trust in me and connected me to a research assistant opportunity that opened the door to a whole new world for me. Over the past few years, he has engaged me in various projects and guided me through the process of conducting different types of educational research. John is exceptionally knowledgeable but very easy to work with. He is always patient, considerate, and respectful. I often think John is not just helping me learn. He makes me a better person. I am also extremely grateful to Dr. Phil Winne for his guidance on experimental design and data analysis. Phil has a keen eye for detail. Each conversation with him was enlightening. As a world-renowned scholar, Phil never ceases to amaze me with his wisdom and acumen. My supervisory committee is small but mighty. It really has been my great pleasure and good fortune to get to know and work with them.

I further express my appreciation to Dr. Kevin O'Neill, Dr. Alyssa Wise, Dr. David Kaufman, and Dr. Cheryl Amundsen. They have been inspirational in many ways. What they taught me prepared me well for this research and helped me thrive at work.

I would like to sincerely thank Dr. Cindy Xin for hand-holding me step by step through the process of doing content analysis and Dr. Olusola Adesope for teaching me how to conduct a meta-analysis when I just started studying at SFU. Their guidance and encouragement meant a lot to a new graduate student and ignited my interest in doing research.

Genuine thanks should also go to Prof. Joan Sharp for recognizing my work and connecting me to multiple exciting opportunities. Her passion for integrating innovative technologies and pedagogical practices into classrooms led me to appreciate the impact of educational research on teaching and learning. I want to give a shout-out to Kenny Teng for setting up the DMap tool for my experiment and offering prompt assistance

during the process of data collection. I feel blessed to have joined the DMap project team and collaborated with a group of amazing people towards a meaningful goal.

I greatly appreciate Dr. Robyn Schell and my other colleagues at the Centre for Educational Excellence (formerly Teaching and Learning Centre) and the Beedie School of Business for supporting me to accomplish a variety of projects that have challenged me to step out of my comfort zone and contributed to my professional growth.

I also want to thank my peers whom I have been studying and working with throughout the years, especially Miwa Watanabe for taking a huge amount of time to help me test the inter-rater reliability. The PhD journey is lonely. Because of them, I gain a sense of belonging.

Last but not least, my sincere gratitude goes to my family. I am forever indebted to my parents for their selfless love and unwavering support that makes this endeavour possible. Since I was very young, they have been teaching me the importance of being diligent, tenacious, genuine, and empathetic. The power of such positive values has really shown and carried me through challenging times. I am also grateful to my husband Garrick Chu for respecting all my decisions, shouldering responsibilities, and always putting my needs above his own. Special thanks to my beloved Amber and Aiden for joining us on the journey. Their arrivals have revolutionized my life, cured my procrastination, and reconnected me with the world of wonder and miracle in which fairies and superheroes are within reach.

Table of Contents

List of Tables

List of Figures

Chapter 1.

Overview of the Research

1.1. Theoretical Underpinning

1.1.1. Retrieval Practice

Over the past decade, ample research has unlocked the potential of retrieval practice for promoting learning (Blunt & Karpicke, 2014; Carpenter, 2009; Roediger & Butler, 2011; Roediger & Karpicke, 2006; Pyc & Rawson, 2010; Wong & Lim, 2019). Its superiority over repeated studying in enhancing long-term retention of previously learned materials has been well established (Adesope, Trevisan, & Sundararajan, 2017; Roediger & Butler, 2011; Rowland, 2014). There are several theoretical accounts proposed to elucidate test-enhanced learning. For instance, the transfer-appropriate processing hypothesis attributes the testing effect to the identical cognitive processing instigated by initial and final tests (Morris, Bransford, & Franks, 1977; Roediger & Karpicke, 2006; Rowland, 2014). The elaborative retrieval hypothesis underscores that retrieval practice is likely to elicit mediating or cue-related information that aids the recall of studied information at a later time (Carpenter, 2009; Carpenter, 2011; Pyc & Rawson, 2010). The retrieval effort theories highlight the positive relationship between effortful retrieval and subsequent memory (Bjork & Bjork, 1992; Pyc & Rawson, 2009) and are consistent with the idea that desirable difficulties promote durable learning (Bjork & Bjork, 2011).

Although retrieval practice is an effective strategy for boosting recall, studies investigating its use with complex educational materials and its effects on meaningful learning have obtained mixed results (Agarwal, 2019; Gate, 1917; Kühn, 1914; Rawson, 2015; van Gog & Sweller, 2015). According to Mayer (2002), meaningful learning occurs when a learner “not only possesses relevant knowledge…[uses] that knowledge to solve problems and understand new concepts…[but also] can transfer her knowledge to new problems and new learning situations” (p. 227).

Some research has failed to detect the efficacy of standard free recall retrieval practice in fostering knowledge transfer (e.g., Hostetter, Penix, Norman, Batsell, & Carr, 2019) and argumentation (e.g., Wong & Lim, 2019). It is likely that free recall does not effectively promote relational/organizational processing or contribute to a transferable understanding of the text to be learned. Theorists have suggested it may not orient learners' attention to relevant information that is critical for achieving the intended learning outcomes (Hostetter et al., 2019; Wong & Lim, 2019). It is therefore imperative to investigate how to optimize retrieval-based learning for more complex educational materials and learning outcomes.

1.1.2. Argumentation-Based Learning

There is increasing evidence for the benefits of engaging students in argumentation-based learning (Andriessen, 2006; Eskin & Ogan-Bekiroglu, 2013; Nussbaum & Schraw; 2007; van Gelder, 2015). Argumentation is a rational and social activity in which individuals come up with a list of propositions to uphold a specific point of view and refute the reasoning of its opponents to strengthen the arguments they put forward (Nussbaum, 2011; Nussbaum & Sinatra, 2003; van Eemeren & Grootendorst, 2004; van Eemeren, Grootendorst, & Henkemans, 1996). Toulmin's model of argumentation and Walton's framework play a dominant role in guiding the analysis, evaluation, and construction of arguments (Nussbaum, 2011; Toulmin, 1958; van Eemeren et al., 1996; Walton, 1989).

Research has found that making effort to produce reasoned arguments for or against a contentious standpoint offers several potential benefits (Andriessen, 2006; Eskin & Ogan-Bekiroglu, 2013). Engaging in argument-based inquiry nurtures scientific thinking (Jonassen & Kim, 2010), promotes deep processing also known as elaborative processing (Dole & Sinatra, 1998), and reinforces the connections of ideas within a learner's cognitive framework (Hogan & Fisherkeller, 2000). It facilitates construction, reconstruction, and use of conceptual knowledge (Ogan-Bekiroglu & Eskin, 2012).

Although argumentation is often required in everyday life and in many fields is essential for academic success, most students find it difficult to construct strong arguments (Jonassen & Kim, 2010; Kuhn & Udell, 2003; Wolfe, Britt, & Butler, 2009). This may be due to insufficiently developed argument schema (Reznitskaya, Anderson,

McNurlen, Nguyen-Jahiel, Archodidou, & Kim, 2001; Wolfe et al., 2009). An argument schema is an abstract structure that represents one's argumentative knowledge as a relational network among its constituent concepts (Reznitskaya & Anderson, 2002). A well-developed argument schema aids argument construction and repair; it facilitates recognition, comprehension, retrieval, and organization of argument-related information and raises awareness of possible objections as well as holes in their own arguments or those of others (Anderson et al., 2001; Reznitskaya & Anderson, 2002). Schemas cue learners to gaps in their knowledge and direct their attention to relevant and critical information (Anderson, Spiro, & Anderson, 1978).

In view of their pivotal role in promoting learning and argumentation, it is important to investigate how to help students attain functional argument schemas. According to previous research, cognitive tools could be of help (Nesbit, Niu, & Liu, 2019; Pakdaman-Savoji, Nesbit, & Gajdamaschko, 2019).

1.1.3. Cognitive Tools for Visually Scaffolding Argumentation

Cognitive tools carry a variety of labels such as cognitive technologies (Pea, 1987), technologies of the mind (Salomon, Perkins, & Globerson, 1991), or mindtools (Jonassen, 1996). They are employed to engage, enhance, and extend a learner's cognitive abilities in the process of learning (Jonassen, 1992; Jonassen & Reeves, 1996; Kim & Reeves, 2007). Operating as intellectual partners, well-designed cognitive tools off-load lower level tasks so that more cognitive capacity is spared for higher-order thinking; beyond that, they engage learners in cognitive activities that they would not have been capable of otherwise (Lajoie, 1993). Taking on part of the cognitive load induced by information processing and modeling how experts learn, cognitive tools are especially beneficial for novice learners (Jonassen & Reeves, 1996; Salomon et al., 1991). Interacting with cognitive tools holds promise for equipping learners with more skills and capabilities for independent learning (Salomon et al., 1991).

An argument visualization tool (AVT) is an example of cognitive tool that scaffolds argument construction, analysis, and evaluation (Nesbit et al., 2019). Based on the argument schema theory, one's argumentative knowledge acquired through past experiences is stored in memory as schemas that affect how argument-related information is organized, retrieved, and processed (Reznitskaya & Anderson, 2002;

Reznitskaya, Anderson, Dong, Li, Kim, & Kim, 2008). AVTs scaffold the activation, strengthening, and enhancement of existing schemas or the development of a new one (Nussbaum & Schraw, 2007).

AVTs apply both verbal and visuospatial representations to unfold the underlying structures of text-based arguments (Dwyer, Hogan, & Stewart, 2012; Dwyer, Hogan, & Stewart, 2013; van Gelder, 2002; van Gelder, 2015). According to dual coding theory, the interconnections between the two cognitive systems provide additional retrieval routes and aid information processing (Nesbit & Adesope, 2006; Paivio, 1990). Having learners construct a verbal-visual framework of text promotes deep learning (Clark & Paivio, 1991). Furthermore, the presence of AVTs potentially reduces the cognitive load associated with analyzing and constructing relational structures of arguments (Hoffmann & Paglieri, 2011) and thus renders greater computational efficiency (Larkin & Simon, 1987; Robinson & Schraw, 1994).

AVTs have been used in different ways to support learning. For instance, some researchers asked students to read pre-made visual knowledge representations (e.g., Dwyer, Hogan, & Stewart, 2010; Dwyer et al., 2013; Liu & Nesbit, 2018) and some instructed students to complete or construct an argument map (e.g., Dwyer et al., 2012; Harrell, 2012; Niu, 2016; Nussbaum, 2008; Nussbaum & Schraw, 2007). Albeit effective, research has found that the advantages of studying with AVTs attenuate over time (Robinson & Schraw, 1994). Niu (2016) held that the effectiveness of AVTs depends on how they are used. Argument mapping seems to offer more robust benefits than studying a pre-constructed visualization. The latter permits easier encoding of relational information but likely sacrifices effortful cognitive processing that secures enduring learning (Liu & Nesbit, 2018).

1.1.4. Need for Cognition

Although learning tasks and conditions drive how students process study materials, one's cognitive motivation, usually referred to as *need for cognition*, also plays a role (Cacioppo, Petty, Feinstein, & Jarvis, 1996). Need for cognition is defined as one's intrinsic preference for cognitively challenging activities (Cacioppo & Petty, 1982). Students with high need for cognition show a tendency to make greater effort in information processing and knowledge construction (Cacioppo et al., 1996; Dai & Wang,

2007; Kardash & Noel, 2000) and perform better on argumentative tasks (Cacioppo, Petty, & Morris, 1983; Mongeau, 1989).

1.2. Research Purpose and Questions

As stated above, students benefit from retrieval practice and studying with AVTs. However, each strategy has manifested needs for optimization. My thesis set out to investigate if retrieval-based argument mapping that integrates the attributes of argument retrieval and construction could better advance retention of information, transfer of learning, and acquisition of argumentation skills. Moreover, it looked into the role of need for cognition in test-enhanced learning. Specifically, this research aims to address the following questions:

1. Are retrieval-based activities more effective than restudy in promoting long-term memory?
2. Can retrieval-based activities better promote transfer of learning than restudy?
3. How do argument-oriented and unstructured retrieval practice interventions differ in promoting learning and transfer?
4. How do individual differences in need for cognition influence the effects of retrieval-based argument mapping?

1.3. Experiment

A sample of 120 university students participated in the experiment consisting of two sessions: an initial learning session and a delayed posttest session. The participants were randomly divided into three treatment groups: a Restudy group, a Retrieval Practice group, and a Dialectical Map group. After participants completed a demographic questionnaire, a pretest on free recall ability, and an 18-item Need for Cognition scale, they read a text about wind power and engaged in treatment-specific activities. The Restudy group reread the text. The Retrieval Practice group completed two cycles of unstructured retrieval practice of the text. Following a free recall practice, they were asked to write any information they had not written the first time. The Dialectical Map group constructed an argument map in the absence of the text, with the aid of an online argument visualization tool called the Dialectical Map (DMap). They were instructed to build a dialectical map using the information from the text they had just read to argue if

the government should encourage the use of wind energy in British Columbia. Following that, they were asked to write any other information that had not been written in the map.

Participants returned within two weeks to complete the outcome tests, including a free recall test, a short-answer test, and an argument essay. The latter two were transfer measures that examined how participants applied what they learned from the wind power text to predict and explain various phenomena such as residents' protesting against a cell tower proposal and the functioning of tidal turbines. They were also required to write an essay arguing for or against the expansion of tidal energy around the world. The learning activities and tests were learner-paced.

1.4. Research Findings

The results indicated that argumentation-based retrieval practice is superior to restudy and standard free recall testing in promoting learning transfer and argumentation. The Dialectical Map group outperformed the other two groups in the short-answer transfer test, after controlling for free recall ability. Furthermore, retrieval-based argument mapping supports the acquisition of argumentation skills. Participants in the Dialectical Map group included more arguments and counterarguments as well as a greater number of warrants and rebuttals in their essays relative to those in the Restudy group and the Retrieval Practice group. In terms of recall, its effectiveness was not statistically detectable. Contrary to the bulk of previous research investigating test-enhanced learning, the Restudy group and the Retrieval Practice group did not differ in any of the outcome achievement measures. Another unexpected finding is that the effects of retrieval practice did not vary as a function of need for cognition.

This research is the first to investigate the combination of retrieval practice and argument construction as a strategy for transferable learning. It provides theoretical and practical implications for research and instruction. The findings cast new light on how the effects of retrieval practice are shaped by constraints placed on and functional requirements of the retrieved information.

Chapter 2.

Literature Review

2.1. Test-Enhanced Learning

2.1.1. Testing Effect

The conventional role of tests is to assess or measure knowledge. An increasing number of studies have suggested that tests can also be tools that facilitate and foster learning (e.g., Carpenter, 2009; Pan & Rickard, 2018; Pyc & Rawson, 2012). According to Roediger and Karpicke (2006), "[t]aking a test on material can have a greater positive effect on future retention of that material than spending an equivalent amount of time restudying the material, even when the performance on the test is far from perfect and no feedback is given on missing information", a phenomenon which is referred to as the *testing effect* (p. 181). To put it another way, retrieving information from memory, under various circumstances, leads to better recall of previously learned materials than repeated studying (Roediger & Butler, 2011). Retrieval practice that involves internally recalling what has been learned establishes the basis of the testing effect (Burdo & O'Dwyer, 2015). In this thesis, the terms testing and retrieval practice are used interchangeably.

A typical procedure for retrieval practice comprises an initial study session followed by a recall test (Endres & Renkl, 2015; Rowland, 2014). Roediger and Karpicke (2006) discussed two types of effects that a test may have on learning: direct and indirect effects. Direct effects are those that promote long-term retention attributed to the act of taking a test itself. Indirect effects of testing are those that benefit mediating processes that make subsequent studying and encoding more effective. My thesis primarily focuses on the direct effects of retrieval practice on learning.

2.1.2. Theoretical Accounts of the Testing Effect

A wealth of research has reported the superiority of retrieval practice over restudy, but the cognitive or neuroscientific mechanisms that would account for it are not

yet clear (Carpenter, 2009; Dunlovsky, Rawson, Marsh, Nathan, & Willingham, 2013; Endres & Renkl, 2015; Pyc & Rawson, 2010; Rawson, Vaughn, & Carpenter, 2015; Rowland, 2014). This section reviews the theories most commonly used to explain the benefits of retrieval practice. These theories are not mutually exclusive, and in some cases may jointly contribute to test-enhanced learning (Rowland, 2014).

2.1.2.1 Transfer-Appropriate Processing

One theory that underlies the testing effect highlights the concept of transfer-appropriate processing. It underscores the critical relation between initial and final test conditions (Morris et al., 1977). Specifically, the testing effect is attributed to the overlap in cognitive processing invoked by the practice and outcome tests. It posits that the degree of similarity between the initial test and the criterial test determines the magnitude of the testing effect (Rowland, 2014). Students' performance on the final test should be best when the types of questions in the final test and those used for initial practice coincide (Endres & Renkl, 2015; Roediger & Karpicke, 2006).

The transfer-appropriate processing perspective has received empirical support. For example, Duchastel and Nungester (1982) gave 125 high school students a brief history passage to read and then asked them to take either a multiple-choice test or a short-answer test. Each test comprised 24 questions same in content but set in different formats. The control group was invited to complete a study habit questionnaire as a filler task. Two weeks later, all students took a final retention test made up of 24 items selected from the initial tests, 12 from the multiple-choice test and 12 from the short-answer test. The results revealed a strong testing effect, which was format-dependent. In the delayed retention test, both test groups outperformed the control group. Compared with the short-answer test group, students who initially engaged in multiple-choice testing scored higher on the multiple-choice items. The two experimental groups did not differ in the case of short-answer questions.

Johnson and Mayer (2009) asked all participants to study a multimedia lesson on lightning formation, which was followed by either an additional study opportunity, a practice-retention test (i.e., "Please write down an explanation of how lightning works."), or a practice-transfer test (i.e., "What could you do to decrease the intensity of lightning?" and "Suppose you see clouds in the sky but no lightning, why not?"). One week later, the participants completed a posttest consisting of the retention question

used in the practice-retention test and four transfer questions including the two questions of the practice-transfer test along with two new questions. The results showed that there was a significant interaction between initial and final test types, such that participants in the practice-retention group did better in the delayed retention test and the delayed transfer test score was higher for the practice-transfer test group, as predicted by the transfer-appropriate processing hypothesis.

Despite the empirical evidence in support of the hypothesis that a good fit of the initial and final test conditions leads to better performance, contradictory findings exist (Rowland, 2014). For instance, Carpenter and DeLosh (2006) applied a mixed design in which the type of initial practice (i.e., restudying, recognition test, cued recall test, or free recall test) was manipulated within participants but the type of final test (i.e., recognition, cued recall, or free recall test) was manipulated between participants and provided empirical evidence inconsistent with the explanation of transfer-appropriate processing. They found that a match between the intervention and final tests did not yield best final test performance. Studying via the free recall practice test produced the highest scores, regardless of the type of final test given. Endres and Renkl (2015) employed a within-subject experimental design consisting of a learning phase (i.e., studying three expository texts dealing with different psychological topics), an intervention phase (i.e., restudy, free recall test, or short-answer test), and a delayed posttest phase (i.e., free recall and short-answer questions for each text). They did not observe that performance on the posttest was better when it had the same format as the initial test.

As shown above, the transfer-appropriate processing hypothesis has received mixed support in literature, indicating that the testing effect might not be fully explained by the overlap in cognitive processing triggered by initial and final tests (Carpenter & DeLosh, 2006).

2.1.2.2 Elaborative Retrieval Theories

Carpenter (2009) proposed an *elaborative retrieval hypothesis* to elucidate how retrieval practice functions to affect learning. When one is presented with a retrieval cue (e.g., education) and tries to search for a target in long-term memory (e.g., school), some information related to that cue (e.g., students, book, exam, teacher) could possibly be activated during the process of retrieval. That is, “education” may activate such words as “students”, “textbook”, “teacher”, and “exam” as a side effect of the search for the

target word “school”. As contended by Kornell and Vaughn (2016), the retrieval attempt activates not only the direct cue-target pair (i.e., education-school) but also the indirect cue-mediator-target connections (e.g., education-students-school, education-teacher-school, education-book-exam-school). The mediating or cue-related information contributes to an elaborative network and provides additional retrieval routes by which to access the target information. According to Carpenter (2009), this kind of elaborative processing is more likely to occur during the retrieval attempt than during restudy that typically involves learning what is currently presented. This might explain why retrieval practice is more effective in facilitating recall.

The elaborative retrieval hypothesis has obtained solid empirical support. For example, Carpenter (2009) conducted two experiments in which students learned cue-target pairs differing in cue strength through testing or restudying the intact word pairs prior to taking a final recall test over the target items. The results suggested significant benefits of retrieval practice on retention and highlighted the advantage of weak cues over strong cues. Although strong cues were effective in facilitating initial recall, they activated a narrower set of elaborative information that aids later retention. Take the strong cue-target pair *Toast-Bread* as an example. The strong retrieval cue may enable quick and easy access to the target item in memory but reduce the likelihood of activating more elaborative information that is beneficial for retention (Carpenter, 2009). This may account for why items recalled from strong cues cannot be retained as well as those recalled from weak cues over time.

The elaborative retrieval hypothesis advocates the importance of generating extra information during initial retrieval to aid recall of the target at a later time. However, it fails to specify the nature of information spontaneously activated during testing and how it contributes to later retention (Carpenter, 2011).

Building on Carpenter (2009)’s perspective, Pyc and Rawson (2010) proposed a *mediator effectiveness hypothesis*. They defined the sort of information (e.g., a word, phrase, or concept) that links a cue to a target as a mediator and assumed that “mediators generated during testing (versus restudy only) are more likely to be subsequently retrieved and decoded, increasing recall of target responses” (p. 335). Two components that might relate to the testing effect are mediator retrieval and mediator decoding. The former means the mediator generated during testing is more likely to be

remembered when prompted with the cue and the latter suggests activation of the mediator can prompt recall of the target.

To test this hypothesis, Pyc and Rawson (2010) initiated an investigation in which 118 participants were randomly assigned into 6 groups, defined by the factorial combination of the type of practice (test-restudy vs. restudy only) and the format of final test (C-cue only, CM-cue plus mediator, or CMR-cue plus mediator recall). Participants in the test-restudy group, after learning 48 Swahili-English word pairs (e.g., wingu-cloud), were asked to take a cued recall test that was immediately followed by restudy. Those in the restudy group were only required to restudy the word pairs after the initial study trial. In typical studies examining the testing effect, participants are just asked to recall the target in response to a cue. They are not required to think aloud mediators during learning (Carpenter, 2011). While in their study, all the participants, during the initial study and restudy phases, were asked to generate and report a keyword mediator for each word pair (e.g., “wing” for “wingu-cloud” because “wing” looks and sounds like the cue “wingu” and is semantically related to the target “cloud”). On the cued-recall posttest, participants in the C group were presented with the cue and asked to recall the target. Those in the CM group were shown not only the cue but also the keyword mediator they themselves had reported during the restudy phase to help them recall the target. In the CMR group, participants were given the cue and then asked to recall the keyword mediator they had generated during restudy before recalling the target. Final test performance of the CMR group showed that the test-restudy group recalled more mediators (51%) than the restudy group did (34%). Final test performance of the CM group revealed that those experiencing the test-restudy practice performed detectably better than those who restudied the intact word pairs. These results support the mediator effectiveness hypothesis that testing is more likely to induce mediator retrieval and decoding that facilitate recall at a later time.

Expanding on this research, Carpenter (2011) investigated how mediators play a role in normal circumstances in which students are not specifically instructed to generate mediators during learning. As most researchers did, she asked participants to study 16 cue-target pairs (e.g., *weapon: knife*) either through testing (*weapon:* _____) or through restudying (*weapon: knife*). In Experiment 1, all participants were given the same final recognition test that involved cues (e.g., *weapon*) and targets (e.g., *knife*), as well as new items that had never been presented during initial learning. Each new item was

either an unrelated word (e.g., *game*) or a semantic mediator that was strongly associated with one of the cues but semantically unrelated to its respective target (e.g., for the cue-target pair *weapon-knife*, the mediator *gun* is strongly related to *weapon* but weakly related to *knife,* according to the association norms; Nelson, McEvoy, & Schreiber, 1998). Participants were asked to indicate whether they had seen each item during the learning phase. The results confirmed her prediction that participants who learned the cue-target pairs through testing exhibited higher false alarm rates to the semantic mediators compared with those who learned the word pairs through restudying because semantic mediators were more likely to be activated during an attempt to recall a target from a cue. In Experiment 2, participants were given a final cued-recall test in which they were asked to recall the target (e.g., *knife*) in response to the same cue as before (e.g., *weapon:* _____) or from a new cue that was either related to the target (e.g., *ax:* _____) or a semantic mediator (e.g., *gun:* _____). The results showed that targets were better recalled from semantic mediator cues than from new cues related to targets for participants in the test condition. However, the effectiveness of semantic mediators was not as significant for participants in the restudy group. The same pattern of results was reported in Rawson et al. (2015). Such findings expand on previous accounts of the testing effect and demonstrate that the process of retrieving a target from a cue elicits mediators semantic in nature and the link between a semantic mediator and a target is more likely to be strengthened through testing relative to restudy, which notion is referred to as the *semantic mediator hypothesis* (Carpenter, 2011).

Another theory that fits in the elaborative retrieval framework is *gist trace processing* proposed by Bouwmeester and Verkoeijen (2011). This hypothesis is grounded on the fuzzy trace theory and proposed that the encoding of presented words leads to the formation of verbatim traces and gist traces. Verbatim traces are “item-specific traces that preserve the surface details of the stimulus”; gist traces are defined as “an abstraction of the property or properties that the studied words have in common, like the sense of meaning that can be derived from a list of words that are associatively related” (p. 33). Bouwmeester and Verkoeijen (2011) provided a Deese-Roediger-McDermott (DRM) list as an example, including *village, place, Amsterdam, houses, crowded, big, traffic,* and *life.* These words are found to associate with the non-

presented theme word *city*, which instantiates the gist trace for this DRM list. Students may develop both verbatim and gist traces while studying such a word list.

According to Bouwmeester and Verkoeijen (2011), taking a test after an initial study session strengthens the gist traces. When learners try to retrieve what they've learned, they rely on additional information (e.g., the word *city*) to reconstruct what is stored in their memory to facilitate retrieval. The activation of those non-presented items is assumed to interfere with verbatim processing as semantically related distractors. This assumption has been empirically supported by prior research investigating false recall (e.g., Brainerd, Payne, Wright, & Reyna, 2003; McDermott, 1996; Payne, Elie, Blackwell, & Neuschatz, 1996). In contrast, restudying reinforces the verbatim traces instead of gist traces that play an insignificant role in repeated studying. Furthermore, verbatim traces are assumed to be forgotten more rapidly as compared to gist traces. As a result, the testing effect tends to be more salient if the memory performance test is administered after a long retention interval when learners need to base their memory on gist traces that are strengthened through retrieval practice.

In summary, the elaborative retrieval hypothesis is a plausible mechanism underlying the testing effect, but it cannot account for all types or conditions of retrieval practice (Rawson et al., 2015). For instance, the effect of elaborative retrieval is larger during earlier stages of learning and may lessen with extended practice or overlearning that permits quick, direct retrieval of target information from the cue (Kole & Healy, 2013). Also, the elaborative retrieval hypothesis gives an insight into verbal learning that involves associative processing in particular (Kang, 2010; Rawson et al., 2015). For other types of learning activities, it may have less explanatory power (Kang, 2010).

