RETRIEVAL-BASED LEARNING AND ELEMENT INTERACTIVITY:
THE ROLE OF PRIOR KNOWLEDGE

Zachary L. Buchin

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Psychology and Neuroscience (Cognitive).

Chapel Hill
2021

Approved by:

Neil W. Mulligan
Jennifer E. Arnold
Kelly S. Giovanello
Patrick R. Harrison
Joseph B. Hopfinger
ABSTRACT

Zachary L. Buchin: Retrieval-Based Learning and Element Interactivity: The Role of Prior Knowledge
(Under the direction of Neil W. Mulligan)

Retrieving information benefits later memory of that information typically to a greater degree than equivalent restudy. This direct effect of retrieval (i.e., the testing effect) has emerged as one of the most robust findings in cognitive psychology and has been shown to generalize across a range of materials, tasks, and contexts, leading many cognitive scientists to advocate for broad implementation in education. However, the transition from principle to practice has been challenging. Not only do educators call for further research on educationally relevant factors (e.g., prior knowledge), but a recent debate has emerged over whether retrieval practice can enhance complex, meaningful learning. Research grounded in cognitive load theory claims that the testing effect disappears or even reverses when learning tasks are complex (or high in element interactivity). Element interactivity considers the complexity of the materials in relation to the learner’s prior knowledge and can be manipulated by changing either the materials or the knowledge of the learner (i.e., it decreases as knowledge increases).

The current study experimentally manipulated element interactivity by holding material complexity constant and randomly assigning participants to prior knowledge conditions via three days of training (i.e., online lessons) in one of two academic domains. After training, participants studied new information from one domain before two rounds of either focused restudy of examples/key ideas or retrieval practice of short-answer questions with elaborative feedback.
(repeated for the second domain). Thus, learning tasks were either lower (trained topics) or higher (untrained topics) in element interactivity. Although rated as more effortful during learning, retrieval practice led to significantly greater overall performance on a final delayed test than restudy. Critically, despite a substantial effect of prior knowledge (and a clear reduction in element interactivity), there was no interaction between learning strategy and prior knowledge. Nearly identical testing effects for trained and untrained topics provides evidence against the idea that prior knowledge and element interactivity represent significant boundary conditions of retrieval-based learning. Students with higher or lower levels of prior knowledge will similarly benefit more from retrieval practice (short-answer questions with feedback) than restudy (detailed examples/key ideas) when learning complex scientific information.
ACKNOWLEDGEMENTS

I would like to express my gratitude to the many people who have provided invaluable assistance during my graduate school career and throughout the completion of this project.

First, I want to thank my advisor, Dr. Neil Mulligan, for his substantial investment in my academic career. None of this would have been possible without his continuous guidance, expertise, and support throughout my time in graduate school.

Second, I would like to acknowledge my dissertation committee members, Dr. Jennifer Arnold, Dr. Kelly Giovanello, Dr. Patrick Harrison, and Dr. Joseph Hopfinger, for their insightful advice and feedback. I am grateful for their sincere dedication to the successful completion of this project.

Finally, I wish to thank my partner, Kaitlyn Oakley, for her unwavering support and encouragement over the past five years.
TABLE OF CONTENTS

LIST OF TABLES ........................................................................................................................ ix
LIST OF FIGURES ....................................................................................................................... x
LIST OF ABBREVIATIONS ....................................................................................................... xi

CHAPTER 1: INTRODUCTION .................................................................................................. 1
  Background and Motivation .............................................................................................. 1
  Disconnect Between Research and Practice .......................................................... 3
  Retrieval-Based Learning .............................................................................................. 5
  Theories .................................................................................................................. 6
  Indirect Effects of Retrieval ................................................................................... 7
  Ecological Validity of the Testing Effect ................................................................... 8
  Summary .............................................................................................................. 18
  Complex, Meaningful Learning ....................................................................................... 19
  Transfer of Retrieval-Based Learning ......................................................................... 20
  Cognitive Load Theory and Element Interactivity ............................................... 22
  Element Interactivity and the Testing Effect .......................................................... 26
  Summary .............................................................................................................. 34

CHAPTER 2: METHODS ........................................................................................................... 39
  Participants ....................................................................................................................... 39
  Materials and Design ...................................................................................................... 40
Prior Knowledge Rating Scales and Pre-Tests .................................................... 41
Training Topic Lessons ........................................................................................ 41
Domain Text Passages .......................................................................................... 42
Retrieval Practice Questions ................................................................................ 43
Restudy Examples/Key Ideas ............................................................................... 44
Mental Effort Rating Scale .................................................................................... 45
Final Test Questions ............................................................................................. 45

Procedure ......................................................................................................................... 46

Training Phase ............................................................................................................. 46
Learning Phase ............................................................................................................. 47
Testing Phase .............................................................................................................. 49

CHAPTER 3: RESULTS ............................................................................................................. 50

Prior Knowledge ........................................................................................................... 50
Subjective Ratings ..................................................................................................... 50
Pre-Test Scores .......................................................................................................... 51

Training Phase ............................................................................................................. 53
Topic Lesson Attempts ............................................................................................ 53
Topic Lesson Duration ............................................................................................. 53
Topic Lesson Knowledge Check Performance ......................................................... 55
Topic Lesson Test Performance ............................................................................... 55

Learning Phase ............................................................................................................. 56
Topic Text Passage Duration .................................................................................... 56
Retrieval Practice Performance ............................................................................... 58
**LIST OF TABLES**

Table 1 - Domain text characteristics split by topic domain ....................................................... 88
Table 2 - Prior knowledge ratings split by topic domain........................................................................ 89
Table 3 - Pre-test proportion correct split by topic domain: Mean (SD).............................................. 90
Table 4 - Training topic lesson attempts split by topic domain: Number of attempts..................... 91
Table 5 - Training topic lesson duration (min) split by topic domain ................................................. 92
Table 6 - Training topic lesson performance (proportion correct) split by topic domain: Mean (SD).......................................................................................................... 93
Table 7 - Learning phase topic text passage duration (s) split by topic domain............................... 94
Table 8 - Learning phase topic text passage duration (s) split by topic training: Mean (SD)............. 95
Table 9 - Descriptive statistics for the untrained topic in the trained domain: Mean (SD)................. 96
Table 10 - Retrieval practice proportion correct split by topic domain: Mean (SD)......................... 97
Table 11 - Subjective mental effort ratings split by topic domain: Mean (SD).................................. 98
Table 12 - Subjective mental effort ratings split by topic training: Mean (SD).................................. 99
Table 13 - Testing phase final topic test duration (s) split by topic domain: Mean (SD)............... 100
Table 14 - Testing phase final topic test duration (s) split by topic training: Mean (SD).............. 101
LIST OF FIGURES

Figure 1 - Sample learning phase retrieval and restudy materials ............................................. 102
Figure 2 - Sample learning and testing phase near transfer materials ....................................... 103
Figure 3 - Overview of entire experimental procedure.................................................................. 104
Figure 4 - Overview of learning phase for the A-B domain order condition ............................ 105
Figure 5 - Final test proportion correct (+/- SE) as a function of learning condition and element interactivity (prior knowledge), averaging across question type.................. 106
Figure 6 - Final test proportion correct (+/- SE) on retention questions as a function of learning condition and element interactivity (prior knowledge).............................. 107
Figure 7 - Final test proportion correct (+/- SE) on near transfer questions as a function of learning condition and element interactivity (prior knowledge)............................ 108
Figure 8 - Final test proportion correct (+/- SE) on far transfer questions as a function of learning condition and element interactivity (prior knowledge)............................ 109
Figure 9 - Final test testing effect (+/- SE) as a function of element interactivity (prior knowledge) and question type ................................................................. 110
LIST OF ABBREVIATIONS

HPK  High Prior Knowledge
LPK  Low Prior Knowledge
WMC  Working Memory Capacity
CHAPTER 1: INTRODUCTION

Background and Motivation

In the past decade, the testing effect has not only emerged as one of the most robust findings in cognitive psychology but has also been shown to have important educational implications (e.g., Karpicke, 2017; Roediger, 2013; Roediger & Pyc, 2012). Specifically, the testing effect refers to the widely documented finding that retrieving information on a test improves later memory to a greater degree than simply restudying or rereading the same material (e.g., Roediger & Butler, 2011; Roediger & Karpicke, 2006). Retrieval practice can also benefit learning through a number of indirect processes, such as informing the learner of potential gaps in their knowledge (e.g., Roediger, Putnam, & Smith, 2011), reducing rates of proactive interference (e.g., Szpunar et al., 2007, 2008), enhancing the benefits of subsequent restudy (e.g., Arnold & McDermott, 2013), and helping instructors see what their students do and do not know (e.g., McDaniel et al., 2007). Importantly, the benefits of retrieval-based learning have been documented in both highly-controlled lab studies as well as in actual classroom settings (for recent reviews, see Adesope et al., 2017; Karpicke, 2017; Pan & Rickard, 2018; Rowland, 2014; Yang et al., 2021). In a large-scale review of effective learning strategies, Dunlosky et al. (2013) rated retrieval practice high in both utility and adaptability, advocating for broad implementation in education.

However, as with any lab-based principle, the transition to classroom-ready technique is quite challenging (e.g., Daniel, 2012; Daniel & Poole, 2009; Roediger & Pyc, 2012). Although testing effects have been found using a variety of materials, including foreign-language word
pairs (e.g., Carpenter et al., 2008), text passages (e.g., Agarwal et al., 2008), spatial maps (e.g. Carpenter & Pashler, 2007), and multimedia presentations (e.g., Johnson & Mayer, 2009), recent research has sparked a debate over the ability of retrieval practice to enhance “complex, meaningful learning” (e.g., Chan et al., 2018; Leahy & Sweller, 2019; van Gog et al., 2015; van Gog & Sweller, 2015; cf. Karpicke & Aue, 2015; Rawson, 2015). Specifically, it has been claimed that the testing effect decreases (or even disappears) when attempting to learn complex materials, specifically those that are high in element interactivity (van Gog & Sweller, 2015). Element interactivity, discussed in more detail later, refers to the number of elements in the to-be-learned information that are logically related and must be processed simultaneously in working memory (e.g., Chen et al., 2018; Sweller, 2010; van Gog & Sweller, 2015). Importantly, this concept differs from other measures of learning complexity because it not only considers the characteristics of the to-be-learned information but also the prior knowledge of the learner (i.e., as prior knowledge levels increase, the level of element interactivity of the same information decreases; Leahy & Sweller, 2019). Since complex (or high element interactivity) learning tasks are most relevant to education (e.g., van Gog & Sweller, 2015), an understanding of this potential boundary condition is critical to convincing wary educators and students of the benefits of testing.

For several reasons, I believe it is important to take these recent critiques of retrieval-based learning seriously and conduct additional research to assess their validity. First, if true and as mentioned above, this would represent a critical boundary condition of retrieval-based learning. Second, if not true, the claim adds confusion and ambiguity to the educational recommendations of cognitive scientists. As detailed below, there is already a large disconnect between the recommendations of researchers and the strategies used by both students and
practitioners (e.g., Ambrose et al., 2010; Weinstein et al., 2018). Importantly, research on retrieval-based learning has been implicated as a potential bridge between cognitive science and education (e.g., Agarwal et al., 2012; Karpicke & Blunt, 2011; Karpicke & Grimaldi, 2012; McDaniel & Little, 2019; Nosofsky & McDaniel, 2019; Nunes & Karpicke, 2015; Yang et al., 2021). Third, Sweller and colleagues are prominent educational psychologists who have developed one of the most influential and widely cited theories of instruction – cognitive load theory (Sweller et al., 1998). To help bridge the principle-to-practice gap, it is critical to experimentally evaluate the effects of element interactivity on retrieval-based learning through the lens of cognitive load theory. The following section presents an overview of how students and educators currently view and utilize retrieval-based learning and other evidence-based learning strategies.

**Disconnect Between Research and Practice**

Blasiman et al. (2017) assessed students’ intended and actual use of specific study strategies and found less than encouraging results. Specifically, while students intended to rely most on rereading, copying notes, using flashcards, and practice testing, the latter two ended up being some of the most infrequently used strategies (Blasiman et al., 2017). The reported overuse of rereading and underuse of practice testing is consistent with previous surveys on study strategy usage (e.g., Gurung, 2005; Karpicke et al., 2009). Even those students who do use retrieval-based activities may do so for assessment, rather than learning, purposes. For example, although Morehead and colleagues (2016) found that students regularly used both practice testing and rereading strategies, only 31% of students believed that retrieval practice was more beneficial to learning. Earlier survey studies found even lower rates of endorsement, specifically 18% (Kornell & Bjork, 2007) and 27% (Hartwig & Dunlosky, 2012). Similarly, Blasiman et al.
(2017) asked students to rate the effectiveness of a variety of study strategies and found that while rereading notes was the highest rated strategy, practice testing was one of the lowest.

Clearly, students need instruction on the relative effectiveness of certain strategies (e.g., the mnemonic benefits of retrieval). However, research also suggests that educators also hold beliefs that are inconsistent with findings from cognitive and educational research (e.g., Glogger-Frey et al., 2018). For example, Morehead et al. (2016) found that 91% of surveyed instructors believed that students have different learning styles (e.g., visual vs. auditory) and 77% tailored their instruction to accommodate those differences, despite little-to-no supportive evidence of their existence (e.g., Kirschner & van Merriënboer, 2013; Pashler et al., 2008). Interestingly, only 58% of students endorsed the concept of learning styles (Morehead et al., 2016). Further, only 19% of instructors, compared to 31% of students, endorsed the idea that retrieval-based activities can enhance learning more than rereading (Morehead et al., 2016). Thus, it is not surprising that 41% of surveyed instructors reported recommending rereading as a study technique (Morehead et al., 2016).

This disconnect between cognitive research and educational practice can also be seen in current textbooks on educational psychology and instructional methods. For instance, Pomerance and colleagues (2016) evaluated the coverage of six fundamental instructional strategies in 48 teacher education textbooks and found a severe lack of discussion on evidence-based strategies. Specifically, no strategy was covered by even half of the textbooks and the use of assessment as a study strategy (i.e., retrieval-based learning) was not mentioned in a single textbook (Pomerance et al., 2016). In a similar evaluation of Dutch and Flemish university teacher education textbooks, Surma et al. (2018) found that 84% of sampled textbooks did not refer to retrieval practice as a learning strategy and even fewer provided prescriptive applications or
discussed primary scholarly evidence. Despite the frequent recommendations over the past
decade or so to implement retrieval-based learning (e.g., Roediger & Pyc, 2012), there is clearly
more work to be done to bridge the gap between principle and practice.

**Retrieval-Based Learning**

The beneficial effects of retrieval on learning have been sporadically documented for the
past 100 years (e.g., Abbot, 1909; Gates, 1917; Spitzer, 1939) but not extensively investigated
until more recently (e.g., Carpenter & DeLosh, 2006; Chan, 2009; Kang et al., 2007; Roediger &
Karpicke, 2006). As previously noted, taking a test can enhance later memory of the retrieved
information (i.e., the direct effects of testing) as well as benefit learning in a number of other
ways (i.e., the indirect effects of testing; e.g., Karpicke, 2017; Roediger, Putnam, & Smith,
2011).

The typical paradigm used to assess the testing effect consists of three general phases: (1)
an initial study phase; (2) a learning (or re-learning) phase that involves either retrieval practice
or some comparison condition (e.g., restudy); and (3) the final test phase. For example, Roediger
and Karpicke (2006) had students initially study a brief educational passage (phase 1) before re-
learning the material either by restudying the text three more times, restudying the text two more
times followed by one free recall period, or freely recalling the text in three consecutive periods
(phase 2). One week later, students took a final free recall test (phase 3). The authors found that
performance was much better for those who took practice free recall tests during phase 2
compared to those who simply restudied the texts (Roediger & Karpicke, 2006). Importantly, this
benefit emerged in the absence of feedback during phase 2, indicating that the act of taking a test
directly enhanced learning of the tested material (Roediger & Karpicke, 2006). This direct effect
of testing on learning has been replicated numerous times using various materials, contexts, and participants (e.g., Karpicke, 2017; McDermott, 2021; Rowland, 2014).

**Theories**

Though many accounts of the testing effect have been proposed over the past few decades, not all of them specify a specific mechanism that may drive retrieval-based learning (for reviews, see Karpicke, 2017; McDermott, 2021). One set of theories focuses on the increased effort required by retrieval practice compared to restudy. For example, the desirable difficulties framework (Bjork, 1994, 2017) generally states that difficult tasks yield more learning than easier tasks. Similarly, the retrieval effort hypothesis (Pyc & Rawson, 2009) also implicates increased effort as important to the testing effect. Specifically, difficult retrieval practice enhances learning more than easier retrieval practice, but both are more beneficial than restudying, which does not involve any effortful retrieval (e.g., Carpenter & DeLosh, 2006; Pyc & Rawson, 2009). Other frameworks have also been proposed to explain and predict the results of testing effect studies, including a focus on the match/mismatch between the type of processing during learning (phase 2) and during testing (phase 3; i.e., transfer-appropriate processing, Morris et al., 1977) and on the different effects of restudy and retrieval practice on the distribution of initial item memory strength (phase 1), with restudy moderately increasing the memory strength of all items and retrieval practice greatly increasing the memory strength of only those items actually retrieved (phase 2; i.e., bifurcation model, Kornell et al., 2011).

Although these frameworks are helpful when describing the effects of retrieval, they do not implicate a specific mechanism driving those effects (Karpicke, 2017; Karpicke et al., 2014).

Two more recent theories do specify a particular mechanism that underlies the mnemonic benefits of retrieval. First, the elaborative retrieval hypothesis (Carpenter, 2009) states that
attempting to retrieve a target results in activation of semantic associates of the cue and target. On a final test, those elaborations/associates can act as additional retrieval routes for accessing the target, resulting in enhanced performance (Carpenter, 2009, 2011; Carpenter & Yeung, 2017). Because restudying does not involve a memory search, no such elaborations are activated to aid in later retrieval (Carpenter, 2009). Second, the episodic context account (Karpicke et al., 2014) argues that successful retrieval updates the information’s context representation to include features of the present test context along with the prior features of the original study context. Restudying information does not cause contextual updating because context reinstatement is not required as it is with retrieval. The resulting composite trace in the retrieval condition provides varied contextual information that is more likely to match whatever contextual cues are used during the final recall test, restricting the search set of candidate information (Karpicke, 2017).

Although both of these theories specify a mechanism underlying the testing effect, no present theory or framework can fully explain the entirety of results in the literature (e.g., see McDermott, 2021). It is possible that multiple mechanisms are responsible for the testing effect, with the potential importance of each depending on the specific nature of the learning task or paradigm. Perhaps trying to explain the broad range of testing effects is too difficult for a theory specifying a single underlying mechanism. It is also important to note that many of these accounts and frameworks are not mutually exclusive and that additional theoretical frameworks continue to be developed (e.g., dual-memory model, Rickard & Pan, 2018).

**Indirect Effects of Retrieval**

As noted above, testing can also indirectly enhance learning through a number of different processes (e.g., Roediger, Putnam, & Smith, 2011). In the classroom, retrieval practice can be used to show both the student and the instructor potential gaps in understanding, acting
like a formative assessment (e.g., Black & William, 2009; McDaniel et al., 2007). Similarly, testing has been shown to benefit the accuracy of learners’ metacognitive monitoring and predictions of future performance (e.g., Barenberg & Dutke, 2018; Roediger, Putnam, & Smith, 2011). Administering frequent low-stakes quizzes motivates students to study as well as reduces their test anxiety (e.g., Agarwal et al., 2014; Roediger, Agarwal et al., 2011).

Two more indirect effects of retrieval practice on learning warrant discussion. First, test-potentiated learning refers to the finding that attempting to retrieve information improves later encoding of that same information relative to restudying (e.g., Arnold & McDermott, 2013). In other words, the learning benefits that arise from restudying some material are enhanced if it follows an initial retrieval attempt of that material (e.g., Arnold & McDermott, 2013). Thus, testing also indirectly benefits overall learning by increasing the mnemonic benefits of later restudying. Second, engaging in a retrieval attempt can also enhance the subsequent learning of new information (e.g., Chan et al., 2018). Whereas test-potentiated learning involves retrieval practice and later encoding of the same material, this indirect effect of testing, dubbed test-potentiated new learning, is seen when retrieval of some old information aids in the later learning of novel information (e.g., Cho et al., 2017; Szpunar et al., 2008). This rather incredible benefit of retrieval may be due to an increase in cognitive resources available to process the new information or possibly because of enhanced integration and binding of the new information (for review, see Chan et al., 2018). Although the specific mechanism(s) behind these indirect effects is debated, it is clear that retrieval practice can benefit learning in many ways.

**Ecological Validity of the Testing Effect**

Cognitive researchers frequently prescribe learning strategies and tactics that educators should incorporate into their classroom (e.g., Dunlosky et al., 2013; Roediger & Pyc, 2012).
However, educators often push back against these recommendations due to a lack of ecologically valid research (e.g., Daniel & Poole, 2009). Specifically, issues often arise because of the high degree of control in lab studies that may not apply to actual learners in actual classrooms as well as a lack of research on factors most relevant to educators (e.g., see Rittle-Johnson, 2019). The following section reviews the current state of research on the generalizability of retrieval-based learning in terms of factors critical to educational implementation (e.g., Daniel, 2012; Daniel & Poole, 2009; Mintzes et al., 2011; Moreira et al., 2019; Rittle-Johnson, 2019).

**Learning Materials and Tasks.** Benefits of retrieval practice have been found using a variety of verbal materials, ranging from simple word lists and word pairs (e.g., Carpenter, 2009; Carpenter & DeLosh, 2006) to more ecologically valid materials, like foreign-language translations (e.g., Carpenter et al., 2008), general knowledge facts and academic concepts (e.g., Carpenter et al., 2009; McDaniel et al., 2012), and educational texts (e.g., Butler, 2010; Kang et al., 2007; McDaniel et al., 2009). Additionally, research has also found test-enhanced learning using visual materials, including face-name pairs (Tse et al., 2010), Chinese characters and Adinkra symbols (Coppens et al., 2011; Kang, 2010), virtual three-dimensional spatial layouts (Carpenter & Kelly, 2012), and educational lecture videos and multimedia presentations (e.g., Butler & Roediger, 2007; Johnson & Mayer, 2009). Thus, it is clear that a wide range of materials can benefit from retrieval-based learning.

Although research has typically compared the learning benefits of retrieval practice to restudying or rereading, this is not always the case. For example, Karpicke and Blunt (2011) compared the effectiveness of two learning strategies: free recall (retrieval practice) and concept mapping (elaborative studying). After initially studying an educational text, students used one of the two strategies before taking a final test one week later. The authors found that free recall
practice consistently led to higher performance on the final test compared to concept mapping
(Karpicke & Blunt 2011; see also Lechuga et al., 2015). Retrieval practice may also be more
effective than other forms of elaborative studying, such as the keyword mnemonic and mediator-
generation (Karpicke & Smith, 2012).

When considering the educational applicability of retrieval-based learning, it is important
to note that retrieval practice can also be incorporated into other learning strategies to potentially
increase their effectiveness. For example, Bae et al. (2019) found that learners benefited more
from generating practice test questions when a free recall task was added to the generation
strategy. Similarly, it was found that the mnemonic benefits of creating concept maps increased
when students were not allowed to simultaneously view the to-be-learned information (Blunt &
Karpicke, 2014). In other words, it seems that already beneficial learning strategies can be made
more effective by adding some form of required recall. Even the learning benefits from
answering practice short-answer questions have been shown to increase if students are not
allowed to use their book/notes while answering (Agarwal et al., 2008).