2.1.2.3 Retrieval Effort Theories

Another class of theoretical explanations builds on the idea that the testing effect is a product of effortful retrieval. The magnitude of the testing effect corresponds to the level of difficulty or retrieval effort induced by initial testing (Roediger & Karpicke, 2006; Rowland, 2014). Key theories in this camp include the new theory of disuse and the retrieval effort hypothesis, an instantiation of the desirable difficulty framework.

The new theory of disuse (Bjork & Bjork, 1992) depicts the adaptive interplay of storage strength and retrieval strength in human memory. Storage strength measures

the degree to which an item is well learned, and retrieval strength refers to how easily the item can be accessed in memory at a given point of time. It is retrieval strength that determines the probability that an item can be recalled in the presence of a cue or set of cues, whereas storage strength has no direct effects on retrieval performance. Generally, higher retrieval strength creates the potential of successfully recalling an item from memory in response to a given cue. An important but seemingly counterintuitive assertion made by Bjork and Bjork is that there is a negative relation between retrieval strength and gains in storage strength. In other words, the lower the retrieval strength (i.e., more effortful retrieval practice), the greater the effect of retrieval on storage strength (i.e., learning). Thus, an attempt to create conditions that reduce retrieval strength is a worthwhile undertaking to promote long-term learning.

In 1994, Bjork proposed another relevant theoretical framework called *desirable difficulty*. It claims conditions that induce rapid increments in initial learning may inhibit long-term learning: Desirable difficulties may appear to hinder or slow down initial learning, but they are beneficial for establishing more enduring retention and transfer. Bjork and Bjork (2011) pointed out that "[d]esirable difficulties, versus the array of undesirable difficulties, are desirable because they trigger encoding and retrieval processes that support learning, comprehension, and remembering. If, however, the learner does not have the background knowledge or skills to respond to them successfully, they become undesirable difficulties" (p. 58). Interleaving instruction on separate topics, spacing practice, providing delayed feedback, and using tests are examples of creating desirable difficulties (Bjork & Bjork, 2011; Roediger & Karpicke, 2006).

Pyc and Rawson (2009) applied the desirable difficulty framework to retrieval practice and postulated that successful but difficult retrievals enhance memory more than successful but easy retrieval attempts. This is referred to as the *retrieval effort hypothesis*.

To test this hypothesis, Pyc and Rawson (2009) asked participants to study 70 Swahili-English word pairs and manipulated two variables that were assumed to influence the difficulty of successful retrieval during practice: interstimulus interval (ISI) and criterion level. Specifically, ISI was a between-subject manipulation, defined as the number of items between each next practice trial with a given item (6 vs. 34 intervening

items), and criterion level was a within-subject manipulation, suggesting 1, 3, 5, 6, 7, 8, or 10 times an item was correctly recalled before dropping from further practice. Participants' performance on the final cued-recall test supported their predictions that: 1) final test performance was greater after retrieval practice involving longer (i.e., more difficult correct retrieval) lags, and 2) the increasing number of times items were correctly recalled (i.e., criterion level), which suggested less difficulty of their next correct retrieval, predicted decreasing incremental benefit for final test performance.

In Experiment 2, first key press latency (i.e., the amount of time between presentation of the cue and the first key pressed by the participant when entering the answer) was recorded as a measure of retrieval difficulty. Results showed that first key press latencies were shorter for correct retrievals involving a shorter lag and first key press latencies decreased as the number of correct retrievals increased, which provided further evidence suggesting the credibility of the findings of Experiment 1 and confirming that retrieval difficulty did vary as a function of ISI and criterion level. Taken the results of the two experiments together, Pyc and Rawson (2009) validated the retrieval effort hypothesis that emphasizes the positive relationship between the difficulty of correct retrievals and subsequent memory.

Despite the plausibility of the retrieval effort hypothesis, it is not without limitation. For instance, this claim just focuses on the benefit of successful retrieval but ignores the role of unsuccessful retrieval attempts that also require mental effort (Kornell & Vaughn, 2016). In addition, it does not explicitly describe the causal mechanisms at play (Rowland, 2014). By mentioning that the effect of difficult retrievals on subsequent memory probably lies in their efficacy in retarding forgetting and/or enhancing encoding variability, Pyc and Rawson (2009) seemed to acknowledge that the retrieval effort hypothesis does not stand on its own but coexists with other theoretical accounts, for example, the elaborative retrieval hypothesis, to shed light on the nature of retrieval practice effects.

2.1.3. The Effects of Retrieval Practice on Learning

Over the past two decades, there has been a surge of interest in exploring the effects of retrieval practice on learning, including retention of studied information, transfer of learning, and argumentation. Abundant empirical evidence is available to

justify the testing effect. Adesope et al. (2017) meta-analyzed 272 independent effect sizes reported in 118 research papers. Results bolstered the overall effectiveness of retrieval practice (g = .61) relative to comparison learning conditions such as restudying defined as additional exposure to learning materials (g = .51) or no/filler activities (g = .93).

2.1.3.1 Retention of Information

As presented in Section 2.1.2, the theoretical accounts of the testing effect were mainly built upon studies looking into how testing promoted retention of word-pair associates. In this respect, its efficacy has been well founded (e.g., Carpenter, 2009; Carpenter, 2011; Carpenter & Yeung, 2017; Karpicke, Blunt, & Smith, 2016; Rawson, 2015).

There is also evidence for the effects of retrieval practice on enhancing retention of educational relevant materials other than word lists or paired associates. For example, Roediger and Karpicke (2006, Experiment 1) involved 120 undergraduate students in their study that compared the effects of restudy and free recall testing after learning a short prose passage (256 or 275 words in length). Repeated studying, relative to testing, led to better performance on the free recall test given after a 5-minute retention interval. The students in the testing group outperformed those in the restudy group in the delayed free recall test given after either a 2-day or a 1-week retention interval. The results indicated that testing promoted long-term retention. Along similar lines, Wong and Lim (2019) asked participants to study an argumentative text that was 470 words in length. The results of the delayed posttest showed that participants who experienced retrieval practice in the form of free recall included detectably more idea units in the argumentation vee diagram than those who repeatedly studied the text, after controlling for prior attitudes and GRE scores. The effect of retrieval practice on promoting long-term retention of complex learning materials was also supported by Hostetter et al. (2019) who asked students to recall stories they read a week ago. Besides verbal information, Carpenter and Kelly (2012) found that testing led to better recall of spatial information (i.e., locations of several objects).

Rowland (2014) meta-analyzed 159 research studies for the overall effects of retrieval practice on retention. The superiority of testing over restudy was corroborated (g = .50). It was found that a majority of research (81%) included in his meta-analysis

employed word lists or associate pairs to examine the benefits of retrieval practice. Only a small number of studies (14%) used prose passages as learning materials. The moderator analysis of the type of learning materials revealed that prose (*g* = .58) and paired associates (*g* = .59) yielded similar effect sizes. Another moderator of interest was stimulus interrelation defined as the relationships between studied information. The results detected no significant heterogeneity across types of learning materials containing semantically unrelated stimuli (e.g., unrelated word lists, *g* = .50), unstructured but semantically themed stimuli (e.g., categorized lists, *g* = .48), or conceptually integrated stimuli (e.g., prose, *g* = .58). These findings to some extent contradicted those discussed below.

Van Gog and Sweller (2015) reviewed 56 studies published between 2006 and 2015 to investigate if the complexity of learning materials might influence the effect of retrieval practice. In line with what was found by Kühn (1914) and Gate (1917), Van Gog and Sweller (2015) concluded that learning materials high in element interactivity (i.e., concepts cannot be learned in isolation but are related to one another) tended to decrease or eliminate the effect of testing. Here are some plausible explanations. Learning complex materials high in element interactivity might motivate students to spend more time and effort to restudy it, leading to increased effectiveness of repeated studying (van Gog & Sweller, 2015). It is possible that semantically themed materials are more meaningful than word lists or paired associates with a lower level of integration, such that they offer students a more reliable standard for restudy – namely the recognition that they understand what the passage is trying to say. Besides, it has been found that repeated study is more likely to induce item-specific processing that strengthens verbatim traces; whereas testing is more likely to engender relational processing that contributes to a stronger conceptual organization of information to be learned (Congleton & Rajaram, 2012). However, the characteristics of the learning materials may moderate the magnitude of the testing effect. Students might benefit less from retrieval practice if they are presented with learning materials that could effectively engage them in gist processing or induce the formation of gist traces during initial study or restudy (Bouwmeester & Verkoeijen, 2011; Delaney, Verkoeijen, & Spirgel, 2010; van Gog & Sweller, 2015).

de Jonge, Tabbers, and Rikers (2015, Experiment 1) asked 64 undergraduate students to study a text about black holes that included 1070 words and 60 sentences.

The text was presented to students one sentence at a time. After 15 minutes, students in the restudy group continued reading this text and those in the retrieval practice group were instructed to type in the missing information from each sentence presented to them. Half of each group received a final fill-in-the-blank test after a Sudoku Puzzle that took up 5 minutes and the other half returned for the same final test one week later. The results revealed that there was no statistically detectable difference between the two treatment groups at both retention intervals. This finding suggested that text features in terms of coherence or connectedness might moderate the magnitude of benefits of testing. To verify this assumption, de Jonge et al. (2015) conducted a second experiment in which students were invited to study the same black hole text that was, however, presented in a scrambled order to reduce the text coherence. It was found that the main effect of learning condition (restudy vs. testing) was not statistically detectable, but the two groups did show different rates of forgetting. The restudy group demonstrated a more significant decline in recall performance across the one-week interval than the testing group. Taken together, the results of the two experiments indicated the interaction between type of study materials and the testing effect, as alleged by Van Gog and Sweller (2015).

2.1.3.2 Beyond Retention

In recent years, researchers have been inquiring into the effects of retrieval practice on learning that demands more than recall of studied information. Karpicke and Blunt (2011) found that free recall by writing down everything one could remember in paragraph format was more effective than drawing up a concept map while reading the text in promoting university students' abilities to make inferences. The retrieval practice group even outdid the concept map group in the posttest that required students to construct concept maps in the absence of texts.

A recent meta-analysis (Pan & Rickard, 2018) claimed that testing yields robust transfer of learning (d = .40, p < .001). Butler (2010) found that repeated testing promoted transfer of concepts introduced in prose passages. It produced greater transfer to new inferential questions within or across (bat vs. aircraft) knowledge domains compared with repeated studying. The result concerning far transfer was replicated by van Eersel, Verkoeijen, Povilenaite, and Rikers (2016) who ascribed its superiority to focused exposure to key information critical for subsequent problem

solving. Wong, Ng, Tempel, and Lim (2019) randomly assigned 60 students into a restudy group engaging in four consecutive study sessions or a retrieval practice group following a study-free recall-study-free recall procedure. During the final test, students were asked to work on a test scenario that differed contextually from but was fundamentally similar to the studied scenario. The results revealed that the two groups performed equally (d = .01) on the analogical-problem-solving test that was administered in 5 minutes, but the advantage of retrieval practice over repeated studying (d = .81) emerged when the test was conducted one week later.

Following the same line of research, Hostetter et al. (2019) asked students to read two stories with identical schematic structures. Each story was followed by a group-specific activity: the restudy group reread the story; the retrieval practice group wrote as much of the story as they could think of; the copy group typed verbatim of the story. Students returned a week later for the posttest. They were required to complete a free recall of the stories they read, describe the similarities between the two stories, and generate solutions to two problems that were seemingly different but shared same underlying structures as the stories they studied. The results showed that students in the retrieval practice group recalled more of the stories than the restudy group. However, the practice of retrieving stories from memory did not help students identify more schematic similarities between the two scenarios or better solve the analogous problems compared with those who reread or copied the stories. Hostetter et al. (2019) argued that free recall testing did not effectively direct learners' attention to analogy-relevant details or critical pieces of information key to problem solving. To test this assumption, they conducted Experiment 2 to explore the effects of retrieval practice in the form of short-answer questions that explicitly prompted students to retrieve information critical for solving the problems. It was found that cued recall on the critical story information helped students identify the schematic similarities of the stories and apply the information to solve novel problems when they were told that there was a connection between the stories they read and the problems to be solved. However, these effects were found to be due to the increase in memory for the critical story information. These findings suggested that retrieval practice on its own is not powerful enough to promote transfer learning and it may be beneficial to pair it with techniques that advance transferable knowledge (Hostetter et al., 2019).

Wong and Lim (2019, Experiment 1) investigated the effect of retrieval practice on fostering integrative argumentation that involves critically evaluating and integrating arguments on both sides. They randomly assigned 59 undergraduate students into either a repeated study group or a retrieval practice group. The participants studied an argumentative text containing 470 words and 51 idea units. It contained arguments for and against daylight saving time that were listed in two columns. Participants in the retrieval practice group studied the text for 7 minutes, engaged in a free recall test for 7 minutes, restudied the text for 7 minutes followed by another free recall test that lasted 7 minutes. Those in the restudy group studied the text four times, each spanning 7 minutes. The posttest took place one week later, which started with a training session on integrative argumentation preceding a 20-minute final test consisting of two parts: (1) recalling and filling out a blank argumentation vee diagram with the arguments and counterarguments introduced in the argumentative text they studied a week ago, and (2) applying the argumentation skills they just learned from the training session to write an integrative conclusion, as measured by the number of integrative stratagems (weighing or design claims) used. The results showed that, even though retrieval practice augmented verbatim recall performance, the two groups did not differ detectably in the number of integrative stratagems reflected in participants' responses, after controlling for prior attitudes and GRE scores. A follow-up experiment (Wong & Lim, 2019, Experiment 2) found that even though retrieval practice by itself failed to better enhance students' integrative argumentation skills in comparison to repeated studying, its benefits became salient when it was paired with JOLs+, a metacomprehension monitoring intervention (e.g., "If you are asked to argue for/against Daylight Saving Time in Japan, how well do you think you can argue that one claim is weaker than another claim?") that prompted learners to attend to the intended learning outcome and oriented their attention to situation model level processing.

In sum, the effect of retrieval practice has been well established in the literature. However, its effects are found to be inconsistent in studies involving more complex study materials or learning outcomes (de Jonge et al., 2015; Eglington & Kang, 2018; Rawson, 2015; van Gog & Sweller, 2015). It is of significance to explore techniques that could empower retrieval-based activities to more reliably induce deeper engagement with ideas and meaningful learning. Argumentation has been reported to be an effective instructional activity that strengthens knowledge storage and retrieval (Means & Voss,

1996). Engaging students in argument-based inquiry conduces to deep processing, elaborative encoding, and metacognitive reflection (Dole & Sinatra, 1998). The following section inquires into the literature of argumentation to probe the potential of coupling retrieval practice with argument construction to reinforce test-enhanced learning.

2.2. Argumentation-Based Approach to Learning

2.2.1. Conceptualization of Argumentation

Argumentation is a broad and polysemic term. Though there is no single definition in the literature (Garcia-Mila & Andersen, 2007), argumentation has been often defined as a social and rational activity in which individuals put forward a series of propositions to justify for a specific point of view (i.e., argument) and react to the reasoning of its opponents (i.e., counterargument), with an intention to increase or decrease the tenability of a contentious standpoint (Nielsen, 2013; Nussbaum, 2011; Nussbaum & Sinatra, 2003; van Eemeren & Grootendorst, 2004; van Eemeren et al., 1996). Argumentation can be an individual activity (e.g., writing an argumentative essay) or can take place within a group of people (e.g., debate).

Kuhn (1992) proposed two types of arguments: rhetorical and dialogic arguments. The rhetorical arguments, also referred to as monological arguments, intend to prove or disprove something in disregard of alternative points of view. An example in case is that “a teacher provides a scientific explanation to a class or to a group of students with the intent of helping them to see it as reasonable” (Driver, Newton, & Osborne, 2000, p. 291). This type of argumentation is essentially one-sided (Kuhn, 1992). Dialogic arguments, also referred to as dialectical or multi-voiced arguments, take different perspectives into account and aim to come up with a resolution by persuading opponents to accept one’s claim or reaching a compromise between multiple points of view. Dialectical argumentation typically takes place within social groups but may also occur within individuals when, for example, one is trying to make a decision (Driver et al., 2000; Jonassen & Kim, 2010). Stein and Albro (2001) contended that children are able to understand and produce an argument by the age of 3. Their argumentation skills in defending their own standpoints develop as they grow. However, the skills in understanding and reacting to the ideas held by the opponents may not necessarily

develop with age. The quantity and quality of arguments increase along with experience of engaging in argumentative activities (Kuhn, Shaw, & Felton, 1997; Ogan-Bekiroglu & Eskin, 2012).

2.2.2. Models of Argumentation

According to Nussbaum (2011), the purposes of referring to models of argumentation could be analytical, normative, and/or descriptive. The purpose is analytical because models make clear the structure of arguments by breaking arguments down into components and presenting the relationship among those components. It is normative in that models can be used to evaluate the quality of arguments and the appropriateness of argumentation moves. Furthermore, models of argumentation can be used to psychologically describe or explain how people incline to argue. A specific model can but not necessarily serve all three purposes.

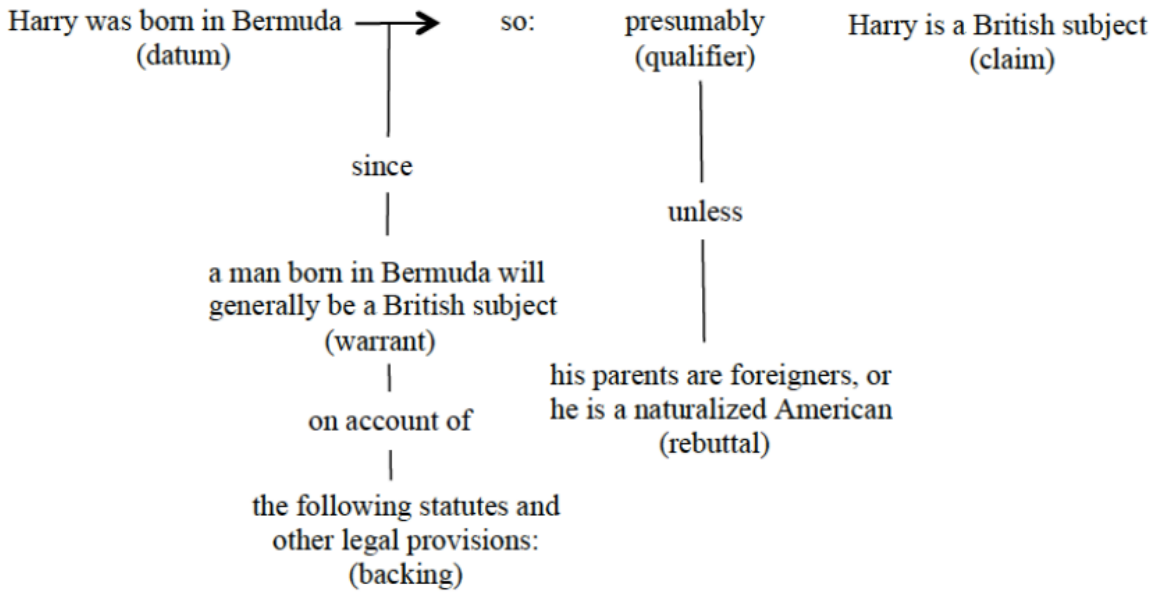
2.2.2.1 Toulmin Model

Toulmin's model of argumentation, developed in 1958, is influential in guiding the analysis, evaluation, and construction of arguments (van Eemeren et al., 1996). It has played a significant role in advancing the study of argumentation in education (Nussbaum, 2011). According to Stephen Toulmin, formal logic is neither a necessary nor a sufficient criterion for evaluating the soundness of argumentation. Given the complexity of argumentation in everyday life and academic disciplines, the evaluation criteria should refer to the nature of the problem at issue and allow for field-dependent and subject-related aspects (Toulmin, 1958; Toulmin, 2003; van Eemeren et al., 1996).

First and foremost, the Toulmin model is analytical. It brings attention to argumentation schemas and unfolds the structure of an argument (Nussbaum, 2011; van Eemeren et al., 1996). This model consists of six core elements: claim, data, warrant, backing, rebuttal, and qualifier. Argumentation starts from expressing an opinion (i.e., claim) on a contentious issue, defending it with facts or evidence in support of the claim (i.e., data), and then specifying the data-claim relationship or how the datum upholds the claim (i.e., warrant). Data tend to be explicitly stated, while warrants remain implicit. The first three steps constitute a simple model or schema of argumentation. The latter three are auxiliary components that need not be present in every argument. In cases where the authority of the warrant is not immediately accepted, a backing should

be included as additional support to the authoritativeness of the warrant. A rebuttal concerns possible objections to the claim. The presence of a rebuttal necessitates the use of a qualifier to suggest the strength of the claim or the degree of certainty to the conclusion. Figure 2.1, adapted from Toulmin (2003, p. 97), shows an example illuminating the six components involved in his model.

Figure 2.1. An Illustration of the Toulmin Model of Argumentation



The validity of argumentation is mainly determined by two factors: the form of argumentation and the authoritativeness of the warrant. The former is field-invariant. Argumentation that takes place in different fields can share the same procedure and structure. The latter, however, varies from one field to another. A warrant may refer to or be backed by, for instance, legal provisions, moral norms, or evolutionary principles to corroborate argumentation in different domains (van Eemeren et al., 1996).

Despite a dominant framework that guides argumentation research in education, the Toulmin model is not without defects. It is useful for analyzing arguments but has less power in evaluating the strength and quality of students' arguments that are tied to domain-specific standards (Nussbaum, 2011). This model does not account for how people argue and therefore contributes little to the construction of psychological models of how students generate and process arguments and how argumentation promotes learning (Nussbaum, 2011; Stein & Albro, 2001). van Eemeren et al. (1996) pointed out that some aspects of the Toulmin model are not rigorously defined. For instance, it's

difficult to identify warrants (Warren, 2010). The explicit-implicit distinction is not a robust criterion for distinguishing data from warrants (van Eemeren et al., 1996). The difference between warrants and backings is equivocal as well (Hample, 1992). Furthermore, the dialectical features of argumentation are underdeveloped (Andriessen, 2006; Nielsen, 2013). The Toulmin model understated the importance of developing and integrating counterarguments (Nussbaum, 2008). Toulmin (1958) loosely defined rebuttals as conditions of exception that can be used to rebut a warranted conclusion, a warrant's applicability or authority (Erduran, 2007). But the relationship among the three kinds of rebuttals is not elaborated; neither are their effects on the quality of arguments (Verheij, 2005). Given such concerns, researchers have proposed various other frameworks to assist the analysis, evaluation, and construction of arguments. One of the most systematic schemes is Walton's (Nussbaum, 2011).

2.2.2.2 Walton's Framework

Douglas Walton brought up the concept of argument as dialogue involving an interchange and interaction of arguments from two or more parties to achieve a collective goal (Walton, 1989). The dialogue theory underscores a dialectical approach to arguing and proposed an abstract, normative model that directs attention to not only the content but also the quality of students' arguments (Nussbaum, 2011; Walton, 1989). Driven by specific goals, there are different types of argument dialogues serving persuasion, inquiry, discovery, negotiation, information-seeking, deliberation, and eristic purposes (Walton, 2013). Each model of dialogue consists of three main stages: the opening stage, the argumentation stage, and the closing stage. For example, persuasion dialogue results from a conflict of opinions that needs to be resolved. The burden of persuasion (i.e., requirement of producing arguments strong enough to convince the other party of the acceptability of its proposition) is set at the opening stage and applies over the entire dialogue. During the argumentation stage, each party produces a series of arguments to support the thesis it commits to and attacks or critically questions the arguments put forward by the others. The dialogue reaches the closing stage when one party has met the burden of persuasion, marking the resolution of conflict (Walton, Atkinson, Bench-Capon, Wyner, & Cartwright, 2010). In short, persuasion dialogue takes place with the goal to "reveal the strongest arguments on both sides by pitting one against the other to resolve the initial conflict posed at the opening stage" (Walton, 2013, p. 9). Unlike persuasion dialogue that is adversarial, deliberation dialogue arises from a

request for a rational choice about which course of action should be taken to tackle a problem affecting all of the parties involved. Deliberation can be a collaborative process or unfolds in a solitary, internal dialogue in which the pros and cons of a possible solution is critically examined (Walton et al., 2010). Those engaged with the deliberation dialogue do not bear the burden of persuasion. There is no need for them to champion a particular course of action. Instead, they can vote for the proposal put forward by some other party after evaluating the strengths and weaknesses of each proposal in light of the goals and circumstances related to decision-making (Walton, 2013; Walton et al., 2010). Despite attempts to achieve different goals, an argument may shift from one type of dialogue to another as it progresses to strengthen the chain of argumentation (Walton, 2013).

In Walton's framework, dialogue type operates at the macrolevel as a general context in which argumentation takes place; argumentation schemes outline the microstructure of argumentation (Nussbaum, 2011). An argumentation scheme represents a stereotypical pattern of reasoning that is commonly used in such contexts as conversational, legal, or scientific argumentation. Each scheme reflects a specific type of argument (e.g., argument from expert opinion, argument from analogy, or argument from example), accompanied by a corresponding set of critical questions (Walton, 2013; Walton & Macagno, 2015). The arguments are subject to defeat as new information or evidence that can potentially refute them comes in (Walton, 2013).

Walton has identified several dozen argumentation schemes over the past two decades. One of the most frequently used schemes is argument from expert opinion. Below lists the basic logic of this scheme, including two premises and a conclusion (Walton, Reed, & Macagno, 2008, p. 14):

> **Major Premise:** Source *E* is an expert in subject domain *S* containing proposition *A*.
> **Minor Premise:** *E* asserts that proposition *A* is true.
> **Conclusion:** *A* is true.

Given one's natural tendency to trust and accept what experts contend, an argument appealing to an expert's opinion seems to be plausible. However, experts are not always right. It is necessary to evaluate the tenability of the argument by asking the critical questions associated with each argumentation scheme. In this example, the respondent could ask if *A* is consistent with what other experts believe, whether *E*'s

assertion is evidence-based, and so forth. If such questions are not satisfactorily answered, refutations would surface and the presumption that the conclusion is true would be undermined. In other words, the argument is defeated. Defeasibility of arguments is a central concept underlying Walton's argumentation schemes as "it provides a foundation for warranting arguments" (Nussbaum, 2011, p. 87).

The Toulmin model and its alternatives like Walton's framework have been frequently applied to support the teaching and learning of argumentation skills. Such models shed light on what components should be included in an argument and how to establish the strength of one's arguments. Learning to argue helps students attain a skill that plays an essential role in everyday life and professional contexts (Asterhan & Schwarz, 2016). On top of that, the process of arguing advances learning.

2.2.3. Argumentation and Learning

Research has found that an attempt to produce reasoned arguments for or against one's own and others' viewpoints promotes learning (Andriessen, 2006; Eskin & Ogan-Bekiroglu, 2013). Incorporating argumentation as an instructional activity fosters scientific thinking and the construction of a situation model that facilitates inference generation, problem solving, and deep learning (Jonassen & Kim, 2010; Kintsch, 1993; Means & Voss, 1996). Hogan and Fisherkeller (2000) pointed out that engaging in argumentation is conducive to reinforcement of connections of ideas within a learner's cognitive framework. It facilitates construction, reconstruction, and use of conceptual knowledge (Ogan-Bekiroglu & Eskin, 2012). Furthermore, argumentation entails a reasoning process that nurtures critical thinking (Ogan-Bekiroglu & Eskin, 2012; Veerman, Andriessen, & Kanselaar, 2002). According to Dole and Sinatra (1998), argumentation in which students "think deeply about the arguments and counterarguments related to the message" induces high engagement featured as "deep processing, elaborative strategy use and significant metacognitive reflection" (p. 121). Such cognitive processing scaffolds conceptual change (Andriessen, 2006; Baker, 2003; Eskin & Ogan-Bekiroglu, 2013).

Asterhan and Schwarz (2007) conducted two studies investigating the effects of argumentation-eliciting interventions on students' learning of evolutionary theory. In Study 1, 76 undergraduate students were randomly paired up to collaboratively answer

questions on evolution. Half of the dyads were prompted to engage in an argumentative discussion; those in the control group were merely asked to collaborate. Study 2 involved 42 students seated with a confederate who read the instructions and questions from a booklet but reacted neutrally to the participants' answers. The participants in the experimental group were prompted to discuss and critically evaluate their own answers and the confederate's. Those in the control group read aloud their solutions to each other without discussing them further. The results of both experiments revealed that dialectical argumentation advanced understanding of evolutionary concepts. The experimental groups showed greater learning gains than the control groups. These learning gains attained during the intervention phases sustained for those engaging in argumentation, as measured by the delayed posttests a week later.

According to Ogan-Bekiroglu and Eskin (2012), students' prior knowledge about the concepts to be learned affects their engagement in scientific argumentation. It was found that those with higher prior knowledge contributed more to argumentative activities, quantitatively and qualitatively.

2.2.4. Schema Theory

Although argumentation is essential to any discipline, many students have difficulty grasping the fundamentals of argumentation or constructing strong arguments (Jonassen & Kim, 2010; Kuhn & Udell, 2003; Reznitskaya et al., 2001; Wolfe et al., 2009). This phenomenon applies to students from any grade level, including college students (Nussbaum & Schraw, 2007; Reznitskaya, Anderson, & Kuo, 2007; Wolfe & Britt, 2008). Kuhn (1991) proposed five essential components that define a strong argument, consisting of supportive theory, evidence, alternative theory, counterarguments, and rebuttal. According to Jonassen and Kim (2010), the most common shortcoming in argumentation is that students provide reasons and evidence to support their own claim but ignore others' points of view. To put it another way, students are generally inept at producing counterarguments and rebuttals, which, however, establish effective argumentation (Wolfe et al., 2009). Their weakness in argumentation may reflect a lack of knowledge on argumentation structure or a defective argument schema stored in memory (Reznitskaya, et al., 2001; Wolfe et al., 2009).

According to Wolfe et al. (2009), argumentative writing “requires the engagement and coordination of several cognitive processes such as retrieving a schema and encoding information from sources” (p. 184). In 1932, Bartlett introduced the term *schema* as an unconscious mental structure that represents generic knowledge gained from past experiences. A schema comprises a network of interrelations among its constituent units, each of which is a schema as well (Rumelhart & Norman, 1976). Rumelhart (1980) defined schemas as building blocks of cognition, on the foundation of which information processing takes place. Schemas represent and are involved in learning both conceptual and procedural knowledge (Nesbit et al., 2019; Rumelhart, 1980).

A schema is not static but modifiable (Anderson et al., 1978; Bartlett, 1932). Each schema is composed of semi-fixed, structural elements and variable elements functioning as slots or placeholders into which new information can fit (Nesbit et al., 2019). Rumelhart and Norman (1976) proposed three modes of learning in a schema-based system: accretion, tuning, and reconstructing. Accretion involves the accumulation of knowledge (e.g., fact learning), that is, adding new data structures to the data base of knowledge without altering its organizational structure. The assumption underlying accretion is that the existing schemas are adequate to account for incoming information. New input is easily assimilated when the information aligns well with the previously available schemas. When discrepancy exists, learning may be induced by tuning or reconstructing. Tuning denotes an evolution of existing schemas. Changes to the relational structure of a schema are minor. Reconstructing yields new schemas primarily through patterned generation, that is, patterning a new schema on a preexisting one with appropriate modification (e.g., creating the schema for “rhombus” by modifying the schema for “parallelogram”). Patterned generation of schema underlies the use of analogies, metaphors, or models as teaching devices.

Schema theory casts light on how knowledge is organized, encoded, and retrieved in the process of learning and provides implications for how to help students learn. Schemas contribute to the construction of cognitive representations by integrating new information into existing mental structure and serve as retrieval mechanisms for subsequent learning (Anderson et al., 1978; Reznitskaya & Anderson, 2002). Schema construction involves amalgamating and consolidating previously scattered schemas

that will then be activated simultaneously in a single chunk, by which means less cognitive load is imposed on working memory (Nesbit et al., 2019).

Remembering bears on schemas in that people tend to retain the interpretations or the gist of a text instead of the text itself. Recollection is goal-oriented, involving information seeking and information interpretation. The search is not random but guided by schemas that map out a search path through memory and account for the memorial fragments stored in memory (Rumelhart, 1980). In a similar fashion, comprehension involves a process of identifying and verifying a configuration of schemas to make sense of the text or situation to be understood (Rumelhart, 1980). Different schemas may induce disparate interpretations of the same situation (Rumelhart & Norman, 1976). Furthermore, schemas allow room for inferences (Reznitskaya & Anderson, 2002). Exposed to fragmentary or incomplete information, one's cognitive structures enable good guesses about those unuttered aspects and drive how the incoming information is interpreted to match the expectations (Reed, 1993).

In the parlance of Anderson et al. (1978), schemas function as ideational scaffolding. Information that fits slots in the schema can be readily remembered and learned, while information that does not is likely to be underrated or ignored. Activating schemas stored in long-term memory facilitates the integration of prior and new knowledge for meaningful learning. Schemas imply what are important to learn and direct learners' attention to seemingly relevant and significant information that, as a result, is better learned (Anderson et al., 1978). The allocation of attention and cognitive resources is schema-driven (West, Farmer, & Wolff, 1991).

2.2.5. Argument Schema

In the context of argumentation, an argument schema refers to an abstract structure that represents one's argumentative knowledge, which comprises "the rhetorical structure and the inferential rules of reasoning, as well as other cognitive and social practices appropriate for argumentation" (Reznitskaya & Anderson, 2002, p. 321).

A well-developed argument schema facilitates learning and acquisition of argumentative knowledge by (a) pointing one's attention to argument-related information; (b) scaffolding comprehension and retrieval of argument-relevant information; (c)

empowering one to effectively organize argument-related information; (d) supporting argument construction and repair; (e) raising awareness of possible objections; and (f) providing the basis for identifying holes in one's own arguments and those of others (Anderson et al., 2001; Reznitskaya & Anderson, 2002). The activation of an argument schema encourages learners to critically interact with text rather than mere encoding of information (Reznitskaya & Anderson, 2002). An argument schema is abstract, so it can be transferred to varied contexts (Anderson et al., 2001; Reznitskaya et al., 2008).

Given its vital role in promoting learning and argumentation, it is imperative to figure out how to help students acquire and develop a well-rounded argument schema. Collaborative reasoning has been reported to be an effective pedagogical strategy (Anderson et al., 2001; Reznitskaya & Anderson, 2002; Reznitskaya et al., 2008). In a small group, students discuss a controversial issue derived from the stories they read. Open participation is encouraged, in which students decide what to discuss and when to talk, without being nominated or dominated by the teacher. Collaborative reasoning fosters dialogical thinking and provides a context where students take a position on the issue, support it with reasons and evidence, and argue against alternative perspectives with rebuttals. Reznitskaya et al. (2008) summarized the results of four studies that employed the same posttest-only, quasi-experimental design and reported that elementary school students who experienced collaborative reasoning tended to present more arguments, counterarguments, and rebuttals in their essays than those who did not. Importantly, they found that argumentative knowledge acquired through collaborative reasoning transferred to argumentative writing tasks performed individually. Despite its effectiveness, collaborative reasoning calls for social interaction and genuine dialogues among a group of people. In many cases, students learn alone. It is necessary to devise instructional interventions that support the development of a well-rounded argument schema through individual or independent learning. Cognitive tools have been found to serve the purpose (Nesbit et al., 2019; Pakdaman-Savoji et al., 2019).

2.3. Cognitive Tools

A variety of labels have been used to signify the term of cognitive tools, such as cognitive technologies (Pea, 1987), technologies of the mind (Salomon et al., 1991), or mindtools (Jonassen, 1996). Despite different names, the shared connotation is cognitive tools as technologies that engage, enhance, and extend one's cognitive

powers in the process of learning (Jonassen, 1992; Jonassen & Reeves, 1996; Kim & Reeves, 2007). Cognitive tools can be either tangible or intangible. Examples include calculators, written language, mathematical concepts, semantic networks, just to name a few. Due to advances in computer technology and its increasing impact on human learning, recent discussions pertaining to cognitive tools focus on computers or computer-based instruments (Jonassen & Reeves, 1996; Kim & Reeves, 2007).

Lajoie (1993) recapitulated a range of functions that cognitive tools serve, which are not mutually exclusive. Specifically, cognitive tools (a) scaffold cognitive (e.g., memory) and metacognitive (e.g., monitoring) processes; (b) off-load lower level cognitive tasks so that more cognitive capacity is spared for higher-order thinking; (c) engage learners in cognitive activities that they would not have been capable of otherwise; and (d) provide a context for hypothesis generating and testing. Kim and Reeves (2007) ascribed the efficacies of cognitive tools to distributed cognition. The act of drawing a diagram on the paper, for example, represents both symbolic and physical distributions of cognition. The symbolic and/or physical tools tend to change the nature of cognitive processing, reflect outcomes of thinking, and eventually transform one's mental structure and processes. Cognitive tools are meant to induce knowledge construction rather than knowledge reproduction or effortless learning. They intend to uncover and amplify one's thinking, activate complex learning strategies, and engage students in critical thinking and deep learning (Jonassen & Reeves, 1996).

Informed by Salomon et al. (1991), there are two kinds of effects that cognitive tools might have on human intellectual ability and performance: effects *with* and *of* cognitive tools. The former denotes an intellectual partnership. Cognitive tools take on part of the cognitive load induced by information processing (Jonassen & Reeves, 1996) and are beneficial for beginning learners in particular (Salomon et al., 1991). In partnership with cognitive tools, novices might be able to complete the same task faster and with less effort. Learning with cognitive tools requires mindful engagement so as to take full advantage of their affordances (Jonassen & Reeves, 1996; Kim & Reeves, 2007; Salomon et al., 1991).

The effects of a cognitive tool are what happens once the learner is working independently of the tool, specifically, the changes in one's cognitive capacities (i.e., the cognitive residue) that result from interacting with the tool. Cognitive residues hold

promise for equipping the learner with more skills and capabilities for independent learning (Salomon et al., 1991). Below is an example that illustrates the distinction between effects with and of cognitive tools:

> On the one hand, students might write better while writing with [an intelligent word processor]; on the other hand, writing with such an intelligent word processor might teach students principles about the craft of writing that they could apply widely when writing with only a simple word processor (Salomon et al., 1991, p. 3).

According to Vygotsky (1978), cognitive tools play a role in advancing learners' cognitive abilities and performance. Piotr Gal'perin and his colleagues operationalized Vygotsky's ideas and further specified that such changes in cognitive functioning were contingent on the characteristics of cognitive tools that learners interact with in the course of instruction, such as criteria, models, and schemas (Arievitch & Stetsenko, 2000). In the context of argumentation, argument visualization tools (AVTs) have been reported to be effective cognitive tools that are instrumental in scaffolding argument analysis, construction, and evaluation (Nesbit et al., 2019).

2.3.1. Argument Visualization Tools (AVTs)

Argument visualization tools (AVTs) apply diagrammatic techniques to visually display arguments. One example of AVTs is an argument map that in its typical form is a box-and-arrow diagram with boxes presenting propositions and arrows (and semantic cues) suggesting inferential relationships (Dwyer et al., 2013; van Gelder, 2002; van Gelder, 2015). According to McCrudden and Rapp (2017), well-designed visualization tools support the selection and organization of critical information to be learned, activate prior knowledge for integrative inferences, and increase processing efficiency.

2.3.1.1 Theoretical Underpinnings of the Effects of AVTs

Argument Schema Theory

The argument schema theory buttresses the use of AVTs in education. As stated in Section 2.2.5, argumentative knowledge is organized and stored in memory symbolically as schemas (Reznitskaya et al., 2008). Argument schemas are acquired and improved though dialogic interaction or collaborative reasoning; well-developed argument schemas facilitate the processing, retrieval, and generation of argument-

relevant information (Reznitskaya & Anderson, 2002; Reznitskaya et al., 2008). According to Nussbaum and Schraw (2007), the presence of AVTs could potentially help students “activate, strengthen, and refine their existing schemas or develop new ones” (p. 65). Well-designed AVTs simultaneously uncover the criteria of a good argument and facilitate the mapping or planning processes (Nussbaum & Schraw, 2007). They effectively prompt students to interact with the text in a critical way (Reznitskaya & Anderson, 2002).

Dual Coding Theory

The effectiveness of graphic displays hinges on how much cognitive processing is required to interpret and integrate information (Vekiri, 2002). AVTs employ both verbal and visuospatial modalities to represent arguments (Dwyer et al., 2012). According to Paivio’s dual coding theory (Paivio, 1986), there are two distinct but interconnected mental systems specialized for storing and processing logogens (e.g., verbal information) and imagens (e.g., visual objects), respectively. It is postulated that three levels of processing operate within these two cognitive systems: representational processing, referential processing, and associative processing. Representational processing refers to the activation of logogens directly triggered by verbal stimuli or the activation of imagens directly triggered by visual stimuli. Referential processing involves both verbal and visual systems. The activation of logogens in the verbal system ignites the activation of imagens in the visual system, or the other way around. Differently, associative processing is an intra-system activity. Specifically, verbal cues may activate related information in the verbal system and visual stimuli may trigger related representations in the visual system. For instance, the word “shark” may potentiate recall of such words as “dolphin” and “whale” and an image of a “moon” may evoke a repertoire of related images like “star” and “sun” stored in one’s visual system.

A learning task may engage students in any or all of the three levels of processing (Paivio, 1986). AVTs capitalize on visual representations for verbal associative structures to facilitate encoding, assimilation, and recall of arguments (Dwyer et al., 2010; Dwyer et al., 2013). The interconnections between the two cognitive systems provide additional retrieval paths that aid long-term retention. An attempt to construct verbal and visuospatial associative structures of text fosters deep learning (Clark & Paivio, 1991; McCrudden, McCormick, & McTigue, 2011).

Computational Efficiency through Visual Argument

The visual argument hypothesis holds that visual displays are more effective than text in communicating information. They are less likely to overload one's working memory as graphic representations convey information through both their individual elements and the spatial arrangement of those elements (Vekiri, 2002; Waller, 1981). This phenomenon is also known as *perceptual advancement* proposed by Larkin and Simon (1987) who attributed the strengths of visual representations to computational efficiency.

Although informationally equivalent, diagrammatic and sentential representations render different computational efficiency pertaining to the ease and speediness of inference as they "differ in their capabilities for recognizing patterns, in the inferences they can carry out directly, and in their control strategies (in particular, the control of search)" (Larkin & Simon, 1987, p. 65). Diagrammatic displays hold appeal since they are believed to scaffold information search, recognition, and inference. In a diagram, information is indexed by location with relevant elements placed close to each other. In this way, information search could be more efficient than searching linearly down the sentential structures. Furthermore, diagrammatic representations typically make explicit the information that is implicit in sentential representations that embody sequential propositions. Learners can therefore draw inferences through visual argument and acquire the intended messages without substantial computation (Larkin & Simon, 1987; Robinson & Schraw, 1994).

Visual representations of text-based information have been found to promote knowledge acquisition (Keller, Gerjects, Scheiter, & Garsoffky, 2006; Robinson & Kiewra, 1995; Robinson & Schraw, 1994). Dwyer et al. (2010, 2013) argued that learning text-based arguments is cognitively demanding. It requires a high degree of attention switching to distinguish and link statements that support or refute claims scattered throughout texts. Displaying information in the form of an argument map creates encoding environments that reduce the level of attention switching and scaffold the construction of arguments (Dwyer et al., 2010; Dwyer et al., 2012). The visualization of arguments facilitates mental modeling of relational structures of arguments and therefore reduces the extraneous cognitive load associated with deciphering and analyzing text-based arguments (Hoffmann & Paglieri, 2011).

It's been found that the advantages of learning with a visual display are salient in immediate tests when the acquired knowledge is still active in working memory. However, its effectiveness significantly decreases after a delay (Robinson & Schraw, 1994). One possible explanation is that learning with visual displays requires little computation but likely sacrifices encoding effort that contributes to the durability of the memory traces. This warrants an investigation of how to induce effortful encoding and retrieval to enhance the effects of learning with AVTs.

2.3.1.2 The Effects of AVTs on Learning

There is abundant research investigating the effects of AVTs on promoting learning. Liu and Nesbit (2018) recruited 120 participants who demonstrated misconceptions about the motion of objects to compare the effects of studying a refutational map, a refutational text, and a non-refutational text on conceptual change. A refutational map is a type of argument map that explicitly presents both scientific concepts and misconceptions. The results showed that studying the refutational map advanced memory and knowledge transfer. The refutational map group outperformed the other two groups on a free recall test. Participants who studied the refutational map performed detectably better than the non-refutational text group on a short-answer transfer test. Dwyer et al. (2010) compared the effects of learning with a text or an argument map on recall and comprehension of arguments that were tested immediately after a 10-minute study session. They found that the participants who read argument maps, either colour or black-and-white maps, outperformed those who studied a standard text in the fill-in-the-blank cued-recall test. The argument structures including 30 propositions induced better recall performance than those with 50 propositions, no matter how arguments were presented. However, reading argument maps did not contribute to better understanding of the relationships among propositions. This pattern of results was echoed in their later investigations (Dwyer et al., 2013, Experiment 1 and Experiment 2). Furthermore, Dwyer et al. (2013) found that the superiority of studying argument maps over standard text does not hinge on topics studied. However, its beneficial effects on the immediate recall test were not transferred to the delayed recall test that took place one week after the study session.

In addition to studying pre-made visual representations, the effects of constructing an argument map on learning have been extensively examined. Dwyer et al.

(2012) conducted an 8-week study with 74 first year psychology students who were randomly assigned to either an experimental group in which critical thinking was taught through argument mapping or a control group receiving no critical thinking interventions. Students who practiced constructing argument maps showed a detectably larger increase in critical thinking ability as measured by the Halpern Critical Thinking Assessment than those in the control group. van Gelder (2015) did a meta-analytic review and obtained a large effect size favoring argument mapping-based instruction relative to other forms of critical thinking instruction or just being at college. A strong correlation between the intensity of argument mapping activities and the amount of gain in critical thinking was found.

Harrell (2012) found that students who were taught and directed to engage in argument mapping activities during a semester-long introductory philosophy course showed statistically greater gains in argument analysis skills (i.e., identifying the key components of an argument and their relationships) than those who were not. In Barstow, Fazio, Lippman, Falakmasir, Schunn, and Ashley (2017)'s study, students enrolled in a psychology research methods course were assigned into three treatment groups, receiving either no argument diagramming support, diagramming support with a domain-general framework, or diagramming support with a domain-specific framework. The results revealed that both types of diagramming activity induced higher quality writing. Students who engaged in argument diagramming activities included more relevant citations and evidence opposing the hypotheses in their research introductions than those in the comparison group. The domain-specific framework elicited more statements pertaining to the validity of both supporting and opposing citations.

Nussbaum and Schraw (2007) compared the effects of argument mapping and criteria instruction on argumentative writing. The argument map used in this study consisted of ovals where students could fill in an argument with supporting reasons, a counterargument with supporting reasons, and a final conclusion placed at the top of the map. Students in the criteria instruction group were instructed what defines a good argument. The results showed that both interventions resulted in essays with more counterarguments in comparison to the control group. The argument mapping group included more rebuttals in their essays but did not do as well as the criteria instruction group in integrating arguments and counterarguments indicated by synthesizing a compromise or creative solution or weighing advantages against disadvantages.

Following the same line of research, Nussbaum (2008) devised an argument visualization tool called argumentation vee diagram (AVD) and conducted a 4-week experiment investigating its effectiveness in supporting argument-counterargument integration. Different from the argument map tested in Nussbaum and Schraw (2007), the AVDs place the conclusion box at the bottom of the diagram and include two critical questions (i.e., “Which side is stronger, and why?” and “Is there a compromise or creative solution?”) to cue the integration strategies. Also, the AVDs allow students to draw an arrow between an argument and a counterargument opposing each other to develop refutations. Results showed that learning with AVDs enhanced argument-counterargument integration but the effect disappeared when students were asked to write on a new topic without the aid of AVDs.

Despite being a useful argument visualization tool, the AVD is not without defects. For example, there is no place for warrants. Learners are not prompted to evaluate the strengths of arguments, a step crucial for synthesizing the conclusion. In addition, the substructure of arguments in which evidence can either support or oppose a reason is not explicitly represented. To scaffold more sophisticated argumentation, Niu (2016) designed an interactive web-based argument visualization tool called the Dialectical Map (DMap) where students can type in pro reasons (i.e., arguments) and con reasons (i.e., counterarguments) relevant to the claim displayed at the top of the map. They are prompted to provide evidence for or against each reason and come up with warrants explaining how they relate to each other. Each reason (pro or con) can be rated according to its significance and reordered to link directly opposing arguments. The conclusion box is placed at the bottom of the map (see Section 3.3.2 for further details). Niu (2016) randomly assigned 125 university students into one of the three groups: (a) the DMap group who received argumentation training followed by studying text with the DMap, (b) the Argue group who received argumentation training followed by studying the text without the DMap, or (c) the Control group who studied the text without argumentation training or DMap. It was found that the DMap group had a stronger tendency to project argumentative thinking into non-argumentative writing tasks. For example, when writing summaries the DMap learners were more likely to apply an introduction-body-conclusion structure that is similar to the organization of the DMap tool. They tended to include more argument markers (e.g., because and however) in

their summaries and fewer neutral ideas (i.e., ideas that contributed to neither side of argument).

Niu (2016) maintained that the effectiveness of AVTs is dependent upon how they are used. Engaging students in effortful information encoding and retrieval while interacting with a visual display is beneficial for enhancing long-term learning (McCrudden & Rapp, 2017). As evidenced by the studies discussed above, there are mixed results pertaining to the effects of reading pre-made argument maps on learning. When students actively create or complete an argument diagram, its efficacy is more robust. According to Easterday, Aleven, and Scheines (2007), argument mapping engages students in three cognitive operations comprising comprehension, construction, and interpretation; whereas studying a given argument map just requires learners to interpret the diagram. Even though the latter allows easier encoding of relational information, it may deprive learners of effortful information processing that contributes to durable learning (Liu & Nesbit, 2018). The beneficial effects of constructing an argument map can be strengthened if the more difficult aspects of the task are scaffolded by a cognitive tool like the DMap which cues construction of counterarguments, provision of warrants, estimating the strength of arguments, and synthesizing arguments and counterarguments to form a conclusion (Niu, 2016).

2.4. Need for Cognition

As discussed above, free recall tests and argument-construction activities encourage active learning with text. Beyond such interventions, students' cognitive motivation, usually referred to as *need for cognition*, plays a role in their engagement in information-processing activities (Cacioppo et al., 1996). In other words, the individual difference factor of need for cognition may influence the effect of retrieval-based interventions on learning.

According to Cacioppo and Petty (1982), need for cognition is a stable dispositional variable that reflects one's intrinsic preference for complex thinking. Research has found that need for cognition is negatively related to individuals' preference for order (Petty & Jarvis, 1996; Webster & Kruglanski, 1994) and structure (Neuberg & Newsome, 1993; Petty & Jarvis, 1996). Students low in need for cognition tend to ignore, avoid, or distort new information (Venkatraman, Marlino, Kardes, & Sklar,

1990). In contrast, individuals high in need for cognition would be more likely to pay attention to an ongoing cognitive task (Osberg, 1987), search and use relevant information to solve problems (Berzonsky & Sullivan, 1992), attend to the quality of arguments (Cacioppo et al., 1983), process conflicting information (Kardash & Scholes, 1996), generate task-relevant thoughts after reading inconsistent information (Lassiter, Briggs, & Bowman, 1991), base judgements on rational considerations (Leary, Sheppard, MaNeil, Jenkins, & Barnes, 1986), be open to ideas (Berzonsky & Sullivan, 1992), be curious (Olsen, Camp, & Fuller, 1984), expect information about various aspects of the world and strive to maximize learning gains (Sorrentino, Bobocel, Gitta, Olson, & Hewitt, 1988). People with high need for cognition incline to enjoy and make great effort in cognitively challenging activities (Cacioppo et al., 1996). Liu and Nesbit (2014) found that need for cognition is a predictor of academic performance.

Research has revealed a positive correlation between need for cognition and recall. For example, Cacioppo et al. (1983) showed that, after reading an editorial containing a series of argumentative statements, college students with high need for cognition remembered more arguments presented in the message. Kardash and Noel (2000) reported a positive relation between need for cognition and recall of information from expository text that followed a problem-solution structure. They related need for cognition to the levels of processing theory (Craik & Lockhart, 1972) and proposed that people high in need for cognition may cognitively engage in elaboration rehearsal and attend to information at a deeper level of processing than those low in need for cognition.

Beyond retention of information, need for cognition has been reported to correlate with text comprehension. Taking Dai and Wang (2007) as an example, college students with higher need for cognition did better than those with lower need for cognition in a comprehension test regardless of whether the studied passage was narrative or expository. Referring to Kintsch's model, Dai and Wang (2007) contended that high need for cognition indicates a tendency to actively construct both the textbase level of comprehension and a well-integrated situation model.

In addition, Cacioppo et al. (1983) and Mongeau (1989) found that individuals with high need for cognition performed better on argumentative tasks. They tend to better evaluate the quality of arguments in persuasive messages, differentiate between

strong and weak arguments, and generate pertinent and task-relevant thoughts (Cacioppo et al., 1996); whereas, in light of the elaboration likelihood model of persuasion (Petty & Cacioppo, 1981, 1986), individuals with low need for cognition tend to avoid actively processing message arguments but depend on such peripheral cues as the attractiveness or credibility of the message source (Petty & Cacioppo, 1986; Petty, Cacioppo, & Goldman, 1981) and the sheer number of arguments presented in the message (Petty & Cacioppo, 1986).

2.5. Research Purpose and Questions

As mentioned in the previous sections, free recall has been found to be one of the most effective retrieval-based approaches for promoting long-term retention of learning materials, vocabulary lists in particular. However, conflicting results exist for studies that used more complex educational materials or aimed at more complex learning outcomes such as problem solving and argumentation (Eglington & Kang, 2018; van Gog & Sweller, 2015). It is thus inferred that free recall is more liable to induce surface learning such that students accept new concepts introduced in the learning material uncritically and perceive them as less related and more isolated than they might be otherwise perceived (Biggs, 1999). It is important to explore strategies that have potential to enhance the effects of retrieval practice on complex learning.

The efficacy of argument-based inquiry on learning has been abundantly demonstrated (Andriessen, 2006; Eskin & Ogan-Bekiroglu, 2013; Nussbaum & Schraw; 2007; van Gelder, 2015). Ample evidence is available to corroborate the effectiveness of AVTs in scaffolding the process of argumentation and advancing meaningful learning (e.g., Dwyer et al., 2013; Liu & Nesbit, 2018; Niu, 2016). Although efficacious, research has found that the benefits of AVTs are associated with how learners interact with the tools that might instigate varying levels of encoding or computation effort crucial for durable learning (Niu, 2016; Robinson & Schraw, 1994). In view of this concern, there is a need to investigate techniques that could induce effortful encoding to reinforce the effects of learning with AVTs.