**Context and Setting.** Though the adaptability of retrieval practice is impressive, a
critical question remains – does it work in a real classroom? Recent meta-analyses of testing
effect classroom research indicate that it does. For example, Adesope et al. (2017) found similar
moderately large weighted mean effect sizes for classroom studies ($g = 0.67$) and lab studies ($g =
0.62$), and Yang et al. (2021) found a medium effect of classroom quizzing on overall academic
achievement ($g = 0.499$). Schwieren and colleagues (2017) conducted a meta-analysis of studies
assessing the testing effect in psychology classrooms and found an overall benefit of testing on
learning outcomes ($d = 0.56$). In a systematic review of test-enhanced learning in health
professions education, Green et al. (2018) found that 21 of 23 retention outcomes and 7 of 7
transfer outcomes favored test-enhanced learning over restudying (the difference between retention and transfer is discussed in a later section).

Retrieval-based learning has been implemented in the classroom in a number of different ways, such as frequent practice quizzes (e.g., Lyle & Crawford, 2011; McDaniel et al., 2012), self-questioning (i.e., students generate and answer relevant questions; King, 1992; Webster & Hadwin, 2012), and quizzing computer software (e.g., Butler et al., 2014; Pennebaker et al., 2013). One major implementation of retrieval practice can be seen in audience response systems (e.g., clicker questions), the frequent use of which has been shown to benefit student learning (e.g., McDaniel, Agarwal et al., 2011; Roediger, Agarwal et al., 2011). Of course, students can also use retrieval practice on their own, such as through flashcards, recitation, and answering adjunct questions (e.g., Dunlosky et al., 2013; Fiorella & Mayer, 2015, 2016; Kornell & Bjork, 2007).

**Learner Characteristics.** In addition to applied contexts, educators often point to individual differences as a critical factor that must be thoroughly explored before a principle can be incorporated into practice. Although research on retrieval practice and certain individual differences is unquestionably wanting, beneficial retrieval effects have been shown to generalize across various populations. The effectiveness of test-enhanced learning has been observed with students of all education levels, including preschool students (e.g., Fritz et al., 2007), elementary school students (e.g., Bouwmeester & Verkoeijen, 2011; Goossens et al., 2014; Lipowski et al., 2014), middle school students (e.g., McDaniel et al., 2011; Roediger, Agarwal et al., 2011), high school students (e.g., Marsh et al., 2009), college students (e.g., Roediger & Karpicke, 2006), and medical students (Kromann et al., 2009). Similarly, the benefits of retrieval practice also seem to extend across the entire age range, from early childhood to older adulthood (e.g., Coane,
Beneficial effects have also been observed with clinical populations, including multiple sclerosis patients (Sumowski, Chiaravalloti, & DeLuca, 2010), dementia sufferers (Haslam et al., 2011), traumatic brain injury patients (Pastötter et al., 2013; Sumowski, Wood et al., 2010), and HIV+ patients (Avci et al., 2017).

The research is less clear about other individual differences that may moderate the mnemonic benefits of retrieval. For example, some studies have found no difference in the size of retrieval practice benefits between students who differed on measures of vocabulary (Coane, 2013; Goossens et al., 2014), processing speed (Coane, 2013; Karpicke et al., 2016), and reading comprehension (Karpicke et al., 2016). On the other hand, studies have also reported larger benefits of multiple-choice pretesting (Pyburn et al., 2014) and guiding questions (Stiegler-Balfour & Benassi, 2015) for learners with lower reading comprehension. Further, although Brewer and Unsworth (2012) found larger benefits for learners with lower fluid intelligence and episodic memory ability, no differences were found in later studies (Jonsson et al., 2020; Pan et al., 2015; Robey, 2019).

Working memory capacity (WMC) has also been explored as a potential moderator of the testing effect (e.g., Brewer & Unsworth, 2012). WMC is defined as a learner’s ability to simultaneously process tasks and maintain to-be-learned information in working memory while also accessing and retrieving relevant information from long-term memory (e.g., Lusk et al., 2009; Unsworth & Engle, 2007). Research suggests that high-WMC students perform better than those with lower WMC on a variety of complex mental tasks and cognitive performance measures, such as long-term memory activation (Cantor & Engle, 1993), reading and language comprehension (Daneman & Carpenter, 1980; Just & Carpenter, 1992), lecture note taking
(Kiewra & Benton, 1988), mnemonic strategy effectiveness (Gaultney et al., 2005), as well as tasks requiring recall and application (e.g., Lusk et al., 2009). Interestingly, the majority of retrieval practice research has found no relationship between WMC and the size of the testing effect (Bertilsson et al., 2020; Bertilsson et al., 2017; Brewer & Unsworth, 2012; Coane, 2013; Jonsson et al., 2020; Pan et al., 2015; Wiklund-Hörnqvist et al., 2014). However, this is not always the case, with some studies reporting larger benefits for high- versus low-WMC learners (Medina et al., 2017; Tse & Pu, 2012) and other studies reporting the opposite pattern (i.e., larger learning benefits from retrieval practice for low- versus high-WMC learners; Agarwal et al., 2017). Taken together, different studies find that WMC has a positive, negative, or null relationship with the mnemonic benefits of retrieval, clearly representing an individual difference that requires further research.

**Prior Knowledge.** Even less certainty surrounds the limited research on the potential moderating influence of learners’ prior knowledge on the mnemonic benefits of testing (e.g., Dunlosky et al., 2013; Murphy & Pavlik, 2018). As previously noted, prior knowledge is an important factor in cognitive load theory because it, along with the complexity of the learning materials, determines the level of element interactivity of a learning situation or task (e.g., Leahy & Sweller, 2019).

Though many different classifications and types of prior knowledge exist, subject-matter domain knowledge, which involves the breadth and scope of one’s declarative, procedural, and conceptual knowledge of a field of study, is most relevant to the present study (e.g., Alexander, 1997; Dinsmore & Alexander, 2016). The importance of a learner’s prior knowledge can be seen in an oft-cited quote by Ausubel, “If I had to reduce all of educational psychology to just one principle, I would say this: the most important single factor influencing learning is what the
learner already knows. Ascertain this and teach [them] accordingly.” (1978, p. 235). Prior knowledge has been shown to affect many, if not all, learning processes, including comprehension (e.g., Bransford & Johnson, 1972, 1973; Stevens, 1980), recall (e.g., Langer & Nicolich, 1981), and transfer of learning (e.g., Wittrock & Cook, 1975). For example, activating students’ prior knowledge before instruction has been shown to facilitate integration, benefit later recall, and enhance learning of novel material (e.g., Chi et al., 1994; Wetzels et al., 2011).

Prior knowledge activation involves moving knowledge from long-term memory into working memory in order to integrate it with new information into an expanded or more accurate knowledge structure (Gagne, 1985). Importantly, research has shown that even if learners do possess sufficient relevant prior knowledge, it will only facilitate new learning if it is actually activated (e.g., Ambrose et al., 2010). Various strategies have been shown to effectively activate prior knowledge, ranging from instructor-provided reminders and minor prompts (e.g., Bransford & Johnson, 1972, 1973; Gick & Holyoak, 1980) to more complex strategies, like having students generate self-explanations (e.g., Chi et al., 1994) or answer elaborative interrogation “why” questions (e.g., Martin & Pressley, 1991; Woloshyn et al., 1992). Thus, retrieval-based activities clearly play a critical role in activating prior knowledge (e.g., Ambrose et al., 2010). However, just as certain learning strategies affect the extent to which prior knowledge is activated, the amount of prior knowledge a learner has can also impact the effectiveness of certain learning strategies (e.g., Kalyuga, 2007).

The moderating influence of prior domain knowledge on the benefits of certain learning strategies can be seen in aptitude-treatment interactions (e.g., Cronbach, 1957; Cronbach & Snow, 1977; Tobias, 1976, 1989) and the expertise-reversal effect (e.g., Kalyuga, 2007, 2011, 2014; Lee & Kalyuga, 2014). Specifically, a type of instruction or a certain learning strategy that
is effective for low-prior knowledge (LPK) learners (i.e., novices) may be ineffective or even counterproductive for high-prior knowledge (HPK) learners (i.e., experts; e.g., Jiang et al., 2017; Kalyuga, 2011). For example, novices may benefit from strategies that guide the construction of new schemas, whereas experts may benefit from instruction that guides the retrieval and use of already acquired schemas (Kalyuga, 2014). Whereas the benefits of some learning strategies, like concept mapping and self-explanation, are greater for LPK students, the benefits of other strategies, such as imagining and enacting (and to a lesser extent – summarizing and drawing), seem to be greater for HPK students (e.g., Ambrose et al., 2010; Fiorella & Mayer, 2015, 2016; McNamara, 2004).

Much less is known about the potential moderating influence of prior knowledge on the benefits of retrieval practice, though some preliminary ideas can be drawn from educational research on similar strategies. For example, Wetzels et al. (2011) found that the benefits of notetaking from memory were larger for high- vs. low-prior knowledge learners. Similarly, Kalyuga and colleagues (1998) observed an increase in the mnemonic benefits of answering problem-solving practice questions as the expertise of the learner increased. Thus, it seems reasonable to predict that the benefits of retrieval practice would be larger for learners with greater prior knowledge, though the limited research on this has produced mixed results.

The testing effect literature has operationalized prior knowledge in a variety of ways. Some studies use measures of academic achievement as a proxy for learners’ prior knowledge, such as grade point averages (GPAs), standardized test scores (e.g., SAT, ACT), and previous relevant course experience and performance (Broadbent & Poon, 2015). Pretests have also been used as a more objective measure of prior knowledge, though typically included to check for equivalent baseline knowledge between groups. However, some studies have included these
measures as covariates to determine if testing effects remain after controlling for differences in prior knowledge (e.g., Pan & Rickard, 2017). In general, the advantage of retrieval practice over restudy continues to emerge when controlling for GPA (e.g., Lloyd et al., 2018), relevant prior experience (e.g., Lloyd et al., 2018; Pan & Rickard, 2017; Weinstein et al., 2016), and pretest scores (e.g., Spreckelsen & Jünger, 2017). However, less research has explicitly compared the benefits of retrieval practice between groups of HPK and LPK learners.

In one recent study, Xiaofeng et al. (2016) had participants learn lists of psychology words for a later free recall test either through retrieval practice or elaborative study. Half of the participants were considered domain-novices (i.e., minimal psychology course experience) and half were considered domain-experts (i.e., substantial psychology course experience; Xiaofeng et al., 2016). On the final test, domain-experts outperformed domain-novices and learners who engaged in retrieval practice outperformed those who engaged in elaboration (Xiaofeng et al., 2016). Critically, the authors found no significant difference between experts and novices in the retrieval practice condition but found an expertise advantage in the elaboration condition (Xiaofeng et al., 2016). This suggests that the benefits of retrieval practice are not moderated by prior knowledge, at least when defined as previous relevant course experience (see also, Carroll et al., 2007). Xiaofeng et al. (2016) interpreted the differential effects of prior knowledge on the benefits of elaboration and retrieval practice as support for the episodic context account of the testing effect (e.g., Karpicke et al., 2014), which unlike the elaborative retrieval hypothesis (e.g., Carpenter, 2009), does not implicate semantic associations as a critical mechanism behind the testing effect.

Not all studies have found equivalent benefits between high and low prior knowledge learners. For example, Carpenter et al. (2016) found contrasting results in a recent classroom
study – students with higher course performance benefited more from retrieving than copying information, while the lower performers benefited more from copying than retrieving. This result aligns with a classic recall study conducted by Spitzer (1939), who found larger recall benefits for learners who scored higher on an initial pretest (see also, Marsh et al., 2009). As stated by Carpenter et al. (2016), this pattern is consistent with research on the expertise-reversal effect (e.g., Kalyuga, 2007; Lee & Kalyuga, 2014) and likely reflects the benefits of possessing greater baseline knowledge when learning via retrieval practice. However, the authors also note that high performers differ from their lower performing peers on many other factors (e.g., motivation, interest, and familiarity with retrieval-based learning strategies) which could not be disentangled from prior knowledge (Carpenter et al., 2016).

The finding that HPK students benefit more from retrieval practice than LPK students was replicated in a very recent classroom study using initial pre-test scores as a measure of prior knowledge (Francis et al., 2020). The authors compared final test performance on information that was practiced via multiple-choice quizzes, practiced via concept mapping, or not practiced at all (Francis et al., 2020). Most relevant to the current study, Francis and colleagues (2020) found that while students in the HPK group had greater performance on initially tested information compared to untested information, students in the LPK group did not. Although there were methodological issues that prevent unambiguous conclusions (e.g., did not experimentally manipulate prior knowledge; no restudy control condition; no control over, or measure of, at-home studying behavior), the results align with Carpenter et al. (2016), Spitzer (1939), and Marsh et al. (2009), and suggest that prior knowledge is an important boundary condition for retrieval-based learning (Francis et al., 2020).
However, other classroom studies have reported the opposite pattern of results, with repeated testing benefiting lower performers more than higher performers categorized by standardized test scores (Hernick, 2015) and course performance (Hattikudur & Postle, 2011; Hernick, 2015; see also Spreckelsen & Jünger, 2017). Another recent classroom study found that within individual students, the benefits of practice testing were greater for lower prior knowledge topics compared to higher prior knowledge topics (in terms of topic pretest scores; Cogliano et al., 2019). Clearly, there are conflicting findings regarding the effects of prior knowledge on the learning benefits of retrieval practice. These discrepancies may be due to the use of inconsistent prior knowledge measures (e.g., course performance, prior experience, student achievement, and pre-test scores), no restudy control condition, and/or a lack of experimental control. To truly assess the causal influence of prior knowledge on learning strategy effectiveness, and in turn the influence of element interactivity, it is necessary to experimentally manipulate learners’ prior knowledge by randomly assigning participants to different knowledge induction conditions, which is the main focus of the present study (e.g., Fyfe & Rittle-Johnson, 2016; Rey & Buchwald, 2011; Tobias, 2010).

**Summary**

Retrieval practice can both directly and indirectly benefit learning and clearly generalizes over many different educationally-relevant factors (e.g., learning materials and contexts, Dunlosky et al., 2013), though educators and researchers alike continue to cite prior knowledge as a critical individual difference that must be explored in future research (e.g., Dunlosky & Rawson, 2019; Fiorella & Mayer, 2015, 2016; Mayer, 2017; Murphy & Pavlik, 2018). Despite this need for additional research, the wide applicability of retrieval practice has sparked a number of articles recommending increased educational implementation (e.g., Agarwal et al.,
2012; Dunlosky et al., 2013; Karpicke & Blunt, 2011; Karpicke & Grimaldi, 2012; Nunes & Karpicke, 2015; Roediger, 2013; Roediger, Putnam, & Smith, 2011; Roediger & Pyc, 2012). However, others have questioned whether these recommendations have gone through the necessary contextual vetting and if retrieval practice can really promote complex, meaningful learning (e.g., Chen et al., 2018; Daniel, 2012; Daniel & Poole, 2009; Leahy et al., 2015; Mintzes et al., 2011; Talkhabi & Nouri, 2012; van Gog et al., 2015; van Gog & Sweller, 2015).

**Complex, Meaningful Learning**

The main assertion by Sweller and colleagues is that the majority of testing effect studies have not assessed the complex, meaningful learning that takes place in most educational settings (van Gog & Sweller, 2015). The distinction between “rote-learning” and “meaningful learning” is common in education research (e.g., Ausubel, 1977; Mayer, 2002), having taken many forms over time, such as “memorizing” and “understanding” (e.g., Bransford et al., 2000) as well as “surface/shallow processing” and “deep processing” (e.g., Dinsmore & Alexander, 2012; Marton & Säljö, 1976). Meaningful learning can be thought of as an active process in which a learner: (1) attends to the relevant to-be-learned information (i.e., selecting); (2) generates relations between different parts of the information (i.e., organizing); and (3) generates relations between the information and prior knowledge (i.e., integrating; Mayer, 2002, 2010, 2014; Wittrock, 1974, 1989).

Transfer tests act as an assessment of meaningful learning by requiring learners to make inferences and apply their knowledge using organized and integrated mental models (Wittrock, 1989; Mayer, 2014). Transfer, broadly defined as the productive use of prior learning in a novel context, is often associated with meaningful learning and is a paramount goal of education (e.g., Dudai et al., 2007; Karpicke, 2017; Son & Rivas, 2016; for review, see Pan & Rickard, 2018).
Thus, it is important to understand if the learning benefits of retrieval practice extend to these complex and educationally valid tests.

**Transfer of Retrieval-Based Learning**

In a recent meta-analysis, Pan and Rickard (2018) reviewed the research on transfer of test-enhanced learning and found an overall medium effect size ($d = .40, 95\% \text{ CI } [.31, .50]$) across all studies when testing was compared to a non-testing re-exposure condition (e.g., restudy). Despite finding overall benefits, there still were certain instances in which transfer did not occur (e.g., Hinze & Wiley, 2011; Pan et al., 2016; Pan & Rickard, 2017, 2018). Transfer seemed to be greatest across different test formats (e.g., practice test: fill-in-the-blank, cued-recall; final test: multiple-choice, recognition), with application (e.g., analysis or evaluation, compare and contrast, and prediction questions) and inference questions (e.g., bridging or integrating multiple pieces of information and uncovering an underlying concept), and when the final test used novel mediator or related word cues (Pan & Rickard, 2018). On the other hand, transfer was weakest to rearranged stimulus-response items (e.g., practice test: recall a definition when given the term; final test: recall a term when given the definition) and to initially studied items that were not practice-tested (Pan & Rickard, 2018). Interestingly, transfer of test-enhanced learning to problem-solving questions was observed on tests involving medical diagnoses (e.g., Larsen et al., 2013) but not on tests involving worked examples (e.g., van Gog et al., 2015; van Gog & Sweller, 2015).

Pan and Rickard (2018) also identified three factors that were critical to transfer of retrieval-based learning. First, the likelihood of positive transfer increased under elaborated retrieval practice, which involves the use of broad encoding methods (e.g., broad retrieval instructions and explanatory recall) and elaborative feedback (e.g., providing detailed,
explanatory feedback and thorough restudying). Second, transfer effects were more likely when performance on the initial practice test was high (Pan & Rickard, 2018).

Pan and Rickard (2018) identified response congruency as the third factor important to transfer of test-enhanced learning. Specifically, transfer was more likely for studies that had the same or substantially overlapping answers on the initial and final tests. This held for some of the transfer categories (i.e., across test formats, medical diagnoses problem-solving skills, and mediator and related word cues) but not for others (i.e., stimulus-response rearrangement, untested materials seen during initial study, application and inference questions, and worked examples problem-solving skills). Transfer can certainly be assessed with tests that have strong response congruency (e.g., recall a term from its definition on a practice test and then provide that same response for a novel application question on a final test), but it is likely that most application, inference, and problem-solving tests in educational settings contain a variety of questions that differ in response congruency strength.

Taken together, retrieval practice does seem to yield transferrable (or meaningful) learning, as long as certain factors are taken into account (e.g., initial practice test performance and elaborative feedback). However, Sweller, van Gog, and colleagues argue that the complexity of the learning task moderates the benefits of retrieval-based learning (e.g., Hanham et al., 2017; van Gog et al., 2015; van Gog & Sweller, 2015; cf. Karpicke & Aue, 2015; Rawson, 2015). Specifically, van Gog and Sweller (2015) claim that the testing effect is eliminated or even reverses with complex learning tasks that are more aligned with the ultimate goals of education (e.g., problem-solving). Importantly, the authors note that these null effects are not a consequence of the procedural nature of problem-solving worked example tasks but are instead due to the meaningfulness and complexity of the task in terms of element interactivity (van Gog
& Sweller, 2015). Thus, their conclusion is not simply that test-enhanced learning won’t occur on problem-solving tasks. Rather, their conclusion is much more important and over-arching – the benefits of retrieval practice are limited to low element interactivity learning tasks, which educators emphasize less than higher element interactivity learning tasks (e.g., Chen et al., 2018; van Gog et al., 2015; van Gog & Sweller, 2015). Clearly, and as suggested by van Gog and Sweller (2015), further discussion and research is needed on this potential boundary condition of element interactivity on test-enhanced learning. The following section provides a brief overview of the most relevant aspects of cognitive load theory before focusing specifically on element interactivity and the role of prior knowledge.

**Cognitive Load Theory and Element Interactivity**

In the late 1980’s, Sweller and colleagues first developed cognitive load theory to incorporate research on human cognitive architecture into instructional design (e.g., Sweller, 1988; Sweller & Chandler, 1994; Sweller et al., 1998). Two key components of that architecture were a limited capacity working memory, drawing on the work of Miller (1956), Baddeley (1992), and others, as well as an effectively unlimited long-term memory, incorporating research on expertise (De Groot, 1965; Ericsson & Kintsch, 1995) and schema development (Chi et al., 1982). According to cognitive load theory, when learning novel information, learners must first hold and process the elements of the to-be-learned information in working memory before it can be stored in long-term memory (e.g., Sweller, 2016). If the working memory load associated with these elements is too great, learning is hindered, and the information may not be successfully transferred into long-term memory. Though the exact limits of working memory are debated (e.g., 7 +/- 2 chunks, Miller, 1956; 4 +/- 1 chunks, Cowan, 2001, 2010; or up to ~20-30
seconds, Cowan, 1988), there is little debate that acute limitations in some form, capacity or duration, exist.

However, these limitations seem to be much less restrictive when the information is activated in or recalled from long-term memory, rather than encoded from the environment (e.g., Ericsson & Kintsch, 1995). In fact, large amounts of organized information, or schemas, can be activated in long-term memory (or long-term working memory; Ericsson & Kintsch, 1995) and used without adhering to the typical working memory constraints (e.g., Cowan, 2001, 2014; Lewandowsky et al., 2007; Paas & Ayres, 2014; Sweller, 2016). Thus, a major focus of cognitive load theory is to optimize learning and decrease cognitive load by taking advantage of this relationship between working memory and knowledge structure activation in long-term memory.

Sweller and colleagues define cognitive load as the demand on working memory resources for a specific learner engaged in a specific learning task (e.g., Kalyuga & Singh, 2016; Sweller et al., 2019). The amount of load experienced by a learner is determined by the level of element interactivity in the current learning situation and has been measured in a variety of ways, including subjective rating scales, physiological measures, and secondary task performance (for reviews, see Paas et al., 2003; Sweller et al., 2019; van Gog & Paas, 2008). For example, the most frequently used subjective rating scale, introduced by Paas (1992), asks learners to indicate their “perceived amount of mental effort” ranging from 1: “very, very low mental effort” to 9: “very, very high mental effort”, and has been shown to be highly reliable (e.g., Paas et al., 2003; van Gog & Paas, 2008; cf. de Jong, 2010). However, Sweller (2010) recommends these measures of cognitive load be used in conjunction with an analysis of element interactivity.

Element interactivity is used to approximate the complexity of to-be-learned information in relation to the learner’s prior knowledge (e.g., Chen et al., 2017; Sweller, 2010). Thus, it is
necessary to consider several factors, in addition to the complexity of the materials, when determining the level of element interactivity. Elements can be defined as anything that needs to be or has been learned, such as symbols, concepts, words, or phrases (e.g., Sweller, 2010). Any elements that must be processed simultaneously in working memory are said to be interactive (e.g., Chen et al., 2017). In other words, if individual elements can be learned without reference to other elements, then the material is considered to be low in interactivity; alternatively, if individual elements cannot be learned or understood in isolation, the material is considered high in interactivity (e.g., Sweller, 2010). For example, although learning the symbols of the periodic table is difficult, it is not considered high in element interactivity because each chemical-symbol can be learned without reference to any other (e.g., you can learn hydrogen-H independently of learning copper-Cu; Chen et al., 2017). On the other hand, solving algebra equations is considered to be high in element interactivity because each symbol must be processed simultaneously (e.g., $8 - x = 3$, solve for $x$, requires the learner to simultaneously consider each individual element as well as the relations among them; Chen et al., 2017).