Given the strengths and vulnerabilities of each instructional strategy, my thesis investigated if retrieval-based argument mapping could optimize the features of retrieval practice and argument construction to more reliably induce deep processing that

strengthens retention of information and transfer of learning. In addition, this research explores if individual differences in need for cognition moderate the effects of argumentation-based retrieval practice. An experiment was conducted to address the following questions:

1. Are retrieval-based activities more effective than restudy in promoting long-term memory?
2. Can retrieval-based activities better promote transfer of learning than restudy?
3. How do argument-oriented and unstructured retrieval practice interventions differ in promoting learning and transfer?
4. How do individual differences in need for cognition influence the effects of retrieval-based argument mapping?

Chapter 3.

Method

3.1. Pilot Study

A pilot study with 9 participants was conducted prior to the main experiment to improve the experimental design. Four of them were randomly assigned to either the restudy group (n = 2) or the retrieval practice group (n = 2). The rest of the participants (n = 5) went through the retrieval-based argument mapping intervention as it involved additional training on how to use the Dialectical Map.

My observations and the feedback collected from the participants led me to identify a few issues. For instance, the participants were initially required to write at least 300 words for the argument essay as one of the posttest measures. Some of them commented that there was no easy way to check how many words they had written. Given this concern, I deliberately adjusted the size of the text entry window and asked the participants to write as much as they could to fill the box.

People held divergent opinions on the speed of the video tutorial that demonstrated how to use and create a dialectical map. Some complained it was running too fast and some wanted to speed it up, which showed the need to give the participants more control over the pacing of the tutorial.

One participant who reported high prior knowledge in renewable energy read much faster than the others and obtained a high score on the posttest. Two questions asking for participants' prior knowledge of wind power and renewable energy were therefore included in the demographic questionnaire as it might have an impact on the effects of treatment activities.

Other suggestions included adding a progress bar indicating how much work remained and resizing the text field for each short-answer question to make the tasks look less demanding. All these concerns were addressed for the main experiment. The students participating in the pilot study were excluded from the main study.

3.2. Participants

A total of 124 students from Simon Fraser University volunteered to take part in the two experimental sessions: a learning session and a delayed posttest session. Each participant was randomly assigned into one of the three groups: a Restudy group, a Retrieval Practice group, or a Dialectical Map group. There were 4 students who did not return to complete the posttest, 3 from the Retrieval Practice group and 1 from the Dialectical Map group. The students who agreed to participate signed a consent form and were each paid $15 and $10 upon completion of session 1 and session 2, respectively.

Of the 120 students who participated in both sessions, 49 were male and 71 were female, with ages ranging from 17 to 37 (M = 21.52, SD = 4.11). There were 105 undergraduate students and 15 students pursuing a graduate degree. Among the participants, 69 spoke English as their first language; of the remaining 51 subjects, 26 had been studying in English-speaking countries for more than 10 years. The participants came from 31 academic programs representing a broad spectrum of disciplines, as shown in Table 3.1.

Table 3.1. Areas of Study

Area of Study	Frequency	Percent
STEM[1]	48	40.00%
Arts, Humanities, and Social Sciences	48	40.00%
Business	15	12.50%
Other[2]	9	7.50%

[1] Science, Technology, Engineering, and Mathematics

[2] Responses included "Undecided", "General Studies", or "Interdisciplinary Studies"

3.3. Materials and Instruments

3.3.1. Reading Text

The participants were invited to study a passage (Appendix A) about wind energy that was adapted from an article posted in National Geographic by Morse and Turgeon (2012). I removed extraneous information and examples (e.g., different types of windmills) and manipulated the distribution of supporting and opposing information. The main purpose of adapting the text was to reduce explicit argumentative features of the

text while ensuring it could serve as a rich and balanced source of evidence supporting and opposing the use of wind turbines. The text contains 1,509 words, 79 sentences, and 151 distinct propositions. The Flesch-Kincaid reading grade level was 10.2, suggesting that the text was understandable for average 10th grade students.

Of the 151 propositions, there were 43 propositions (e.g., “wind will be produced as long as the sun continues to shine”) that could be used by the proponents of wind energy and 44 propositions (e.g., “wind turbines cannot operate at all wind speeds”) that could be borrowed by its opponents to support their positions. The remaining 64 propositions were neutral statements (e.g., “wind is the movement of air across the earth’s surface”) but 14 of them could potentially contribute to either side of the debate. For example, the fact that “many wind turbines are built to be very tall”, on the one hand, explains why wind turbines could operate with great efficiency by accessing stronger and more constant wind; on the other hand, it is one of the reasons that local residents object to erecting wind turbines in their neighbourhood.

Because the presentation format of the learning material might interact with how participants benefit from retrieval-based activities, all of the sentences were presented simultaneously rather than one at a time so as not to obstruct relational processing, as suggested by Eglington and Kang (2018).

3.3.2. Dialectical Map (DMap)

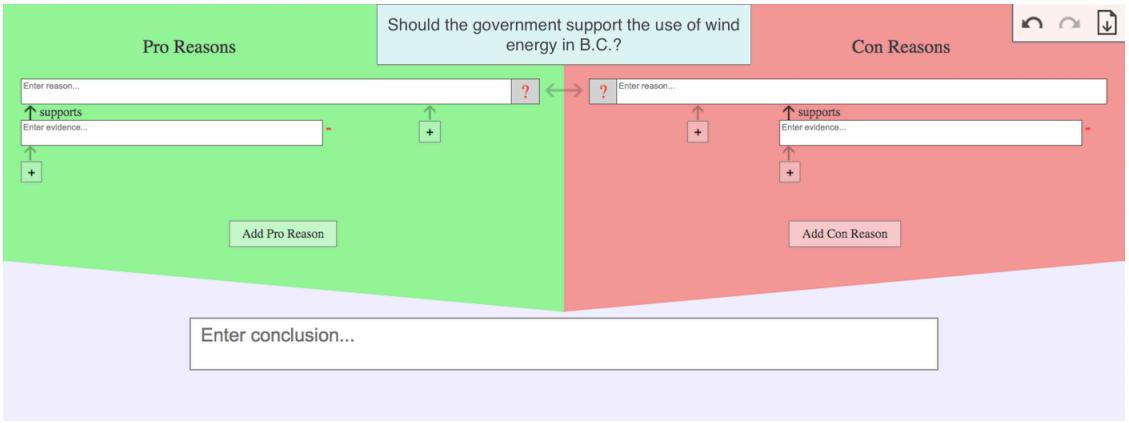
The Dialectical Map (DMap) is a web-based visualization tool that scaffolds argumentation (see Figure 3.1). It was designed by Niu (2016) and hosted in an online learning platform called nStudy that provides a rich collection of tools and features to support self-regulated learning (Winne & Hadwin, 2013; Winne, Nesbit, & Popowich, 2017).

The DMap facilitates the process of constructing arguments for a specific claim, refuting counterarguments, and weighing conflicting perspectives. An instructor or researcher enters a statement or a question at the top of the DMap (e.g., “Should the government support the use of wind energy in B.C.?”), and the DMap visually prompts students to come up with reasons both for and against the statement and support those reasons using evidence and warrants. In addition, the DMap allows students to rate the

strength of each argument, link the arguments and counterarguments that directly oppose each other, and label if a piece of evidence is supporting or opposing the corresponding reason. After weighing the arguments on both sides, students come to a conclusion, a synthesis of the two sides, and type it in a text box placed at the bottom of the map. The DMap was used in this study to scaffold the retrieval-based argument mapping intervention.

The participants in the Dialectical Map group received a training session on how to argue with the DMap. They watched a video tutorial that demonstrated how each feature embedded in the DMap worked. Balloon comments popped up along the timeline of the video to introduce the roles that pro reasons, con reasons, evidence, and warrants played in argumentation and how they related to each other. Examples were provided to foster understanding. The tutorial lasted 11 minutes. It was not fully self-paced, but the students were allowed to pause, fast-forward, or rewind the video whenever necessary. They were instructed to watch the video tutorial attentively so they could use the tool for a learning task assigned later in this study.

Figure 3.1. Initial Student Interface of the DMap



3.3.3. Demographic Questionnaire

A questionnaire was distributed to collect a range of background information about age, gender, highest level of education (in progress), major, first language, and years of studying in English-speaking schools if English was not their first language. In addition, there were two questions that asked the participants to self-evaluate their knowledge about wind power and renewable energy.

3.3.4. Pretest on Free Recall Ability

The participants were invited to take a free recall test as a pretest (Appendix B) in that one's ability to retain and retrieve information might affect the benefits of retrieval-based interventions. They read a passage introducing the concept of natural gas (255 words and 29 key propositions) that was an excerpt of the text posted by Farris (2012) in EnergyBC. After reading this short passage, participants played two unscored "spot the differences" games. The games were an interpolated task introduced to reduce the possible use of rehearsal strategies by some participants. Two minutes later, they were instructed to write down what they could remember from the natural gas text. Their responses were scored following a propositional scoring method (see Section 4.2). The pretest scores suggested individual differences in free recall ability.

3.3.5. Need for Cognition Scale (NCS)

This study used the 18-item Need for Cognition Scale (NCS) to assess one's tendency to engage in and enjoy cognitively challenging activities (Cacioppo, Petty, & Kao, 1984). The participants rated the extent to which they agreed with statements like "I prefer my life to be filled with puzzles that I must solve." and "I only think as hard as I have to." using a scale from 1 (*strongly disagree*) to 5 (*strongly agree*). Of the 18 items, 9 were reverse scored. A higher score suggested the student was more likely to engage in effortful thinking (Cacioppo et al., 1996; Cacioppo et al., 1984; Cacioppo & Petty, 1982).

Previous research has established the validity and reliability of the NCS (Cacioppo et al., 1996). Need for cognition scores were found to correlate with such constructs as intrinsic motivation (Amabile, Hill, Hennessey, & Tighe, 1994; Olson et al., 1984), cognitive innovativeness (Venkatraman & Price, 1990), need to evaluate (Jarvis & Petty, 1996), and academic achievement (Liu & Nesbit, 2014; Tolentino, Curry, & Leak, 1990), but were not influenced by sex (Sadowski, 1993; Spotts, 1994; Tolentino et al., 1990) or potential response biases such as test anxiety (Cacioppo & Petty, 1982) and social desirability (Cacioppo & Petty, 1982; Fletcher, Danilovics, Fernandez, Peterson, & Reeder, 1986; Olson et al., 1984; Petty & Jarvis, 1996). Osberg (1987) examined the convergent and discriminant validity of the scale as evidence further supporting its construct validity (Cacioppo et al., 1996; Osberg, 1987). In addition, Sadowski and

Gulgoz (1992) revealed a test-retest correlation (r = .88) over a 7-week period with a sample of 71 undergraduates. The internal consistency reliability of this instrument was reported to be high (α = .90, Cacioppo et al., 1984). In the present study, the Cronbach's alpha for the 18 questions was .83.

3.3.6. Outcome Achievement Measures

There was a delayed posttest (Appendix C) including free recall and transfer measures. The free recall test asked the participants to write everything they could remember from the wind power text they studied. To more clearly present their understanding on how the idea units related to each other, they were encouraged to write in complete sentences but not to worry about spelling or grammar. This test aimed to detect the effects of treatment-specific activities on long-term retention.

The transfer measures included five short-answer questions and an argument essay. The short-answer questions examined how participants applied what they learned from the wind power text (e.g., the working mechanism of a wind turbine, the impact of erecting a wind turbine, and factors to consider when locating a wind farm) to predict and/or explain various phenomena such as residents' protesting a cell tower proposal (Q1) and how tidal turbines work (Q2-Q5).

The argument essay test asked the participants to take a side on whether or not tidal energy farms should be expanded in many locations around the world and then write an argument essay to defend the position they selected. To avoid confounding effects, the instruction did not highlight what made a strong argument essay because such directions might influence one's performance on argumentation (Nussbaum & Kardash, 2005). The participants were instructed to write as much as they could to fill a textbox accommodating around 300 words. The purpose was to encourage adequate input that could more reliably reflect their knowledge and argumentation skills.

3.4. Research Design and Procedure

Table 3.2 presents an overview of the research procedure and how the treatment groups differed. Participants interacted with the learning materials and tasks in FluidSurvey. All learning tasks and tests were learner-paced, except the interpolated

activities that took up 2 minutes and 10 minutes for the free recall pretest and the learning strategy session, respectively. After signing the consent form, each participant was randomly assigned into one of the three treatment groups and completed the demographic questionnaire, the free recall pretest, and the 18-item NCS, consecutively.

Table 3.2. Overview of Research Treatments and Procedure

Group	Questionnaires and Pretest	Treatment-Specific Activities								Delayed Posttest
		DMap Training	Wind Power Text	10-min Interpolation Task	Reread	Free Recall Practice	Retrieval-Based DMapping	10-min Interpolation Task	All Other Recall Practice	
Restudy	√		√	√	√					√
Retrieval Practice	√		√			√		√	√	√
Dialectical Map	√	√	√				√	√	√	√

3.4.1. Treatment-Specific Activities

Upon completion of the questionnaires and the free recall pretest, the Restudy group and the Retrieval Practice group immediately started reading the wind power text, but those in the Dialectical Map group first received a training session on how to use the DMap prior to reading the text. All participants were encouraged to read the text carefully and attend to the relationships among concepts because they would later be tested without access to the reading passage. After they finished reading, the participants clicked the "*I'VE FINISHED READING THE WIND POWER TEXT, NEXT"* button to proceed to the next stage of intervention.

The Restudy group was instructed to reread the text after a 10-minute "Sketch Me!" task. The purpose of introducing an interpolated task was to, according to the spacing effect, induce greater focus and processing effort when restudying (Dempster, 1989).

After reading the wind power text, the Retrieval Practice group engaged in a free recall practice and a 10-minute "Sketch Me!" activity in succession. To further strengthen

the retrieval practice effect, the participants were given an additional opportunity to recall after the interpolation task. Apart from what they had written during initial retrieval practice, they were asked to think hard and write all other information from the wind power text they could think of. The participants were provided with an empty textbox, without access to what they had written in the first recall of the text.

The participants in the Dialectical Map group engaged in a retrieval-based dialectical mapping practice using the information in the wind power text they just read to answer a question (i.e., Should the government encourage the use of wind energy in B.C.?). The structure of the DMap served as cues that prompted the retrieval of information relevant to this task. It has been found that effortful encoding and retrieval of the overall material might be hampered by strong cues since students are likely to focus on and memorize items that are cued but neglect those that are not cued (Carpenter, 2009; Roediger & Karpicke, 2006). To address this concern, the participants in the Dialectical Map group were asked to take a follow-up recall exercise after the 10-minute "Sketch Me!" activity. They were told to write down any other information that had not been written in the map. The instruction underscored that the additional information need not be related to the question of "Should the government encourage the use of wind energy in B.C.?" During the second recall practice, the participants did not have access to the map they had constructed.

Summing up the key elements of the treatment conditions, each group had an initial opportunity to read the wind power text and each group performed a 10-minute interpolation task. Beyond these common activities, the Restudy group had one opportunity to reread the text, the Retrieval Practice group engaged in one free recall plus one 'all other' recall practice, and the Dialectical Map group took part in one recall activity cued by the DMap plus one 'all other' recall.

3.4.2. Delayed Posttest

After completing the first research session, each participant received $15 and signed up for the posttest. To avoid fatigue and examine if any intervention effects persisted, a delayed posttest was scheduled. The participants were asked to return within two weeks, according to their availability, to complete the outcome tests. The posttest consisted of a free recall test, five short-answer questions, and an argument

essay. Participants were allowed to take as much time as they wanted to answer the questions. Upon submission of their responses, the participants were thanked and given $10 for completing the second research session.

Chapter 4.

Results

4.1. Overview of the Types of Data Collected

This study collected the following types of data: (1) demographic data including age, gender, educational level, academic major, first language, and years of studying in English-speaking schools if English was not their first language; (2) self-reported prior knowledge about wind power and renewable energy; (3) free recall pretest scores indicating a participant's ability to retain and retrieve information; (4) need for cognition scores that reflect a participant's tendency to engage in cognitively challenging tasks; (5) performance scores for retrieval practice and retrieval-based dialectical mapping activities during the treatment phase, including the number of idea units and new ideas (i.e., ideas that were relevant but not mentioned in the learning material) presented in each cycle of unstructured recall practice for the Retrieval Practice group and the number of idea units and new ideas presented in the dialectical map and the follow-up recall response for the Dialectical Map group; (6) time spent on text reading, each intervention activity (i.e., rereading, retrieval practice, and retrieval-based dialectical mapping), and the delayed posttest; (7) interval between the learning session and the posttest session; (8) scores on posttest measures including a free recall test, a short-answer transfer test, and an argument essay transfer test.

4.2. Scoring Free Recall Reponses

A propositional scoring method was applied to score the free recall responses (Bovair & Kieras, 1985; Chmielewski & Dansereau, 1998; Holley, Dansereau, McDonald, Garland, & Collins, 1979). The texts were broken into a set of idea units, each containing a single piece of information. As mentioned in previous sections (i.e., 3.3.1 and 3.3.4), the natural gas passage used for the pretest on free recall ability contained 29 key propositions (e.g., "biogenic gas is released directly into the atmosphere") and the wind energy text is composed of 151 key propositions (e.g., "wind turbines emit no carbon dioxide"). All participants engaged in two free recall measures included in the pre- and

posttests. Those in the Retrieval Practice group and the Dialectical Map group took part in an additional recall activity as a retrieval practice intervention.

Blind to treatment conditions, I scored the free recall responses by comparing the participants' inputs with the propositions listed in the scoring protocol. The propositions that were completely and accurately stated were assigned 1 and those incomplete or partially accurate were given 0.5. Irrelevant or inaccurate propositions received 0. It was noticed that many participants recalled extraneous ideas, which I call new ideas, that were relevant but not introduced or discussed in the original wind power text. A dichotomous variable indicating whether new ideas were presented was coded for the free recall posttest to investigate if retrieval practice induced the mobilization of prior knowledge for information processing. I used a dichotomous variable because it was rare (n = 5) that participants presented more than one new idea in their free recall responses. Scoring reliabilities were established by having a second rater score 12 randomly selected samples of the free recall responses, r = .87. A paired-samples t-test detected no difference between the two raters' mean scores, $t(11) = 1.36$, p = .20.

4.3. Scoring Short-Answer Questions

Based on the model answers, a rubric was developed to aid scoring short-answer questions. As shown in the example below (Table 4.1), the answers were divided into key components that were assigned different weightings according to their importance and pertinence. Prior to grading, I read over all the participants' responses to each question to identify and incorporate unanticipated cases into the rubric.

There were five short-answer questions, with 40 points in total (8, 5, 7, 7, and 13 for questions 1 to 5, respectively). The participant's score for each question was determined by how well the answer aligned with the rubric. Because including more than one new idea was rare (Q1: n = 3; Q2: n = 0; Q3: n = 5; Q4: n = 0; Q5: n = 14), the new idea variable for each short-answer question was dichotomously coded. The sum of the five short-answer new idea variables represented the number of questions to which the participant gave an answer that included an extraneous idea.

All the answers to a given question were scored before proceeding to the next one to enhance the reliability of grading. A second scorer using the same rubric graded

12 randomly selected samples of short-answer questions, r = .94 for Q1, r = .86 for Q2, r = .91 for Q3, r = .84 for Q4, and r = .87 for Q5. Paired-samples t-tests detected no differences between the two raters' mean scores on Q1, $t(11) = -1.40$, p = .19, Q2, $t(11) = 1.34$, p = .21, Q3, $t(11) = -.58$, p = .57, Q4, $t(11) = 1.82$, p = .10, and Q5, $t(11) = 1.03$, p = .33.

Table 4.1. Excerpt of the Scoring Rubric

Question		
David is an energy consultant. He and his colleagues are assigned by their company to help the government search for optimal locations in Nova Scotia to build tidal power stations. To fulfill this task, what aspects or issues do they need to consider?		
Score		Component
2		The optimal locations should have sufficient flow rate and tidal range.
2		Assess the impact of building tidal turbines on local marine ecosystem
2	1	Investigate at what times of day and what times of year the peak rate of tidal flow is most likely to occur.
	1	Finding a site where the tidal flow coincides with times of highest electricity demand is helpful in integrating tidal energy into the power grid.
2	1	NIMBYism
	1	Tidal farms should be built away from densely populated areas to expedite the permitting process.
2	1	Proximity and accessibility
	1	To keep down costs it is helpful to find sites that are not difficult to access utilizing existing logging or mine roads and in close proximity to existing transmission lines.
2	1	Tidal farms should be strategically located away from busy harbors and shipping routes.
	1	This will help avoid marine accidents.
1		Refer to wind farm site selection.
New Ideas? (Yes/No)		New ideas are those not mentioned in the wind power text but related to this question (e.g., its impact on existing economic activities such as fishery and tourism).

Since the five short-answer test items applied different scoring scales, I standardized each item before creating a composite score to allow equal weighting of test items (Song, Lin, Ward, & Fine, 2013). Each item score for an individual participant was first converted to a z-score. Given the inconvenience of interpreting negative values as outcome scores, I transformed the z-scores into standard scores with a mean of 5 and a standard deviation of 1.

Cronbach's α for the 5 short-answer questions was .61. According to Tavakol and Dennick (2011), "if a low alpha is due to poor correlation between items then some should be revised or discarded" (p. 54). Table 4.2 presents the inter-item correlations among those questions. It was observed that the correlations between Q4 and the other items (Q1, Q3, and Q5 in particular) were extremely weak, and removal of Q4 increased Cronbach's α to an acceptable level (α = .65). A total score for the four questions (Q4 deleted) was computed to represent performance on the short-answer transfer test.

Table 4.2. Inter-Item Correlation Matrix

	Short-Answer Transfer Test				
	Q1	Q2	Q3	Q4	Q5
Q1	-				
Q2	.30	-			
Q3	.24	.25	-		
Q4	.00	.30	.07	-	
Q5	.36	.36	.38	.09	-

4.4. Scoring Argument Essays

A coding scheme comprising 5 variables was used to score the argument essays, as shown below. Finding that, in many cases, evidence could not be reliably distinguished from reasons, these two components were combined into one variable to increase reliability of grading. Each statement (or a group of statements) pertinent to any of these variables was assigned 1 point. The overall argument essay score for each participant was calculated by summing the scores for each variable.

1) A ***pro argument*** is a reason or supporting evidence used to defend the position one takes.

2) A ***warrant (pro)*** is a statement, included in a pro argument, that justifies why or how the evidence supports the reason.

3) A ***con argument*** is a reason or evidence against one's proposed claim.

4) A ***warrant (con)*** is used in a con argument to bridge the evidence and the reason.

5) A ***rebuttal*** is evidence or reasoning presented to show why a con argument is problematic.

It was observed that a number of participants referred to the case of wind power while arguing for or against the expansion of tidal energy farms around the world. Therefore, ***referring to wind power*** was dichotomously coded (Yes/No) to investigate if this indicator of analogical reasoning was associated with learning transfer. Also, plenty of participants brought up ideas that were not mentioned in the wind power text but relevant to their arguments, so the number of ***new ideas*** each participant included in the essay was separately coded to explore if dialectical mapping was more likely to induce prior knowledge activation and integration. These two variables were not counted in the overall score of argument essay.

Scoring reliabilities were established by having a second rater score 12 randomly selected samples of the argument essay responses, r = .88 for pro argument, r = .76 for pro warrant, r = .86 for con argument, r = .78 for con warrant, r = .84 for rebuttal, and r = .95 for the new idea variable. Paired-samples t-tests detected no differences between the two raters' mean scores on pro argument, $t(11) = .49$, $p = .63$, pro warrant, $t(11) = -.43$, $p = .67$, con argument, $t(11) = 1.47$, $p = .17$, con warrant, $t(11) = 1.00$, $p = .34$, rebuttal, $t(11) = -1.00$, $p = .34$, and new idea, $t(11) = 1.48$, $p = .17$.

Below is an example showing how an excerpt of the essay was coded. This short paragraph earned 10 points.

> I believe tidal energy should be widely used around the world. Since fossil fuel is going to be depleted **[pro argument]**, people started to look for renewable energy **[pro argument]** like wind **[referring to wind energy]** and water as alternatives. Tides are created by the gravitational pull of the moon and the sun **[pro argument]**. They are available as long as the sun and the moon exist **[warrant (pro)]**. Tidal energy is clean **[pro argument].** It does little damage to the ozone layer **[pro argument; new idea]**. But tidal turbines might cause health issues **[con argument]**. They create low frequency noise **[con argument]** that likely causes hearing problems **[warrant (con)]**, so they should be built away from residential areas **[rebuttal]**.

4.5. Data Screening

Prior to data analyses, all variables were screened for outliers. Among continuous variables, cases with z scores exceeding ±3.29 were identified to be potential univariate outliers (Tabachnick & Fidell, 2007) that were adjusted using the

recommendation of “one unit larger (or smaller) than the next most extreme score in the distribution” (Tabachnick, Fidell, & Ullman, 2019, p. 67).

4.6. Test of Equivalence

Table 4.3 presents the distribution of gender and participants’ English proficiency across the three treatment groups. It also shows the means and standard deviations (in parentheses) for demographic and individual differences variables including age, prior knowledge, free recall ability, and need for cognition (NFC). Since participants’ self-reported prior knowledge about wind power and renewable energy were strongly correlated (r = .76), the two variables were added to indicate how much they knew about the topic of study before learning the wind power text.

A one-way MANOVA detected no difference across three treatment groups, Wilks’ Λ = .95, F (8, 228) = .70, p = .69, η_p^2 = .02. Follow-up ANOVAs showed that the three groups did not detectably differ in age, F (2, 117) = 1.25, p = .29, prior knowledge, F (2, 117) = 1.07, p = .35, need for cognition, F (2, 117) = .42, p = .66, and free recall ability, F (2, 117) = .72, p = .49.

Table 4.3. Demographic and Individual Differences Data

Group	Female	English Proficiency[1]	Age	Prior Knowledge	Pretest on Free Recall Ability	NFC
Restudy	57.5%	34	21.55 (3.71)	5.85 (1.70)	7.45 (4.46)	62.75 (9.96)
Retrieval Practice	52.5%	27	20.78 (3.29)	5.30 (1.70)	6.71 (3.70)	60.80 (8.74)
Dialectical Map	67.5%	34	22.23 (5.09)	5.60 (1.66)	7.80 (4.24)	62.33 (11.18)

[1] Number of native speakers and participants studying in English-speaking countries for 10 years or longer

4.7. Time-On-Task

Table 4.4 presents the means and standard deviations (in parentheses) for time-on-task measures including time spent studying the wind power text prior to the treatment activities, engaging in rereading or retrieval-based interventions, and completing the posttest. The table also shows mean days between the learning and posttest sessions. The days between the two sessions ranged from 2 to 8 days.

One-Way ANOVAs revealed that there was no statistically detectable difference among three groups in time spent studying the wind power text, F (2, 117) = .68, p =. 51, or the number of days between the learning and posttest sessions, F (2, 117) = 2.21, p = .12. However, the treatment groups differed in posttest completion time, F (2, 117) = 5.02, p = .01. Post hoc comparisons using the Bonferroni test showed that participants in the Dialectical Map group (M = 2181.08, SD = 933.07) spent more time working on the posttest questions than those in the Restudy group (M = 1904.83, SD = 794.00) and the Retrieval Practice group (M = 1822.65, SD = 820.44). The difference in posttest completion time between the Retrieval Practice group and the Restudy group was not statistically detectable.

With regard to time spent engaging in the treatment-specific activities, the homogeneity of variance assumption was seriously violated, F (2, 117) = 17.34, p < .001. A Kruskal-Wallis H test showed that group difference was statistically detectable for this time measure, χ^2 (2) = 82.11, p < .001, with a mean rank of 22.43 for the Restudy group, 67.10 for the Retrieval Practice group, and 91.97 for the Dialectical Map group. Pairwise comparisons were performed using Dunn's (1964) procedure with a Bonferroni correction for multiple comparisons. Adjusted p-values are presented. This post hoc analysis revealed statistically detectable differences in time spent engaging in treatment activities between the Restudy (Median = 221.01) and Retrieval Practice (Median = 888.83) groups (p < .001), between the Restudy and Dialectical Map (Median = 1395.39) groups (p < .001), and between the Retrieval Practice and Dialectical Map groups (p = .004).

Table 4.4. Means (Standard Deviations) for Time-On-Task Data

Group	Text Reading	Intervention Activity	Posttest	Interval
Restudy	534.30s (291.90)	235.64s (157.89)	1904.83s (794.00)	3.33 days (1.12)
Retrieval Practice	540.14s (243.39)	966.81s (527.09)	1822.65s (820.44)	3.95 days (1.66)
Dialectical Map	597.21s (262.24)	1538.50s (560.00)	2381.08s (933.07)	3.38 days (1.60)

The correlations between time-on-task and outcome achievement measures indicated to some extent the complex cognitive effects of the study strategies students engaged in (see the table in Appendix D). For example, time spent studying the wind power text was detectably correlated with free recall performance for the Restudy group

(r = .51) and the Retrieval Practice group (r = .60), but not for the Dialectical Map group (r = .12). It is possible that intervening treatment activity moderated the effect of length of exposure to the learning material on free recall performance because students who spent less time in the study phase gained more from retrieval-based dialectical mapping than those who spent greater time in the study phase. Another example is time spent engaging in intervention activities was not detectably correlated with short-answer transfer test performance for the Restudy group (r = .09) and the Retrieval Practice group (r = .28), but the correlation was statistically detectable for the Dialectical Map group (r = .44). The results suggested that, with regard to transfer, spending more time restudying or engaging in retrieval practice was subject to diminishing returns, whereas spending more time on the retrieval-based dialectical mapping activity returned measurable benefit.

4.8. Initial Retrieval Performance

Each participant in the Retrieval Practice and Dialectical Map groups engaged in two retrieval exercises during the treatment phase. The first retrieval exercise was either free recall or creating a dialectical map. The second retrieval exercise asked them to recall all other information they could remember from the wind power text. Table 4.5 shows the means and standard deviations (in parentheses) of the number of idea units and new ideas (ideas not in the text) presented in each retrieval test for the Retrieval Practice group and the Dialectical Map group. The number of new ideas was not counted in the overall number of idea units presented in the recall responses.

An independent-samples t-test was conducted to determine if there was a difference in initial retrieval performance between the Retrieval Practice group and the Dialectical Map group. Welch's t-tests were run for the variables that violated the assumption of homogeneity of variance (p < .01), as shown in Table 4.5. The results revealed that the difference in the number of idea units produced at the first attempt between the Retrieval Practice group and the Dialectical Map group was not statistically detectable, t(63.658) = -.19, p = .85, 95% CI [-5.53, 4.56], d = .04. However, there was a statistically detectable difference in the number of idea units recalled during the follow-up retrieval practice, with the Dialectical Map group scoring higher than the Retrieval Practice group, t(68.779) = -2.31, p = .02, 95% CI [-5.97, -.43], d = .52.

Furthermore, it was found that during the first round of retrieval activity, participants in the Dialectical Map group developed a greater number of new ideas than the Retrieval Practice group, $t(46.011) = -4.24$, $p < .001$, 95% CI [-2.25, -.80], $d = .94$, but there was no statistically detectable difference between the two groups in the number of new ideas presented in the second round of retrieval, $t(78) = -.55$, $p = .58$, 95% CI [-.35, .20], $d = .11$.

Table 4.5. Means Scores (Standard Deviations) of Initial Retrieval Tests

	1st Retrieval Attempt		2nd Retrieval Attempt	
Group	Idea Unit*	New Idea*	Idea Unit*	New Idea
Retrieval Practice	22.88 (13.73)	.33 (.66)	6.91 (4.94)	.28 (.51)
Dialectical Map	23.35 (8.19)	1.85 (2.18)	10.11 (7.25)	.35 (.70)

* Variables that violated the assumption of homogeneity of variance ($p < .01$).

4.9. Correlations

Table 4.6 presents the Pearson correlations among prior knowledge, free recall ability, need for cognition (NFC), and posttest performance. The correlations between free recall ability and the outcome achievement measures were consistently significant ($p < .003$). This warranted the use of free recall ability as a covariate in subsequent analyses.

Table 4.6. Correlation Matrix of Individual Differences Variables and Posttest Measures

	1	2	3	4	5	6
1. Prior knowledge	-					
2. Free recall ability	.162	-				
3. NFC	.262**	.292**	-			
4. Posttest free recall score	.100	.562**	.253**	-		
5. Posttest short-answer score	.018	.473**	.191*	.567**	-	
6. Posttest argument essay score	.015	.270**	.052	.323**	.460**	-

*Correlation is statistically detectable at the .05 level (2 tailed).
** Correlation is statistically detectable at the .01 level (2 tailed).

4.10. Analyses of Outcome Achievement Measures

A one-way MANCOVA was conducted to investigate how the study strategy groups performed on the overall outcome test including a free recall test, a short-answer transfer test, and an argument essay transfer test as three dependent variables. Given the strong correlations between the free recall pretest score and the outcome measures, adjustment was made for the covariate of free recall ability. Table 4.7 shows the means and standard deviations (in parentheses) and adjusted means and standard errors (in parentheses) of the dependent variables for each treatment group.

Table 4.7. Means (Standard Deviations) and Adjusted Means (Standard Errors) of Each Outcome Measure

	Outcome Measures					
	Free Recall Score		Short-Answer Score		Argument Essay Score	
Group	*M (SD)*	M_{adj} *(SE)*	*M (SD)*	M_{adj} *(SE)*	*M (SD)*	M_{adj} *(SE)*
Restudy*	18.10 (11.52)	17.92 (1.36)	19.34 (2.74)	19.30 (.36)	8.23 (3.97)	8.19 (.60)
Retrieval Practice*	14.96 (10.34)	15.80 (1.36)	19.12 (2.58)	19.30 (.36)	7.60 (3.69)	7.77 (.60)
Dialectical Map*	19.75 (8.84)	19.09 (1.36)	21.54 (2.43)	21.40 (.36)	15.68 (4.20)	15.54 (.60)

* N=40

No serious violations were detected in the test of preliminary assumptions. Five univariate outliers ($z > 3.29$), 2 on the free recall and 3 on the argument essay scores, were adjusted as per Tabachnick et al. (2019). There were no multivariate outliers as assessed by the Mahalanobis distance for each case on the three dependent variables ($p > .001$). According to the Shapiro-Wilk's test, the three dependent variables were normally distributed within each treatment group ($p \geq .01$). Meyers, Gamst, and Guarino (2016) recommended "the .01 level as a suitably stringent alpha level with these tests because of their sensitivity to any normality departures and particularly with small sample sizes" (p. 114). There was no multicollinearity as suggested by Pearson correlations among the three dependent variables ranging from .323 to .567. Tabachnick and Fidell (2007) regarded a correlation of .90 or above as an indicator of multicollinearity. There was a linear relationship between each pair of dependent variables within each study strategy group. The relationship between the covariate, free recall ability, and each dependent variable was approximately linear as well. The data

met the assumption of homogeneity of regression slopes as assessed by the interaction term between free recall ability and group, F (6, 224) = .47, p = .83. Box's M test was run to confirm there was homogeneity of variances and covariances, p = .42.

A one-way MANCOVA revealed a statistically detectable difference between the treatment groups on the overall outcome test performance after controlling for free recall ability, Wilks' Λ = .49, F (6, 228) = 16.19, p < .001, partial η^2 = .30. Follow-up one-way ANCOVAs were performed to investigate the impact of intervening treatment activities on each outcome measure. A Bonferroni adjustment was applied such that an effect was statistically detected when the returned p-value was less than .0167.

4.10.1. Free Recall Test

The number of idea units included in their responses was used to suggest how well participants performed on the free recall test. The Levene's test of homogeneity reported that no difference in the variances across groups was detected (p = .18). A one-way ANCOVA found no statistically detectable differences in the adjusted mean for the free recall test after controlling for free recall ability, F (2, 116) = 1.49, p = .23, partial η^2 = .03. In other words, participants from the three study strategy groups performed approximately equally well on the free recall test, as shown in Table 4.8.

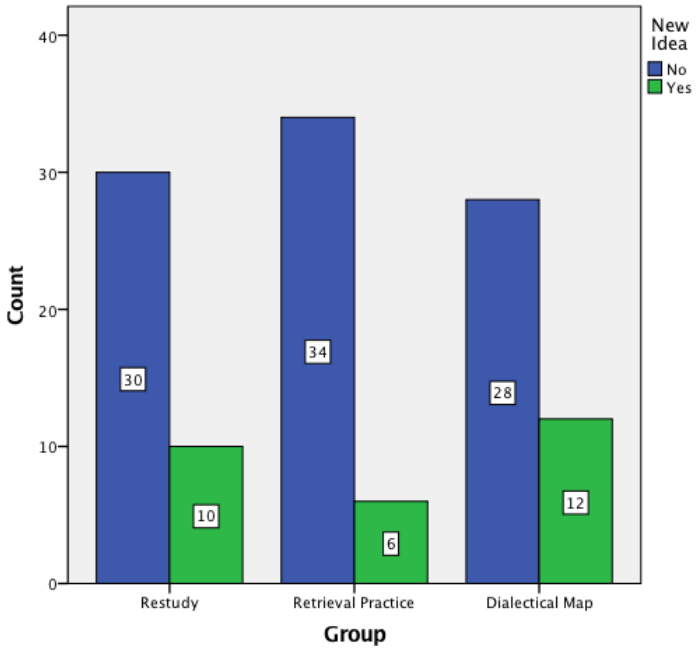
Table 4.8. Pairwise Contrasts for Adjusted Means of Each Outcome Measure

Outcome Measure	Differences in Adjusted Means (95% CI)		
	Restudy vs. Retrieval Practice	Restudy vs. Dialectical Map	Retrieval Practice vs. Dialectical Map
Free Recall Test	2.12 (-2.56, 6.81)	-1.17 (-5.84, 3.50)	-3.29 (-7.99, 1.41)
Short-Answer Transfer Test	.01 (-1.24, 1.25)	-2.09 (-3.33, -.85)*	-2.10 (-3.34, -.85)*
Argument Essay Transfer Test	.42 (-1.65, 2.49)	-7.35 (-9.42, -5.28)*	-7.77 (-9.85, -5.69)*

* Statistically detectable difference (p < .0167) based on Bonferroni adjustment

A χ^2 test for association was conducted between treatment and the dichotomous variable denoting whether a new idea was presented in one's free recall response. All expected cell frequencies were greater than five. There was no statistically detectable association between these two variables, χ^2 (2) = 2.61, p = .27. The strength of correlation was weak, φ = .15, p = .27. Figure 4.1 is a bar chart that illustrates the differences in the number of participants who presented new idea(s) in their free recall responses across the three treatment groups.

Figure 4.1. Distribution of Participants Who Presented New Ideas Across Treatment Groups



4.10.2. Short-Answer Transfer Test

Short-answer questions were developed to examine if participants could apply what they learned from the wind power text to address problems in different contexts. A total score of four questions (Q4 deleted) functioned as an indicator of one's performance on this transfer measure.

The Levene's test of homogeneity detected no difference in the variances across the groups (p = .40). A one-way ANCOVA found that the three treatment groups differed detectably in the adjusted mean for the short-answer transfer test after controlling for free recall ability, F (2, 116) = 11.20, p < .001, partial η_p^2 = .16. As shown in Table 4.8, pairwise comparisons using a Bonferroni post hoc test revealed that participants in the Dialectical Map group (*Adjusted M* = 21.40, *SE* = .36) scored better than those in the Restudy (*Adjusted M* = 19.30, *SE* = .36) and Retrieval Practice (*Adjusted M* = 19.30, *SE* = .36) groups in the short-answer test, but there was no statistically detectable difference between the Restudy group and the Retrieval Practice group.

A Kruskal-Wallis *H* test was conducted to determine if the study strategy groups differed in the number of questions that participants put new ideas on because the data failed the Shapiro-Wilk's test for normality ($p < .001$). The results showed that group difference was statistically detectable, $\chi^2(2) = 12.59$, $p = .002$. Pairwise comparisons were performed using Dunn's (1964) procedure with a Bonferroni correction for multiple comparisons. Adjusted *p*-values are presented. This post hoc analysis revealed that the difference between the Restudy group (mean rank = 46.58) and the Dialectical Map group (mean rank = 71.95) was statistically detected ($p = .001$). The difference between the Restudy group and the Retrieval Practice group (mean rank = 62.98) was considered statistically detectable ($p = .07$) because the estimated probability of type I error was close enough to the conventional and arbitrary threshold of $p = .05$. The Retrieval Practice group and the Dialectical Map group did not differ in this new idea measure ($p = .65$).

4.10.3. Argument Essay Transfer Test

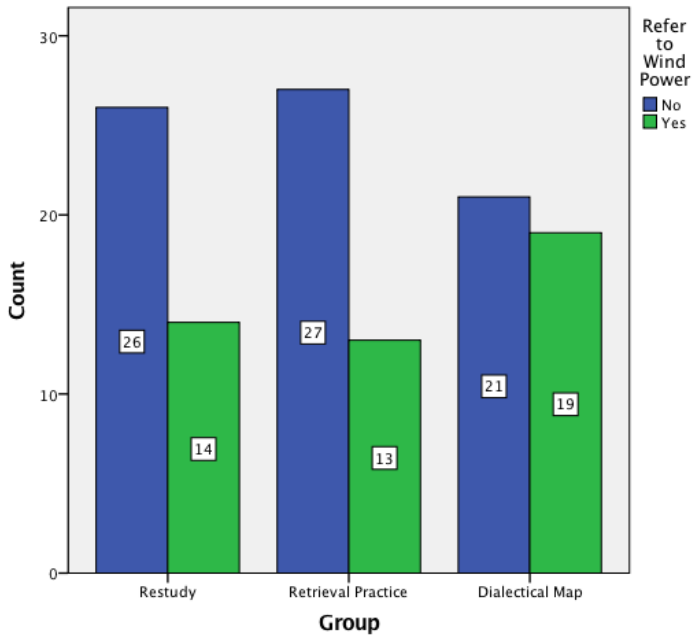
The argument essay was another measure of learning transfer. Table 4.9 summarizes the means and standard deviations (in parentheses) of the number of ideas generated in each category of the coding scheme and the percentage of participants who explicitly mentioned wind power in their argument essays. The results revealed that the Dialectical Map group tended to produce more ideas of each coded type in comparison to the Restudy and Retrieval Practice groups.

Table 4.9. Mean (Standard Deviation) Ideas for Each Coded Argument Essay Variable

	Restudy	Retrieval Practice	Dialectical Map
Pro Argument	6.05 (2.96)	5.48 (2.85)	10.38 (4.07)
Pro Warrant	.23 (0.53)	.33 (0.62)	1.80 (1.91)
Con Argument	1.28 (1.62)	1.48 (1.43)	2.75 (2.11)
Con Warrant	.03 (0.16)	.03 (0.16)	.23 (0.62)
Rebuttal	.85 (1.05)	.30 (0.52)	1.00 (1.13)
New Idea	.93 (0.89)	.95 (1.01)	2.40 (2.06)
Referred to Wind Power	35.00%	32.50%	47.50%

A χ^2 test for association was conducted between treatment and the dichotomous variable denoting whether wind energy was referred to in one's argument essay. All expected cell frequencies were greater than five. There was no statistically detectable association between these two variables, χ^2 (2) = 2.19, p = .34. The strength of correlation was weak, φ = .14, p = .34. Figure 4.2 is a bar chart that illustrates the differences in the number of participants who referred to wind energy in their argument essays across the three treatment groups.

Figure 4.2. Distribution of Participants Who Referred to Wind Power Across Treatment Groups



The participants were encouraged to write as much as they could for the argument essay test. An overly brief response might not reliably reflect their knowledge and argumentation skill. Table 4.10 gives an overview of the lengths of argument essays participants submitted. Levene's test indicated equal variances across groups (p = .24). A one-way ANOVA showed that participants engaging in different intervention activities tended to generate essays of detectably different lengths, F (2, 117) = 16.93, p < .001, partial η_p^2 = .22. A Bonferroni post hoc test found that the difference between the Restudy group and the Retrieval Practice group in the number of words included in their essays was not statistically detectable. The Dialectical Map group produced more words than the other two groups, as shown in Table 4.10.

Table 4.10. Argument Essay Length

Group	Mean (N of Words)	Std. Deviation
Restudy	192.92	113.90
Retrieval Practice	167.32	86.20
Dialectical Map	300.98	123.55
Total	220.41	122.77

A one-way ANCOVA was conducted to investigate if the treatment groups differed in their performance on the argument essay test. As stated in Section 4.4, the sum of the number of pro arguments, pro warrants, con arguments, con warrants, and rebuttals included in the essay was used as the dependent variable. A group difference in the adjusted mean for the argument essay test was statistically detectable, F (2, 116) = 52.36, $p < .001$, partial $\eta_p^2 = .47$. The Levene's test of homogeneity detected no difference in the variances across the groups ($p = .28$). Pairwise comparisons using a Bonferroni post hoc test suggested that the Dialectical Map group (*Adjusted M* = 15.54, *SE* = .60) did detectably better than the Restudy group (*Adjusted M* = 8.19, *SE* = .60) and the Retrieval Practice group (*Adjusted M* = 7.77, *SE* = .60). However, the difference between the Restudy group and the Retrieval Practice group was not statistically detectable, as shown in Table 4.8.

4.11. Interaction with Need for Cognition (NFC)

In addition to the main effect of argumentation-based retrieval practice, this study also investigated if the treatment effect was moderated by need for cognition (NFC), an individual's tendency to engage in cognitively demanding tasks. Hierarchical multiple regressions were conducted to explore how group differences in each outcome measure might vary as a function of NFC after controlling for free recall ability.

No significant outliers were identified. The assumptions of a hierarchical multiple regression including linearity, homoscedasticity, normality, and absence of multicollinearity were all met.

4.11.1. Dummy Coding

A dummy coding procedure was applied to represent categorical variables describing three groups. The Dialectical Map group was designated as the reference group because of a special interest in comparing the other two treatment groups with the Dialectical Map group in terms of posttest achievement. Two dummy variables were created, as depicted in Table 4.11. In this coding system, the first dummy variable (*D1*) compared the Restudy group with the Dialectical Map group that was assigned a value of 0. The second dummy variable (*D2*) compared the Retrieval Practice group with the Dialectical Map group.

Table 4.11. Dummy Variables with Dialectical Map as the Reference Group

Group	D1	D2
Restudy	1	0
Retrieval Practice	0	1
Dialectical Map	0	0

4.11.2. Mean Centering

To make a 0-value meaningful and ameliorate problems with multicollinearity, the continuous moderator variable of NFC was mean centered by subtracting the mean (M = 61.96) from the NFC score of each individual (Aiken & West, 1991; Jaccard & Turrisi, 2003; Tabachnick & Fidell, 2007). As a result, the mean score for the new variable NFC_c became 0, instead of 61.96. The covariate free recall ability and the outcome variable included in each multiple regression model were not centered.

4.11.3. Interaction Effect in Multiple Regression

In a hierarchical multiple regression, sets of variables are added to a regression equation in a specified order to determine how much the newly introduced independent variable(s) adds to the prediction of the dependent variable (Tabachnick & Fidell, 2007). Following the instructions by Aiken and West (1991) and Jaccard and Turrisi (2003), two product terms ($D1*NFC_c$ and $D2*NFC_c$) were created to explore the interaction between treatment and NFC, as shown in the regression equation below:

$$y = a + b_1D1 + b_2D2 + b_3NFC_c + b_4(D1*NFC_c) + b_5(D2*NFC_c) + b_6FreeRecallAbility + e$$

In the first block, free recall ability was entered as a covariate; in the second block, D1, D2, and NFC_c were simultaneously entered; in the third block, the interaction terms, D1* NFC_c and D2* NFC_c, were entered as the primary variables of interest.

4.10.3.1 Free Recall Test

The three-step hierarchical multiple regression was performed to determine if the addition of the treatment × NFC interaction terms improved the prediction of free recall performance when prior free recall ability was statistically controlled. Table 4.12 presents the results of the hierarchical multiple regression analysis on each model. The predictive powers of all three models were statistically detectable ($p < .001$). It was found that free recall ability was a statistically detectable predictor of posttest free recall scores, explaining 32% of the variance ($p < .001$). The addition of D1, D2, and NFC_c explained an additional 3% of the variance. The increase was not statistically detectable, F (3, 115) = 1.44, $p = .24$. None of the newly introduced variables were statistically detectable predictors of free recall performance ($B = -1.23$, $p = .52$ for D1, $B = -3.22$, $p = .10$ for D2, and $B = .10$, $p = .25$ for NFC_c).

Table 4.12. Hierarchical Multiple Regression Predicting Free Recall Performance

	Variable	R^2	R^2_{adj}	F	ΔR^2	ΔF	B	SE_B	β	t
Model 1	Constant	.32	.31	54.56*	.32	54.56	7.26	1.61		4.52*
	Free Recall Ability						1.41	.19	.56	7.39*
Model 2	Constant	.34	.32	14.87*	.03	1.44	9.49	2.06		4.61*
	Free Recall Ability						1.31	.20	.52	6.56*
	D1						-1.23	1.92	-.06	-.64
	D2						-3.22	1.93	-.15	-1.67
	NFC_c						.10	.08	.09	1.15
Model 3	Constant	.35	.32	10.18*	.01	.88	9.08	2.09		4.35*
	Free Recall Ability						1.37	.20	.54	6.68*
	D1						-1.15	1.93	-.05	-.60
	D2						-2.97	1.94	-.14	-1.53
	NFC_c						.06	.13	.05	.44
	D1* NFC_c						-.06	.19	-.03	-.34
	D2* NFC_c						.21	.20	.10	1.05

* $p < .001$

The moderating effect of NFC was not statistically detectable. Adding the interaction terms to the model did not result in a statistically detectable increase in total variation explained, $\Delta R^2 = .01$, F (2, 113) = .88, p = .42. The coefficients of the D1* NFC_c (B = -.06, p = .74) and D2* NFC_c (B = .21, p = .30) interaction terms were not statistically detectable.

A new set of dummy variables D3 and D4, as shown in Table 4.13, were created with the Restudy group as the reference group. In this way, the Restudy group and the Retrieval Practice group were compared.

Table 4.13. Dummy Variables with Restudy as the Reference Group

Group	D3	D4
Restudy	0	0
Retrieval Practice	1	0
Dialectical Map	0	1

A hierarchical multiple regression was rerun with D3 and D4 representing group comparisons. Same pattern of results was produced. The difference in posttest free recall performance between the Restudy group and the Retrieval Practice group was not statistically detectable, as evidenced by the coefficient of D3 at Step 2 (B = -1.99, p = .31). The addition of the interaction terms (D3* NFC_c and D4* NFC_c) did not lead to a salient increase in total variation explained, $\Delta R^2 = .01$, F (2, 113) = .88, p = .42. The coefficients of the D3* NFC_c (B = .27, p = .21) and D4* NFC_c (B = .06, p = .74) interaction terms were not statistically detectable.

4.10.3.2 Short-Answer Transfer Test

A hierarchical multiple regression was conducted to investigate the moderating effect of NFC on short-answer transfer test performance when free recall ability was held constant. Table 4.14 presents the results of each regression model. The predictive powers of all three models were statistically detectable (p < .001). In Model 1, free recall ability was found to be a statistically detectable predictor of short-answer scores, accounting for 22% of the variance (p < .001). The addition of D1, D2, and NFC_c explained an additional 13% of the variance. The increase was statistically detectable, F

(3, 115) = 7.61, $p < .001$. Both D1 ($B = -2.10$, $p < .001$) and D2 ($B = -2.09$, $p < .001$) were statistically detectable predictors of short-answer transfer test performance, indicating that the Dialectical Map group outperformed the Restudy group and the Retrieval Practice group on the short-answer test. NFC was not a statistically detectable predictor of this transfer measure ($B = .02$, $p = .47$).

Table 4.14. Hierarchical Multiple Regression Predicting Short-Answer Transfer Test Performance

	Variable	R^2	R^2_{adj}	F	ΔR^2	ΔF	B	SE_B	β	t
Model 1	Constant	.22	.22	34.02*	.224	34.02	17.67	.46		38.46*
	Free Recall Ability						.32	.06	.47	5.83*
Model 2	Constant	.35	.33	15.64*	.129	7.61	19.29	.55		35.20*
	Free Recall Ability						.29	.05	.43	5.42*
	D1						-2.10	.51	-.36	-4.12*
	D2						-2.09	.51	-.35	-4.06*
	NFC_c						.02	.02	.06	.73
Model 3	Constant	.36	.32	10.37*	.003	.24	19.35	.56		34.63*
	Free Recall Ability						.28	.06	.42	5.13*
	D1						-2.12	.52	-.36	-4.11*
	D2						-2.12	.52	-.36	-4.08*
	NFC_c						.02	.03	.07	.56
	D1* NFC_c						.01	.05	.03	.27
	D2* NFC_c						-.03	.05	-.05	-.47

* $p < .001$

The moderating effect of NFC was not statistically detectable. Adding the interaction terms to the regression model did not increase substantially in its predictive power, $\Delta R^2 = .003$, F (2, 113) = .24, $p = .79$. The coefficients of the D1* NFC_c ($B = .01$, $p = .79$) and D2* NFC_c ($B = -.03$, $p = .64$) interaction terms were not statistically detectable.

A hierarchical multiple regression analysis with D3 and D4 representing group comparisons yielded a similar pattern of results as the regression model involving D1 and D2. At Step 2, the predictive power of D3 was not statistically detectable ($B = .02$, $p = .97$), indicating that the Restudy group and the Retrieval Practice group performed equally on the short-answer test. The addition of the interaction terms (D3* NFC_c and D4* NFC_c) did not lead to a statistically detectable increase in total variation explained,

ΔR^2 = .003, F (2, 113) = .24, p = .79. The coefficients of the D3* NFC_c (B = -.04, p = .50) and D4* NFC_c (B = -.01, p = .79) interaction terms were not statistically detectable.

4.10.3.3 Argument Essay Transfer Test

A hierarchical multiple regression was conducted to investigate if NFC moderated the effects of retrieval practice on argument essay performance when free recall ability was statistically controlled. Table 4.15 presents the results of the hierarchical multiple regression analysis on each model. The predictive powers of all three models were statistically detectable (p < .003). As shown in Model 1, free recall ability was a statistically detectable predictor of argument essay scores, accounting for 7.3% of the variance (p = .003). The addition of D1, D2, and NFC_c explained an additional 44.1% of the variance. The increase was statistically detectable, F (3, 115) = 34.75, p < .001. Both D1 (B = -7.34, p < .001) and D2 (B = -7.79, p < .001) were statistically detectable predictors of argument essay performance, suggesting that the Dialectical Map group achieved detectably higher scores than the Restudy group and the Retrieval Practice group in the argument essay test. The predictive power of NFC was not statistically detectable (B = -.02, p = .63).