However, element interactivity cannot be estimated without an understanding of the learner’s relevant domain knowledge. A learning situation that is high in interactivity for a domain novice might be considered low in interactivity for a domain expert (e.g., Leahy & Sweller, 2019). Specifically, learners who have already acquired schemas relevant to the learning task can leverage their prior knowledge to reduce working memory load by integrating multiple interacting elements into fewer elements or even into a single, higher-order element (e.g., Chen et al., 2017). Experts have well-organized domain-specific schemas that can facilitate the creation of mental representations (e.g., Chi et al., 1981; Chi & Glaser, 1985; Chiesi et al., 1979). For example, research on chess experts, who demonstrate superior mental updating and memory
for chess positions compared to novices, implicates domain-specific knowledge structures and learned memory skills as critical to their success (e.g., Chase & Simon, 1973; Ericsson et al., 2018; Ericsson & Kintsch, 1995). Sweller and colleagues emphasize this point – the ability to process domain-specific information in working memory is greatly enhanced by large amounts of organized relevant information in long-term memory because previously learned individual elements can be recalled as a single, higher-order element (e.g., procedure or schema; Chen et al., 2017; Rey & Buchwald, 2011). Thus, holding the to-be-learned information constant, the level of element interactivity associated with that information is predicted to decrease as the learner’s expertise increases (e.g., Chen et al., 2017; Sweller et al., 2019).

Proponents of cognitive load theory use the concept of element interactivity to generate predictions about the potential effectiveness of an instructional technique or learning strategy in a specific situation (e.g., Chen et al., 2017; Leahy & Sweller, 2019; Sweller et al., 2019). To reiterate, the central premise is that learning will be hindered if the level of element interactivity imposes a load that exceeds the processing capabilities of working memory (e.g., Chen et al., 2017). In these situations, learning can be enhanced if an instructional technique or learning strategy reduces the number of elements that the learner must simultaneously process (e.g., Chen et al., 2017).

A number of studies have found general support for these predictions. For example, studies have shown larger learning benefits from studying worked-out examples compared to solving equivalent practice problems (e.g., van Gog et al., 2015). Worked examples outline and describe the steps of a procedure that is needed to solve a specific type of problem. Again, this benefit has been explained by a reduction of element interactivity – learners do not need to hold multiple alternative problem states in working memory because those states and subsequent
moves are provided in the worked example (e.g., Sweller, 2010). However, this is only true in the case of domain-novices who lack the relevant schema in long-term memory. The pattern of results reverses for domain-experts – the worked example becomes redundant with the already learned problem schema, leading to unnecessary processing and therefore larger benefits from problem-solving practice (i.e., an expertise-reversal effect; Chen et al., 2017; Kalyuga, 2007, 2011, 2014; Lee & Kalyuga, 2014).

Proponents of cognitive load theory now consider the expertise-reversal effect to be a specific instance of the more general element interactivity effect (Chen et al., 2017). Specifically, novice-learners benefit more from studying worked-examples than from problem-solving practice because the former reduces the number of interacting elements that need to be simultaneously processed in working memory. Because novice-learners lack the relevant prior knowledge (or schema) necessary to efficiently solve the problems, engaging in repeated problem-solving practice imposes a much heavier load than does studying worked examples (Chen et al., 2017). On the other hand, expert-learners benefit more from problem-solving practice than studying worked examples because the relevant, previously learned schema can be recalled from long-term memory as a single, higher-order element in order to generate the solution. Studying worked examples is less beneficial for expert-learners because the provided procedures and solutions are redundant with those already retrieved from long-term memory and held in working memory (or activated in long-term working memory; e.g., Chen et al., 2017; Ericsson & Kintsch, 1995; Sweller et al., 2019).

**Element Interactivity and the Testing Effect**

In terms of the testing effect, cognitive load theory uses the same logic discussed above to make the following explicit predictions (e.g., van Gog & Sweller, 2015). If the to-be-learned
If the to-be-learned information is low in element interactivity (e.g., individual words, foreign language word pairs, or isolated facts) then retrieval practice will benefit learning more so than restudy. If the to-be-learned information is high in element interactivity (e.g., passages with interrelated idea units, electrical circuit diagnostic steps, or instructional texts on the mechanics of a hydraulic brake system) then restudy will benefit learning more so than retrieval practice (e.g., Chen et al., 2018). Importantly, the relative benefit of retrieval practice over restudy in both situations is predicted to increase as the learner’s level of expertise increases, which would decrease the overall level of element interactivity.

Taken together, cognitive load theory predicts that restudying information will be more beneficial to learning under higher element interactivity situations, but retrieval practice will be more beneficial under lower element interactivity situations (e.g., Chen et al., 2018; van Gog & Sweller, 2015). Element interactivity can be altered by changing one of two major factors while holding the other constant. First, holding learner prior knowledge constant, the level of element interactivity is lower when learning isolated vs. interrelated materials. Second, holding material complexity constant, the level of element interactivity is lower for learners with high vs. low prior knowledge. Regardless of which factor it is changed, the relative advantage of retrieval practice over restudy is predicted to decrease as the level of element interactivity increases (e.g., Chen et al., 2018; Hanham et al., 2017; van Gog et al., 2015; van Gog & Sweller, 2015).

Are the benefits of retrieval practice moderated by the level of element interactivity of a learning task? Van Gog and Sweller (2015) say yes, citing multiple studies that found no, or even a reverse, testing effect in high element interactivity learning situations (e.g., Leahy et al., 2015; van Gog et al., 2015; van Gog & Kester, 2012). A majority of those studies compared the learning benefits of studying worked examples to problem-solving practice (note that these are
the same studies reviewed by Pan and Rickard, 2018, that found no transfer of test-enhanced learning. Though not perfect proxies for restudy and retrieval practice, research on worked examples and problem-solving often does resemble the typical testing effect paradigm. Learners study some information, usually presented as a worked example (phase 1), before either reading through additional worked examples of similar problems or solving similar problems that require retrieval of the previously learned procedure (phase 2), and then take a final problem-solving test (phase 3).

To further support their claim that test-enhanced learning does not occur under high element interactivity situations, van Gog and Sweller (2015) also rated the level of element interactivity in prior testing effect studies. They note that most of the research that did find positive testing effects used materials low in element interactivity (van Gog & Sweller, 2015). Further, those testing effect studies that did use high element interactivity materials often used final tests that did not tap into high element interactivity knowledge (van Gog & Sweller, 2015). For example, final retention tests (e.g., cued-recall of foreign language word pairs or idea unit fill-in-the-blank questions) do not tap into as high a level of element interactivity as do final transfer tests (e.g., applying a previously learned procedure to a novel problem or integrating previously learned isolated idea units in order to make deductive inferences; van Gog & Sweller, 2015). Thus, even if the learning materials themselves are high in element interactivity, it is important to use a final test that taps into that level of interactivity to determine whether meaningful learning, as opposed to rote memorization, has occurred.

1 In response, Karpicke and Aue (2015) point to several studies that found test-enhanced learning under complex, seemingly high element interactivity situations, that were not included in van Gog and Sweller’s (2015) review (e.g., Butler, 2010; Chan, 2009; McDaniel et al., 2009).
However, certain issues with the cognitive load research on test-enhanced learning create ambiguity over the potential moderating role of element interactivity. One major challenge involves the objectivity and clear operationalization of element interactivity (for a detailed critique, see Karpicke & Aue, 2015). It is difficult to estimate the number of interacting elements in a specific learning situation, largely because the number is determined by both the complexity of the to-be-learned information as well as the learner’s prior knowledge (e.g., Chen et al., 2017). Although there is no absolute quantitative metric for measuring element interactivity, it can still be used as a relative assessment within an experiment, as long as influential factors such as prior knowledge are taken into account. However, this has not always been the case in the cognitive load research on the testing effect. For example, two studies that did find a reverse testing effect did not include a prior knowledge assessment (Hanham et al., 2017; Leahy et al., 2015). In other cases, pre-tests were used to check for equivalent baseline knowledge between groups (van Gog et al., 2015; van Gog & Kester, 2012), though scores were only included as a covariate in van Gog et al. (2011).

If individual differences such as prior knowledge are not examined, differences in average learning strategy effectiveness may mask and miss important effects (Snow, 1996). Finding no significant difference in mean pretest scores between groups does not mean that prior knowledge will not affect the outcome (e.g., Gruijters, 2016; van Breukelen & van Dijk, 2007). Further, if it is known a priori that a certain measure is associated with the dependent variable or that a substantial treatment by measure interaction may emerge (i.e., learning strategy by prior knowledge), that measure should be included as a covariate in the final analyses (e.g., Leppink, 2018, 2019). For example, and generally in line with the predictions of cognitive load theory, the effect of prior knowledge on the benefits of retrieval practice may be more pronounced than the
effect of prior knowledge on the benefits of restudy. Specifically, if the testing effect does not emerge under high element interactivity situations, then learners with higher prior knowledge within the retrieval practice group should benefit more than those with lower prior knowledge, and this difference should be larger than the same comparison within the restudy group. Thus, even if two groups have equivalent average pretest scores, learners’ prior knowledge can still affect learning, possibly even to different degrees depending on the specific learning strategy group, which can be masked by comparing averages (e.g., Leppink, 2018, 2019).

Additionally, including pretest scores as a covariate would increase power and thus the likelihood of finding a significant effect, if an effect truly exists (Gruijters, 2016; Leppink, 2018). For example, van Gog et al. (2015) conducted multiple studies that found no significant benefit of retrieval practice over restudying worked examples. However, there was a numerical advantage in favor of retrieval practice in seven of eight total comparisons (van Gog et al., 2015). Because the authors did not conduct a power analysis to determine adequate sample sizes, low power could have contributed to the null testing effect results (van Gog et al., 2015; for further discussion on this point, see Karpicke & Aue, 2015).

Along with prior knowledge, it seems critical to the predictions of element interactivity, and to cognitive load theory in general, to assess mental effort or load during learning. In terms of the testing effect, cognitive load theory predicts less load during worked example study (or restudy) than during problem-solving (or retrieval practice), although this was only found in some instances (i.e., Hanham et al., 2017, Experiments 1 and 2; van Gog et al., 2015, Experiment 3; van Gog et al., 2011). Interestingly, and in line with a desirable difficulties framework, Hanham et al. (2017) found significant testing effects only in those experiments in which participants rated the retrieval practice learning task as more difficult than the worked example
study task (Experiments 1 and 2). Additionally, two of the cognitive load studies that did find a reverse testing effect did not include mental effort measures (i.e., Leahy et al., 2015; van Gog & Kester, 2012). Thus, the claim that increased load or effort during learning reduces the size of the testing effect has not been adequately tested or substantiated.

In addition to problematic issues associated with element interactivity, other methodological features of some of the cognitive load studies have been found to limit the benefits of testing. For example, research on the direct effects of retrieval frequently observe larger testing effects when assessed on a delayed final test (e.g., days or weeks later) compared to an immediate final test, which has been said to be a consequence of low initial test performance (e.g., Karpicke, 2017; Roediger & Karpicke, 2006; Rowland & DeLosh, 2015). Therefore, it is important to assess initial test performance, especially when using an immediate final test.

Leahy et al. (2015), who did not report initial test performance, found a reverse testing effect on an immediate final test but not on a delayed final test. Initial test performance was also not reported by van Gog et al. (2011) nor Hanham et al. (2017). Further, although van Gog and Kester (2012) claim to have found a reverse testing effect on a delayed final test, both the restudy and retrieval practice groups took an immediate test before returning for a delayed final test, confounding the benefits of restudy with retrieval practice. No differences were found on the immediate final test, possibly because of low initial practice test performance (< 50%; van Gog & Kester, 2012). As previously discussed, Pan and Rickard (2018) found that high initial test performance, as well as elaborative feedback, are critical to observing transfer of test-enhanced learning.
Most of the cognitive load studies did not provide feedback to the retrieval practice groups during learning (i.e., van Gog et al., 2011; van Gog et al., 2015; van Gog & Kester, 2012), with Leahy et al. (2015) being the sole exception. Although Hanham et al. (2017) did not provide explicit feedback, the previously studied worked example was available and could be consulted during problem-solving/retrieval practice. Thus, it is not clear how much participants truly needed to recall during retrieval practice. Small sample sizes (n ≤ 10) also cloud the interpretability of their results (Hanham et al., 2017, Experiments 3, 4, and 5).

Taken together, inconsistencies in the prior cognitive load research on the testing effect warrant further investigation. First, the studies that did find a reverse testing effect have certain methodological limitations, such as: using an immediate test without reporting initial practice test performance (Leahy et al., 2015, Experiments 1 and 2; van Gog et al., 2011); confounding the benefits of restudy and retrieval practice (van Gog and Kester, 2012); and providing the to-be-retrieved procedure during retrieval practice (Hanham et al., 2017; Experiments 4 and 5). Second, no study provided feedback after retrieval practice and used a delayed final test (factors important to the testing effect) while also including a prior knowledge pretest and a mental effort assessment (factors important to element interactivity). Finally, and perhaps most importantly, none of the studies experimentally manipulated element interactivity within a single experiment (see Karpicke & Aue, 2015).

Before introducing the present study, it should be noted that some recent testing effect studies have started to include predictions, analyses, and/or discussions of element interactivity. For example, Eglington and Kang (2018) investigated the potential role of element interactivity in the ability of retrieval practice to benefit deductive inference, building upon earlier research by Tran et al. (2015). In four experiments, participants first studied sentences one at a time and
then either restudied or engaged in fill-in-the-blank retrieval practice (Tran et al., 2015). Although testing increased performance relative to restudying on a final sentence memory test, it did not benefit participants’ ability to answer inference questions that required integration across multiple sentences (Tran et al., 2015). van Gog and Sweller (2015) interpreted these results as providing additional evidence that retrieval practice is not beneficial under high element interactivity situations, such as when integrating information to make later inferential deductions.

Eglington and Kang (2018) conducted three experiments to understand why retrieval practice might not benefit deductive inference. Overall, they found that both fill-in-the-blank testing and free recall could improve later deductive reasoning, as long as the to-be-integrated information was presented simultaneously during retrieval practice (Eglington & Kang, 2018). In other words, if a final test requires the integration of previously learned discrete units of information, it is critical that relational processing not be hindered during practice (Eglington & Kang, 2018). Importantly, the authors concluded that this explained their results, and those by Tran et al. (2015), better than attenuation due to high element interactivity.

The role of element interactivity has been discussed in other recent retrieval-based learning studies, though the conclusions are mixed. For example, Yeo and Fazio (2019) assessed the benefits of retrieval practice when learning flexible procedures and proposed that different learning goals may explain the contrasting testing effect results better than moderation by element interactivity (e.g., Koedinger et al., 2012). On the other hand, Peterson and Wissman (2018) assessed transfer of test-enhanced learning in the domain of analogical problem-solving and used the concept of element interactivity to generate initial predictions that were ultimately supported. Roelle and Berthold (2017) concluded that the best way to explain the variable effects of incorporating retrieval into complex learning tasks is to not only consider cognitive, retrieval-
oriented theories (e.g., the episodic-context account; Karpicke et al., 2014), but also educational, knowledge construction-oriented theories (e.g., the generative theory of learning; Wittrock, 1989). The authors further note that learners’ prior knowledge (or mental representations) represents a key factor in determining when retrieval practice will and will not benefit learning (Roelle & Berthold, 2017; see also Roelle & Nückles, 2019).

Summary

Retrieval practice can clearly benefit learning, usually to a greater degree than other learning strategies. However, uncertainty remains over the potential moderating role of element interactivity. Since much of the learning in educational contexts is considered to be high in element interactivity (e.g., van Gog & Sweller, 2015), this is a critical factor to explore in relation to retrieval practice application. As noted by Rittle-Johnson, “instead of trying to apply our research to practice, we need to do research that is inherently relevant to and driven by the needs of practice.” (2019, p. 140). The present study aims to help bridge this principle-to-practice gap by: (1) assessing the impact of prior knowledge on the benefits of retrieval practice, which is a very understudied individual difference within the testing effect literature; (2) using cognitive load theory, one of the most influential educational psychology learning theories, to generate predictions of the effectiveness of specific learning strategies in educationally-relevant situations; and (3) manipulating element interactivity by experimentally inducing learners’ prior knowledge, while holding other aspects of learning complexity constant, in order to bring this educationally-relevant construct under experimental control.

Overview of the Present Study

The major goal of the present study was to assess the moderating influence of element interactivity on the benefits of test-enhanced learning. Element interactivity is usually
manipulated by varying the complexity of the to-be-learned information. As noted by Sweller and colleagues (2019), because element interactivity depends so heavily on the learner’s prior knowledge, it is difficult to obtain a precise measure of the number of interactive elements in a specific set of learning materials (e.g., Chen et al., 2017; Leahy & Sweller, 2019; van Gog and Sweller, 2015). This introduces uncertainty and subjectivity when manipulating element interactivity by varying the complexity of the learning materials such that one has a greater number of interactive elements than the other (e.g., Karpicke & Aue, 2015). Therefore, in the present study, the complexity of the learning materials was held constant and learners’ prior knowledge was experimentally manipulated. Not only does this circumvent the problematic issue of defining what is and isn’t an interactive element, it also provides the most theoretically relevant assessment of the predictions of element interactivity and cognitive load theory in general (e.g., Leahy & Sweller, 2019; Sweller et al., 2019).

Experimentally inducing prior knowledge by randomly assigning participants to different knowledge conditions is necessary to establish causal evidence of any potential moderating influence of prior knowledge on strategy effectiveness (e.g., Fyfe & Rittle-Johnson, 2016; Rey & Buchwald, 2011; Tobias, 2010). This is especially important because of problematic confounding variables (e.g., motivation, interest, WMC, and intelligence) inherent in methods that assume differences in prior knowledge based on education level and similar general characteristics (e.g., Sentz & Stefaniak, 2019), student achievement and course performance (e.g., Carpenter et al., 2016), and prior relevant experience (e.g., Fyfe & Rittle-Johnson, 2016; Rey & Buchwald, 2011). Even using pretest scores, arguably a more direct measure of prior knowledge, to categorize participants into different groups does not alleviate the issue of random assignment (e.g., Fyfe & Rittle-Johnson, 2016; Rey & Buchwald, 2011; Tobias, 2010).
Some prior studies have used prefamiliarization techniques to induce prior knowledge by randomly exposing some learners to relevant material but not others (Fyfe & Rittle-Johnson, 2016; Petersen & McNeil, 2013; Rey & Buchwald, 2011). However, random assignment of prior knowledge is rare. A recent meta-analysis of prior knowledge and general learning research found that only nine studies randomized participants into different prior knowledge manipulation groups, and only three of those also included a manipulation check (Simonsmeier et al., 2018). Therefore, the present study manipulated element interactivity by experimentally inducing prior knowledge in order to establish casual evidence and avoid the problematic issues associated with grouping by student achievement, course performance, prior experience, and pretest scores (e.g., Fyfe & Rittle-Johnson, 2016; Rey & Buchwald, 2011; Sentz & Stefaniak, 2019; Tobias, 2010).

Participants were trained in one of two academic domains: Sensation and Perception (Domain A) or Historical Geology (Domain B). These domains both fall within larger disciplines (i.e., psychology and geology, respectively) and are generally less well-known than some other domains within these disciplines. Additionally, beneficial effects of prior knowledge on learner comprehension have been well documented with scientific texts (e.g., Chi et al., 1981; Kendeou & van den Broek, 2007). After being trained, all participants initially studied and then either restudied or retrieved information from multiple topics within both domains. Two days later, participants took two final tests containing retention and transfer questions that assessed the previously studied and restudied/retrieved information.

If the benefits of retrieval practice over restudy depend on low element interactivity learning situations, as hypothesized by cognitive load theory (e.g., van Gog & Sweller, 2015), then the testing effect should be larger for topics within the trained, than untrained, domain. By increasing the level of learners’ prior knowledge and holding all other factors constant, the
element interactivity of the learning task will decrease. Thus, cognitive load theory predicts an interaction between prior knowledge and learning condition, such that the benefits of retrieval practice over restudy will be larger for HPK topics (i.e., lower in element interactivity) than for LPK topics (i.e., higher in element interactivity; e.g., Kalyuga, 2007; Lee & Kalyuga, 2014; van Gog & Sweller, 2015). This would also align with prior studies that found larger benefits of retrieval-based learning for higher performing students (Carpenter et al., 2016; Francis et al., 2020; Marsh et al., 2009; Spitzer, 1939).

However, it may be that retrieval-based learning enhances later performance to a greater degree than restudy regardless of the level of element interactivity. If so, the benefits of retrieval practice should be similar in size for trained (lower element interactivity) and untrained (higher element interactivity) topics (e.g., Carroll et al., 2007; Xiaofeng et al., 2016). This result would contrast with the claim that the testing effect disappears or even reverses as the level of element interactivity increases (e.g., Chen et al., 2018; Leahy & Sweller, 2019; van Gog & Sweller, 2015). Additionally, finding that the benefits of test-enhanced learning are unaffected by the degree of relevant background information would support testing effect theories that propose an alternative mechanism to semantic association and relational processing (e.g., the episodic context account; Karpicke et al., 2014).

A third possibility is that retrieval-based learning is even more effective for topics within the untrained domain (higher element interactivity) than those within the trained domain (lower element interactivity; e.g., Cogliano et al., 2019). Not only would this provide evidence against the claim that the benefits of retrieval practice emerge only under low levels of element interactivity (e.g., van Gog & Sweller, 2015), but it would also indicate that retrieval practice is clearly a viable strategy for meaningful learning of complex information (e.g., Karpicke & Aue,
2015). Although this finding would have clear educational implications, there is no obvious theoretical reason to expect this pattern of results.
CHAPTER 2: METHODS

Participants

To determine the needed sample size for each between-subjects learning condition, a power analysis was conducted using G*Power (Faul et al. 2007). Because the final test contained both retention and transfer questions, multiple meta-analyses were consulted. The average weighted mean effect size of the testing effect was $g = .69$ (Rowland, 2014, for retention intervals > 1-day), $g = .63$ (Adesope et al., 2017), and $d = .68$ (Pan & Rickard, 2018) on retention, and $g = .53$ (Adesope et al., 2017) and $d = .40$ (Pan & Rickard, 2018) on transfer. As a compromise, an effect size of $d = .60$ was used in the power analysis, which indicated that 60 participants were needed in each condition to detect an effect of that size with power = .90 and $\alpha = .05$ (Faul et al., 2007). Thus, a total of 128 participants were needed to satisfy counterbalancing conditions (see Appendix A for a breakdown of all counterbalancing conditions).

Undergraduate students from the University of North Carolina at Chapel Hill (UNC-CH) were either recruited through a participant pool and received PSYC 101 course credit ($n = 97$) or through email advertisements and received monetary compensation at a rate of $10 per hour ($n = 31$). Of the 31 participants recruited through email advertisements, 15 were trained in Sensation and Perception (seven in the restudy condition and eight in the retrieval condition) and 16 were trained in Historical Geology (eight in each learning condition).²

² Participants recruited through the participant pool were similar to those recruited through email advertisements. Specifically, both groups had similar prior knowledge ratings, $ts < 0.30$, $ps > .75$, pre-test scores, $ts < 1.3$, $ps > .20$, and final test performance, $ts < 0.75$, $ps > .48$. When recruitment group was added as a factor in the final analyses, none of the critical results changed and the main effect of recruitment method was non-significant, $Fs < 1.20$, $ps > .30$. Taken together, participants recruited through either method were similar in all critical comparisons.
Due to the multiple sessions, additional participants were recruited to account for attrition \((n = 79)\) and replacement due to non-compliance or performance issues \((n = 19)\).\(^3\) Half of the remaining 128 total participants were randomly assigned to be trained in Sensation and Perception \((n = 32\) per learning condition) and half in Historical Geology \((n = 32\) per learning condition). The majority of participants \((n = 89)\) identified as female (31 as male, one as non-binary, and one as trans-female; gender demographic information was missing from six participants). Most participants were between 18-21 years old \((M = 19.61, SD = 1.60, Median = 19, Range = 18-27;\) age demographic information was missing from nine participants).

**Materials and Design**

There were two major independent variables: (1) prior knowledge (high vs. low), manipulated within-subjects and (2) learning condition (restudy vs. retrieval), manipulated between-subjects. The prior knowledge manipulation occurred in the first phase of the experiment – the Training Phase. Participants were randomly assigned to be trained in three of four topics within one of two domains. The four topics within Sensation and Perception were: (1) Color Perception; (2) Auditory Perception; (3) Cutaneous Senses; and (4) Chemical Senses. The four topics within Historical Geology were: (1) Geologic Time; (2) Minerals; (3) Rocks; and (4) Isotopic Dating. These topics were chosen based on a variety of introductory textbooks to reflect

\(^3\) A total of 226 participants completed the initial part of the experiment, which consisted of the informed consent form, experiment instructions, and the pre-training prior knowledge ratings and pre-tests. Of those 226 participants, 79 were lost due to attrition, either dropping out from the experiment after completing: only the initial part \((n = 40)\); only one or two of the training lessons \((n = 19)\); only the training phase \((n = 17)\); or only the training and learning phases \((n = 3)\). Of the remaining 147 participants, 19 were replaced either before \((n = 8)\) or after completing the entire experiment \((n = 11)\). Reasons for replacement included: submitting one or more phases late \((n = 3)\); rushing (e.g., reading multiple paragraphs of text in less than 10 seconds) and/or skipping multiple entire sections/pages of questions during one or more phases \((n = 8)\); indicating that they took/used notes and/or looked info up online during one or more phases \((n = 5)\); computer/software issues that prevented them from continuing with one or more phases \((n = 2)\); and completing one or more phases on a mobile device instead of a desktop computer \((n = 1)\). Of these 19 replaced participants, 10 were trained in Sensation and Perception (4 restudy and 6 retrieval) and 9 were trained in Historical Geology (3 restudy and 6 retrieval).
major topics within each domain. Information for the domain pretests, training topic lessons, learning tasks, and final retention and transfer tests were taken from these same textbooks to reflect what would be learned in a typical college course.\textsuperscript{4}

\textbf{Prior Knowledge Rating Scales and Pre-Tests}

Before starting each pre-test, participants were asked to rate their own prior knowledge in each of the four topics using a 1 (no prior knowledge/novice) to 10 (very high prior knowledge/expert) rating scale. Each domain pretest was created using material from the relevant textbooks and consisted of five, 4-alternative multiple-choice questions from each of the four topics, for a total of 20 questions per domain (see Appendix B for example pre-test questions). This number of questions was similar to pretests in previous research (e.g., Cogliano et al., 2019; Fyfe & Rittle-Johnson, 2015; van Gog et al., 2015). Prior to the experiment, multiple sets of questions were pilot tested with participants not involved in the current study to ensure that they could not be answered correctly without the relevant prior knowledge.

\textbf{Training Topic Lessons}

The training phase for each participant consisted of three online training topic lessons that were completed over a three-day period. Each training lesson was specific to one of the topics within their assigned trained domain. Pilot testing indicated that each lesson took about 45 minutes to fully complete. All lessons followed a similar format and were created using information from the relevant textbooks. Each topic lesson contained multiple text passages interspersed with relevant graphics and knowledge check multiple-choice questions with

\textsuperscript{4} To create the materials in the Sensation and Perception domain, permission was obtained to use information from the following textbooks: \textit{Introduction to Psychology} (Kalat, 2011, 2017); \textit{Psychology: Themes and Variations} (Weiten, 2017); and \textit{Sensation and Perception} (Goldstein & Brockmole, 2017). To create the materials in the Historical Geology domain, permission was obtained to use information from the following textbooks: \textit{An Introduction to Geology} (Johnson et al., 2017); \textit{Physical Geology} (Earle, 2015); and \textit{Historical Geology: Evolution of Earth and Life Through Time} (Wincander & Monroe, 2016).
feedback. Each lesson ranged from approximately 4,000-4,500 words, excluding the knowledge check questions. The final section of each lesson involved a 15 multiple-choice question test (without feedback) that covered information throughout the lesson. In order to fully complete a training topic lesson, participants had to score 80% or higher (i.e., answer no more than 3 questions wrong) on the training test. Participants who earned less than an 80% on the topic test were required to repeat the entire training lesson and retake the test. No multiple-choice questions from the domain pre-tests were used as questions in the training phase (though similar information was covered). See Appendix C for an example training topic lesson (including the graphics, knowledge check questions, and final topic test).

**Domain Text Passages**

During the learning phase, participants first read four topic text passages within a single domain text passage before engaging in two rounds of restudy or retrieval practice (domain order was counterbalanced across participants). After the second round of restudying or retrieving information from all four topics, participants moved on to initially study a second set of four topic text passages within the other domain. As before, this was followed by two rounds of either restudy or retrieval practice, depending on the participant’s specific learning condition.

Each domain text passage contained four smaller topic text passages (see Appendix D for an example topic text passage). Coh-Metrix was used to assess the characteristics and complexity of the domain texts used in the present study (Graesser et al., 2011; McNamara et al., 2014; see Table 1). The Sensation and Perception domain text was 3,159 words and had a Flesch Reading Ease score of 52.12 (Flesch, 1948) and a Flesch-Kincaid Grade Level of 10.21 (Kincaid et al., 1975). The individual topic texts were each approximately 790 words ($M = 789.75$, $SD = 34.12$), with an average Flesch Reading Ease score of 52.07 ($SD = 6.87$) and a Flesch-Kincaid Grade
Level of 10.22 ($SD = 0.93$). The Historical Geology domain text was 3,088 words and had a Flesch Reading Ease score of 42.84 and a Flesch-Kincaid Grade Level of 11.48. The topic texts were each approximately 770 words ($M = 772.00$, $SD = 22.11$), with an average Flesch Reading Ease score of 42.34 ($SD = 6.13$) and a Flesch-Kincaid Grade Level of 11.64 ($SD = 0.92$).⁵

One of the main criticisms made by van Gog and Sweller (2015) was that the majority of testing effect studies do not use materials that resemble the complex learning required in educational settings. The authors provide some examples of these high element interactivity materials, including instructional texts on scientific phenomena (van Gog & Sweller, 2015). Although the materials used in the present study clearly belong to that category, it is important to objectively assess their level of complexity. Two testing effect studies that van Gog and Sweller (2015) rated high in learning material element interactivity (i.e., de Jonge et al., 2015; Tran et al., 2015) used textual information with Flesch Reading Ease scores ranging from 62.8 to 83.3 and Flesch-Kincaid Grade Levels ranging from 3.8 to 8.8 (see Karpicke & Aue, 2015). Taken together, the learning phase materials used in the present study seem comparable to, if not more complex than, the materials van Gog and Sweller (2015) gave as examples of complex, high element interactivity scientific texts. Therefore, if any cognitive load effects are not observed in the present study, it seems unlikely that low element interactivity materials are to blame.

**Retrieval Practice Questions**

Each topic had a total of eight short answer, free-response questions, for a total of 32 questions per domain. Participants were shown each set of eight questions for a total of four

---

⁵ A high level of difficulty was expected due to the educational nature of the materials and the introduction of technical terms that are defined in the text. To assess this, all technical terms in both domain texts were replaced with more common synonyms and then both were re-examined. The Sensation and Perception text Reading Ease score rose to 57.00 and the Grade Level dropped to 9.54, and the Historical Geology text Reading Ease score rose to 54.59 and the Grade Level dropped to 9.65 (Flesch, 1948; Kincaid et al., 1975). These reductions in difficulty indicate that the technical terms did contribute to the high ratings, especially for the Historical Geology text.
minutes, with two minutes to type their answers and two minutes to read and respond to feedback (see Figure 1 for an example). Feedback for each question consisted of: (1) the correct answer; (2) the participant’s answer; and (3) an explanation/detailed answer containing relevant information from the text. To ensure participants were attending to the feedback, they were asked to rate the correctness of their answer by selecting one of three choices: full credit (“my answer was correct and contained all of the important information”); partial credit (“my answer was close and/or contained most of the important information”); or no credit (“my answer was wrong and/or was missing most of the important information”). After completing the questions for all four topics within a domain, the entire process was repeated for a second time.

Restudy Examples/Key Ideas

Participants in the restudy condition were presented with the same four sets of eight items in each domain. However, the questions were re-formatted into pseudo-worked examples such that the answer and any critical relevant information was provided (i.e., almost identical to the explanation/detailed answer in the retrieval condition; see Figure 1 for an example). This form of restudy was chosen over general rereading for two main reasons. First, it is more similar to the typical worked examples vs. problem-solving paradigm used in cognitive load research (e.g., van Gog & Sweller, 2015). Second, although many testing effect studies compare retrieval practice to rereading the initially studied information, this likely favors the retrieval condition by focusing participants on the specific information that is likely to be assessed on the final test (e.g., Kornell et al., 2012; van Eersel et al., 2016). For example, van Eersel et al. (2016) found a reduced, yet still significant, testing effect on transfer when retrieval practice was compared to focused restudying than when it was compared to general rereading. Taken together, focused restudying
of detailed examples provides a more robust test of the benefits of retrieval practice as well as the predictions of cognitive load theory and element interactivity in general.

Participants were given four minutes to read over, study, and rate their understanding of the eight examples. Similar to the retrieval condition, participants were asked to rate their understanding of each example by selecting one of three choices: full understanding (“I fully understand the above information/example”); partial understanding (“I understand some, but not all, of the above information/example”); or no understanding (“I do not understand the above information/example”).

*Mental Effort Rating Scale*

Mental effort during the restudy or retrieval portion of the learning phase was measured using the nine-point subjective rating scale frequently used in cognitive load research (Paas, 1992). The scale ranged from: (1) “very, very low effort” to (9) “very, very high effort”.

Participants rated their mental effort a total of 16 times throughout the learning phase, once after restudying or retrieving information related to each topic.

*Final Test Questions*

After a two-day retention interval, all participants completed the two final domain tests. Each test had a total of 48 short answer questions, with 12 questions per topic. There were three types of questions on the final test: (1) retention questions identical to those restudied or retrieved during the learning phase; (2) near transfer questions similar to those restudied or retrieved (e.g., rearranged questions and answers or isomorphic versions of prior questions with different specific values; see Figure 2 for an example); and (3) far transfer questions that assessed information that was initially studied but not subsequently restudied or retrieved.
Procedure

The entire experiment was conducted online using Qualtrics software and required ~6 hours of total participation time per participant, spread out over 5-6 days. There were three major phases: (1) the Training Phase; (2) the Learning Phase; and (3) the Testing Phase (see Figure 3 for an overview of the entire procedure). Participants were given a basic overview of the experiment and told that they would be trained in a specific academic domain over a 3-day period before learning additional related and unrelated information that would be assessed on a final test two days later. After agreeing to participate, participants were reminded not to take or use notes or search for any answers online throughout the entire experiment. Next, participants completed the prior knowledge rating scales and pre-tests for each topic within the Sensation and Perception domain and then for each topic within the Historical Geology domain. Upon completion, participants were given a PDF containing instructions, a schedule, and links to their three specific training topic lessons that needed to be completed over the next three days. If all three topic lessons were not completed within the three-day time period, participants were not able to move on to the Learning Phase of the experiment.

Training Phase

At the start of each training topic lesson, participants were told about the general structure of the lesson (i.e., text passages interspersed with graphics and knowledge check multiple choice questions) and the subsequent topic test that they would need to score 80% or higher on to fully complete that lesson. Participants were also reminded to take their time, fully read through the text passages before answering the corresponding knowledge check questions, and not look up any answers online or take/use any notes. After completing the topic test, participants were asked if they looked up any answers or used any notes and told that their
answer would not affect the compensation they earn for that part of the experiment. Participants could complete the three training topic lessons at any point over the 3-day period, as long as all three were fully completed before starting the Learning Phase.

Learning Phase

Three days after submitting their initial Qualtrics survey, participants were automatically sent a link to the ~2-hour Learning Phase of the experiment. However, participants could only start this phase if they had already completed their three specific topic lessons.

The Learning Phase resembled the first two phases of the typical testing effect paradigm. Participants first read text passages for initial study, before either restudying or retrieving information from those passages in preparation for the final test two days later (see Figure 4 for an overview of the Learning Phase). Domain order was counterbalanced across participants, such that half studied and restudied/retrieved information from the Sensation and Perception domain before studying and restudying/retrieving information from the Historical Geology domain and half did the reverse (i.e., Historical Geology before Sensation and Perception).

The initial study portion for each domain presented participants with four sets of text passages with interspersed graphics. Participants could read through the text passages at their own pace but were required to spend a minimum of 5 minutes (300 s) reading each topic text passage (a duration determined to be sufficient via pilot testing). After reading a topic text passage, participants were given a 30 second break before moving on to the next text passage.

After studying the fourth topic text passage in a given domain, participants started the restudy/retrieval portion of the Learning Phase. Regardless of their specific learning condition, participants were told that they would have another opportunity to learn the just-studied information in preparation for the final test. Participants in the restudy condition were instructed
that they would have a total of four minutes to restudy eight examples/key ideas from each topic text passage. After four minutes of restudying, they were given 15 seconds to rate the amount of mental effort needed to complete the previous learning task. Participants then repeated this process for the remaining three topics within that domain. After restudying information from all four topics, participants completed a second round of restudy for each topic, one at a time.

Participants in the retrieval learning condition were told that they would have two minutes to answer a set of eight short-answer questions from the topic they just studied. After two minutes, participants were shown the correct answers, their answers, and elaborative/detailed feedback for each question for an additional two minutes. Then they were asked to rate their invested mental effort on the previous learning task. This was repeated for each of the remaining topics. As in the restudy condition, participants then completed this entire round of retrieval practice for a second time for each topic, just as before.

After the second round of restudy/retrieval practice, participants were given an optional five-minute break before moving on to the second domain. The entire procedure was repeated for the second domain (i.e., initial study of each topic text passage, followed by two rounds of restudy/retrieval practice). When the second round of restudy/retrieval practice was completed for the second domain, participants were asked if they looked up any answers online or took/used any notes (as before, they were told that their answer would not affect their compensation for that phase). Finally, participants were thanked for their time and effort and reminded that they would need to complete the final Testing Phase of the experiment in exactly two days. The entire Learning Phase took about 2 hours to complete: a minimum of 20 minutes of initial study followed by ~35 minutes of restudy/retrieval practice for each domain.
**Testing Phase**

Two days after completing the Learning Phase, participants were sent a link to the final Testing Phase of the experiment. Participants were instructed to complete two final tests, one per domain, each containing a total of 48 short answer questions (12 per topic). For each participant, domain order was the same as in the Learning Phase. As before, participants were asked not to look up any answers online or take/use any notes.

Participants were given a minimum of five minutes and a maximum of seven minutes to complete each set of 12 questions per topic. After answering the four sets of 12 questions for the first domain, participants moved on to the corresponding four sets of 12 questions in the second domain. When the fourth topic test in the second domain was completed, participants were asked if they took/used any notes or looked up any info online, debriefed, thanked for their participation, and compensated for their time. The entire Testing Phase took just under 1 hour to complete: 5-7 minutes to answer each set of 12 questions, for a total of 20-28 minutes for all four topics within each of the two domains.
CHAPTER 3: RESULTS

All retrieval practice and final test questions were scored using two scales: an all-or-nothing scale and a partial credit scale in increments of .20. Due to the open-ended nature of the free-response questions, results using the latter scale are discussed below (though the critical results were identical between scales, see Appendix E). Each answer could earn either 0 points, .2 points, .4 points, .6 points, .8 points, or 1 point, depending on the completeness of the response. Two raters independently scored the first 20% of responses and then compared scores to determine interrater reliability. Reliability between raters was very high ($r = .985$, CI 95% [.979, .991]; $ICC = .992$, CI 95% [.992, .993]) and one rater scored the remaining responses.

Prior Knowledge

Subjective Ratings

Participants rated their prior knowledge for each of the eight topics on a scale from 1-10, with larger values indicating greater prior knowledge. Across both domains, the average prior topic knowledge rating was 2.7 ($M = 2.70$, $SD = 1.67$, $Median = 2.25$). Participants rated their prior knowledge for topics within the Sensation and Perception domain ($M = 3.19$, $SD = 1.71$, $Median = 3.00$) significantly higher than topics within the Historical Geology domain ($M = 2.20$, $SD = 1.47$, $Median = 1.75$), $t(127) = 5.05$, $SE = 0.20$, $p < .001$, $d = 0.45$.

It was also important to verify that there were no a priori differences in prior knowledge between the four main experimental conditions: (1) training in Sensation and Perception and learning via restudy; (2) training in Sensation and Perception and learning via retrieval; (3) training in Historical Geology and learning via restudy; and (4) training in Historical Geology
and learning via retrieval. The four groups had similar subjective prior knowledge ratings for topics within the Sensation and Perception domain, $F(3, 124) = 0.37, p = .78$, and for topics within the Historical Geology domain, $F(3, 124) = 0.61, p = .61$.

Two 4 (main experimental condition) x 4 (topic) ANOVAs, one per academic domain, were used to compare ratings between experimental groups for the individual topics within each domain (Table 2). For Sensation and Perception topics, there was no main effect of experimental group, $F(3, 124) = 0.37, p = .78$, nor an interaction between experimental group and topic, $F(5.7, 234.4) = 0.69, p = .69$. However, the main effect of topic was significant, $F(1.9, 234.4) = 59.78$, $MS_e = 2.09, p < .001, \eta^2_p = .32$, with participants giving higher ratings for Color Perception and Auditory Perception than Chemical Senses and Cutaneous Senses and higher ratings for Chemical senses than Cutaneous Senses ($t_s > 3.3, p_{bonf} < .01$).

A similar ANOVA was used to assess ratings for the Historical Geology topics. As before, there was no significant difference between the experimental groups, $F(3, 124) = 0.61, p = .61$, nor was there an interaction between experimental group and topic, $F(7.0, 288.5) = 0.47, p = .86$. The main effect of topic was significant, $F(2.3, 288.5) = 3.33, MS_e = 0.71, p = .030, \eta^2_p = .03$, due to significantly higher ratings for the Rocks topic than the Geologic Time topic, $t = 3.11, SE = 0.09, p_{bonf} = .012, d = .28$. No other comparisons were significant ($t_s < 2.0, p_{bonf} > .32$).

Taken together, there were no differences in subjective prior knowledge ratings between the main experimental conditions, although some topics were rated as more familiar than others.

**Pre-Test Scores**

Participants also answered 40 multiple-choice questions (20 per domain, 5 per topic) to objectively assess their prior topic knowledge. Average pre-test scores did not differ between

---

6 When sphericity was violated, the Greenhouse-Geisser correction was applied.
topics within the Sensation and Perception ($M = .23, SD = .12$) and Historical Geology ($M = .24, SD = .12$) domains, $t(127) = 1.03, p = .30$, both of which were also not significantly greater than chance performance (i.e., .25), $ts < 0.01, ps > .70$. When comparing pre-test scores to chance performance for each topic individually, only the Color Perception topic ($M = .29, SD = .22$), $t(127) = 1.84, p = .034$, within the Sensation and Perception domain, and the Isotopic Dating topic ($M = .31, SD = .24$), $t(127) = 2.94, p = .002$, within the Historical Geology domain had performance significantly greater than chance (for all other topics, $ts < 1.2, ps > .12$).

As with the subjective prior knowledge ratings, it was also important to assess pre-test scores for the four major experimental conditions (i.e., Sensation and Perception-Restudy, Sensation and Perception-Retrieval, Historical Geology-Restudy, and Historical Geology-Retrieval). First, average pre-test score for each individual experimental group was not significantly above chance level performance for Sensation and Perception topics, $ts < 0.01, ps > .68$, nor for Historical Geology topics, $ts < 1.3, ps > .10$. A second set of two 4 (main experimental condition) x 4 (topic) ANOVAs, one per academic domain, were conducted to compare performance between groups across individual topics (Table 3). For Sensation and Perception topics, both the main effect of experimental group, $F(3, 124) = 0.50, p = .68$, and the interaction between group and topic, $F(9, 372) = 1.54, p = .13$, were non-significant. There was a significant main effect of topic, $F(3, 372) = 12.79, MSE = 0.04, p < .001, \eta^2_p = .09$, due to significantly greater scores on the Color Perception and Cutaneous Senses questions than on the Auditory Perception and Chemical Senses questions, $ts > 2.7, p_{bonf} < .01$.

For the Historical Geology topics, there was also no significant main effect of experimental group, $F(3, 124) = 1.81, p = .15$, nor an interaction between experimental group and topic, $F(9, 372) = 0.71, p = .70$. However, the main effect of topic was again significant, $F(3,$
372) = 7.61, $MS_e = 0.04, p < .001, \eta^2_p = .06$, due to higher scores on the Isotopic Dating questions compared to the other three topics ($ts > 2.8, p_{bonf} < .05$), which did not significantly differ from each other, $ts < 1.7, p_{bonf} > .55$. Overall, despite some differences in pre-test scores between topics, the critical result is the lack of a difference between the four main experimental conditions and no interaction with topic for either domain. In summary, pre-test scores indicate that participants had very little knowledge of the topics and the knowledge they did have was similar across experimental groups, as expected based on random assignment.

**Training Phase**

**Topic Lesson Attempts**

On average, participants completed each training lesson on their first attempt for topics within both Sensation and Perception ($M = 1.03, SD = 0.12, Median = 1.00$) and Historical Geology ($M = 1.07, SD = 0.17, Median = 1.00$), which did not significantly differ, $t(126) = -1.61, p = .11$. Across all training topic lessons, there were only 19 instances in which a participant did not pass a training topic lesson on their first try (i.e., ~5% of the 384 total lessons; see Table 4).  

**Topic Lesson Duration**

Pilot testing indicated that each training topic lesson would take about 45 minutes to complete. Although Qualtrics does provide a total duration measure, it is not a perfect measure of the actual time spent on each training topic lesson. For example, some participants had durations greater than seven hours because they clicked a Qualtrics link (starting the survey and

---

7 When broken down by later learning condition (Table 4), there was no significant difference in the number of attempts on Sensation and Perception topics between the restudy ($M = 1.05, SD = 0.15, Median = 1.00$) and retrieval ($M = 1.01, SD = 0.06, Median = 1.00$) conditions, $t(62) = -1.47, p = .15$. Although not significant, there was a trend for greater attempts on Historical Geology topics for the restudy ($M = 1.12, SD = 0.22, Median = 1.00$) vs. retrieval ($M = 1.03, SD = 0.10, Median = 1.00$) condition, $t(62) = -1.97, SE = 0.04, p = .053, d = -0.49$. This was driven by numerically, though not significantly, greater attempts on the Geologic Time topic lesson for the restudy ($M = 1.29, SD = 0.55, Median = 1.00$) vs. retrieval ($M = 1.13, SD = 0.34, Median = 1.00$) condition, $t(46) = -1.27, p = .21$. 