Table 4.15. Hierarchical Multiple Regression Predicting Argument Essay Transfer Test Performance

	Variable	R^2	R^2_{adj}	F	ΔR^2	ΔF	B	SE_B	β	t
Model 1	Constant	.07	.07	9.25*	.073	9.25	7.93	.97		8.18*
	Free Recall Ability						.35	.12	.27	3.04*
Model 2	Constant	.51	.50	30.36*	.441	34.75	13.43	.92		14.65*
	Free Recall Ability						.29	.09	.22	3.25*
	D1						-7.34	.86	-.65	-8.59*
	D2						-7.79	.86	-.69	-9.06*
	NFC_c						-.02	.04	-.03	-.48
Model 3	Constant	.52	.49	19.97*	.001	.12	13.36	.94		14.28*
	Free Recall Ability						.30	.09	.23	3.25*
	D1						-7.33	.86	-.64	-8.49*
	D2						-7.75	.87	-.68	-8.90*
	NFC_c						-.03	.06	-.05	-.46
	D1* NFC_c						-.01	.08	-.01	-.09
	D2* NFC_c						.04	.09	.04	.41

* p < .003

The moderating effect of NFC was not statistically detectable. Introducing the interaction terms to the regression model did not result in a salient increase in total variation explained, $\Delta R^2 = .001$, F (2, 113) = .12, p = .89. The coefficients of the D1* NFC_c (B = -.01, p = .93) and D2* NFC_c (B = .04, p = .68) interaction terms were not statistically detectable.

A hierarchical multiple regression analysis with D3 and D4 representing group comparisons yielded a similar pattern of results as the regression model involving D1 and D2. At Step 2, the coefficient of D3 was not statistically detectable (B = -.45, p = .60), indicating that the Restudy group and the Retrieval Practice group did not perform statistically differently on the argument essay test. The addition of the interaction terms (D3* NFC_c and D4* NFC_c) did not lead to a statistically detectable increase in total variation explained, $\Delta R^2 = .001$, F (2, 113) = .12, p = .89. The coefficients of the D3* NFC_c (B = .05, p = .64) and D4* NFC_c (B = .01, p = .93) interaction terms were not statistically detectable.

Chapter 5.

General Discussion and Conclusion

Prior research indicates retrieval practice is an effective strategy for promoting retention of studied information, but when adopted for the purpose of fostering meaningful learning of realistic educational materials, its advantages are not robust (Eglington & Kang, 2018; Rawson, 2015; van Gog & Sweller, 2015). My thesis investigated if coupling retrieval practice with argument construction could more reliably elicit deep processing for long-term memory and learning transfer. The results suggested that argumentation-based retrieval practice did not contribute to superior free recall performance in the delayed posttest after controlling for initial free recall ability. However, it helped participants score higher on the short-answer transfer test and the argument essay transfer test relative to free recall testing and restudy. Unexpectedly, though, participants who engaged in free recall practice and those who reread the text performed similarly on all three posttest measures. Retrieval practice effects were not observed in this comparison. Another unexpected finding was that the correlations between need for cognition and the outcome measures tended to be low. Need for cognition did not moderate the effects of retrieval-based interventions on learning.

5.1. Discussion of the Results

In this study, a posttest that took place two to eight days after the initial study session was conducted to examine long-term learning. A pretest of free recall ability was included in analyses as a covariate because of its strong correlations with the outcome measures. Major findings of my thesis are discussed below.

5.1.1. The Effects of Retrieval-Based Activities on Long-Term Memory

The results showed that the three groups achieved similar scores on the delayed free recall test that asked participants to write as much of the wind power text as they could remember, after controlling for initial free recall ability. According to the elaborative retrieval hypothesis, retrieval practice contributes to an elaborative network, strengthens gist traces, and provides additional retrieval pathways that aid recall of information

(Bouwmeester & Verkoeijen, 2011; Carpenter, 2009; Pyc & Rawson, 2010). Furthermore, both free recall and retrieval-based argument mapping are believed to induce effortful retrieval attempts, which, as suggested by the retrieval effort theories, lead to greater recall performance (Bjork, 1994; Pyc & Rawson, 2009). Therefore, the finding that participants who reread the text did as well as those engaging in either free recall or retrieval-based argument mapping in the memory test is least expected as it contradicts the findings of much previous research on the testing effect (e.g., Hostetter et al., 2019; Roediger & Karpicke, 2006; Wong & Lim, 2019).

During the study strategy intervention session, the Dialectical Map group and the Retrieval Practice group took part in retrieval activities. After studying the wind power text, participants in the Retrieval Practice group first worked on a free recall trial and in a second trial were instructed to write down all other information they could remember from the wind power text. Those in the Dialectical Map group first completed a retrieval-based argument mapping trial followed by a recall trial that asked them to write down any additional information from the wind power text they could think of. According to the transfer-appropriate processing hypothesis, the magnitude of the testing effect is determined by the degree of similarity between the initial and final test conditions. Optimal performance is achieved when the two conditions involve identical cognitive processing. The results of this study at least partially contradicted this hypothesis. On the free recall posttest, neither the Retrieval Practice group nor the Dialectical Map group outperformed the Restudy group who reread the text during the intervention phase. Echoing some earlier research (e.g., Carpenter & DeLosh, 2006; Endres & Renkl, 2015; Rowland, 2014), this study indicated that the testing effect could not be fully explained by the transfer-appropriate processing hypothesis focusing on the overlap in cognitive processing induced by initial and criterial tests.

Although this finding is incongruent with the conclusions of many previous studies that found retrieval practice was more effective than repeated studying in promoting long-term retention (e.g., Dunlosky et al., 2013; Roediger & Butler, 2011; Roediger & Karpicke, 2006; Rowland, 2014), there have been some results paralleling those reported here.

In research by de Jonge et al. (2015, Experiment 1), students were either asked to reread the learning material or instructed to work on an initial fill-in-the-blank test after

reading a lengthy science text comprising 1070 words and 60 sentences. No group difference was found in the delayed fill-in-the-blank final test. Kühn (1914), as cited in Gate (1917), concluded that "[i]t appears that the advantage of recitation differs considerably according to the kind of material being studied; the more senseless and less connected the material, the greater the advantage of recitation over reading… the superiority of recitation [is] rather small in the learning of verses, about twice as great for learning series of words, and larger still for learning nonsense syllables" (p. 7). This contention obtained support from van Gog and Sweller (2015). Their review paper suggested that the effect of retrieval practice might decrease or disappear as the complexity of learning materials increases. They defined complex learning materials as "containing various information elements that are related and must therefore be processed simultaneously in working memory" (p. 248). According to van Gog and Sweller (2015), the difference becomes indiscernible as students in the restudy group benefit from studying highly structured materials. On the one hand, complex study materials might motivate students to spend more time and effort to restudy it so as to develop a solid understanding of what the text is trying to say. Increasing motivation likely abates the effect of study strategy enacted and therefore decreases the benefit of testing over restudy (Kang & Pashler, 2014). On the other hand, there is less need for relational/organizational processing that is more likely to occur during testing relative to restudy. To put it another way, students might benefit less from retrieval-based activities if they are presented with learning materials that could effectively engage them in relational/organizational processing during initial study or restudy (Bouwmeester & Verkoeijen, 2011; Congleton & Rajaram, 2012; Delaney, Verkoeijen, & Spirgel, 2010; van Gog & Sweller, 2015). In the present research, participants were invited to read a text containing 1509 words and 151 propositions that were interrelated. It could be counted as a complex learning material high in element interactivity, as per van Gog and Sweller (2015)'s definition. This might explain why no group difference was statistically detectable.

Another possible reason relates to the unique design of the restudy intervention. The present study asked students in the Restudy group to take part in a 10-minute interpolated activity prior to rereading the text. According to the spacing effect, "the subjects [tend] to devote more attention or processing effort to spaced repetitions than to

massed repetitions" that "[likely] inspire a false sense of knowing and confidence" (Dempster, 1989, p. 318-319).

5.1.2. The Effects of Retrieval-Based Activities on Learning Transfer

Classic studies on test-enhanced learning primarily focused on its benefits for retention and recall of information (e.g., Carpenter, 2009; Rawson, 2015; Roediger & Karpicke, 2006; Rowland, 2014; Pyc & Rawson, 2010). In recent years, there has been an increasing interest in researching the effect of retrieval practice on learning that requires more than recall of information (Hostetter et al., 2019; Karpicke & Aue, 2015; Wong et al., 2019). My thesis examined the effect of retrieval practice on transfer of learning, as measured by participants' short-answer test and argument essay performance. Overall, the results support an interpretation that retrieval-based argument mapping enhanced transfer. However, the superiority of free recall practice over restudy was not found in either test. In other words, knowledge transfer was not strengthened by the challenge of freely recalling information in paragraph format. This finding is inconsistent with that of previous research highlighting the effect of retrieval practice on learning transfer (e.g., Butler, 2010; Butler, Black-Maier, Raley, & Marsh, 2017; Pan & Rickard, 2018; Wong et al., 2019).

5.1.2.1 Short-Answer Test

The short-answer questions included in this study required students to not only recall the relevant information but also apply it to address issues or explain phenomena (e.g., residents' protesting a cell tower proposal or how tidal turbines work) that differed contextually from what they studied but shared similar fundamental concepts. The three treatment groups differed in the short-answer transfer test scores after controlling for initial free recall ability. The Dialectical Map group outperformed the other two groups, but the difference between the Retrieval Practice group and the Restudy group was not statistically detectable. In the literature, opposing views are expressed on the effect of free recall exercise on learning transfer. My thesis contributes to the body of evidence that can be cited to argue against its efficacy.

There is an ongoing debate about whether building a foundation of factual knowledge precedes and enables higher-order learning (Hirsch, 1996; Ravitch, 2009) or higher-order learning can be advanced by directly engaging students in complex tasks

during the initial learning stage (Cuban, 1984; Mehta, 2018). In response to this controversy, Agarwal (2019) compared the effects of retrieval practice with fact questions and higher-order questions falling into *apply*, *analyze*, *evaluate*, and *create* categories as listed in the revised Bloom's taxonomy on long-term learning (Anderson et al., 2001). It was found that delayed higher-order test performance was enhanced by retrieval practice with high-order questions; fact questions failed to promote higher-order learning. Agarwal (2019) therefore concluded that engaging students in retrieval practice that encourages higher-order thinking is more effective than fact-based retrieval practice for higher-order learning. She posited that students working on fact quizzes were not able to transfer factual knowledge to the higher-order questions because they were unaware of the underlying connections. This assumption was echoed in research by Hostetter et al. (2019), which did not detect the benefit of free recall tests on analogical problem solving. They contended that, while retrieval practice by itself is not potent enough to foster higher-order learning, its effect could be strengthened by add-on techniques that boost students' understanding of what they are learning (Hostetter et al., 2019).

The research findings of Agarwal (2019) and Hostetter et al. (2019) are consistent with the result of my thesis that practicing free recall did not benefit students' performance on the delayed short-answer transfer test beyond that produced by restudy, but the advantage of retrieval practice in the form of argument construction was salient. It appears that standard free recall practice tends to induce surface learning such that students retrieve fragmented information from memory and make little effort to consider how the fragments of information relate. If this is the case, then standard retrieval practice might not be effective in helping learners refine or develop a schema that facilitates meaningful learning (Rumelhart, 1980). Combining recall with retrieval-based argument mapping seems to have mitigated the shortcomings of standard retrieval practice. The dialectical map (DMap) may have challenged learners to retrieve the gist of a text rather than isolated items. In other words, free recall testing was more likely to induce knowledge reproduction but retrieval-based argument mapping with the aid of a DMap promoted knowledge construction. The former strengthens a textbase mental representation and the latter a situation model that is crucial for complex learning (Kintsch, 1993).

5.1.2.2 Argument Essay

Despite that participants read a text about wind power during the learning phase, they were instructed to write an essay arguing for or against the expansion of global tidal energy in the posttest. They were not told to but could borrow ideas presented in the wind power text to construct arguments in that wind power and tidal energy are schematically similar in both mechanism and how they impact society. The results showed that the three treatment groups performed differently on this argument essay test. The Restudy group and the Retrieval Practice group did not differ detectably in their argument essay scores, after controlling for initial free recall ability. The Dialectical Map group greatly outperformed the other groups. Participants in the Dialectical Map group tended to write much longer essays and included more arguments on both sides as well as warrants and rebuttals in their responses in comparison to the other two groups. As well, DMap learners were more likely to touch on new concepts (i.e., prior knowledge) that were not mentioned in the wind power text but related to their arguments. These findings suggested that retrieval practice in the form of argument mapping improved knowledge transfer and argumentation skills.

This result can possibly be explained by the transfer-appropriate processing theory. Filling out the DMap with no access to the source text engaged participants in similar cognitive processing as the delayed argument essay task. Another possible reason is that participants in the Dialectical Map group were prompted to retrieve information critical for evaluating the benefits and shortcomings of utilizing wind power. This may have strengthened their memory of related information that can then be more readily transferred to the argument essay test. Beyond that, many students are not skilled in argumentation (Jonassen & Kim, 2010; Kuhn & Udell, 2003; Reznitskaya et al., 2001; Wolfe et al., 2009), which might be due to a deficient argument schema stored in memory (Reznitskaya et al., 2001; Wolfe et al., 2009). The DMap is a cognitive tool. In comparison to restudy or free recall practice, interacting with the DMap is more likely to help learners acquire a well-rounded argument schema that directs their attention to argument-related information, scaffolds the retrieval and organization of information both for and against the proposed claim, and aids argument construction (Anderson et al., 2001; Reznitskaya & Anderson, 2002).

5.1.3. Comparing Free Recall with Argumentation-Based Retrieval Practice

Echoing the findings of many previous studies (e.g., Asterhan & Schwarz, 2007; Jonassen & Kim, 2010; Nesbit et al., 2019), my thesis found that engaging students in argument-based inquiry is conducive to learning transfer. It demonstrated the superiority of argumentation-based retrieval practice over standard retrieval practice in the form of free recall in promoting meaningful learning measured as learning transfer and argumentation. However, these two interventions were equally effective in enhancing students' free recall performance in the delayed posttest.

To my knowledge, my thesis is the first to explore the effects of retrieval practice in the form of argument construction, but research has been conducted to compare free recall and concept mapping as a retrieval-based activity, which lays a foundation for analyses of how retrieval-based argument mapping supports learning in that both concept mapping and argument mapping are theorized to engage learners in relational/organizational processing (Novak & Gowin, 1984; van Gelder, 2002).

Previous research has found that concept mapping in the absence of text is equally effective as free recall in fostering long-term learning as assessed by recall of verbatim knowledge and making inferences by connecting multiple concepts in the text (Blunt & Karpicke, 2014; Ortega-Tudela, Lechuga, & Gómez-Ariza, 2019). According to Blunt and Karpicke (2014), the relational/organizational processing associated with concept mapping is redundant with the cognitive processing afforded by recalling information in paragraph format. To examine if free recall and retrieval-based concept mapping entail exactly the same cognitive processing, Ortega-Tudela et al. (2019) coupled these two retrieval activities and manipulated the order in which the activities took place: concept mapping followed by free recall or free recall preceding concept mapping. The results revealed that creating a concept map in the absence of text before practicing free recall led to better performance on the verbatim and inference tests than taking part in the same activities the other way around. Contrary to what was maintained by Blunt and Karpicke (2014), Ortega-Tudela et al. (2019) argued that distinct cognitive operations underlie these two retrieval activities; otherwise, the sequences would not matter.

In the present study, students were asked to build an argument map after reading the text. It engaged learners in relational/organizational processing by retrieving and fitting information into slots that were placed in a relational structure. Although both argument mapping and concept mapping are said to engender relational/organizational processing (Novak & Gowin, 1984; van Gelder, 2002), the advantage of retrieval-based argument mapping over free recall appeared to be more pronounced than retrieval-based concept mapping as reported in Blunt and Karpicke (2014). This might be due to the fact that the two studies asked students to construct maps on learning texts of different lengths (259 or 236 words vs. 1509 words). Probably it is more demanding to construct a meaningful cognitive representation of a lengthier text, which heightens the difference between free recall and mapping.

Mintzes et al. (2011) advocated that students' proficiency in concept mapping impacts its benefit as a retrieval activity. Previous research failed to detect the superiority of concept mapping over free recall practice might also be because no cognitive tools were provided to their participants who were possibly not adequately trained on concept mapping (Mintzes et al., 2011). Without appropriate aids, concept mapping in the absence of the learning materials provided as weak retrieval cues as free recall practice. In other words, free concept mapping was functionally the same as free recall. Students in need of guidance on concept mapping might not strategically engage in relational/organizational processing as expected such that its advantage over free recall was undermined. With regard to the present study, most of the participants were presumed to be less skilled in argumentation and argument mapping, so the presence of the DMap tool may have scaffolded relational/organizational processing, eliminated additional cognitive load associated with argument map construction, and fostered elaboration, which led to greater learning transfer.

In addition, a DMap is a cognitive tool contributing to the construction of a well-rounded argument schema (Niu, 2016). The cognitive residue (i.e., the changes in one's cognitive capacities that result from interacting with the DMap tool) may have equipped learners with better argumentation skills, and these were reflected in their argument essay responses (Salomon et al., 1991).

The structure of the DMap prompted learners to search and retrieve relevant information from memory as retrieval cues. Carpenter (2009) highlighted the benefits of

weak cues over strong cues in advancing durable learning. Although the DMap tool provided stronger retrieval cues than free recall tests, the Dialectical Map group, like the Retrieval Practice group, were given an additional retrieval activity in which they were asked to freely write anything they could remember from the wind power text which they had not retrieved in the previous trial. A mix of structured and unstructured retrieval activities is speculated to yield greater learning outcomes than argument mapping alone as it mitigates the issue of cue strength.

Furthermore, engaging in argument mapping before recalling additional information by writing paragraphs may have evoked and strengthened gist-based retrieval that is conducive to more elaborative knowledge representations (Ortega-Tudela et al., 2019). The result might be different if the intervention starts with free recall followed by argument mapping, a sequence which may induce more verbatim-oriented retrieval (Bouwmeester & Verkoeijen, 2011; Ortega-Tudela et al., 2019) because task requirements frame retrieval orientations (Rugg & Wilding, 2000). The findings of my thesis partially supported this hypothesis. It was observed that the Retrieval Practice group and the Dialectical Map group did not differ in the number of idea units recalled at the first retrieval trial, but participants in the Dialectical Map group presented detectably more idea units at the second retrieval trial than those in the Retrieval Practice group during the intervention phase. As contended by Ortega-Tudela et al. (2019), practicing retrieval-based mapping activity prior to retrieving information in a paragraph format orients students to adopt a more effective retrieval strategy, which has a beneficial downstream influence on subsequent retrieval attempts. An alternative explanation could be that in the first recall activity participants in the Dialectical Map group covertly retrieved a greater number of ideas but selectively included only relevant information in the map (Blunt & Karpicke, 2014). The residue of retrieved information triggered the recall of related ideas. As a result, the Dialectical Map group generated more idea units than the Retrieval Practice group in the follow-up unstructured retrieval practice.

It was also found that during the first round of treatment-specific retrieval activity interacting with the DMap elicited a greater number of new ideas than freely recalling studied information from memory. For example, some students touched on the theory of photosynthesis and some discussed the impact of building wind farms on existing economic activities such as fishery and tourism, which were not mentioned in the wind

power text they studied. Correspondingly, participants in the Dialectical Map group showed a stronger tendency to draw on prior knowledge to evaluate a viewpoint and answer questions in the delayed transfer tests. The results indicated that the effect of argumentation-based retrieval practice aided by the DMap tool was more robust than that of free recall practice on inducing elaborative processes that contributed to better learning transfer and argumentation. This finding to some extent contradicts the claim that retrieval practice fosters learning solely by retrieval-specific mechanisms rather than elaboration (Karpicke & Blunt, 2011; Karpicke & Smith, 2012).

Hostetter et al. (2019) and Wong and Lim (2019) argued that practicing free recall is not well suited for complex learning such as problem solving and argumentation because it does not effectively support schematic understanding of the learning materials or direct learners' attention to critical information for the intended learning outcomes. This research suggested that retrieval practice with the aid of a DMap is more effective in inducing situation model level processing than freely recalling information from memory.

5.1.4. Retrieval-Based Argument Mapping Takes More Time

As shown in Table 4.4, the Dialectical Map group spent a greater amount of time than the other two groups working on the intervention activities. Engaging in retrieval-based argument mapping helped the participants score higher in the short-answer and argument essay transfer tests; however, more time-on-task did not contribute to better performance on the delayed free recall test. These findings shed light on the cognitive effects and processes induced by retrieval-based argument mapping, and may also raise concerns about its efficiency in promoting learning valued by practitioners.

The Dialectical Map participants spent more time completing the treatment activities but recalled a similar number of idea units as the other two groups in the free recall test. The benefit of DMapping emerged in the transfer tests. Even though the participants in the Restudy group and the Retrieval Practice group could remember as much information from the wind power text as those in the Dialectical Map group, they were not able to apply it to address problems in different contexts as well as DMappers. It was therefore inferred that time-on-task reflected cognitive engagement or cognitive demand. Likely, retrieval-based argument mapping more effectively engaged learners in

elaborative processing, which was time-intensive, such that they did a better job in tests that required more than verbatim recall. In this sense, the effectiveness of DMapping in promoting meaningful learning is superior to the other treatment conditions.

5.1.5. Does Need for Cognition Moderate the Effect of Retrieval Practice?

The last two decades have witnessed a surge of interest in test-enhanced learning. However, research on who could benefit more from retrieval-based activities is scarce (Adesope et al., 2017; Dunlosky et al., 2013). To fill this gap, this research ventures to explore how individual differences in need for cognition bear on learning through testing. Need for cognition reflects a learner's intrinsic preference for cognitively demanding tasks and is a predictor of academic achievement (Cacioppo et al., 1996; Liu & Nesbit, 2014). It was selected as a potential moderator because research has found that students high or low in need for cognition tend to perform differently on information processing (Cacioppo et al., 1996; Kardash & Noel, 2000) and argumentation-oriented learning activities (Cacioppo et al., 1983; Mongeau, 1989; Nussbaum, 2005). However, the present research found need for cognition was not strongly correlated with the posttest performance. Neither did it moderate the effects of retrieval-based interventions on any of the outcome measures including free recall, transfer, and argumentation, after controlling for initial free recall ability.

Cacioppo et al. (1996) suggested that the distinction between individuals high and low in need for cognition becomes more evident as learning tasks become more difficult, but what does task difficulty mean? Could it be that task characteristics such as complexity or novelty, rather than difficulty per se, are motivating learners who have high need for cognition? It may be difficult to memorize a list of esoteric terms and their definitions, but that task might not be more attractive to those with high need for cognition who expect cognitively complex activities (Schneider, Huff, Egan, Gaines, & Ferrara, 2013). The meta-analysis conducted by Liu and Nesbit (2014) categorized outcome measures as retention, comprehension, transfer, and argumentation, which roughly correspond to the levels of learning tasks from less to more cognitively complex in Bloom's taxonomy of the cognitive domain (Anderson et al., 2001). They did not find consistent differences in the effects of need for cognition across those presumed levels of task complexity. It was therefore argued that whatever type of difficulty turns need for

cognition off and on operates at all levels of Bloom's taxonomy. Even the lowest level of the taxonomy includes some tasks that engage learners with high need for cognition (Liu & Nesbit, 2014). The findings of my thesis are consistent with this idea and indicated that retrieval practice either in the form of free recall or argument construction did not uniquely engage learners with high need for cognition.

Although no previous research has examined the role of need for cognition in retrieval-based learning, Kang and Pashler (2014) investigated if motivation, manipulated using monetary bonuses or time savings, moderated the benefit of retrieval practice. Contrary to their hypotheses, none of the three experiments detected the interaction between extrinsic motivation and retrieval practice. They concluded that "the beneficial effect of retrieval practice is probably not driven by relatively lower attention or engagement in the control condition" (p. 187). It appears that neither intrinsic nor extrinsic motivation is a significant moderator. Is it possible that confounding variables have undermined the role of learning motivation? This research only controlled for individual differences in free recall ability, but there are other variables that might play a role, such as intelligence (Brewer & Unsworth, 2012; Minear, Coane, Boland, Cooney, & Albat, 2018), characteristics of learning materials (Minear et al., 2018; van Gog & Sweller, 2015), and individual interest in the topic to be learned (Krapp, 1999).

5.2. Theoretical Contributions

This research marks the first attempt to investigate if effects of standard retrieval practice could be strengthened by argument construction to foster meaningful learning. The findings suggested that retrieval-based argument mapping effectively facilitates learning of complex materials and promotes knowledge transfer and argumentation. This research provides new insight into the literature of test-enhanced learning.

The area of test-enhanced learning has been well researched but the cognitive mechanisms that would elucidate the testing effect are still under debate (Carpenter, 2009; Dunlovsky et al., 2013; Endres & Renkl, 2015; Pyc & Rawson, 2010; Rowland, 2014). My thesis lends support to the elaborative retrieval hypotheses and indicates that there might be two mechanisms at play: retrieval and elaboration. The coaction of retrieval and elaboration is contingent on such factors as task requirements and study materials (e.g., length, topic, and coherence) that learners interact with. In comparison to

free recall practice, retrieval-based argument mapping with appropriate scaffolding can more reliably turn on both processes that co-act to reinforce the testing effect.

There is ample evidence for the effect of retrieval practice on learning, however, the advantage of freely retrieving information over restudy was not statistically detectable in the present study. This result might be due to a factor such as topic interest (Schiefele & Krapp, 1996) or text structure (van Gog & Sweller, 2015) that could potentially affect learner motivation and information processing. In any case, this unexpected finding contributes to the growing body of research studying what boundary conditions might limit or qualify the testing effect.

This study highlights the importance of cognitive tools for test-enhanced learning. Retrieval-based argument mapping is devised as a desirable difficulty geared toward learners proficient in argumentation. For those lacking argumentative knowledge or skills, it may become an undesirable difficulty (Bjork & Bjork, 2011; Bjork & Bjork, 2019). Under either situation, cognitive tools like the DMap play a beneficial role and help learners get the most out of the cognitive activity as intended. This research lends a fresh perspective to interpreting the effectiveness of test-enhanced techniques, for example, concept mapping.

Although the effects of retrieval practice have been extensively researched, there is a dearth of research investigating its efficacy on improving argumentation skills. This research adds further evidence to extant literature and manifests the potential of retrieval-based activities to advance complex learning. Another area under researched is how individual differences impact the testing effect (Adesope et al., 2017; Cogliano, Kardash, & Bernacki, 2019; Dunlosky et al., 2013). Having examined the role of need for cognition, this research outlines a limit on the relationship between one type of intrinsic motivation and retrieval-based activities.

5.3. Implications for Practice

Testing is not new to teachers or students. This study highlights the importance of using practice testing to promote meaningful learning (Adesope et al., 2017; Dunlosky et al., 2013). When tests are devised as learning tools, the selection of questions should be directed by the intended learning outcomes and the ease of implementation in

classrooms if multiple techniques are proven to be equally effective. While free recall practice, which is easy to implement and requires minimal training, may be an optimal strategy for fostering learning that involves knowledge reproduction, the findings of my thesis suggest retrieval-based argument mapping that strengthens the holistic understanding of the learning materials is well suited for complex learning involving knowledge transfer and argumentation.