---

53
timer) but didn’t actually start working on the lesson until later in the day. Similarly, some participants fully finished a lesson but forgot to hit the final submit button until they noticed it hours later. Although these excessively long durations don’t necessarily reflect problematic performance issues, which can be more easily detected in other measures (e.g., performance on the training topic lesson knowledge check and final test), they do skew the topic lesson duration data itself. Therefore, a 2-step outlier removal process was used to provide a more accurate assessment of average training lesson duration. First, all durations greater than seven hours (420 min) were excluded ($n = 11$) from the entire dataset of 383 (each of the 128 participants provided three separate lesson duration datapoints, but one datapoint was not recorded due to a software error). Second, average duration for each topic lesson was calculated and any durations greater than $3-SD$ above the mean were excluded ($n = 10$), leaving a final subset of 362 total topic lesson duration datapoints after excluding a total of 21 datapoints (~5.5% of the total datapoints). On average, it took participants just over 55 minutes ($M = 56.54, SD = 23.92$) to complete each training topic lesson. A 2 (topic domain) x 2 (later learning condition) ANOVA indicated similar training duration between the main experimental conditions (both main effects and the interaction were non-significant, $Fs < .05, ps > .80$). Specifically, average training lesson duration was similar for participants later assigned to the restudy and retrieval learning conditions for Sensation and Perception topics ($M = 55.94, SD = 14.89$ and $M = 56.15, SD = 22.79$ for the restudy and retrieval conditions, respectively) as well as for Historical Geology topics ($M = 57.11, SD = 36.02$ and $M = 56.95, SD = 22.47$ for the restudy and retrieval conditions, respectively), $ts < .05, ps > .95$. This was also true when lesson durations between

---

8 Of the 21 total excluded duration outliers, 9 were from Sensation and Perception topic lessons (Color Perception = 3, Auditory Perception = 3, Cutaneous Senses = 1, and Chemical Senses = 2) and 12 were from Historical Geology topic lessons (Geologic Time = 3, Minerals = 3, Rocks = 4, and Isotopic Dating = 2).
participants later assigned to the restudy and retrieval conditions were compared separately for each individual topic (Table 5), $t < 1.81, ps > .075$. Thus, participants in both learning conditions spent a similar amount of time on each topic lesson, regardless of the specific domain.

**Topic Lesson Knowledge Check Performance**

For the following training topic lesson performance analyses, only participants’ first attempts were included. On average, participants correctly answered 87% of the knowledge check questions across all training lessons ($M = .87, SD = .06$). There was no significant difference in average knowledge check question proportion correct between the Sensation and Perception topics ($M = .87, SD = .05$) and the Historical Geology topics ($M = .87, SD = .06$), $t(126) = 0.57, p = .570$. Further, average knowledge check performance did not significantly differ between the restudy ($M = .88, SD = .06$) and retrieval ($M = .87, SD = .06$) conditions, $t(126) = -0.99, p = .32$. This was also true when split by specific domain, $ts < 1, ps > .35$, as well as when looking at each training topic lesson individually (Table 6), $ts < 1.2, ps > .25$.

**Topic Lesson Test Performance**

Performance on the final topic test at the end of each training lesson was analyzed to determine whether participants in both learning conditions were similarly and adequately trained in their respective topics. For the following analyses, only participants’ first attempts were included. On average, participants scored a 93% on the training test across all topics ($M = .93, SD = .06$) and there was no significant difference between Sensation and Perception ($M = .94, SD = .05$) and Historical Geology topics ($M = .92, SD = .06$), $t(126) = 1.19, p = .237$. Further, topic test performance did not significantly differ between the restudy ($M = .92, SD = .07$) and retrieval ($M = .94, SD = .04$) conditions, $t(126) = 1.54, p = .125$. This was also true when split by specific domain, $ts < 1.3, ps > .22$, as well as when looking at each topic individually (Table 6),
$t < 1.6, ps > .13$. This indicates that participants in the restudy and retrieval conditions were both adequately and similarly trained in their respective topics.

**Learning Phase**

**Topic Text Passage Duration**

Pilot testing indicated that each learning topic text passage would take about 5 minutes (300 s) to read. Participants were told to read through each text passage one at a time, at their own pace, and were able to advance after a minimum of 5 minutes (300 s). There was no maximum time limit and some participants had excessively long durations (likely due to a software recording error, an interruption, etc.). Therefore, a similar 2-step outlier procedure was used as with the training topic lesson duration data to get a more accurate understanding of average topic text reading duration during the learning phase. First, all durations greater than 20 minutes (1200 s) were excluded ($n = 7$) from the entire dataset of 1024. Second, the average duration for each topic text passage was calculated and any durations greater than $3-SD$ above the mean were excluded ($n = 26$), leaving a final subset of 991 total training topic lesson duration datapoints after excluding a total of 33 datapoints (~3.2% of the total datapoints).\(^9\)

On average, it took participants about six minutes (~360 s) to read each learning phase topic text passage ($M = 366.49, SD = 61.41$). Reading durations were similar for topic text passages within the Sensation and Perception ($M = 365.06, SD = 70.02$) and Historical Geology ($M = 371.06, SD = 82.35$) domains, $t < .90, p > .40$. There was a marginal effect of learning condition on topic text passage reading duration for topics within the Sensation and Perception domain ($M = 354.12, SD = 61.43$ and $M = 376.01, SD = 76.59$ for the restudy and retrieval

---

\(^9\) Of the 33 total excluded duration outliers, 16 were from Sensation and Perception topic text passages (Color Perception = 4, Auditory Perception = 4, Cutaneous Senses = 4, and Chemical Senses = 4) and 17 were from Historical Geology topic text passages (Geologic Time = 5, Minerals = 2, Rocks = 7, and Isotopic Dating = 3).
conditions, respectively), \(t(126) = 1.78, SE = 12.27, p = .077, d = .32\), and within the Historical Geology domain (\(M = 358.13, SD = 67.85\) and \(M = 383.98, SD = 93.41\) for the restudy and retrieval conditions, respectively), \(t(126) = 1.79, SE = 14.43, p = .076, d = .32\). Though non-significant, the numerical pattern indicates that participants in the retrieval condition spent about 20 more seconds reading a given topic text passage than those in the restudy condition. When each topic was examined individually (Table 7), it was apparent that this marginal difference between learning conditions was largely driven by significantly longer reading durations on the Isotopic Dating topic text passage for the retrieval than restudy condition, \(t(123) = 3.28, SE = 9.34, p = .001, d = 0.59\) (for all other topics, \(ts < 1.75, ps > .085\)).

However, the above reading duration analyses ignore the role of prior topic knowledge training. Therefore, a 2 (prior knowledge) x 2 (learning condition) ANOVA was conducted (Table 8), with prior knowledge as a within-subjects factor and learning condition as a between-subjects factor. Note that for all instances in which prior knowledge (HPK vs. LPK) is included as a factor, the HPK condition only includes the three trained topics within the trained domain and thus excludes the untrained topic within the trained domain (the LPK condition includes all four untrained topics within the untrained domain). Performance on the untrained topic within the trained domain is discussed later in the results and general discussion (but see Table 9 for the relevant mean duration, performance, and effort data).

Participants read passages on previously trained topics almost 25 seconds (~6.5%) faster than passages on untrained topics, \(F(1, 126) = 11.34, MS_e = 3351.46, p = .001, \eta^2_p = .08\). The main effect of learning condition was also significant, with the retrieval group spending more
time (20-25 seconds, ~6% longer) than the restudy group, $F(1, 126) = 5.76, MS_e = 7470.90, p = .018, \eta^2_p = .04$. The interaction was non-significant, $F(1, 126) = 1.22, p = .27$.

**Retrieval Practice Performance**

After reading the four topic text passages within a given domain, participants in the retrieval condition completed two rounds of retrieval practice for each set of topics. Specifically, participants answered a set of eight free-response questions (with feedback) for each of the four topics in a domain before repeating the entire round of retrieval practice for a second time. The two rounds of retrieval practice were designed to boost initial retrieval success, an important factor for obtaining testing effects in general, and on transfer questions in particular (e.g., Pan & Rickard, 2018).

Proportion correct on retrieval practice questions (Table 10) was analyzed using a 2 x 2 ANOVA, with both prior knowledge (HPK vs. LPK) and practice round (first vs. second) as within-subject factors. As noted above, the HPK condition excludes the untrained topic within the trained domain. There was a significant effect of prior knowledge, $F(1, 63) = 34.94, MS_e = .02, p < .001, \eta^2_p = .36$, such that performance on questions related to HPK topics was greater than on questions related to LPK topics. There was also a significant effect of practice round number, $F(1, 63) = 665.46, MS_e = .01, p < .001, \eta^2_p = .91$, with greater performance on the second vs. first round of retrieval practice. The interaction between prior knowledge and practice round number was non-significant, $F(1, 63) = 0.88, p = .35$. Thus, initial retrieval success was greater for HPK, than LPK, topics, indicating that participants benefited from experimentally induced

---

10 To ensure any benefit of retrieval over restudy seen on the final test was not simply a byproduct of spending more time on the learning phase text passages, participants’ average topic text duration was added as a covariate to the analyses of final test performance. All results and conclusions were unchanged after controlling for participants’ average learning phase reading duration. This is not surprising given the modest difference (approximately 6%) in study times across conditions.
prior knowledge. Further, performance increased with an additional round of retrieval practice, indicating that participants learned from their first round of retrieval and subsequent feedback.

**Learning Task Mental Effort**

After completing a set of eight restudy examples or retrieval practice questions for each topic, participants reported their mental effort on a scale ranging from 1-9, with larger numbers indicating greater effort (Paas, 1992). Learning information from Sensation and Perception topics ($M = 5.04, SD = 1.38$) was rated as similarly effortful as learning information from Historical Geology topics ($M = 4.96, SD = 1.60$), $t(127) = 0.73, p = .47$. Thus, there were no inherent differences in necessary mental effort between the two sets of materials (Table 11).

Subjective mental effort ratings were analyzed with a 2 x 2 x 2 ANOVA, using learning condition (restudy vs. retrieval) as a between-subjects factor and prior knowledge (high vs. low) and practice round (first vs. second) as within-subject factors (Table 12). First, the main effect of learning condition was significant, $F(1, 124) = 51.85, MS_e = 5.15, p < .001, \eta^2_p = .30$, such that participants in the retrieval condition rated their invested mental effort higher than those in the restudy condition. Second, the main effect of prior knowledge was significant, $F(1, 124) = 72.37, MS_e = 1.34, p < .001, \eta^2_p = .37$, with lower mental effort ratings given to HPK topics than to LPK topics. Third, the main effect of practice round was also significant, $F(1, 124) = 189.67, MS_e = 0.65, p < .001, \eta^2_p = .61$, due to decreased mental effort with an additional round of restudy or retrieval practice. Both the two-way interaction between learning condition and prior knowledge, $F(1, 124) = 0.003, p = .95$, and between prior knowledge and practice round, $F(1, 124) = 1.67, p = .20$, were non-significant. However, there was a significant two-way interaction between learning condition and practice round, $F(1, 124) = 10.01, MS_e = 0.65, p = .002, \eta^2_p = .08$. Finally, the three-way interaction was also significant, $F(1, 124) = 7.28, MS_e = 0.40, p = .008, \eta^2_p = .06$. 

59
Additional analyses were conducted to explore the significant three-way interaction between learning condition, prior knowledge, and practice round on invested mental effort during the learning task. First, two 2 (prior knowledge) x 2 (learning condition) ANOVAs were used to analyze subjective mental effort ratings separately for each practice round and produced similar results. For the first practice round, participants in the retrieval condition rated their mental effort significantly higher than those in the restudy condition, $F(1, 124) = 74.36, MS_e = 2.40, p < .001, \eta^2_p = .38$, and LPK topics were rated as more effortful than HPK topics, $F(1, 124) = 66.44, MS_e = 0.86, p < .001, \eta^2_p = .35$. Similarly, both main effects were also significant for the second practice round, with greater mental effort needed for retrieval practice than restudy, $F(1, 126) = 27.72, MS_e = 3.36, p < .001, \eta^2_p = .18$, and greater mental effort needed for LPK than HPK topics, $F(1, 126) = 47.23, MS_e = 0.87, p < .001, \eta^2_p = .27$. The interaction between learning condition and prior knowledge was non-significant for both round of practice, $F_s < 1.65, ps > .15$.

Thus, retrieval practice was more effortful than restudy and LPK topics were more effortful than HPK topics throughout the two rounds of restudy/retrieval practice, an important manipulation check when assessing cognitive load theory predictions regarding element interactivity.

Learning task mental effort ratings were also analyzed using a 2 (prior knowledge) x 2 (practice round) ANOVA separately for the restudy and retrieval conditions. Participants rated their mental effort higher when restudying examples from LPK topics compared to HPK topics, $F(1, 61) = 37.92, MS_e = 1.28, p < .001, \eta^2_p = .38$, and mental effort ratings significantly decreased with an additional round of restudy, $F(1, 61) = 78.00, MS_e = 0.46, p < .001, \eta^2_p = .56$. A significant interaction, $F(1, 61) = 6.22, MS_e = 0.50, p = .015, \eta^2_p = .09$, indicated that this reduction in mental effort from additional practice was larger for LPK topics, $t(62) = 9.35, SE = 0.11, p < .001, d = 1.18$, than HPK topics, $t(62) = 3.74, SE = 0.14, p < .001, d = 0.47$, though the
reduction was significant in both cases. Likewise, restudying HPK topics was less effortful than
restudying LPK topics during both the first, \( t(61) = -6.18, SE = 0.18, p < .001, d = -0.79, \) and
second, \( t(63) = -4.23, SE = 0.15, p < .001, d = -0.53, \) rounds of practice.

A similar 2 (prior knowledge) x 2 (practice round) ANOVA was used to analyze mental
effort ratings during retrieval practice. Both the main effect of prior knowledge, \( F(1, 63) = 34.66, \)
\( MS_e = 1.40, p < .001, \eta^2_p = .36, \) and practice round, \( F(1, 63) = 113.79, MS_e = 0.83, p < .001, \eta^2_p =
.64, \) were significant. However, there was no significant interaction, \( F(1, 63) = 1.34, p = .25, \)
indicating that HPK topics were rated as less effortful than LPK topics during both rounds of
retrieval practice and invested mental effort decreased with an additional round of practice for
both the HPK and LPK topics.

For completeness, learning task mental effort ratings were also analyzed with two
additional 2 (learning condition) x 2 (practice round) ANOVAs separately for HPK and LPK
topics. For the HPK topics (i.e., when the learning task was lower in element interactivity),
participants in the retrieval condition rated their invested mental effort significantly higher than
those in the restudy condition, \( F(1, 125) = 41.42, MS_e = 3.24, p < .001, \eta^2_p = .25. \) Further, mental
effort ratings significantly decreased when restudying or retrieving for a second time, \( F(1, 125) =
80.98, MS_e = 0.64, p < .001, \eta^2_p = .39. \) The interaction between learning condition and practice
round number was also significant, \( F(1, 125) = 14.83, MS_e = 0.64, p < .001, \eta^2_p = .11, \) due to a
larger difference in mental effort between learning conditions during the first, \( t(125) = -8.33, SE
= 0.22, p < .001, d = -1.48, \) vs. second, \( t(126) = -3.93, SE = 0.27, p < .001, d = -0.70, \) round of
practice, although ratings were significantly higher for retrieval than restudy in both cases.

For LPK topics (i.e., when the learning task was higher in element interactivity), retrieval
practice continued to be rated as significantly more effortful than restudy, \( F(1, 125) = 40.70, MS_e \)
= 3.22, \( p < .001, \eta^2_p = .25 \), and the first round of practice more effortful than the second, \( F(1, 125) = 179.50, MS_e = 0.40, p < .001, \eta^2_p = .59 \). Unlike with the HPK topics, there was no significant interaction, \( F(1, 125) = 0.91, p = .34 \), indicating a similar increase in necessary mental effort for retrieval practice than restudy across both rounds of practice.

### Testing Phase

#### Final Test Duration

Participants had 5-7 minutes (300-420 s) to complete each set of 12 topic-specific questions. Across learning conditions, average duration was similar for Sensation and Perception (\( M = 353.55, SD = 34.90 \)) and Historical Geology (\( M = 353.49, SD = 36.12 \)) topics, \( t(127) = 0.02, p = .98 \), with participants spending just under six minutes (360 seconds) on questions from either domain (Table 13). Final test duration was also analyzed with a 2 (prior knowledge) x 2 (learning condition) ANOVA (Table 14), which revealed shorter durations on the final test for participants in the retrieval than restudy condition, \( F(1, 126) = 6.05, MS_e = 1993.58, p = .015, \eta^2_p = .05 \). The main effect of prior knowledge, \( F(1, 126) = 0.17, p = .68 \), and the interaction between learning condition and prior knowledge, \( F(1, 126) = 0.01, p = .94 \), were both non-significant.

#### Final Test Performance

Overall performance on the final topic tests was first analyzed with a 2 x 2 ANOVA, using prior knowledge (high vs. low) as a within-subjects factor (note that the untrained topic within the trained domain is excluded) and learning condition (restudy vs. retrieval) as a between-subjects factor, averaging across the different question types (i.e., retention, near transfer, and far transfer; Figure 5). Both main effects were significant, indicating greater performance in the retrieval than restudy condition (i.e., a testing effect), \( F(1, 126) = 20.55, MS_e = .03, p < .001, \eta^2_p = .14 \), as well as higher performance on the HPK than LPK topics, \( F(1, 126) \).
= 141.05, \(MS_e = .01, p < .001, \eta^2_p = .53\). Critically, the interaction between learning condition and prior knowledge was non-significant, \(F(1, 126) = 0.003, p = .96\).

Additional Bayesian analyses were conducted using the statistical software program JASP (JASP Team, 2020) to quantify the evidence in favor of the null interaction. Specifically, Bayes Factors (\(BF_{01}\)) were calculated using the default prior settings in JASP (i.e., fixed effect scale factor \(r_a = 0.5\); e.g., Rouder et al., 2017) to compare the relative fit of the data under a model containing only main effects to a model containing the main effects and the interaction term (Wagenmakers et al., 2018). Supporting the ANOVA results, this analysis indicated that the data were 5.56 times more likely under a model with only main effects vs. a model with main effects and an interaction, \(BF_{01} = 5.563\). Thus, the magnitude of the testing effect (i.e., the difference in final test proportion correct in the retrieval condition minus the restudy condition) was similar for learning tasks lower (i.e., HPK topics) and higher (i.e., LPK topics) in element interactivity (+.10 and +.11, respectively).

Next, proportion correct on the final test was analyzed with a 2 x 2 x 3 ANOVA, using learning condition as a between-subjects factor, and prior knowledge and question type (retention vs. near transfer vs. far transfer) as within-subject factors. As in the overall analysis, both the main effects of learning condition, \(F(1, 126) = 20.55, MS_e = .10, p < .001, \eta^2_p = .14\), and prior knowledge, \(F(1, 126) = 141.05, MS_e = .03, p < .001, \eta^2_p = .53\), were significant, and the critical interaction between learning condition and prior knowledge remained non-significant, \(F(1, 126) = 0.003, p = .96\). There was a main effect of question type, \(F(2, 252) = 129.44, MS_e = .01, p < .001, \eta^2_p = .51\), with greater performance on retention than transfer questions and on near than far transfer questions. There were also significant interactions between learning condition and question type, \(F(2, 252) = 78.72, MS_e = .01, p < .001, \eta^2_p = .39\), as well as between prior
knowledge and question type, $F(1.7, 209.9) = 4.50, MS_e = .02, p = .017, \eta^2_p = .04$. Finally, there was no significant three-way interaction between learning condition, prior knowledge, and question type, $F(1.7, 209.9) = 0.30, p = .70$. Subsequent analyses explored the effects of prior knowledge and learning condition on final performance separately for each question type.

**Retention Questions.** A 2 x 2 ANOVA using prior knowledge as a within-subjects factor and learning condition as a between-subjects factor was used to analyze mean proportion correct on retention questions (Figure 6). First, there was a testing effect – participants who used retrieval practice had significantly higher performance than those who used restudy, $F(1, 126) = 93.90, MS_e = .04, p < .001, \eta^2_p = .43$. Second, there was a significant effect of prior knowledge, with greater performance on HPK than LPK topics, $F(1, 126) = 56.59, MS_e = .01, p < .001, \eta^2_p = .31$. Third, the interaction was non-significant, $F(1, 126) = 0.55, p = .46$, with the data being 4.12 times more likely under a model with only the main effects of learning condition and prior knowledge than under a model that also includes their interaction, $BF_{01} = 4.117$.

Although the interaction was non-significant, follow-up comparisons were conducted to assess the similarity of the testing effect at each level of element interactivity. Final test performance was significantly greater in the retrieval than restudy condition when the learning task was lower in element interactivity (i.e., HPK topics), $t(126) = 8.33, SE = 0.03, p < .001, d = 1.47$ (a testing effect of about +.22), and when the learning task was higher in element interactivity (i.e., LPK topics), $t(126) = 8.74, SE = 0.03, p < .001, d = 1.55$ (a testing effect of about +.24). The main effect of prior knowledge was significant for both the restudy, $t(63) = 6.85, SE = 0.02, p < .001, d = 0.86$, and retrieval, $t(63) = 4.25, SE = 0.02, p < .001, d = 0.53$, learning conditions. Taken together, if the benefits of prior knowledge seen on the final test are interpreted as arising from a reduction in element interactivity during learning, then the nearly
identical testing effects under either level of interactivity are particularly noteworthy. Despite a large and beneficial reduction in element interactivity from greater prior knowledge, retrieval practice continued to benefit learning more than restudy similarly in both learning situations (i.e., the testing effect did not decrease under higher levels of element interactivity).

**Near Transfer Questions.** Another 2 (prior knowledge) x 2 (learning condition) ANOVA was used to analyze performance on the near transfer final test questions (Figure 7), which were similar to information restudied/retrieved during the learning phase but with the question and answer rearranged or new specific values/examples (i.e., isomorphic). Again, both the main effects of learning condition, $F(1, 126) = 12.92, MS_e = .05, p < .001, \eta^2_p = .09$, and prior knowledge, $F(1, 126) = 51.26, MS_e = .03, p < .001, \eta^2_p = .29$, were significant, with greater performance in the retrieval than restudy condition and on HPK than LPK topics. Additionally, the interaction between learning condition and prior knowledge was non-significant, $F(1, 126) = 0.06, p = .81$, with the data being 5.14 times more likely under a model with both main effects, but no interaction, than under a model including the interaction term, $BF_{01} = 5.141$.

Specifically, participants in the retrieval condition outperformed those in the restudy condition on near transfer questions when looking at only the HPK topics (i.e., when the learning task was lower in element interactivity), $t(126) = 2.97, SE = 0.03, p = .004, d = 0.53$, and when looking at only the LPK topics (i.e., when the learning task was higher in element interactivity), $t(126) = 2.60, SE = 0.04, p = .010, d = 0.46$ (a testing effect of about +.10 for both). This lack of a difference occurred in the face of robust effects of topic training on final test performance for both the restudy, $t(63) = 4.86, SE = 0.03, p < .001, d = 0.61$, and retrieval, $t(63) = 5.27, SE = 0.03, p < .001, d = 0.66$, conditions. Thus, despite greater performance on near-transfer final test questions for information learned under lower levels of element interactivity (reduced via greater
levels of prior knowledge), the learning benefits of retrieval vs. restudy were virtually identical across tasks with lower and higher levels of element interactivity.

**Far Transfer Questions.** A 2 (prior knowledge) x 2 (learning condition) ANOVA was used to analyze performance on far transfer final test questions (Figure 8), which assessed initially studied information that was not subsequently restudied or retrieved. First, there was no significant main effect of learning condition, $F(1, 126) = 0.80, p = .37$, indicating similar performance on far transfer questions between the restudy and retrieval conditions. Second, the main effect of prior knowledge was significant, $F(1, 126) = 90.13, MSe = .02, p < .001, \eta^2_p = .42$, with greater performance on HPK than LPK topics. Critically, the interaction between learning condition and prior knowledge was non-significant, $F(1, 126) = 0.14, p = .37$. Bayesian analyses indicated that the best fitting model included only the main effect of prior knowledge, $BF_{01} > 12e+10$. The data were 17.27 times more likely under that model than under the model with both main effects and the interaction, $BF_{01} = 17.272$. A second more relevant comparison indicated that the data were 5.01 times more likely under a model with both main effects than under a model with main effects and an interaction, $BF_{01} = 5.01$.

Although the interaction was non-significant, follow-up comparisons were conducted to assess the similarity of the testing effect (or lack thereof) between low and high levels of element interactivity. Performance on far transfer questions was similar between learning conditions when the learning task was both lower (i.e., HPK topics), $t(126) = -0.52, p = .61$, and higher (i.e., LPK topics), $t(126) = -1.02, p = .31$, in element interactivity. Importantly, despite a lack of a testing effect in either condition, the numerical difference between retrieval and restudy was similar (i.e., about -.025). Finally, this occurred when experimentally-induced prior topic
knowledge significantly benefited performance in both the restudy, \( t(63) = 7.24, SE = 0.02, p < .001, d = 0.91 \), and retrieval, \( t(63) = 6.35, SE = 0.03, p < .001, d = 0.79 \), learning conditions.

**The Untrained Topic Within the Trained Domain**

A final set of supplementary analyses were conducted to assess the information related to the untrained topic within the trained domain. For example, if a participant was trained in Color Perception, Auditory Perception, and Cutaneous Senses within the Sensation and Perception domain, those three topics would be the trained domain topics, the Chemical Senses topic would be the untrained topic within the trained domain, and the four topics within the Historical Geology domain would be the untrained domain topics. Table 9 contains all relevant descriptive statistics related to the untrained topic within the trained domain. During the learning phase, retrieval practice performance on the untrained topic within the trained domain increased with an additional round of practice, \( t(63) = 13.12, SE = 0.03, p < .001, d = 1.64 \). When compared to the other prior knowledge conditions, performance on the untrained topic within the trained domain was significantly worse than on trained domain topics (for either round of practice, \( ts > 3.4, ps < .01 \)), but did not significantly differ from performance on untrained domain topics (for either round of practice, \( ts < 0.40, ps > .70 \)).

Subjective mental effort ratings during the learning phase indicated that retrieval practice was significantly more effortful than restudy when learning information related to the untrained topic within the trained domain (both rounds, \( ts > 4.10, ps < .001 \)). Effort ratings were also significantly reduced with an additional round of practice, \( t(108) = 8.24, SE = 0.13, p < .001, d = 0.79 \). Averaging across learning strategy, learning information related to the untrained topic within the trained domain was significantly more effortful than learning information related to trained domain topics (both rounds, \( ts > 3.7, ps < .001 \)), but significantly less effortful than
learning information related to the untrained domain topics (both rounds, ts > 2.4, ps < .05). A similar pattern emerged when looking at each learning condition individually – learning information from the untrained topic within the trained domain was more effortful than information from trained domain topics but less effortful than untrained domain topics (though this latter difference was non-significant within the second round of restudy). Taken together, the mental effort needed to learn information related to the untrained topic within the trained domain was: (1) significantly greater when learning via retrieval practice than restudy; (2) significantly greater than the effort needed for the trained domain topics; and (3) significantly less than the effort needed for the untrained domain topics.

The following analyses examined performance on final test questions related to the untrained topic within the trained domain (see Table 9). A 2 (learning condition) x 3 (question type) ANOVA revealed significant main effects of learning condition and question type ($F$s > 10, $ps < .01), as well as an interaction between learning condition and question type, $F(2, 252) = 4.11, MS_e = .05, p = .017, \eta^2_p = .03$. The interaction was driven by a significant testing effect on retention questions, $t(126) = 4.36, SE = .05, p < .001, d = 0.77$, no testing effect on near transfer questions, $t(126) = 1.06, p = .29$, and a marginal, though non-significant, testing effect on far transfer questions, $t(126) = 1.85, SE = .05, p = .066, d = 0.33$. Thus, unlike performance on the trained domain topics and untrained domain topics, retrieval practice was not more beneficial than restudy for near transfer but was marginally more effective for far transfer.

Final test performance on questions related to the untrained topic within the trained domain was significantly worse than on questions related to trained domain topics, $t(127) = 6.77, SE = 0.02, p < .001, d = .60$, but did not differ from performance on questions related to untrained domain topics, $t(127) = 0.69, p = .49$. An identical pattern was found when assessing
performance on retention, near transfer, and far transfer questions individually – performance on
the untrained topic within the trained domain was significantly worse than on the trained topics,
\( t_s > 3.70, ps < .001 \), but did not differ from performance on the untrained domain topics, \( ts < 0.80, ps > .44 \). When looking at each learning condition individually, the same pattern occurred
for all but the far transfer final test questions. Specifically, although far transfer performance on
questions related to the untrained topic within the trained domain did not significantly differ
from performance on questions related to the trained domain topics for participants in the restudy
condition, \( t(63) = 1.35, p = .18 \), there was a significant difference for those in the retrieval
condition – greater performance on far transfer questions related to the untrained topic within the
trained domain than on questions related to the untrained domain topics, \( t(63) = -2.08, SE = 0.03, 
\( p = .041, d = .26 \). Although interesting, the two most critical results are: (1) the significant testing
effect on overall final test performance and (2) the benefit of topic training (and thus lower
element interactivity) seen in the significantly greater performance on trained domain topics than
on the untrained topic within the trained domain.
CHAPTER 4: DISCUSSION

Although the learning benefits of retrieval practice are well documented, research grounded in cognitive load theory has sparked a recent debate over the boundary conditions of the testing effect in educationally relevant situations (e.g., van Gog & Sweller, 2015; cf. Karpicke & Aue, 2015). Specifically, does retrieval practice continue to benefit learning to a greater degree than restudy when the learning task is high in element interactivity? Or does the level of element interactivity moderate the benefits of retrieval such that it becomes no more effective (or even less effective) than restudy?

One way to answer this question is to compare the size of the testing effect between learning situations that differ in element interactivity. While some research has attempted to do just that, methodological issues (e.g., no experimental manipulation of element interactivity within a single experiment, small sample sizes, providing the to-be-retrieved information during the retrieval task; Leahy et al., 2015; van Gog et al., 2011) and inconsistent results (e.g., Hanham et al., 2017; van Gog et al., 2015; see Karpicke & Aue, 2015) make it difficult to draw concrete conclusions. Therefore, the present study was designed to address this issue by manipulating element interactivity within a single experiment by randomly assigning participants to high prior knowledge (HPK) and low prior knowledge (LPK) conditions via an extensive training procedure, while holding the complexity of the learning materials constant. Because element interactivity measures the complexity of a learning task in relation to the learner’s prior knowledge (e.g., Sweller, 2010), the level of element interactivity during the learning phase of the experiment is reduced when learning information from previously trained, HPK topics. If the
benefits of retrieval practice only emerge under low element interactivity situations (or even if they are simply moderated by element interactivity), there should be an interaction – a larger testing effect when learning information from HPK topics (i.e., lower element interactivity) than when learning information from LPK topics (i.e., higher element interactivity).

Before discussing the critical final test results, it is important to first determine if the prior knowledge manipulation was successful in order to accurately assess the predictions related to element interactivity and the testing effect. First, the effectiveness of the training procedure can be seen when comparing pre-training topic test scores (i.e., proportion correct on the pre-test questions associated with a given topic) to post-training topic test scores (i.e., proportion correct on the test at the end of each training topic lesson). Although these tests were not identical, both used multiple-choice questions that assessed knowledge of the same topic. Post-training test scores were significantly greater than pre-training test scores for Sensation and Perception topics (pre-test: $M = .24, SD = .12$ vs. post-test: $M = .94, SD = .05$) and Historical Geology topics (pre-test: $M = .25, SD = .12$ vs. post-test: $M = .92, SD = .06$), $F_s > 1500, p_s < .001$. This dramatic increase in performance provides evidence that participants both started with little-to-no prior knowledge of these topics and subsequently gained substantial knowledge after completing the training lessons. Second, retrieval practice performance during the learning phase was significantly greater for HPK than LPK topics. Third, performance on all types of final test questions (i.e., retention and both levels of transfer) was significantly greater for HPK than LPK topics. Thus, there is strong evidence that the experimental manipulation of prior knowledge was successful.

It also important to investigate whether the successful prior knowledge manipulation produced results in line with a successful reduction in element interactivity. First, the materials
used in the present study (i.e., scientific text passages) were complex and similar to other materials van Gog and Sweller (2015) gave as examples of high element interactivity learning materials. Further, readability indices (e.g., reading ease and grade level) indicated that these materials were actually more complex than the specific scientific text passages used by de Jonge et al. (2015), which were rated high in element interactivity by van Gog and Sweller (2015). Because the highly complex learning materials were held constant across participants, the theoretical conception of element interactivity necessitates that the level of interactivity will decrease as the learner’s prior knowledge increases (e.g., Sweller, 2010). If the prior knowledge training procedure and complex learning materials used in the present study did effectively manipulate element interactivity, one would expect lower mental effort ratings (used to measure cognitive load; Paas, 1992) when learning information from HPK than LPK topics. This is precisely what was found – participants reported significantly greater mental effort when learning information related to LPK than HPK topics (see Table 12).

A final important manipulation check concerns the mental effort needed to learn via retrieval practice vs. restudy. In line with the predictions of cognitive load theory (e.g., van Gog & Sweller, 2015) and many other testing effect studies (e.g., Pyc & Rawson, 2009), participants consistently rated retrieval practice as more effortful than restudy. When this is coupled with the clear effectiveness of the training procedure and element interactivity’s theoretical dependence on prior knowledge, the more effortful retrieval practice strategy should lead to a significant reduction, elimination, or even reversal, of the testing effect when learning information related to LPK than HPK topics (e.g., Chan et al., 2018; van Gog & Sweller, 2015).

Now that the effectiveness of the prior knowledge manipulation (and thus element interactivity) is clearly established, we can turn to performance on the final test. Across all
question types and levels of prior knowledge, there was a significant testing effect (i.e., the difference in final performance between the retrieval – restudy group), with participants who used retrieval practice scoring about +10% (one full letter grade) higher than those who used restudy. Importantly, this benefit was similar in size for information related to HPK (lower in element interactivity; +10%) and LPK (higher in element interactivity; +11%) topics. When looking at each question type individually, the lack of a difference in the magnitude of the testing effect between HPK and LPK topics continued to emerge, with testing effects of +22% and +24% (retention), +10% and +9% (near transfer), and -2% and -3% (far transfer) for HPK (lower in element interactivity) and LPK (higher in element interactivity) topics, respectively. As previously noted, this occurred alongside consistent and substantial effects of prior knowledge, with participants scoring about +15% higher on the final test on questions related to HPK vs. LPK topics (a benefit that emerged across and within learning conditions and question types).

Taken together, there was no difference in the size of the testing effect under higher (LPK topics) and lower (HPK topics) levels of element interactivity, which can also be seen in the consistent non-significant interaction between prior knowledge and learning condition. The reliability of this null effect was further supported by Bayes Factors indicating that the data were about 4.1 to 5.6 times more likely under a model without an interaction term. This represents moderate evidence against the inclusion of the interaction according to the descriptive classification scheme by Wagenmakers et al. (2018).

The larger overall learning benefits from retrieval practice compared to restudy are particularly noteworthy because of the greater mental effort needed during learning. This supports the idea that retrieval practice is a desirable difficulty – an effortful strategy that produces greater learning than a less effortful strategy (e.g., Bjork, 2017). However, there was no
benefit of retrieval > restudy when looking specifically at the final far transfer questions. These questions were considered to be “far transfer” because they assessed information that was initially studied, but not later restudied or retrieved. On the other hand, questions labeled “near transfer” were similar to those restudied/retrieved but either had the question and answer rearranged or were isomorphic (assessed the same concept but used different specific values/examples). Although retrieval practice benefited near transfer to a greater degree than restudy, there was no such advantage on far transfer.

These results are generally in agreement with the meta-analysis by Pan and Rickard (2018), which found a weak effect at best for transfer across stimulus-response rearrangement and no compelling evidence of transfer to untested materials seen during initial study (though results varied widely across papers). As noted in the introduction, Pan and Rickard (2018) found that transfer was more likely to occur when initial test performance was high, retrieval used broad encoding and/or elaborative feedback, and there was substantial overlap in the response congruency between the initial and final tests. Thus, the testing effect on near transfer found in the present study is likely due to a combination of enhanced initial test performance due to two rounds of retrieval practice as well as the provision of detailed elaborative feedback (which was almost identical to the corresponding restudied information). Further, there was potentially greater partial overlap between initial and final test cues due to the detailed nature of short answer questions compared to less detailed materials like word pairs and key-term definitions that were used by the majority of studies in the stimulus-response rearrangement transfer category in Pan and Rickard (2018). The lack of a testing effect on far transfer may also be explained by the degree of overlap between initial and final test cues. Since none of this
information was on the initial test, the low response congruency likely contributed to the lack of far transfer of test-enhanced learning (Pan & Rickard, 2018).

There was also an interaction between prior knowledge and question type, driven by larger benefits of HPK (or lower element interactivity) on transfer (+16%) than retention questions (+10%). This suggests that the likelihood of transfer increases as the element interactivity of the learning task decreases (at least when decreased via greater prior knowledge). Interestingly, this increased likelihood of transfer was similar between the restudy and retrieval practice conditions, which provides additional evidence against the prediction of no testing effect for complex, high element interactivity learning tasks (e.g., Leahy & Sweller, 2019; Sweller, 2010; van Gog & Sweller, 2015).

Though not as relevant to the main goal of the present study, the untrained topic within the trained domain (Table 9) could also be thought of as a type of transfer – transfer across topics within the same academic domain (e.g., Chan, 2009). Final test performance was greater for the retrieval practice than restudy condition across question types, though this was driven primarily by a large testing effect on retention questions. There was no difference on near transfer questions but a marginal benefit of retrieval over restudy on far transfer questions. Although future studies specifically designed to assess this type of transfer (i.e., across topics within the same academic domain) are needed, these results imply that retrieval is either more beneficial (retention) or just as beneficial (near and far transfer) as restudy when learning information related to a novel topic within a HPK domain.

**Element Interactivity and the Testing Effect**

How can these results be reconciled with prior cognitive load research on retrieval-based learning? In their review of the testing effect, van Gog and Sweller (2015) cite the results of six
studies as evidence against a testing effect with complex learning materials (i.e., de Jonge et al.,
2015; Leahy et al., 2015; Tran et al., 2015; van Gog et al., 2011; van Gog et al., 2015; van Gog
& Kester, 2012). One major difference between these prior studies and the present study
concerns the manipulation, or lack thereof, of element interactivity. As noted by Karpicke and
Aue (2015), none of the cognitive load studies on the testing effect actually manipulated element
interactivity within a single experiment, which is necessary to establish causal effects. However,
it is still beneficial to explore why these studies may have found a lack of test-enhanced learning.

Across the six studies cited by van Gog and Sweller (2015) and two more recent relevant
studies (i.e., Hanham et al., 2017; Leahy & Sweller, 2019), there are a total of 23 experiments
that permit a comparison between the benefits of restudy and retrieval. Of these 37 total
comparisons (averaged over isomorphic and identical problems when applicable, i.e., van Gog et
al., 2015), there was a significant positive testing effect in five (i.e., Hanham et al., 2017,
Experiments 1 and 2; Leahy & Sweller, 2019, Experiments 2 and 3: delayed final test; Tran et
al., 2015, Experiment 4: retention), a significant negative testing effect in six (i.e., Hanham et al.,
2017, Experiments 4 and 5; Leahy et al., 2015, Experiments 1 and 2; van Gog et al., 2011; van
Gog & Kester, 2012), and a null testing effect in the remaining 26 comparisons (i.e., de Jonge et
al., 2015; Hanham et al., 2017, Experiments 3-6; Leahy et al., 2015; Leahy & Sweller, 2019;
Tran et al., 2015; van Gog et al., 2015; van Gog & Kester, 2012). If there truly is no moderating
effect of element interactivity, as the current study suggests, how can these disparate findings be
explained?

Starting with the four studies that found a significant negative testing effect in one or
more comparisons, only one (in van Gog & Kester, 2012) used a delayed final test. However, as
previously discussed, in that study, both the restudy and retrieval groups took an immediate final
test prior to the delayed test, which means both conditions benefited from retrieval-based learning. The other five instances in which a significant negative testing effect was found all occurred on an immediate final test, but only two comparisons involved feedback at retrieval practice (i.e., Leahy et al., 2015, Experiments 1 and 2). Further, none of these studies reported initial test performance (i.e., Hanham et al., 2017, Experiments 4 and 5; Leahy et al., 2015, Experiments 1 and 2; van Gog et al., 2011). Taken together, these observations of a negative testing effect on an immediate final test are not as unprecedented as Sweller and colleagues claim. This pattern often occurs when low initial test performance produces different subsets of item experienced by each learning condition at practice (i.e., the restudy group will reexperience all items but the retrieval group will only reexperience the [potentially small number of] items they successfully recall; e.g., Karpicke, 2017; Kornell et al., 2011; Rowland & DeLosh, 2015). This speaks to the importance of reporting initial retrieval performance and providing feedback during practice. Although Leahy et al. (2015) did provide feedback, they did not report initial recall, nor did they administer a prior knowledge pre-test or assess mental effort during learning.

How can the 26 null testing effect results be understood in light of the current experiment? Out of these 26 instances of no significant difference between restudy and retrieval, 14 were numerically in favor of a positive testing effect (i.e., de Jonge et al., 2015; Hanham et al., 2017, Experiments 3, 4, and 6; Leahy et al., 2015, Experiment 3; van Gog et al., 2015) and 12 were numerically in favor of a negative testing effect. Focusing on the instances in which a numerical negative testing effect was found, six of the 12 comparisons involved sample sizes small enough that they make the results difficult to interpret, with ns of 9-16 per condition (de Jonge et al., 2015; Hanham et al., 2017, Experiments 3 and 5; Leahy & Sweller, 2019, Experiments 2 and 3). Of the remaining six instances in which a numerical negative testing effect
was found, three occurred on an immediate final test (Tran et al., 2015, Experiments 1 and 3; van Gog & Kester, 2012) and three occurred on a delayed final test (Tran et al., 2015, Experiments 2-4). The lack of a testing effect in van Gog and Kester (2012) is likely due to the combination of fairly low initial test performance (i.e., ~20% and ~45% on the two practice test questions) and the lack of feedback during practice.

This leaves the results of Tran and colleagues (2015), who found no difference between restudy and retrieval on an immediate or delayed test of deductive inference. As discussed earlier in the paper, this important finding was explored by Eglington and Kang (2018), who concluded that the specificity of the retrieval practice task hindered relational processing during practice, making it difficult to make later deductions by integrating previously-learned discrete units of information. The authors demonstrated that a testing effect can occur on later deductive reasoning, as long as the to-be-integrated information is presented simultaneously during practice (Eglington & Kang, 2018). Thus, the disparate results from cognitive load research on the testing effect may be due to a variety of factors (e.g., no experimental manipulation of element interactivity, confounding the benefits of retrieval and restudy, no feedback, low or unknown initial test performance, small sample sizes, and no measure of mental effort or prior knowledge) that help temper the conclusion that the testing effect does not emerge under high element interactivity learning situations. Recent meta-analyses of classroom studies on the testing effect further support the idea that students will in fact benefit more from retrieval-based learning than restudy when learning educational material (Agarwal et al., in press; Yang et al., 2021).

Prior Knowledge and the Testing Effect

The few studies that have assessed the role of prior knowledge in relation to the testing effect also warrant discussion. First, it is important to note that, to the best of the author’s
knowledge, no prior testing effect study has experimentally manipulated prior knowledge. The present study is the first to randomly assign participants to high and low knowledge conditions and assess its potential causal effect on the benefits of retrieval practice over restudy. Without an experimental manipulation of prior knowledge, the prior studies can only offer correlational evidence pertaining to its impact on the effectiveness of retrieval-based learning. However, since experimental manipulations are not always possible in classroom studies, it is still beneficial to discuss how the present results align with those of the prior correlational research.

The similar sized benefits of retrieval practice over restudy for learners with high and low prior knowledge found in the present study align with the results of Xiaofeng et al. (2016) who found no interaction between learners’ prior knowledge and the benefits of retrieval practice (see also, Carroll et al., 2007). Because Xiaofeng and colleagues (2016) also observed a moderating influence of prior knowledge on the effectiveness of elaborative study, the authors interpreted their results as evidence in support of the episodic-context account and against the elaborative-retrieval hypothesis. If the testing effect is driven by greater (semantic) elaboration during retrieval practice than restudy, one would predict that learners with greater prior knowledge (and thus greater potential semantic elaborations/associates that could act as additional retrieval routes after being activated during the initial retrieval attempt) would benefit from retrieval practice more than those with less prior knowledge (e.g., Carpenter, 2009, 2011; Carpenter & Yeung, 2017). Further, this benefit of prior knowledge should be even greater than the benefit to restudy, which does not involve a memory search during practice and would thus not benefit as much from the activation of additional semantic information as retrieval. The results of the current study do not support the predictions of the elaborative-retrieval hypothesis.
On the other hand, the results are more compatible with theories that propose an alternative mechanism to semantic association and relational processing – the episodic-context account (e.g., Karpicke et al., 2014). This theory proposes that successful retrieval, unlike restudy, requires context reinstatement, which updates the contextual information of the memory representation to include features associated with the current context (i.e., the context prevailing during retrieval practice). On a final test, it is easier for those in the retrieval condition to reinstate the context necessary for successful retrieval because the varied contextual features are more likely to match the current test’s contextual cues. Because this theory proposes no role for the semantic prior knowledge possessed by the learner, it would predict no moderating influence of prior knowledge on the magnitude of the testing effect. The results of the current study and Xiaofeng et al. (2016) are in agreement with this prediction (see also, Carroll et al., 2007).

How can we reconcile the disparate findings of Carpenter et al. (2016) and Francis et al. (2020), who found that HPK learners benefited more from retrieval practice than their LPK peers (see also, Marsh et al., 2009; Spitzer, 1939), with the results of Cogliano et al. (2019), who found greater benefits of practice testing for information from LPK topics than HPK topics (see also, Hattikudur & Postle, 2011; Hernick, 2015; see also Spreckelsen & Jünger, 2017)? One possible answer can be seen in the different operationalizations of prior knowledge. As previously discussed, categorizing learners into high and low prior knowledge groups based on proxy measures like education level, course performance, or prior experience introduces ambiguity with respect to potential confounding variables. For example, two groups of students who differ in average course performance likely differ in other ways (e.g., motivation, WMC, intelligence, etc.) that could also moderate the effectiveness of a specific learning strategy. As noted by Carpenter et al. (2016) who found larger benefits of retrieval practice for students with higher
course performance, high performers likely differ from their lower performing peers in terms of other key factors, like interest and even familiarity with retrieval-based learning strategies. In other words, finding greater benefits of retrieval practice for students with higher course performance may reflect the beneficial effects of greater prior knowledge, enhanced motivation, greater interest in the material, more familiarity with using similar retrieval strategies, and many other differences that cannot be disentangled from one another (e.g., Carpenter et al., 2016; Fyfe & Rittle-Johnson, 2016; Rey & Buchwald, 2011; Tobias, 2010).