Retrieval-based argument mapping can be incorporated into classroom learning, either as an individual activity or as a collaborative project. It holds promise for game-based learning. After studying the learning materials or listening to a lecture, students could be divided into, for instance, groups of 4 who collaborate to build a map with as many arguments as they can think of. Upon completion of their maps, they exchange and review each other's work based on a peer-review rubric. The group generating most of the relevant arguments within a specified period of time wins the game. Retrieval-based argument mapping could also be used to support ipsative assessment. For example, students could add a newly constructed argument map to their learning portfolios and reflect on if and how their learning is progressing.

Furthermore, this research indicates the critical role of providing support for retrieval practice oriented towards meaningful learning. For students who are less skilled in argumentation or argument mapping, it is helpful to provide prompts or an interface such as a DMap to aid the process of argument construction so they can take full advantage of meaningful retrieval activities. The provision of a DMap resembles the externalization of the argument schema possessed by an expert debater. After long-term interaction with the tool, the argument schema that the DMap takes on is likely to be internalized by students and used spontaneously (Donato, 1994; Nussbaum, 2008; Pakdaman-Savoji et al., 2019; Vygotsky, 1978).

5.4. Limitations and Future Research

In this experiment, the intervention activities were learner-paced. Compared with restudy, retrieval activities induced more time-on-task. More specifically, the Dialectical Map group spent detectably more time (Median = 1395 s) engaging in the treatment-specific tasks than the Retrieval Practice group (Median = 888 s) and the Restudy group (Median = 221 s). That the total amount of time spent completing the treatment activities

was not matched may cloud the interpretation of the research findings (Karpicke & Blunt, 2011). According to the meta-analytic review conducted by Adesope et al. (2017), the differences between practice tests and comparison interventions tended to be greater when the time is not equated.

The findings of my thesis indicate that students who lack argumentation knowledge and skills might benefit more from learning with cognitive tools such as a DMap that could potentially turn an undesirable difficulty into a desirable difficulty. It would be helpful to collect data relating to students' prior knowledge and skills in argumentation to verify this assumption.

Contradicting many previous studies, my thesis failed to find the superiority of practice testing either in the form of free recall or argument mapping over restudy in the memory test. The disparity could be ascribed to the varying complexity or structure of the learning materials studied by the participants (van Gog & Sweller, 2015). In light of inconsistent evidence for the testing effect with complex learning materials (Karpicke & Aue, 2015; Rawson, 2015; van Gog & Sweller, 2015), I recommend that future investigations manipulate the characteristics of learning materials (e.g., element interactivity or text coherence) to explore when and why the effects of retrieval practice might be disrupted (Rawson, 2015).

In this research, the Retrieval Practice group and the Dialectical Map group were instructed to recall all other information that they had not written during the initial free recall practice or the retrieval-based argument mapping exercise at the second retrieval trial, intending to induce effortful recall of previously unretrieved information. But in most classic studies of the testing effects (e.g., Blunt & Karpicke, 2011; Roediger & Karpicke, 2006), researchers approached this differently: They asked the participants to recall the entire text on each successive test. In future research, it might be helpful to include an additional treatment group following the procedure of study-free recall-interpolation task-free recall to investigate if the "all other" follow-up recall test had undermined the relative advantage of retrieval practice over restudy that was found salient in many previous studies.

My thesis examined the role of need for cognition in test-enhanced learning. It is significant to further explore how other individual differences factors might moderate the

effects of retrieval-based activities (Adesope et al., 2017; Cogliano et al., 2019; Dunlosky et al., 2013). Such research findings would suggest who could benefit more from practice tests as instructional tools and therefore guide the selection of appropriate pedagogical strategies to promote student learning.

Roediger and Karpicke (2006) discussed two types of effects induced by retrieval practice: direct and indirect effects. My thesis focused on the direct effect of practicing retrieval on long-term learning. Previous research has found that testing facilitates learning by making subsequent studying and encoding more effective (Arnold & McDermott, 2013; Bahrick & Hall, 2005; Pyc & Rawson, 2012). Additional exposure to the learning material after taking a test may function as feedback for students to recognize their current level of performance, attend to what was not well learned, and adjust learning strategies appropriately (Carpenter, 2009; Karpicke & Roediger, 2006). It has been suggested that testing followed by having students restudy the text is relatively more effective than testing alone (Karpicke & Roediger, 2006). In future research, it is of interest to examine the effects of retrieval-based argument mapping on mediating processes (e.g., optimizing study strategies and allocation of attentional resources during restudy). I hypothesize that the beneficial effects of retrieval-based argument mapping would be synergistically amplified if students were exposed to feedback or invited to restudy the learning materials after testing. It is also worth investigating how retrieval success impacts subsequent learning and the magnitude of the effect of argumentation-based retrieval practice.

Finally, practice testing was rated as *high utility* by Dunlosky et al. (2013) as it "benefit[s] learners of different ages and abilities and [has] been shown to boost students' performance across many criterion tasks and even in educational contexts" (p. 5). This research was conducted with a sample of postsecondary students in a laboratory context. Future research is needed to examine if the effects of retrieval-based argument mapping can be transferred to K-12 and postsecondary classrooms.

5.5. Conclusion

As is evident in prior literature, testing generally enhances learning. My thesis explored the efficacy of an innovative learning strategy that integrates retrieval practice and argument construction to improve test-enhanced learning for complex educational

materials and learning outcomes. Retrieval-based argument mapping with the aid of a DMap was found to more effectively promote knowledge transfer and argumentation on a related topic than standard free recall testing. This research renders theoretical and practical insights into the use of retrieval activities to promote meaningful learning. A research agenda is proposed to further unravel the mystery and unlock the potential of test-enhanced learning.

References

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research, 87*(3), 659-701.

Agarwal, P. (2019). Retrieval Practice & Bloom's Taxonomy: Do Students Need Fact Knowledge Before Higher Order Learning? *Journal of Educational Psychology, 111*(2), 189-209.

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions.* Thousand Oaks, CA, US: Sage Publications, Inc.

Amabile, T. M., Hill, K. G., Hennessey, B. A., & Tighe, E. M. (1994). The work preference inventory: Assessing intrinsic and extrinsic motivational orientations. *Journal of Personality and Social Psychology, 66*(5)*,* 950-967.

Anderson, R. C., Nguyen-Jahiel, K., McNurlen, B., Archodidou, A., Kim, S. Y., Reznitskaya, A., Tillmanns, M., & Gilbert, L. (2001). The snowball phenomenon: Spread of ways of talking and ways of thinking across groups of children. *Cognition and Instruction*, *19*(1), 1-46.

Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., Raths, L., & Wittrock, M. C. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives* (abridged ed.). New York, NY: Addison Wesley Longman.

Anderson, R. C., Spiro, R. J., & Anderson, M. C. (1978). Schemata as scaffolding for the representation of information in connected discourse. *American Educational Research Journal*, *15*(3), 433-440.

Andriessen, J. (2006). Arguing to learn. In R. K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (pp. 443-460). New York: Cambridge University Press.

Arievitch, I. M., & Stetsenko, A. (2000). The quality of cultural tools and cognitive development: Gal'perin's perspective and its implications. *Human Development, 43*(2), 69-92.

Arnold, K. M., & McDermott, K. B. (2013). Free recall enhances subsequent learning. *Psychonomic Bulletin and Review, 20*(3), 507-513.

Asterhan, C. S. C., & Schwarz, B. B. (2007). The effects of monological and dialogical argumentation on concept learning in evolutionary theory. *Journal of Educational Psychology, 99*(3), 626-639.

Asterhan, C. S., & Schwarz, B. B. (2016). Argumentation for learning: Well-trodden paths and unexplored territories. *Educational Psychologist*, *51*(2), 164-187.

Bahrick, H. P., & Hall, L. K. (2005). The importance of retrieval failures to long-term retention: A metacognitive explanation of the spacing effect. *Journal of Memory and Language, 52*(4), 566-577.

Baker, M.J. (2003). Computer-mediated Argumentative interactions for the co-elaboration of scientific notions. In J. Andriessen, M.J. Baker & D. Suthers (Eds.), *Arguing to Learn: Confronting Cognitions in Computer-Supported Collaborative Learning Environments* (pp. 47-78). Dordrecht, The Netherlands : Kluwer Academic Publishers.

Barstow, B., Fazio, L., Lippman, J., Falakmasir, M., Schunn, C. D., & Ashley, K. D. (2017). The impacts of domain-general vs. domain-specific diagramming tools on writing. *International Journal of Artificial Intelligence in Education*, *27*(4), 671-693.

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology.* Cambridge, UK: Cambridge University Press.

Berzonsky, M. D., & Sullivan, C. (1992). Social-cognitive aspects of identity style: Need for cognition, experiential openness, and introspection. *Journal of Adolescent Research, 7*(2)*,* 140-155.

Biggs, J. (1999). What the student does: Teaching for enhanced learning. *Higher Education Research & Development*, *18*(1), 57-75.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185-205). Cambridge, MA: MIT Press.

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. Healy, S. Kosslyn, & R. Shiffrin (Eds.), *From learning processes to cognitive processes: Essays in honor of William K. Estes* (Vol. 2, pp. 35-67). Hillsdale, NJ: Erlbaum.

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56-64). New York: Worth Publishers.

Bjork, R. A., & Bjork, E. L. (2019). Forgetting as the friend of learning: implications for teaching and self-regulated learning. *Advances in Physiology Education*, *43*(2), 164-167.

Blunt, J. R., & Karpicke, J. D. (2014). Learning with retrieval-based concept mapping. *Journal of Educational Psychology*, *106*(3), 849-858.

Bouwmeester, S., & Verkoeijen, P. P. (2011). Why do some children benefit more from testing than others? Gist trace processing to explain the testing effect. *Journal of Memory and Language*, *65*(1), 32-41.

Bovair, S., &Kieras, D. E. (1985). A guide to propositional analysis for research on technical prose. In B. K. Britton & J. B. Black (Eds.),*Understanding expository text* (pp. 315-362). Hillsdale, NJ: Erlbaum.

Brainerd, C. J., Payne, D. G., Wright, R., & Reyna, V. F. (2003). Phantom recall. *Journal of Memory and Language*, *48*(3), 445-467.

Brewer, G. A., & Unsworth, N. (2012). Individual differences in the effects of retrieval from long-term memory. *Journal of Memory and Language*, *66*(3), 407-415.

Burdo, J., & O'Dwyer, L. (2015). The effectiveness of concept mapping and retrieval practice as learning strategies in an undergraduate physiology course. *Advances in physiology education*, *39*(4), 335-340.

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(5), 1118-1133.

Butler, A. C., Black-Maier, A. C., Raley, N. D., & Marsh, E. J. (2017). Retrieving and applying knowledge to different examples promotes transfer of learning. *Journal of Experimental Psychology: Applied*, *23*(4), 433-446.

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology, 42*(1), 116-131.

Cacioppo, J. T., Petty, R. E., Feinstein, J. A., & Jarvis, W. B. (1996). Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. *Psychological Bulletin, 119*(2), 197-253.

Cacioppo, J. T., Petty, R. E., & Kao, C. F. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment*, *48*(3), 306-307.

Cacioppo, J. T., Petty, R. E., & Morris, K. J. (1983). Effects of need for cognition on message evaluation, recall, and persuasion. J*ournal of Personality and Social Psychology, 45*(4), 805-818.

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(6), 1563-1569.

Carpenter, S. K. (2011). Semantic information activated during retrieval contributes to later retention: Support for the mediator effectiveness hypothesis of the testing effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(6), 1547-1552.

Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition*, *34*(2), 268-276.

Carpenter, S. K., & Kelly, J. W. (2012). Tests enhance retention and transfer of spatial learning. *Psychonomic Bulletin & Review*, *19*(3), 443-448.

Carpenter, S. K., & Yeung, K. L. (2017). The role of mediator strength in learning from retrieval. *Journal of Memory and Language, 92*, 128-141.

Chmielewski, T. C., & Dansereau, D. F. (1998). Enhancing the recall of text: Knowledge mapping training promotes implicit transfer. *Journal of Educational Psychology, 90*(3), 407-413.

Clark, J. M., & Paivio, A. (1991). Dual coding theory and education. *Educational psychology review*, *3*(3), 149-210.

Cogliano, M., Kardash, C. M., & Bernacki, M. L. (2019). The effects of retrieval practice and prior topic knowledge on test performance and confidence judgments. *Contemporary Educational Psychology*, *56*, 117-129.

Congleton, A., & Rajaram, S. (2012). The origin of the interaction between learning method and delay in the testing effect: The roles of processing and conceptual retrieval organization. *Memory & Cognition, 40*, 528-539.

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11*(6), 671–684.

Dai, D. Y., & Wang, X. (2007). The role of need for cognition and reader beliefs in text comprehension and interest development. *Contemporary Educational Psychology, 32*(3), 332-347.

Cuban, L. (1984). Policy and research dilemmas in the teaching of reasoning: Unplanned designs. *Review of Educational Research, 54*(4), 655-681.

de Jonge, M., Tabbers, H. K., & Rikers, R. M. (2015). The effect of testing on the retention of coherent and incoherent text material. *Educational Psychology Review*, *27*(2), 305-315.

Delaney, P. F., Verkoeijen, P. P., & Spirgel, A. (2010). Spacing and testing effects: A deeply critical, lengthy, and at times discursive review of the literature. In *Psychology of learning and motivation* (Vol. 53, pp. 63-147). Academic Press.

Dempster, F. N. (1989). Spacing effects and their implications for theory and practice. *Educational Psychology Review*, *1*(4), 309-330.

Dole, J. A., & Sinatra, G. M. (1998). Reconceptalizing change in the cognitive construction of knowledge. *Educational psychologist*, *33*(2/3), 109-128.

Donato, R. (1994). Collective scaffolding in second language learning. In J. P. Lantolf, & G. Appel (Eds.), *Vygostkian approaches to second language research* (pp. 33-56). New Jersey: Ablex.

Driver, R., Newton, P., & Osborne, J. (2000). Establishing the norms of scientific argumentation in classrooms. *Science Education, 84*(3), 287-312.

Duchastel, P. C., & Nungester, R. J. (1982). Testing effects measured with alternate test forms. *Journal of Educational Research, 75*(5), 309-313.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, *14*(1), 4-58.

Dwyer, C., Hogan, M., & Stewart, I. (2010). The evaluation of argument mapping as a learning tool: Comparing the effects of map reading versus text reading on comprehension and recall of arguments. *Thinking Skills and Creativity, 5*(1), 16-22.

Dwyer, C. P., Hogan, M. J., & Stewart, I. (2012). An evaluation of argument mapping as a method of enhancing critical thinking performance in e-learning environments. *Metacognition and Learning, 7*(3), 219-244.

Dwyer, C. P., Hogan, M. J., & Stewart, I. (2013). An examination of the effects of argument mapping on students' memory and comprehension performance. *Thinking Skills & Creativity, 8*, 11-24.

Easterday, M. W., Aleven, V., & Scheines, R. (2007). Tis better to construct than to receive? The effects of diagram tools on causal reasoning. *Frontiers in Artificial Intelligence and Applications*, *158*, 93-100.

Eglington, L. G., & Kang, S. H. (2018). Retrieval practice benefits deductive inference. *Educational Psychology Review*, *30*(1), 215-228.

Endres, T., & Renkl, A. (2015). Mechanisms behind the testing effect: an empirical investigation of retrieval practice in meaningful learning. *Frontiers in psychology*, *6*(1054), 1-6.

Erduran, S. (2007). Methodological foundations in the study of argumentation in science classrooms. In *Argumentation in science education* (pp. 47-69). Springer, Dordrecht.

Eskin, H., & Ogan-Bekiroglu, F. (2013). Argumentation as a Strategy for Conceptual Learning of Dynamics. *Research in Science Education, 43*(5), 1939-1956.

Farris, A. (2012). Natural gas [Web article]. Retrieved from http://www.energybc.ca/naturalgas.html

Fletcher, F. J. O., Danilovics, P., Fernandez, G., Peterson, D., & Reeder, G. D. (1986). Attributional complexity: An individual difference measure. *Journal of Personality and Social Psychology, 51*(4)*,* 875-884.

Garcia-Mila, M., & Andersen, C. (2007). Cognitive foundations of learning argumentation. In *Argumentation in science education* (pp. 29-45). Springer, Dordrecht.

Gates, A. I. (1917). Recitation as a factor in memorizing. *Archives of Psychology, 6*(40), 1-104.

Hample, D. (1992). The Toulmin model and the syllogism. In W. L. Benoit, D. Hample, & P. J. Benoit (Eds.), *Readings in argumentation* (pp. 225–238). Dordrecht: Foris.

Harrell, M. (2012). Assessing the efficacy of argument diagramming to teach critical thinking skills in introduction to philosophy. *Inquiry: Critical Thinking Across the Disciplines*, *27*(2), 31-39.

Hirsch, E. D. (1996). *The schools we need and why we don't have them.* New York, NY: Doubleday.

Hogan, K. & Fisherkeller, J. (2000). Dialogue as data: Assessing students' scientific reasoning with interactive protocols. In J. J. Mintzes, J. H. Wandersee & J. D. Novak (Eds.), *Assessing science understanding: A human constructivist view* (pp. 95-127). San Diego: Academic.

Hoffman, M., & Paglieri, F. (2011). Cognitive effects of argument visualization tools. In *Proceedings of the 9th international conference of the Ontario Society for the Study of Argumentation*, 1-12. Windsor: OSSA.

Holley, C. D., Dansereau, D. F., McDonald, B. A., Garland, J. C, & Collins, K. W. (1979). Evaluation of hierarchical mapping techniques as an aid to prose processing. *Contemporary Educational Psychology, 4*(3), 227-237.

Hostetter, A. B., Penix, E. A., Norman, M. Z., Batsell Jr, W. R., & Carr, T. H. (2019). The role of retrieval practice in memory and analogical problem-solving. *Quarterly Journal of Experimental Psychology*, *72*(4), 858-871.

Jaccard, J., & Turrisi, R. (2003). *Interaction effects in multiple regression*. Thousand Oaks, CA: SAGE Publications, Inc.

Jarvis, W. B. G., & Petty, R. E. (1996). The need to evaluate. *Journal of Personality and Social Psychology*, *70*(1), 172-194.

Johnson, C. I., & Mayer, R. E. (2009). A testing effect with multimedia learning. *Journal of Educational Psychology*, *101*(3), 621-629.

Jonassen, D. (1992). What are cognitive tools? In P. A. M. Kommers, D. H. Jonassen, & J. T. Mayes (Eds.), *Cognitive Tools for Learning* (pp. 1-6). Berlin: Springer-Verlag.

Jonassen, D. (1996). *Computers in the classroom: Mindtools for critical thinking.* Englewood Cliffs, N.J: Merrill.

Jonassen, D. H., & Kim, B. (2010). Arguing to learn and learning to argue: Design justifications and guidelines. *Educational Technology Research and Development, 58*(4), 439-457.

Jonassen, D., & Reeves, T. (1996). Learning with technology: Using computers as cognitive tools. In D. Jonassen (Ed.), *Handbook of research on educational communication and technology* (pp. 693-719). New York: Macmillan.

Kang, S. H. (2010). Enhancing visuospatial learning: The benefit of retrieval practice. *Memory & Cognition*, *38*(8), 1009-1017.

Kang, S. H., & Pashler, H. (2014). Is the benefit of retrieval practice modulated by motivation?. *Journal of Applied Research in Memory and Cognition*, *3*(3), 183-188.

Kardash, C. M., & Noel, L. K. (2000). How organizational signals, need for cognition, and verbal ability affect text recall and recognition. *Contemporary Educational Psychology, 25*(3), 317-331.

Kardash, C. M., & Scholes, R. J. (1996). Effects of pre-existing beliefs, epistemological beliefs, and need for cognition on interpretation of controversial issues. *Journal of Educational Psychology, 88*(2)*,* 260-271.

Karpicke, J. D., & Aue, W. R. (2015). The testing effect is alive and well with complex materials. *Educational Psychology Review*, *27*(2), 317-326.

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science*, *331*(6018), 772-775.

Karpicke, J. D., & Smith, M. A. (2012). Separate mnemonic effects of retrieval practice and elaborative encoding. *Journal of Memory and Language, 67*(1), 17-29.

Karpicke, J. D., Blunt, J. R., & Smith, M. A. (2016). Retrieval-based learning: positive effects of retrieval practice in elementary school children. *Frontiers in Psychology*, *7*, 1-9.

Keller, T., Gerjets, P., Scheiter, K., & Garsoffky, B. (2006). Information visualizations for knowledge acquisition: The impact of dimensionality and color coding. *Computers in Human Behavior*, *22*(1), 43-65.

Kim, B., & Reeves, T. C. (2007). Reframing research on learning with technology: In search of the meaning of cognitive tools. *Instructional Science, 35*(3), 207-256.

Kintsch,W. (1993). Information accretion and reduction in text processing: Inferences. *Discourse Processes*, *16*(1-2), 193-202.

Kole, J. A., & Healy, A. F. (2013). Is retrieval mediated after repeated testing? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(2), 462-472.

Kornell, N., & Vaughn, K. E. (2016). How retrieval attempts affect learning: A review and synthesis. In *Psychology of learning and motivation* (Vol. 65, pp. 183-215). Academic Press.

Krapp, A. (1999). Interest, motivation and learning: An educational-psychological perspective. *European Journal of Psychology of Education*, *14*(1), 23-40.

Kühn, A. (1914). Über Einprägung durch Lesen und durch Rezitieren [On imprinting through reading and reciting]. *Zeitschrift für Psychologie, 68*, 396-481.

Kuhn, D. (1991). *The skills of argument.* New York: Cambridge University Press.

Kuhn, D. (1992). Thinking as argument. *Harvard Educational Review, 62*(2), 155-178.

Kuhn, D., Shaw, V., & Felton, M. (1997). Effects of dyadic interaction on argumentive reasoning. *Cognition and instruction*, *15*(3), 287-315.

Kuhn, D., & Udell, W. (2003). The development of argument skills. *Child Development*, *74*(5), 1245-1260.

Lajoie, S. (1993). Computer environments as cognitive tools for enhancing learning. In S. Lajoie & S. Derry (Eds.), *Computers as cognitive tools* (pp. 261-288). Hillsdale, NJ: Lawrence Erlbaum Associates.

Larkin, J. H., & Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science*, *11*(1), 65-100.

Lassiter, G. D., Briggs, M. A., & Bowman, R. E. (1991). Need for cognition and the perception of ongoing behavior. *Personality and Social Psychology Bulletin, 17*(2)*,* 156-160.

Leary, M. R., Sheppard, G. A., McNeil, M. S., Jenkins, T. B., & Barnes, B. D. (1986). Objectivism in information utilization: Theory and measurement. *Journal of Personality Assessment, 50*(1)*,* 32-43.

Liu, Q., & Nesbit, J. C. (2014). *The relation between need for cognition and academic achievement: A meta-analysis.* Paper presented at American Educational Research Association. Philadelphia, Pennsylvania.

Liu, Q., & Nesbit, J. C. (2018). Conceptual change with refutational maps. *International Journal of Science Education*, *40*(16), 1980-1998.

McCrudden, M. T., McCormick, M. K., & McTigue, E. M. (2011). Do the spatial features of an adjunct display that readers complete while reading affect their understanding of a complex system?. *International Journal of Science and Mathematics Education*, *9*(1), 163-185.

McCrudden, M. T., & Rapp, D. N. (2017). How visual displays affect cognitive processing. *Educational Psychology Review*, *29*(3), 623-639.

McDermott, K. B. (1996). The persistence of false memories in list recall. *Journal of Memory and Language*, *35*(2), 212-230.

Means, M. L., & Voss, J. F. (1996). Who reasons well? Two studies of informal reasoning among children of different grade, ability, and knowledge levels. *Cognition and Instruction*, *14*(2), 139-178.

Mehta, J. (2018, January 4). A pernicious myth: Basics before deeper learning [Blog post]. Retrieved from http://blogs.edweek.org/edweek/learning_deeply/2018/01/a_pernicious_myth_basics_before_deeper_learning.html

Meyers, L. S., Gamst, G., & Guarino, A. J. (2016). *Applied multivariate research: Design and interpretation*. Sage publications.

Minear, M., Coane, J. H., Boland, S. C., Cooney, L. H., & Albat, M. (2018). The benefits of retrieval practice depend on item difficulty and intelligence. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 44*(9), 1474-1486.

Mintzes, J. J., Canas, A., Coffey, J., Gorman, J., Gurley, L., Hoffman, R., McGuire, S. Y., Miller, N., Moon, B., Trifone, J., & Wandersee, J. H. (2011). Comment on "retrieval practice produces more learning than elaborative studying with concept mapping". *Science*, *334*(6055), 453-453.

Mongeau, P. (1989). Individual differences as moderators of persuasive message processing and attitude-behavior relations. *Communication Research Reports, 6*(1), 1-6.

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*(5), 519-533.

Morse, E., & Turgeon, A. (2012). Putting wind to work [Web article]. Retrieved from https://www.nationalgeographic.org/news/putting-wind-work/

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). *The University of South Florida Word Association, Rhyme, and Word Fragment Norms.* Retrieved from http://w3.usf.edu/FreeAssociation/

Nesbit, J. C., & Adesope, O. O. (2006). Learning with concept and knowledge maps: A meta-analysis. *Review of Educational Research, 76*(3), 413-448.

Nesbit, J., Niu, H., & Liu, Q. (2019). Cognitive tools for scaffolding argumentation: Maximizing student engagement, motivation, and learning. In: O. Adesope & A. Rud (Eds.), *Contemporary Technologies in Education* (pp. 97-117). Palgrave Macmillan, Cham.

Neuberg, S. L., *&* Newsom, J. T. (1993). Personal need for structure: Individual differences in the desire for simple structure. *Journal of Personality and Social Psychology, 65*(1)*,* 113-131.

Nielsen, J. A. (2013). Dialectical features of students' argumentation: A critical review of argumentation studies in science education. *Research in Science Education*, *43*(1), 371-393.

Niu, H. (2016). *Pedagogical efficacy of argument visualization tools* (Unpublished doctoral dissertation), Simon Fraser University, Burnaby, BC, Canada.

Novak, J.D., & Gowin, D.B. (1984). *Learning how to learn.* New York: Cambridge University Press.

Nussbaum, E. M. (2005). The effect of goal instructions and need for cognition on interactive argumentation. *Contemporary Educational Psychology*, *30*(3), 286-313.

Nussbaum, E. M. (2008). Using argumentation vee diagrams (AVDs) for promoting argument-counterargument integration in reflective writing. *Journal of Educational Psychology*, *100*(3), 549-565.

Nussbaum, E. M. (2011). Argumentation, dialogue theory, and probability modeling: Alternative frameworks for argumentation research in education. *Educational Psychologist*, *46*(2), 84-106.

Nussbaum, M. E., & Kardash, C. A. M. (2005). The effects of goal instructions and text on the generation of counterarguments during writing. *Journal of Educational Psychology*, *97*(2), 157-169.

Nussbaum, E. M., & Schraw, G. (2007). Promoting argument-counterargument integration in students' writing. *Journal of Experimental Education, 76*(1), 59-92.

Nussbaum, E. M., & Sinatra, G. M. (2003). Argument and conceptual engagement. *Contemporary Educational Psychology, 28*(3), 384-395.

Ogan-Bekiroglu, F., & Eskin, H. (2012). Examination of the relationship between engagement in scientific argumentation and conceptual knowledge. *International Journal of Science and Mathematics Education*, *10*(6), 1415-1443.

Olson, *K.,* Camp, C., & Fuller, D. (1984). Curiosity and need for cognition. *Psychological Reports, 54*(1)*,* 71-74.