Although pre-tests do not alleviate all of the confounding issues discussed above (e.g., differences in motivation or interest), they are arguably a more direct measure of relevant prior knowledge. Both Cogliano et al. (2019) and Francis et al. (2020) used pre-tests to assess prior knowledge but found opposite patterns of results. However, neither study used a restudy control condition – both compared memory for practice-tested information to untested information. This is problematic because the learning benefits of retrieval can’t be disentangled from the effects of reexperiencing or exposure and does not allow for an assessment of the testing effect itself (i.e., calculated as the difference in final performance of a retrieval group minus a restudy group). Additional methodological limitations inherent in most classroom studies may have also contributed to the disparate results, such as a lack of control over, or measure of, student studying behavior outside of class (which is important because the final tests were high stakes and counted towards students’ final course grade; Cogliano et al., 2019).

Turning specifically to Cogliano et al. (2019), what other differences may explain their finding of a larger advantage of practice-testing on for LPK topics than HPK topics? Cogliano et al. (2019) did not observe a main effect of prior knowledge in any comparison, which makes it hard to assess the effect of prior knowledge on the effectiveness of a specific learning strategy.
(in fact, final performance was numerically higher on LPK tested information than in any other condition). The current study also differed from Cogliano et al. (2019) in terms of the questions themselves – whereas Cogliano et al. (2019) used the same questions for the pre-test, practice test, and the final test, the current study used different questions on the pre-test, training lessons, and practice test (and on the transfer portion of the final test).

The quasi-experimental study conducted by Francis and colleagues (2020) also had similar issues that cloud direct comparisons of the results. In addition to those discussed above, Francis et al. (2020) used a median split to create the LPK and HPK groups, resulting in a LPK group with pre-test scores between 20-40% and a HPK group with pre-test scores between 44-76%. This is problematic because a student with a pre-test score of 75% would be considered to have equal prior knowledge as a student with a pre-test score of 45% (both of which would be in the HPK group), although the learning benefits from 75% prior knowledge are likely not the same as the benefits from only 45% (Francis et al., 2020). This also means that those two students are assumed to be more similar in terms of prior knowledge than are two students with pre-test scores of 40% (in the LPK group) and 45% (in the HPK group).

However, there is a more problematic issue pertaining to the pre-test prior knowledge classification (Francis et al., 2020). Specifically, the pre-test contained a total of 25 multiple-choice questions: 10 from four topics that would be later practiced via retrieval-based concept-mapping; 10 from four topics that would be later practiced via standard practice quizzing; and five from two topics that would not be practiced. Critically, prior knowledge was not calculated separately for the specific set of topics associated with each learning condition. This means that a HPK participant who scored a 60% on the pre-test could have correctly answered 90%, 10%, and 100% of the pre-test questions related to topics that would be later concept-mapped, quizzed, and
not practiced, respectively. Alternatively, that same score could have been earned by a participant who correctly answered 50%, 90%, and 20% of the questions related to concept-mapped, quizzed, and non-practiced topics, respectively. In other words, it is unknown if a given HPK participant actually had low (1/10) or high (9/10) prior knowledge of the specific topics that would actually be quizzed and thus benefit from retrieval-based learning. To truly assess the influence of prior knowledge on strategy effectiveness, knowledge of the set of topics specific to each learning condition would need to be analyzed.

Taken together, there are several reasons why the current study found a different pattern of results than the prior research that used proxy measures of prior knowledge (e.g., course performance) or pre-tests to categorize participants into knowledge groups. Most importantly, the conflicting results seen in the prior correlational research demonstrates the importance of experimental manipulations of prior knowledge when attempting to draw clear conclusions about the effect of prior knowledge on the testing effect. Although not always possible, future research on this issue should experimentally manipulate prior knowledge instead of using proxy measures or pre-test scores to estimate differences in prior knowledge. If experimental manipulations are not possible (e.g., in classroom studies that are not conducive to random assignment), future correlational research should: include a restudy control condition; use different questions on the pre-test, practice test, and final test; avoid using a median split to categorize learners into knowledge groups; and ensure that the information being learned in each learning condition is within or related to the specific topic that was used to determine their level of prior knowledge.

**Educational Implications**

The results of the present study should come as welcome news to educators and practitioners. Learners randomly assigned to different training conditions either benefited more
from retrieval practice than restudy (on retention and near transfer questions) or to a similar
degree (on far transfer questions). Although additional supporting research is needed, the current
results imply that teachers can utilize retrieval-based learning strategies without worrying about
potentially disparate benefits for learners with high or low levels of prior knowledge. Further, the
learning materials were not simply words or word pairs – a common critique of testing effect lab
studies. All of the to-be-learned materials and the questions used to assess that learning were
constructed from college-level textbook chapters, implying that a similar pattern of results would
be found with analogous educationally relevant material. Compared to typical lab-based memory
studies, the learning and testing phases of the current study are also more in line with actual
student behavior and testing in the classroom – learning was spread over multiple days before a
final delayed test containing retention and transfer questions.

The results pertaining to far transfer questions are also practically relevant. Even though
retrieval practice did not lead to worse performance than restudy on questions assessing
information that was initially studied, but not restudied or retrieved, there was still no benefit of
retrieval over restudy. This suggests that teachers who want to see the largest benefits of
retrieval-based learning should make sure that the to-be-tested information is included within, or
at least overlaps with, the specific retrieval practice task or questions. Although many features of
the present study are relevant to education, additional research using students in an actual
classroom, and an experimental manipulation of prior knowledge, is needed to fully vet the
similar benefits of retrieval and restudy for learners with different levels of prior knowledge.

Limitations and Future Directions

Although the current study is the first to randomly assign learners to different prior
knowledge conditions to assess the causal effect on retrieval-based learning, it is not without its
limitations. For example, all phases of the experiment were completed online, from the participant’s home, and on their own time. This makes it difficult to be certain that each participant was fully attending to each part of the entire task throughout the multiple phases. However, each page of the Qualtrics surveys was timed and contained questions that required a response from the participant, allowing for a rough estimation of task focus. Participants were also asked if they took or used any notes and/or looked up any information online at the end of each phase of the experiment. Before giving their answer, participants were told that their response would not affect the compensation they receive for already completed phases. This, coupled with the continuous required responses (acting like attention checks) and previous research showing similar results on cognitive tasks between online and in-person participants (e.g., Crump et al., 2013; Germine et al., 2012), increases confidence in the validity of the results in terms of matching actual effortful student learning and studying behavior. If more control over participant behavior is necessary, future research could use a similar procedure but bring participants back into the lab for each phase to observe their progress.

In addition to limitations inherent to online experiments, two additional points related to the generalizability of the current results warrant discussion. First, all participants were undergraduate college students who either completed the experiment for partial credit or monetary compensation. Thus, the claim that the testing effect continues to emerge under high element interactivity learning situations should be assessed with other populations that differ in age, education level, motivation, etc. One important avenue for future research would be to replicate these results with other student populations to be certain that these findings are not limited to college students who may be better at initiating and regulating their own learning.
Second, although the current study used highly relevant materials constructed from multiple textbook chapters, the results may not extend to other academic domains outside of Psychology and Geology. Future research should also assess the potential interaction between retrieval practice (or desirable difficulties in general) and prior knowledge (or element interactivity specifically) using other measures of mental effort. The majority of cognitive load research uses and advocates for the 9-point subjective mental effort rating scale developed by Paas (1992), which is why it was used in the current study. However, recent cognitive load research has expanded to other assessments of mental effort that enhance objectivity (e.g., pupil responses) and discriminate between different types of cognitive load (e.g., see Leppink, 2017). Despite these limitations, the current study is still highly relevant to education and, by randomly assigning participants to different knowledge conditions, represents the most internally valid manipulation element interactivity within a single testing effect experiment to date.

Concluding Remarks

Research on the effectiveness of retrieval-based learning has exploded over the past decade, leading many cognitive scientists to advocate for its broad implementation in the classroom (e.g., Dunlosky et al., 2013; Karpicke, 2017; Yang et al., 2021). However, these recommendations have been met with some push back from both educators (e.g., Daniel, 2012), who call for additional research on the ecological validity of the testing effect across learning contexts and individual differences critical to education, and educational psychologists (e.g., van Gog & Sweller, 2015), who question the ability of retrieval practice to enhance complex, meaningful learning (i.e., learning tasks high in element interactivity). The current study sought to tackle both of these critiques by experimentally manipulating element interactivity through the random assignment of participants to different prior knowledge conditions to assess its causal
effect (and thus the effect of element interactivity) on the benefits of retrieval-based learning. Regardless of the learning task’s level of element interactivity (i.e., learning information from trained or untrained topics), retrieval practice benefited learning more than equivalent restudy on a final delayed test of retention and near transfer, despite being rated as more effortful during learning. Further, this occurred alongside substantial effects of prior knowledge induction – for both learning conditions, information related to trained topics was easier to learn (i.e., higher performance and lower mental effort) than information related to untrained topics. Therefore, prior knowledge (and thus element interactivity) does not seem to be a critical boundary condition of the testing effect, suggesting that educators can feel confident using retrieval-based learning strategies in the classroom with students of varying levels of background knowledge.
Table 1.

<table>
<thead>
<tr>
<th>Domain text characteristics split by topic domain.</th>
<th>Words</th>
<th>Average words per Sentence</th>
<th>Flesch Reading Ease</th>
<th>Flesch-Kincaid Grade Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensation and Perception</td>
<td>3,159</td>
<td>16.98</td>
<td>52.12</td>
<td>10.21</td>
</tr>
<tr>
<td>Color Perception</td>
<td>799</td>
<td>16.31</td>
<td>59.24</td>
<td>9.05</td>
</tr>
<tr>
<td>Auditory Perception</td>
<td>760</td>
<td>17.67</td>
<td>54.97</td>
<td>9.98</td>
</tr>
<tr>
<td>Cutaneous Senses</td>
<td>766</td>
<td>15.96</td>
<td>43.10</td>
<td>11.21</td>
</tr>
<tr>
<td>Chemical Senses</td>
<td>834</td>
<td>18.13</td>
<td>50.96</td>
<td>10.66</td>
</tr>
<tr>
<td>Historical Geology</td>
<td>3,088</td>
<td>16.87</td>
<td>42.84</td>
<td>11.48</td>
</tr>
<tr>
<td>Geologic Time</td>
<td>799</td>
<td>15.08</td>
<td>43.32</td>
<td>10.96</td>
</tr>
<tr>
<td>Minerals</td>
<td>756</td>
<td>14.54</td>
<td>40.56</td>
<td>11.21</td>
</tr>
<tr>
<td>Rocks</td>
<td>752</td>
<td>18.8</td>
<td>35.39</td>
<td>12.99</td>
</tr>
<tr>
<td>Isotopic Dating</td>
<td>781</td>
<td>20.56</td>
<td>50.11</td>
<td>11.38</td>
</tr>
</tbody>
</table>
Table 2.

Prior knowledge ratings split by topic domain.

<table>
<thead>
<tr>
<th>Topic Domain</th>
<th>Trained in Sensation and Perception</th>
<th>Trained in Historical Geology</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Restudy</td>
<td>Retrieval</td>
</tr>
<tr>
<td>Sensation and Perception</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td>Mean (SD)</td>
<td>Median</td>
</tr>
<tr>
<td>3.13</td>
<td>3.23 (1.68)</td>
<td>3</td>
</tr>
<tr>
<td>Color Perception</td>
<td>4</td>
<td>4.09 (2.04)</td>
</tr>
<tr>
<td>Auditory Perception</td>
<td>4</td>
<td>3.84 (1.87)</td>
</tr>
<tr>
<td>Cutaneous Senses</td>
<td>2</td>
<td>2.31 (1.79)</td>
</tr>
<tr>
<td>Chemical Senses</td>
<td>2</td>
<td>2.69 (1.99)</td>
</tr>
<tr>
<td>Historical Geology</td>
<td>1.63</td>
<td>2.14 (1.56)</td>
</tr>
<tr>
<td>Geologic Time</td>
<td>1</td>
<td>1.91 (1.33)</td>
</tr>
<tr>
<td>Minerals</td>
<td>2</td>
<td>1.94 (1.23)</td>
</tr>
<tr>
<td>Rocks</td>
<td>2</td>
<td>2.09 (1.38)</td>
</tr>
<tr>
<td>Isotopic Dating</td>
<td>1</td>
<td>2.19 (1.91)</td>
</tr>
</tbody>
</table>
Table 3.

Pre-test proportion correct split by topic domain: Mean (SD).

<table>
<thead>
<tr>
<th></th>
<th>Trained in Sensation and Perception</th>
<th>Trained in Historical Geology</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Restudy</td>
<td>Retrieval</td>
</tr>
<tr>
<td><strong>Sensation and Perception</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Color Perception</td>
<td>.24 (.13)</td>
<td>.24 (.11)</td>
</tr>
<tr>
<td>Auditory Perception</td>
<td>.29 (.25)</td>
<td>.31 (.20)</td>
</tr>
<tr>
<td>Cutaneous Senses</td>
<td>.28 (.21)</td>
<td>.23 (.20)</td>
</tr>
<tr>
<td>Chemical Senses</td>
<td>.18 (.16)</td>
<td>.19 (.17)</td>
</tr>
<tr>
<td><strong>Historical Geology</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Geologic Time</td>
<td>.28 (.13)</td>
<td>.22 (.10)</td>
</tr>
<tr>
<td>Minerals</td>
<td>.24 (.19)</td>
<td>.24 (.21)</td>
</tr>
<tr>
<td>Rocks</td>
<td>.16 (.20)</td>
<td>.19 (.18)</td>
</tr>
<tr>
<td>Isotopic Dating</td>
<td>.28 (.29)</td>
<td>.36 (.24)</td>
</tr>
</tbody>
</table>
Table 4.

*Training topic lesson attempts split by topic domain: Number of attempts.*

<table>
<thead>
<tr>
<th></th>
<th>Restudy</th>
<th>Retrieval</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Participants per Attempt</td>
<td>Average Attempts</td>
<td>Participants per Attempt</td>
<td>Average Attempts</td>
</tr>
<tr>
<td></td>
<td>One</td>
<td>Two</td>
<td>Mean (SD)</td>
<td>One</td>
</tr>
<tr>
<td><strong>Sensation and Perception</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Color Perception</td>
<td>21</td>
<td>3</td>
<td>1.13 (0.34)</td>
<td>24</td>
</tr>
<tr>
<td>Auditory Perception</td>
<td>23</td>
<td>1</td>
<td>1.04 (0.20)</td>
<td>23</td>
</tr>
<tr>
<td>Cutaneous Senses</td>
<td>24</td>
<td>0</td>
<td>1.00 (0.00)</td>
<td>24</td>
</tr>
<tr>
<td>Chemical Senses</td>
<td>23</td>
<td>1</td>
<td>1.04 (0.20)</td>
<td>24</td>
</tr>
<tr>
<td><strong>Historical Geology</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Geologic Time</td>
<td>18</td>
<td>6*</td>
<td>1.29 (0.55)</td>
<td>21</td>
</tr>
<tr>
<td>Minerals</td>
<td>24</td>
<td>0</td>
<td>1.00 (0.00)</td>
<td>24</td>
</tr>
<tr>
<td>Rocks</td>
<td>21</td>
<td>3</td>
<td>1.13 (0.34)</td>
<td>24</td>
</tr>
<tr>
<td>Isotopic Dating</td>
<td>23</td>
<td>1</td>
<td>1.04 (0.20)</td>
<td>24</td>
</tr>
</tbody>
</table>

Note. * = One of the six participants needed three attempts.
<table>
<thead>
<tr>
<th></th>
<th>Restudy</th>
<th>Retrieval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>Mean (SD)</td>
</tr>
<tr>
<td>Sensation and Perception</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Color Perception</td>
<td>23</td>
<td>59.35 (19.73)</td>
</tr>
<tr>
<td>Auditory Perception</td>
<td>22</td>
<td>60.00 (24.64)</td>
</tr>
<tr>
<td>Cutaneous Senses</td>
<td>23</td>
<td>45.00 (13.73)</td>
</tr>
<tr>
<td>Chemical Senses</td>
<td>22</td>
<td>58.18 (24.57)</td>
</tr>
<tr>
<td>Historical Geology</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Geologic Time</td>
<td>23</td>
<td>60.87 (46.63)</td>
</tr>
<tr>
<td>Minerals</td>
<td>22</td>
<td>57.05 (35.28)</td>
</tr>
<tr>
<td>Rocks</td>
<td>22</td>
<td>65.00 (36.15)</td>
</tr>
<tr>
<td>Isotopic Dating</td>
<td>23</td>
<td>32.39 (10.54)</td>
</tr>
</tbody>
</table>

*Note.* First attempts only. Excludes 21 duration datapoints that were greater than seven hours (420 min) or were 3-SD above the mean duration for a given topic.
Table 6.

*Training topic lesson performance (proportion correct) split by topic domain: Mean (SD).*

<table>
<thead>
<tr>
<th>Topic Lesson Test</th>
<th>Knowledge Check Questions</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Restudy</td>
<td>Retrieval</td>
<td>Restudy</td>
<td>Retrieval</td>
</tr>
<tr>
<td>Sensation and Perception</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Color Perception</td>
<td>.88 (.05)</td>
<td>.86 (.05)</td>
<td>.90 (.09)</td>
<td>.93 (.07)</td>
</tr>
<tr>
<td>Auditory Perception</td>
<td>.88 (.07)</td>
<td>.87 (.05)</td>
<td>.94 (.08)</td>
<td>.96 (.06)</td>
</tr>
<tr>
<td>Cutaneous Senses</td>
<td>.89 (.07)</td>
<td>.88 (.09)</td>
<td>.94 (.06)</td>
<td>.93 (.06)</td>
</tr>
<tr>
<td>Chemical Senses</td>
<td>.88 (.07)</td>
<td>.86 (.09)</td>
<td>.94 (.07)</td>
<td>.95 (.05)</td>
</tr>
<tr>
<td>Historical Geology</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Geologic Time</td>
<td>.90 (.06)</td>
<td>.90 (.06)</td>
<td>.85 (.12)</td>
<td>.88 (.09)</td>
</tr>
<tr>
<td>Minerals</td>
<td>.89 (.07)</td>
<td>.88 (.07)</td>
<td>.95 (.05)</td>
<td>.97 (.03)</td>
</tr>
<tr>
<td>Rocks</td>
<td>.83 (.10)</td>
<td>.85 (.07)</td>
<td>.92 (.12)</td>
<td>.96 (.05)</td>
</tr>
<tr>
<td>Isotopic Dating</td>
<td>.87 (.10)</td>
<td>.83 (.13)</td>
<td>.95 (.08)</td>
<td>.93 (.08)</td>
</tr>
</tbody>
</table>

*Note.* First attempts only.
### Table 7.

*Learning phase topic text passage duration (s) split by topic domain.*

<table>
<thead>
<tr>
<th>Topic Domain</th>
<th>Restudy</th>
<th>Retrieval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>Mean (SD)</td>
</tr>
<tr>
<td><strong>Sensation and Perception</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Color Perception</td>
<td>63</td>
<td>343.81 (58.68)</td>
</tr>
<tr>
<td>Auditory Perception</td>
<td>63</td>
<td>349.68 (58.72)</td>
</tr>
<tr>
<td>Cutaneous Senses</td>
<td>63</td>
<td>363.25 (104.82)</td>
</tr>
<tr>
<td>Chemical Senses</td>
<td>62</td>
<td>353.39 (84.31)</td>
</tr>
<tr>
<td><strong>Historical Geology</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Geologic Time</td>
<td>62</td>
<td>377.34 (109.25)</td>
</tr>
<tr>
<td>Minerals</td>
<td>63</td>
<td>357.54 (83.26)</td>
</tr>
<tr>
<td>Rocks</td>
<td>61</td>
<td>364.18 (89.94)</td>
</tr>
<tr>
<td>Isotopic Dating</td>
<td>63</td>
<td>326.75 (35.41)</td>
</tr>
</tbody>
</table>

*Note.* Excludes 33 duration datapoints that were greater than 20 minutes (1200 s) or were 3-SD above the mean duration for a given topic.
### Table 8.

**Learning phase topic text passage duration (s) split by topic training: Mean (SD).**

<table>
<thead>
<tr>
<th></th>
<th>Restudy</th>
<th>Retrieval</th>
</tr>
</thead>
<tbody>
<tr>
<td>HPK Topics (lower element interactivity)</td>
<td>336.09 (40.56)</td>
<td>370.03 (78.32)</td>
</tr>
<tr>
<td>LPK Topics (higher element interactivity)</td>
<td>368.46 (74.75)</td>
<td>386.46 (90.99)</td>
</tr>
</tbody>
</table>

*Note.* Excludes 33 duration datapoints that were greater than 20 minutes (1200 s) or were 3-$SD$ above the mean duration for a given topic. HPK: High Prior Knowledge. LPK: Low Prior Knowledge.
Table 9.

*Descriptive statistics for the untrained topic within the trained domain: Mean (SD).*

<table>
<thead>
<tr>
<th></th>
<th>Restudy</th>
<th>Retrieval</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Learning Phase</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Topic Text Passage Duration (s)</td>
<td>336.09 (40.56)</td>
<td>370.03 (78.32)</td>
</tr>
<tr>
<td>Retrieval Practice Proportion Correct: 1&lt;sup&gt;st&lt;/sup&gt; Round</td>
<td>-</td>
<td>.31 (.18)</td>
</tr>
<tr>
<td>Retrieval Practice Proportion Correct: 2&lt;sup&gt;nd&lt;/sup&gt; Round</td>
<td>-</td>
<td>.67 (.23)</td>
</tr>
<tr>
<td>Subjective Mental Effort Ratings: 1&lt;sup&gt;st&lt;/sup&gt; Round</td>
<td>4.66 (1.52)</td>
<td>6.22 (1.59)</td>
</tr>
<tr>
<td>Subjective Mental Effort Ratings: 2&lt;sup&gt;nd&lt;/sup&gt; Round</td>
<td>3.88 (1.72)</td>
<td>5.13 (1.63)</td>
</tr>
<tr>
<td><strong>Testing Phase</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Final Test Duration (s)</td>
<td>360.39 (47.08)</td>
<td>346.25 (43.57)</td>
</tr>
<tr>
<td>Final Test Proportion Correct: All</td>
<td>.47 (.21)</td>
<td>.58 (.20)</td>
</tr>
<tr>
<td>Final Test Proportion Correct: Retention</td>
<td>.52 (.28)</td>
<td>.73 (.26)</td>
</tr>
<tr>
<td>Final Test Proportion Correct: Near Transfer</td>
<td>.48 (.30)</td>
<td>.54 (.31)</td>
</tr>
<tr>
<td>Final Test Proportion Correct: Far Transfer</td>
<td>.40 (.25)</td>
<td>.48 (.25)</td>
</tr>
</tbody>
</table>
Table 10.

Retrieval practice proportion correct split by topic training: Mean (SD).