Ortega-Tudela, J. M., Lechuga, M. T., & Gómez-Ariza, C. J. (2019). A specific benefit of retrieval-based concept mapping to enhance learning from texts. *Instructional Science*, *47*(2), 239-255.

Osberg, T. (1987). The convergent and discriminant validity of the Need for Cognition Scale. *Journal of Personality Assessment, 51*(3)*,* 441-450.

Pakdaman-Savoji, A., Nesbit, J. C., & Gajdamaschko, N. (2019). The conceptualisation of cognitive tools in learning and technology: A review. *Australasian Journal of Educational Technology*, *35*(2), 1-24.

Paivio, A. (1986). *Mental representations: A dual coding approach.* Oxford: Oxford University Press.

Pan, S. C., & Rickard, T. C. (2018). Transfer of test-enhanced learning: Meta-analytic review and synthesis. *Psychological bulletin*, *144*(7), 710-756.

Payne, D. G., Elie, C. J., Blackwell, J. M., & Neuschatz, J. S. (1996). Memory illusions: Recalling, recognizing, and recollecting events that never occurred. *Journal of Memory and Language*, *35*(2), 261-285.

Pea, R. (1987). Cognitive technologies for mathematics education. In A. H. Schoenfeld (Ed.), *Cognitive science and mathematics education* (pp. 89-122). Hillsdale, NJ: Lawrence Erlbaum.

Petty, R. E., & Cacioppo, J. T. (1981). *Attitudes and persuasion: Classic and contemporary approaches.* Dubuque, IA: William C. Brown.

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In L. Berkowitz (Ed.), *Advances in experimental social psychology (Vol. 19*, 123-205). New York: Academic Press.

Petty, R. E., Cacioppo, J. T, & Goldman, R. (1981). Personal involvement as a determinant of argument-based persuasion. *Journal of Personality and Social Psychology, 41*(5)*,* 847-855.

Petty, R. E., & Jarvis, W. B. G. (1996). An individual differences perspective on assessing cognitive processes. In N. Schwarz & S. Sudman (Eds.), *Answering questions: Methodology for determining cognitive and communicative processes in survey research* (pp. 221-257). San Francisco, CA, US: Jossey-Bass.

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language, 60*(4), 437-447.

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science*, *330*(6002), 335-335.

Pyc, M. A., & Rawson, K. A. (2012). Why is test–restudy practice beneficial for memory? An evaluation of the mediator shift hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*(3), 737-746.

Rawson, K. A. (2015). The status of the testing effect for complex materials: Still a winner. *Educational Psychology Review, 27*(2), 327-331.

Rawson, K. A., Vaughn, K. E., & Carpenter, S. K. (2015). Does the benefit of testing depend on lag, and if so, why? Evaluating the elaborative retrieval hypothesis. *Memory & Cognition*, *43*(4), 619-633.

Ravitch, D. (2009, September 15). *Critical thinking? You need knowledge.* The Boston Globe. Retrieved from http://archive.boston.com/bostonglobe/editorial_opinion/oped/articles/2009/09/15/critical_thinking_you_need_knowledge/

Reed, S. K. (1993). A schema-based theory of transfer. In D. K. Detterman & R. J. Sternberg (Eds.), *Transfer on trial: Intelligence, cognition, and instruction* (pp. 39-67). Westport, CT, US: Ablex Publishing.

Reznitskaya, A., & Anderson, R. (2002). The argument schema and learning to reason. In C. Block & M. Pressley (Eds.), *Comprehension instruction: Research-based best practices* (pp. 319-334). New York: The Guilford Press.

Reznitskaya, A., Anderson, R. C., Dong, T., Li, Y., Kim, I.-H., & Kim, S.-Y. (2008). Learning to think well: Application of argument schema theory to literacy instruction. In C. C. Block & S. R. Parris (Eds.), *Solving problems in the teaching of literacy. Comprehension instruction: Research-based best practices* (pp. 196-213). New York, NY, US: The Guilford Press.

Reznitskaya, A., Anderson, R. C., & Kuo, L. J. (2007). Teaching and learning argumentation. *The Elementary School Journal*, *107*(5), 449-472.

Reznitskaya, A., Anderson, R., McNurlen, B., Nguyen-Jahiel, K., Archodidou, A., & Kim, S. (2001). Influence of oral discussion on written argument. *Discourse Processes, 32*(2-3), 155-175.

Robinson, D. H., & Kiewra, K. A. (1995). Visual argument: Graphic organizers are superior to outlines in improving learning from text. *Journal of Educational Psychology*, *87*(3), 455-467.

Robinson, D. H., & Schraw, G. (1994). Computational efficiency through visual argument: Do graphic organizers communicate relations in text too effectively?. *Contemporary Educational Psychology*, *19*(4), 399-415.

Rowland, C. A. (2014). The effect of testing versus restudy on retention: a meta-analytic review of the testing effect. *Psychological Bulletin, 140*(6), 1432-1463.

Roediger III, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*(1), 20-27.

Roediger III, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*(3), 249-255.

Rugg, M. D., & Wilding, E. L. (2000). Retrieval processing and episodic memory. *Trends in Cognitive Science, 4*(3), 108-115.

Rumelhart, D. (1980). Schemata: The building blocks of cognition. In R. Spiro, B. Bruce, & W. Brewer (Eds.), *Theoretical issues in reading comprehension* (pp. 33-58). Hillsdale, N.J.: Erlbaum.

Rumelhart, D. E., & Norman, D. A. (1978). Accretion, tuning, and restructuring: Three modes of learning. In J. W. Cotton & R. Klatzky (Eds.), *Semantic factors in cognition* (pp. 37-53). Hillsdale, N J: Erlbaum.

Sadowski, C. J. (1993). An examination of the short Need for Cognition Scale. *Journal of Psychology, 127*(4)*,* 451-454.

Sadowski, C. J., & Gulgoz, S. (1992). Internal consistency and testretest reliability of the Need for Cognition Scale. *Perceptual and Motor Skills, 74*(2)*,* 610-610.

Salomon, G., Perkins, D. N., & Globerson, T. (1991). Partners in cognition: Extending human intelligence with intelligent technologies. *Educational Researcher, 20*(3), 2-9.

Schiefele, U., & Krapp, A. (1996). Topic interest and free recall of expository text. *Learning and Individual Differences, 8*(2), 141-160.

Schneider, M. C., Huff, K. L., Egan, K. L., Gaines, M. L., & Ferrara, S. (2013). Relationships among item cognitive complexity, contextual demands, and item difficulty: Implications for achievement-level descriptors. *Educational Assessment, 18*(2), 99-121.

Song, M. K., Lin, F. C., Ward, S. E., & Fine, J. P. (2013). Composite variables: when and how. *Nursing Research*, *62*(1), 45-49.

Sorrentino, R. M., Bobocel, D. R., Gitta, M. Z., Olson, J. M., & Hewitt, E. C. (1988). Uncertainty orientation and persuasion: Individual differences in the effects of personal relevance on social judgments. *Journal of Personality and Social Psychology, 55*(3)*,* 357-371.

Spotts, H. (1994). Evidence of a relationship between need for cognition and chronological age: Implications for persuasion in consumer research. *Advances in Consumer Research, 21*(1)*,* 238-243.

Stein, N., & Albro, E. (2001). The origins and nature of arguments: Studies in conflict understanding, emotion, and negotiation. *Discourse Processes, 32*(2-3), 113-133.

Tabachnick, B., & Fidell, L. (2007). *Using multivariate statistics* (5th ed.). Boston, MA: Allyn and Bacon.

Tabachnick, B., Fidell, L., & Ullman, J. (2019). *Using multivariate statistics* (7th ed.). New York: Pearson.

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, *2*, 53-55.

Tolentino, E., Curry, L., & Leak, G. (1990). Further validation of the short form of the Need for Cognition Scale. *Psychological Reports, 66*(1),321-322.

Toulmin, S. E. (1958). *The uses of argument.* Cambridge, Eng: University Press.

Toulmin, S. E. (2003). *The uses of argument* (updated edition). Cambridge: Cambridge University Press.

van Eemeren, F. H., & Grootendorst, R. (2004). *A systematic theory of argumentation: The Pragma-Dialectical Approach*. Cambridge: Cambridge University Press.

van Eemeren, F., Grootendorst, R., & Henkemans, F. (1996). *Fundamentals of argumentation theory: A handbook of historical backgrounds and contemporary developments.* Mahwah, N.J.: L. Erlbaum.

van Eersel, G. G., Verkoeijen, P. P., Povilenaite, M., & Rikers, R. (2016). The testing effect and far transfer: The role of exposure to key information. *Frontiers in Psychology*, *7*, 1-11.

Van Gelder, T. (2002). Argument mapping with reason! able. *The American Philosophical Association Newsletter on Philosophy and Computers*, *2*(1), 85-90.

van Gelder, T. (2015). Using argument mapping to improve critical thinking skills. In: M. Davies, R. Barnett (Eds.), *The Palgrave Handbook of Critical Thinking in Higher Education* (pp. 183-92). Palgrave Macmillan, Basingstoke.

Van Gog, T., & Sweller, J. (2015). Not new, but nearly forgotten: the testing effect decreases or even disappears as the complexity of learning materials increases. *Educational Psychology Review*, *27*(2), 247-264.

Veerman, A., Andriessen, J., & Kanselaar, G. (2002). Collaborative argumentation in academic education. *Instructional Science*, *30*(3), 155-186.

Verheij, B. (2005). Evaluating arguments based on Toulmin's scheme. *Argumentation*, *19*(3), 347-371.

Vekiri, I. (2002). What is the value of graphical displays in learning? *Educational Psychology Review, 14*(3), 261-312.

Venkatraman, M. P., Marlino, D., Kardes, F. R., & Sklar, K. B. (1990). Effects of individual difference variables on response to factual and evaluative ads. *Advances in Consumer Research, 17,* 761-765.

Venkatraman, M. P., & Price, L. L. (1990). Differentiating between cognitive and sensory innovativeness: Concepts, measurement, and implications. *Journal of Business Research*, *20*(4), 293-315.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Waller, R. (1981). Understanding network diagrams. Paper presented at the Annual Meeting of the American Educational Research Association, Los Angeles, USA.

Walton, D. (1989). *Informal logic: A handbook for critical argumentation.* Cambridge: Cambridge University Press.

Walton, D. (2013). *Methods of argumentation.* New York, NY: Cambridge University Press.

Walton, D., Atkinson, K., Bench-Capon, T., Wyner, A. and Cartwright, D. (2010) Argumentation in the framework of deliberation dialogue. In C. Bjola and M. Kornprobst (Eds.), *Arguing Global Governance* (pp.201-230). London: Routledge.

Walton, D., & Macagno, F. (2015). A classification system for argumentation schemes. *Argument & Computation*, *6*(3), 219-245.

Walton, D., Reed, C., & Macagno, F. (2008). *Argumentation schemes.* Cambridge, New York: Cambridge University Press.

Warren, J. E. (2010). Taming the warrant in Toulmin's model of argument. *English Journal, 99*(6), 41-46.

Webster, D. M., & Kruglanski, A. W. (1994). Individual differences in need for cognitive closure. *Journal of Personality and Social Psychology, 67*(6)*,* 1049-1062.

West, C., Farmer, J. and Wolff, P. (1991). *Instructional design: Implications from cognitive science*. Englewood Cliffs: Prentice Hall.

Winne, P. H., & Hadwin, A. F. (2013). nStudy: Tracing and supporting self-regulated learning in the Internet. In *International handbook of metacognition and learning technologies* (pp. 293-308). Springer, New York, NY.

Winne, P. H., Nesbit, J. C., & Popowich, F. (2017). nStudy: A system for researching information problem solving. *Technology, Knowledge and Learning*, *22*(3), 369-376.

Wolfe, C. R., & Britt, M. A. (2008). Locus of the myside bias in written argumentation. *Thinking & Reasoning, 14*(1), 1-27.

Wolfe, C. R., Britt, M. A., & Butler, J. A. (2009). Argumentation schema and the myside bias in written argumentation. *Written Communication, 26*(2), 183-209.

Wong, S. S. H., & Lim, S. W. H. (2019, April 25). From JOLs to JOLs+: Directing Learners' Attention in Retrieval Practice to Boost Integrative Argumentation. *Journal of Experimental Psychology: Applied*. Advance online publication. http://dx.doi.org/10.1037/xap0000225

Wong, S. S. H., Ng, G. J. P., Tempel, T., & Lim, S. W. H. (2019). Retrieval practice enhances analogical problem solving. *The Journal of Experimental Education*, *87*(1), 128-138.

Appendix A.

Wind Power Text

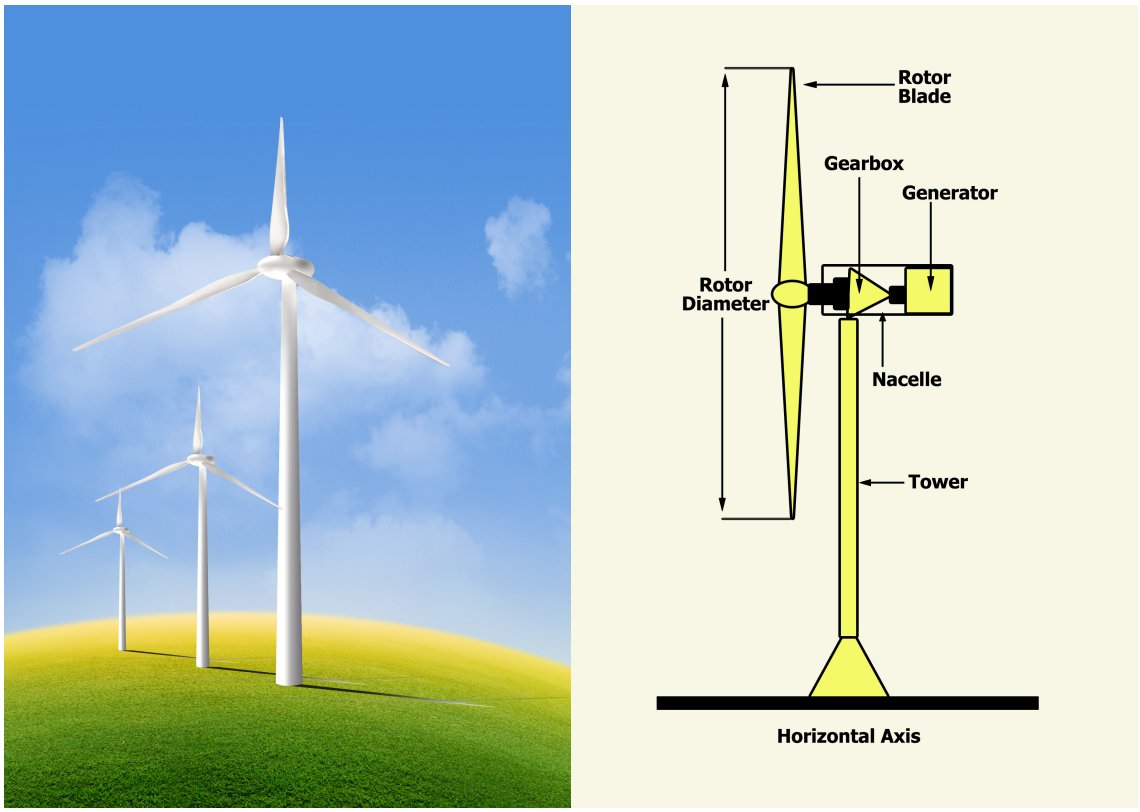
Wind Power

Wind is the movement of air across the earth's surface. Differences in air pressure are created because the sun heats the surface of the earth unevenly. When the heated air rises cooler air moves in to fill the void, causing wind to flow from high pressure areas to low pressure areas. Wind will be produced as long as the sun continues to shine. Humans have exploited wind power for thousands of years, using sails to propel ships and windmills to pump water. However, wind energy fell out of favor in the 20th century because it is inherently unpredictable. Wind is often thought to be too unreliable to provide a constant minimum or "baseload" power supply. As a result, modern societies tend to rely intensively on the fossil fuels of coal, oil, and natural gas to provide heat and power for their homes and industries. In recent decades, wind has started to make a comeback. There are several reasons for this shift: fossil fuels are the primary driver of global climate change; they are finite resources that will be depleted; many countries are worried about the security risks of depending on imported fossil fuels that are subject to volatile pricing.

Wind Turbines

Electricity can be generated by wind turbines that use large, specially-shaped blades to harness moving air. About 90% of the wind turbines now in use are horizontal-axis wind turbines (HAWTs). As you can see from Figure A1, they look like the medieval windmills that were used to grind flour. The blades rotate on an axis or shaft that is parallel to the ground and supported by a tower. Wind pushing on the blades causes the entire blade assembly, called the rotor, to spin around a central hub, called the nacelle. Fixed at the top of the support tower, the nacelle contains a gearbox which converts the low-speed rotational force of the rotor into high-speed rotational force that is powerful enough to run an electrical generator also housed in the nacelle. Other wind turbines work in the same way, except the rotors spin around a central vertical tower and generate power for a gearbox and generator located at the base of the tower. They are called vertical-axis wind turbines (VAWTs).

Figure A1. Horizontal-Axis Wind Turbines



A small wind turbine can provide enough power for a single home, and a large one can power 600 homes under optimum conditions. The diameter of the rotor, including its blades, is the most important determinant of generating capacity. The larger the rotor, the more wind (and therefore energy) is captured by the turbine.

A small amount of carbon dioxide can be released during manufacture and maintenance of wind turbines, but wind turbines emit no carbon dioxide (CO_2) or other greenhouse gases while they are producing electricity. After they have been put into operation, wind turbines do not require the burning of fuel and generate no emissions.

Because wind turbines convert the kinetic energy of moving air into electrical energy, higher wind speeds generally allow them to produce more electrical power. However, wind turbines cannot operate at all wind speeds. The optimal wind speed for generating electricity is usually between 13 and 90 kilometers per hour. Wind turbines can be damaged if the winds are too strong.

NIMBYism

Many wind turbines are built to be very tall so they can access the stronger and more constant winds that blow at higher elevations. They also have long blades so they can operate with greater efficiency by having a larger area interact with the wind. The largest wind turbine in the world, the Enercon E-126, is 198 metres tall, which is as high as a 50-storey building. The most common type of wind turbine is between 40 and 60 metres tall. There are also smaller-scale wind turbines which are less than 25 metres tall. They are usually put on rooftops for residential use.

Local residents often have objections to erecting wind turbines near human habitations, resulting in delays in getting building permits and legal battles. One intensely debated aspect of wind energy is NIMBYism—which stands for Not In My Back Yard. It describes the attitude of people who may support an idea like renewable energy in principle, but are opposed to developments happening near where they live. With wind energy, NIMBYism most commonly takes the form of complaints about the noise wind turbines create and their visual impact on the landscape.

Standing tens or hundreds of metres tall, wind turbines can visually disfigure the surrounding landscape. Another concern for local residents is the noise created by wind turbines, both inaudible (i.e., low-frequency noise) and audible (i.e., high-frequency noise) to humans. Some doctors have raised the concern of “wind-turbine syndrome” caused by the low-frequency noise. Its symptoms are thought to include high-blood pressure, ringing in the ears, migraines, and other stress-related illnesses. However, wind turbine syndrome has not been scientifically studied. It has been suggested that anti-wind development activists coined this term to block the expansion of wind energy.

Wind Farm Sites

Wind farms, sometimes consisting of several hundred wind turbines, are built to generate a substantial amount of electricity. A large wind farm covers huge tracts of land, over hundreds of acres.

Finding the right geographical location for a wind farm needs attention to factors, especially wind speed. Sufficient wind speed is a necessity. Some of the major determinants of wind speed are pressure gradients, frictional forces, and elevation. A pressure gradient is caused by difference in air pressure between two adjacent locations.

The friction that the earth exerts on the air slows the wind speed. Surface features such as trees and mountains increase frictional force. Elevation also plays a role in wind speed because it determines the amount of friction exerted on wind. At high enough elevations there are no obstacles to slow down the wind.

These factors make certain areas the best candidates of wind farm sites such as hilly areas and open oceans. In most cases, trees have to be cut when proposing to construct a wind farm in a mountainous area. This may cause irreversible damage to the habitats of local species and even results in food web collapse. Another area particularly suitable for constructing a wind farm is the open ocean, where the stronger, more frequent, and more predictable winds develop as a result of the interaction between cooler ocean breezes and warmer continental winds. Given that the seafloor must be drilled to erect a wind turbine, injury or destruction of some sea creatures is inevitable. It may disturb the marine ecosystem. To avoid marine accidents, wind farms are strategically located away from busy harbors and shipping routes, but vessels are still put at risk during violent storms.

Agricultural areas are also ideal for siting wind farms. In the United States and Australia where crop belts usually overlap with wind belts, wind turbines are often constructed in or near agricultural areas. Farmers or ranchers are paid for leasing out sections of their land. In addition to collecting rents, wind turbines, as suggested by some scientists, help the surrounding crops grow. The rotating blades stir up the air and push more carbon dioxide (CO_2) to the crops in the field. Empowered by the advancement of technology, scientists and engineers are creating wind turbines at extremely high altitudes where the blades are able to interact with fast-moving winds like jet streams that blow through the stratosphere around 10 kilometers above the surface. Such a wind turbine would look like a kite tethered to the ground but float thousands of meters in the air to harness jet streams.

Apart from wind speed, other factors need to be taken into account when choosing a site for a wind farm. For example, is the supply of electricity able to meet the demand? Investigating at what times of day and what times of year the wind is most likely to blow is necessary before integrating wind energy into the power grid. In most areas, electricity is most demanded in the late afternoon and evening. People in Canada use more energy for home heating during winter than air-conditioning in summer.

Given the power of NIMBYism and the frequent objections of local residents, wind farms should be built away from densely populated areas to speed up the permitting process. To cut down costs, wind farms ought to be built in areas that can be accessed through existing logging or mine roads. Proximity to existing transmission lines is another factor that influences the selection of wind farm sites because constructing new transmission lines takes time and money. This concern looms larger for offshore wind farms and becomes an impediment to the growth of that sector.

Such requirements exclude a considerable number of potential sites for wind farms. It's estimated that about 75% of the investment needed to develop a wind farm consists of start-up, one-time costs. The high cost of getting the turbines up and running makes it difficult to finance wind farm projects. For this reason, deliberate planning and investigation of where to construct a wind farm is a worthwhile undertaking.

Appendix B.

Pretest on Free Recall Ability

Like the other fossil fuels, coal and oil, natural gas is formed from the decayed remains of plant and animal life. The organic material is covered, compacted and pressurized by layers of sand and rock over tens of millions of years. In the case of natural gas, as well as oil, it is usually algae and zooplankton layered on the ocean's floors that formed the basis of the thick organic deposits that are the oil and gas we use today.

If the organic matter descends far enough into the Earth's crust and reaches a temperature of 120˚C, it begins to cook. Eventually the carbon bonds in the organic matter break down and fossil fuels are formed. Natural gas formed in this way is called *primary gas*. If oil, once it has formed, continues cooking for millions more years, it too can degenerate into natural gas. This is known as *secondary gas.* This implies that the deeper the fossil deposits, the more heat and pressure they are subjected to, and the more likely they are to be natural gas, instead of oil. It also means oil fields are usually accompanied by natural gas deposits.

Micro-organisms called methanogens can create gas. Methanogens work to break organic matter down into methane and are found in places devoid of oxygen, such as beneath the Arctic permafrost, or in the stomachs of animals (i.e. cows). Gas produced this way is called *biogenic gas*. As this occurs very near the Earth's surface, biogenic gas is usually released directly into the atmosphere.

Question:

Write down everything you can remember from the natural gas text you have just read. Try to write in complete sentences but do not worry about spelling or grammar.

Appendix C.

Delayed Posttest Questions

Free Recall Question

Write down everything you can remember from the WIND POWER text you previously studied in this experiment. Try to write in complete sentences but not to worry about spelling or grammar.

Short-Answer Questions

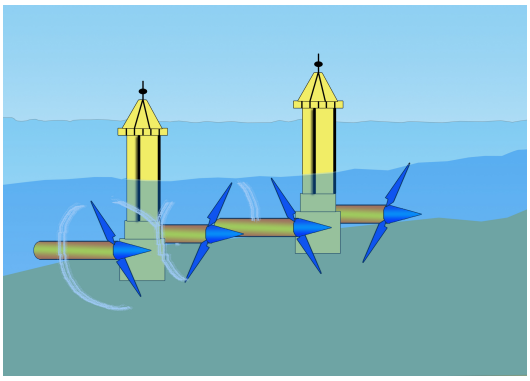
Q1:

To improve cell phone communication in West Kelowna, Rogers planned to build a 20-metre cell phone tower in the vineyard of a small winery. However, the proposal was opposed by many nearby residents. Close to 200 names were collected on a neighborhood petition opposing the proposed tower. Why do you think the residents objected to the proposal?

Q2-Q5:

Tidal energy, which harnesses the ebb and flow of the tides to produce power, is an alternative energy to fossil fuels. Tides are created by the gravitational pull of the moon and sun, combined with the rotation of the earth. The tide's rise and fall move on a predictable, daily schedule (once or twice a day depending on location). One way to exploit tidal energy is by sinking turbines to the sea floor, as shown below. Tidal turbines are in many cases set up in groups to allow more energy production.

Tidal Energy Turbines



(Q2) Can you imagine how a tidal turbine works to produce power? What is its basic principle of operation?

(Q3) Which organization is most likely to resist the building of tidal farms? Why?

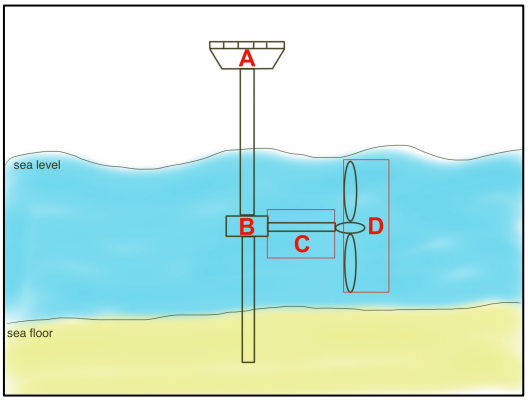
A. Children's Future International

B. Habitat Conservation

C. Pollution Probe

D. Food and Agriculture Organization

(Q4) Which part of the tidal turbine is the most important determinant of generating capacity? Why?



(Q5) David is an energy consultant. He and his colleagues are assigned by their company to help the government search for optimal locations in Nova Scotia to build tidal power stations. To fulfill this task, what aspects or issues do they need to consider?

Argument Essay

Select the claim most consistent with your belief:

1) Tidal energy farms should be expanded in many locations around the work.

2) Tidal energy farms should not be expanded.

Write an argument essay that defends the position you select. Please write as much as you can to fill the text box below.

Appendix D.

Correlations between Time-On-Task and Outcome Measures

Time-On-Task	Group	Free Recall Score	Short-Answer Score	Argument Essay Score
Text Reading	Restudy	.51*	.34*	-.14
	Retrieval Practice	.60*	.14	.01
	Dialectical Map	.12	-.07	-.11
Intervention Activity	Restudy	.44*	.09	.001
	Retrieval Practice	.77*	.28	.26
	Dialectical Map	.44*	.44*	.10
Interval	Restudy	-.24	-.20	-.36*
	Retrieval Practice	-.15	-.25	-.05
	Dialectical Map	-.13	-.42*	-.14
Posttest	Restudy	.57*	.39*	.25
	Retrieval Practice	.67*	.34*	.33*
	Dialectical Map	.20	.07	.36*

*Correlation is statistically detectable at the .05 level (2 tailed).