<table>
<thead>
<tr>
<th></th>
<th>First Round</th>
<th>Second Round</th>
</tr>
</thead>
<tbody>
<tr>
<td>HPK Topics (lower element interactivity)</td>
<td>.41 (.14)</td>
<td>.75 (.15)</td>
</tr>
<tr>
<td>LPK Topics (higher element interactivity)</td>
<td>.31 (.13)</td>
<td>.67 (.16)</td>
</tr>
<tr>
<td>Table 11.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subjective mental effort ratings (1-9) split by topic domain: Mean (SD).</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td></td>
<td>Round 1</td>
<td>Round 2</td>
</tr>
<tr>
<td></td>
<td>Restudy</td>
<td>Retrieval</td>
</tr>
<tr>
<td><strong>Sensation and Perception</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Color Perception</td>
<td>4.58 (1.61)</td>
<td>5.76 (1.18)</td>
</tr>
<tr>
<td>Auditory Perception</td>
<td>4.70 (1.70)</td>
<td>5.92 (1.39)</td>
</tr>
<tr>
<td>Cutaneous Senses</td>
<td>5.07 (1.76)</td>
<td>6.75 (1.43)</td>
</tr>
<tr>
<td>Chemical Senses</td>
<td>4.75 (1.48)</td>
<td>6.62 (1.44)</td>
</tr>
<tr>
<td><strong>Historical Geology</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Geologic Time</td>
<td>4.44 (1.73)</td>
<td>6.33 (1.49)</td>
</tr>
<tr>
<td>Minerals</td>
<td>4.64 (1.46)</td>
<td>6.18 (1.52)</td>
</tr>
<tr>
<td>Rocks</td>
<td>5.16 (1.68)</td>
<td>6.81 (1.34)</td>
</tr>
<tr>
<td>Isotopic Dating</td>
<td>4.30 (1.86)</td>
<td>6.03 (1.51)</td>
</tr>
</tbody>
</table>
Table 12.

*Subjective mental effort ratings (1-9) split by topic training: Mean (SD).*

<table>
<thead>
<tr>
<th></th>
<th>Round 1</th>
<th>Round 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Restudy</td>
<td>Retrieval</td>
</tr>
<tr>
<td>HPK Topics (lower element interactivity)</td>
<td>4.03 (1.40)</td>
<td>5.87 (1.07)</td>
</tr>
<tr>
<td>LPK Topics (higher element interactivity)</td>
<td>5.15 (1.38)</td>
<td>6.67 (1.21)</td>
</tr>
</tbody>
</table>
Table 13.

*Testing phase final topic test duration (s) split by topic domain: Mean (SD).*

<table>
<thead>
<tr>
<th></th>
<th>Restudy</th>
<th>Retrieval</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sensation and Perception</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Color Perception</td>
<td>359.14 (48.60)</td>
<td>342.89 (47.90)</td>
</tr>
<tr>
<td>Auditory Perception</td>
<td>336.72 (42.19)</td>
<td>330.47 (37.83)</td>
</tr>
<tr>
<td>Cutaneous Senses</td>
<td>381.56 (46.44)</td>
<td>365.23 (48.18)</td>
</tr>
<tr>
<td>Chemical Senses</td>
<td>367.81 (43.98)</td>
<td>344.53 (41.72)</td>
</tr>
<tr>
<td><strong>Historical Geology</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Geologic Time</td>
<td>377.11 (50.33)</td>
<td>362.11 (43.76)</td>
</tr>
<tr>
<td>Minerals</td>
<td>345.15 (48.67)</td>
<td>331.88 (38.41)</td>
</tr>
<tr>
<td>Rocks</td>
<td>359.30 (48.57)</td>
<td>351.64 (48.03)</td>
</tr>
<tr>
<td>Isotopic Dating</td>
<td>356.56 (46.86)</td>
<td>344.14 (40.49)</td>
</tr>
</tbody>
</table>
Table 14.

*Testing phase final topic test duration (s) split by topic training: Mean (SD).*

<table>
<thead>
<tr>
<th></th>
<th>Restudy</th>
<th>Retrieval</th>
</tr>
</thead>
<tbody>
<tr>
<td>HPK Topics (lower element interactivity)</td>
<td>359.58 (38.43)</td>
<td>346.07 (37.23)</td>
</tr>
<tr>
<td>LPK Topics (higher element interactivity)</td>
<td>361.06 (36.90)</td>
<td>347.11 (31.24)</td>
</tr>
</tbody>
</table>
Figures

A

**Question:** Theoretically, paint that appears green must absorb what wavelength(s) of light?

**Answer:**

**Correct Answer:** Short and long wavelengths

**Your Answer:**

**Explanation/Detailed Answer:** Paints reflect some wavelengths but absorb others, and any wavelengths absorbed are not reflected into the eye. For example, paint that appears green must reflect only medium wavelengths and thus absorb both short and long wavelengths.

- Full Credit: My answer was correct and contained all of the important information
- Partial Credit: My answer was close and/or contained most of the important information
- No Credit: My answer was wrong and/or was missing most of the important information

B

**Explanation/Detailed Answer:** Paints reflect some wavelengths but absorb others, and any wavelengths absorbed are not reflected into the eye. For example, theoretically, paint that appears green must reflect only medium wavelengths and thus absorb both short and long wavelengths.

- Full Understanding: I fully understand the above information/example
- Partial Understanding: I understand some, but not all, of the above information/example
- No Understanding: I do not understand the above information/example

*Figure 1.* Sample learning phase retrieval/restudy materials. A: Retrieval practice question and feedback. B: Corresponding restudy example/key idea.
Figure 2. Sample learning and testing phase near transfer materials. A: Retrieval practice question and feedback. B: Corresponding final test near transfer question.
Figure 3. Overview of entire experimental procedure.
Figure 4. Overview of learning phase for the A-B domain order condition.
Figure 5. Final test proportion correct (+/- SE) as a function of learning condition and element interactivity (prior knowledge), averaging across question type.
Figure 6. Final test proportion correct (+/− SE) on retention questions as a function of learning condition and element interactivity (prior knowledge).
Figure 7. Final test proportion correct (+/- SE) on near transfer questions as a function of learning condition and element interactivity (prior knowledge).
Figure 8. Final test proportion correct (+/- SE) on far transfer questions as a function of learning condition and element interactivity (prior knowledge).
Figure 9. Final test mean testing effect (+/- SE) as a function of element interactivity (prior knowledge) and question type.
### APPENDIX A: COUNTERBALANCING CONDITIONS

#### Table A.1

<table>
<thead>
<tr>
<th>Domain Training</th>
<th>Training Lesson 1</th>
<th>Training Lesson 2</th>
<th>Training Lesson 3</th>
<th>Learning Condition</th>
<th>Learning and Testing Phase Order</th>
</tr>
</thead>
<tbody>
<tr>
<td>S&amp;P</td>
<td>Color</td>
<td>Aud</td>
<td>Cut</td>
<td>Restudy</td>
<td>S&amp;P – Geo</td>
</tr>
<tr>
<td>S&amp;P</td>
<td>Color</td>
<td>Aud</td>
<td>Cut</td>
<td>Restudy</td>
<td>Geo – S&amp;P</td>
</tr>
<tr>
<td>S&amp;P</td>
<td>Color</td>
<td>Aud</td>
<td>Cut</td>
<td>Retrieval</td>
<td>S&amp;P – Geo</td>
</tr>
<tr>
<td>S&amp;P</td>
<td>Color</td>
<td>Aud</td>
<td>Cut</td>
<td>Retrieval</td>
<td>Geo – S&amp;P</td>
</tr>
<tr>
<td>S&amp;P</td>
<td>Aud</td>
<td>Cut</td>
<td>Chem</td>
<td>Restudy</td>
<td>S&amp;P – Geo</td>
</tr>
<tr>
<td>S&amp;P</td>
<td>Aud</td>
<td>Cut</td>
<td>Chem</td>
<td>Restudy</td>
<td>Geo – S&amp;P</td>
</tr>
<tr>
<td>S&amp;P</td>
<td>Aud</td>
<td>Cut</td>
<td>Chem</td>
<td>Retrieval</td>
<td>S&amp;P – Geo</td>
</tr>
<tr>
<td>S&amp;P</td>
<td>Aud</td>
<td>Cut</td>
<td>Chem</td>
<td>Retrieval</td>
<td>Geo – S&amp;P</td>
</tr>
<tr>
<td>S&amp;P</td>
<td>Cut</td>
<td>Chem</td>
<td>Color</td>
<td>Restudy</td>
<td>S&amp;P – Geo</td>
</tr>
<tr>
<td>S&amp;P</td>
<td>Cut</td>
<td>Chem</td>
<td>Color</td>
<td>Restudy</td>
<td>Geo – S&amp;P</td>
</tr>
<tr>
<td>S&amp;P</td>
<td>Chem</td>
<td>Color</td>
<td>Aud</td>
<td>Restudy</td>
<td>S&amp;P – Geo</td>
</tr>
<tr>
<td>S&amp;P</td>
<td>Chem</td>
<td>Color</td>
<td>Aud</td>
<td>Restudy</td>
<td>Geo – S&amp;P</td>
</tr>
<tr>
<td>S&amp;P</td>
<td>Chem</td>
<td>Color</td>
<td>Aud</td>
<td>Retrieval</td>
<td>S&amp;P – Geo</td>
</tr>
<tr>
<td>S&amp;P</td>
<td>Chem</td>
<td>Color</td>
<td>Aud</td>
<td>Retrieval</td>
<td>Geo – S&amp;P</td>
</tr>
<tr>
<td>Geo</td>
<td>Time</td>
<td>Min</td>
<td>Rocks</td>
<td>Restudy</td>
<td>S&amp;P – Geo</td>
</tr>
<tr>
<td>Geo</td>
<td>Time</td>
<td>Min</td>
<td>Rocks</td>
<td>Restudy</td>
<td>Geo – S&amp;P</td>
</tr>
<tr>
<td>Geo</td>
<td>Time</td>
<td>Min</td>
<td>Rocks</td>
<td>Retrieval</td>
<td>S&amp;P – Geo</td>
</tr>
<tr>
<td>Geo</td>
<td>Time</td>
<td>Min</td>
<td>Rocks</td>
<td>Retrieval</td>
<td>Geo – S&amp;P</td>
</tr>
<tr>
<td>Geo</td>
<td>Min</td>
<td>Rocks</td>
<td>Iso</td>
<td>Restudy</td>
<td>S&amp;P – Geo</td>
</tr>
<tr>
<td>Geo</td>
<td>Min</td>
<td>Rocks</td>
<td>Iso</td>
<td>Restudy</td>
<td>Geo – S&amp;P</td>
</tr>
<tr>
<td>Geo</td>
<td>Min</td>
<td>Rocks</td>
<td>Iso</td>
<td>Retrieval</td>
<td>S&amp;P – Geo</td>
</tr>
<tr>
<td>Geo</td>
<td>Min</td>
<td>Rocks</td>
<td>Iso</td>
<td>Retrieval</td>
<td>Geo – S&amp;P</td>
</tr>
<tr>
<td>Geo</td>
<td>Rocks</td>
<td>Iso</td>
<td>Time</td>
<td>Restudy</td>
<td>S&amp;P – Geo</td>
</tr>
<tr>
<td>Geo</td>
<td>Rocks</td>
<td>Iso</td>
<td>Time</td>
<td>Restudy</td>
<td>Geo – S&amp;P</td>
</tr>
<tr>
<td>Geo</td>
<td>Rocks</td>
<td>Iso</td>
<td>Time</td>
<td>Retrieval</td>
<td>S&amp;P – Geo</td>
</tr>
<tr>
<td>Geo</td>
<td>Rocks</td>
<td>Iso</td>
<td>Time</td>
<td>Retrieval</td>
<td>Geo – S&amp;P</td>
</tr>
<tr>
<td>Geo</td>
<td>Iso</td>
<td>Time</td>
<td>Min</td>
<td>Restudy</td>
<td>S&amp;P – Geo</td>
</tr>
<tr>
<td>Geo</td>
<td>Iso</td>
<td>Time</td>
<td>Min</td>
<td>Restudy</td>
<td>Geo – S&amp;P</td>
</tr>
<tr>
<td>Geo</td>
<td>Iso</td>
<td>Time</td>
<td>Min</td>
<td>Retrieval</td>
<td>S&amp;P – Geo</td>
</tr>
<tr>
<td>Geo</td>
<td>Iso</td>
<td>Time</td>
<td>Min</td>
<td>Retrieval</td>
<td>Geo – S&amp;P</td>
</tr>
</tbody>
</table>

*Note. N = 4 per row. S&P: Sensation and Perception domain; Geo: Historical Geology domain; Color: Color Perception topic; Aud: Auditory Perception topic; Cut: Cutaneous Senses topic; Chem: Chemical Senses topic; Time: Geologic Time topic; Min: Minerals topic; Rocks: Rocks topic; Iso: Isotopic Dating topic.*
APPENDIX B: SAMPLE PRE-TEST QUESTIONS

Domain A: Sensation and Perception

Topic A.1: Color Perception
Question: What two structures of the human eye focus incoming light onto the visual receptors?
   a) Cornea and Retina  
   b) Cornea and Lens  
   c) Pupil and Lens  
   d) Pupil and Retina

Topic A.2: Auditory Perception
Question: Which of the following ear components separates the middle ear from the inner ear?
   a) The Oval Window  
   b) The Tympanic Membrane (or Eardrum)  
   c) The Auditory Canal  
   d) The Basilar Membrane

Topic A.3: Cutaneous Senses
Question: Merkel receptors are most associated with which of the following perceptions?
   a) Movement across the skin  
   b) Fine details and shape  
   c) Dull pain and cold temperature  
   d) Fine textures and rapid vibration

Topic A.4: Chemical Senses
Question: Which of the following types of papillae does not contain taste buds?
   a) Filiform Papillae  
   b) Fungiform Papillae  
   c) Circumvallate Papillae  
   d) Foliate Papillae
Domain B: Historical Geology

Topic B.1: Geologic Time
Question: Which of the following represents the correct ordering of Earth’s major eons?
   a) The Hadean -> The Proterozoic -> The Archean -> The Phanerozoic
   b) The Archean -> The Hadean -> The Phanerozoic -> The Proterozoic
   c) The Hadean -> The Archean -> The Proterozoic -> The Phanerozoic
   d) The Archean -> The Proterozoic -> The Hadean -> The Phanerozoic

Topic B.2: Minerals
Question: What is the most abundant element in the Earth’s crust, making up almost 50%?
   a) Silicon
   b) Oxygen
   c) Iron
   d) Carbon

Topic B.3: Rocks
Question: Which of the following pairs of rocks and rock types is incorrect?
   a) Pumice – Igneous
   b) Gneiss – Metamorphic
   c) Granite – Igneous
   d) Slate – Sedimentary

Topic B.4: Isotopic Dating
Question: What type of radioactive decay occurs when Uranium-238 decays into Thorium-234?
   a) Alpha Decay
   b) Beta Decay
   c) Half-Life Decay
   d) Electron Capture
APPENDIX C: SAMPLE TRAINING TOPIC LESSON

Color Perception: Introduction

Color is one of the most obvious and pervasive qualities in our environment. We interact with it every time we note the color of a traffic light, choose clothes that are color coordinated, or appreciate the colors of a painting. We pick favorite colors (blue is the most favored), we associate colors with emotions (we turn purple with rage, red with embarrassment, green with envy, and feel blue), and we imbue colors with special meanings (for example, in many cultures red signifies danger; purple, royalty; green, ecology). But for all of our involvement with color, we sometimes take it for granted, and just as with our other perceptual abilities, we may not fully appreciate color unless we lose our ability to experience it.

Color serves important signaling functions, both natural and contrived by humans. The natural and human-made world provides many color signals that help us identify and classify things: we know a banana is ripe when it has turned yellow, and we know to stop when the traffic light turns red.

In addition to its signaling function, color helps facilitate perceptual organization, by which similar elements become grouped together and objects are segregated from their backgrounds. Color’s role in perceptual organization is crucial to the survival of many species. Consider, for example, a monkey foraging for fruit in the forest or jungle. A monkey with good color vision easily detects red fruit against a green background, but a color-blind monkey would find it more difficult to find the fruit. Color vision thus enhances the contrast of objects that, if they didn’t appear colored, would be more difficult to perceive. This link between good color vision and the ability to detect colored food has led to the proposal that monkey and human color vision may have evolved for the express purpose of detecting fruit.

![Image](image_url)

1. Which of the following is true about color?
   a. It serves important signaling functions
   b. It facilitates perceptual organization
   c. It may have evolved for the express purpose of detecting fruit
   d. All of the above

2. A monkey with good color vision _____.
   a. would have difficulty with figure-ground segregation
   b. would have a better chance of surviving than a color-blind monkey
   c. would have an equal chance of survival as a color-blind monkey
   d. would be very atypical, since most monkeys are color-blind

**Vision**

The ability to see a tree, or any other object, depends on light being reflected from that object into the eye. Information about the object is carried in light reflected from the object and into the eye. When this light reaches the receptors in the retina it becomes transformed into electrical signals that contain information about the object, which are transmitted to the brain. Eventually, these electrical signals become transformed into a perception of the object.

1. What happens when light reflects off an object into our eye?
   a. We are blinded.
   b. *We see the object.*
   c. We are unable to focus on the object.
   d. None of the above
The Human Eye

When we see an object, light reflected from that object passes through the pupil, an adjustable opening in the eye. The pupil widens and narrows to control how much light enters the eye. The iris is the colored structure on the surface of the eye surrounding the pupil and controls the pupil’s size based on the amount of incoming light. It is the structure we describe when we say someone has brown, green, or blue eyes. Light passing through the pupil travels through the vitreous humor (a clear jellylike substance) to strike the retina, a layer of visual receptors covering the back surface of the eyeball.

Focusing Light: The Cornea and the Lens

Light reflected from an object into the eye is focused onto the retina by a two-element optical system: the cornea and the lens. The cornea, the transparent covering of the front of the eye, accounts for about 80% of the eye’s focusing power, but like the lenses in eyeglasses, it is fixed in place so it can’t adjust its focus. The lens, which supplies the remaining 20% of the eye’s focusing power, can change its shape to adjust the eye’s focus for objects located at different distances. This change in shape is achieved by the action of ciliary muscles, which increase the focusing power of the lens (its ability to bend light) by increasing its curvature.

The adjustable lens, which controls a process called accommodation, comes to the rescue to help prevent blurring. Accommodation is the change in the lens’s shape that occurs when the ciliary muscles at the front of the eye tighten and increase the curvature of the lens so that it gets thicker. This increased curvature increases the bending of the light rays passing through the lens so the focus point is pulled back to A to create a sharp image on the retina. This means that as you look around at different objects, your eye is constantly adjusting its focus by accommodating, especially for nearby objects.
The Retina and the Visual Receptors

Focusing light onto the retina activates two types of visual receptors—the rods and the cones—which absorb light and transmit the information as an electrical signal. The rods and cones differ in many ways, including their appearance, locations, and functions. The receptors get their names because of their rod- and cone-shaped outer segments. The rod and cone receptors are also distributed differently across the retina. Of all the visual receptors in the human retina, about 5% are cones. Although 5% may not sound like much, the cone-rich parts of the retina send more signals to the visual cortex in the brain than do the rod-rich areas. The fovea (FOE-vee-uh), the central area of the human retina, is adapted for detailed vision and only contains cones. Of all retinal areas, the fovea has the greatest density of receptors. When we look directly at an object, the object’s image falls on the fovea.

The peripheral retina, which includes all of the retina outside of the fovea, contains both rods and cones. It is important to note that although the fovea has only cones, there are also many cones in the peripheral retina. The fovea is so small (about the size of this “o”) that it contains only about 1 percent, or 50,000, of the 6 million cones in the retina. The peripheral retina contains many more rods than cones because there are about 120 million rods and only 6 million cones in the retina.

The table below summarizes key functional differences between the rods and cones.
1. What structure is responsible for the majority of the eye’s focusing power?
   a. The iris
   b. The pupil
   c. The cornea
   d. The lens

2. The ciliary muscles change the shape of the ______.
   a. lens
   b. pupil
   c. cornea
   d. iris

3. Which structure focuses light to the same degree, regardless of the distance to the object?
   a. The pupil
   b. The cornea
   c. The iris
   d. The lens

4. What eye structure provides about 20% of the eye’s focusing power?
   a. The iris
   b. The pupil
   c. The cornea
   d. The lens
5. The process in which the lens adjusts its shape depending on the distance between the eye and the object viewed in order to project a clear image onto the retina is _____.
   a. accommodation
   b. focusing
   c. constriction
   d. dilation

6. Where are the rods and cones?
   a. In the cornea
   b. In the retina
   c. In the pupil
   d. In the fovea

7. Which part of the retina, if any, has the best color vision?
   a. The periphery
   b. The fovea
   c. The area surrounding the blind spot
   d. All areas have equally good color vision

8. Which of the following is true about the visual receptors?
   a. Cones are more sensitive to light than rods
   b. Cones have better detailed vision/acuity than rods
   c. There are many more cones than rods in the retina
   d. All of the above are true

**Light Wavelengths**

Vision is based on visible light. What we call visible light is part of the electromagnetic spectrum, which is a continuum of electromagnetic energy that is produced by electric charges and is radiated as waves. The energy in this spectrum can be described by its wavelength – the distance between the peaks of the electromagnetic waves. Wavelengths in the electromagnetic spectrum range from gamma rays and x-rays with very short wavelengths, through ultraviolet, visible light, and infrared, to radio and TV transmissions with very long wavelengths.
What makes light visible? The answer is our receptors, which respond to wavelengths from 400 to 700 nanometers (nm). With different receptors, we would see a different range of wavelengths. In fact, many insects and birds see ultraviolet wavelengths, which are invisible to humans. For humans and some other animals, the wavelength of visible light is associated with the different colors of the spectrum, with short wavelengths appearing blue, middle wavelengths green, and long wavelengths yellow, orange, and red. Additionally, light waves with higher amplitudes/intensities are perceived as brighter. The amplitude of a wave is similar to the height of the wave. We will return to the relationship between color and our receptors later.

1. The electromagnetic spectrum is a continuum of electromagnetic energy that is produced by _____ and is radiated as _____.
   a. electric charges; waves
   b. magnetism; waves
   c. electric charges; magnetism
   d. magnetism; electric charges

2. What is a light wave’s wavelength?
   a. The distance between two wave peaks
   b. The distance between a wave peak and the next wave trough
   c. The time it takes for one full wave cycle
   d. The height of the wave peak compared to the wave trough
3. Which of the following pairs is most incorrect?
   a. Visible light = 400-700 nanometers (nm)
   b. Radio waves > 700 nanometers (nm)
   c. Infrared light < 400 nanometers (nm)
   d. X-rays < 400 nanometers (nm)

4. Why does all light have a wavelength between about 400 and 700 nm?
   a. Wavelengths outside that range cannot travel very far through the air.
   b. We define light as the wavelengths that stimulate our receptors.
   c. Wavelengths outside that range travel more slowly.
   d. The air acts as a prism to convert other wavelengths toward this range.

5. The wavelength of light mainly affects our perception of _____.
   a. color
   b. brightness
   c. saturation
   d. light purity

**Light: Reflection, Transmission, and Absorption**

The colors of light in the spectrum are related to their wavelengths, but what about the colors of objects? The colors of objects are largely determined by the wavelengths of light that are reflected from the objects into our eyes. Chromatic colors, such as blue, green, and red, occur when some wavelengths are reflected more than others, a process called selective reflection. A sheet of paper that reflects long wavelengths of light and absorbs short and medium wavelengths would appear red because only the long wavelengths reach our eyes.

Achromatic colors, such as white, gray, and black, occur when light is reflected equally across the spectrum. A sheet of paper that reflects all wavelengths of light equally appears white.

Individual objects don’t usually reflect a single wavelength of light, however. The figure below shows reflectance curves that plot the percentage of light reflected from lettuce and tomatoes at each wavelength in the visible spectrum. Notice that both vegetables reflect a range of wavelengths, but each selectively reflects more light in one part of the spectrum. Tomatoes predominantly reflect long wavelengths of light into our eyes, whereas lettuce principally reflects medium wavelengths. As a result, tomatoes appear red whereas lettuce appears green.
You can also contrast the reflectance curves for the lettuce and tomato with the curves for the achromatic (black, gray, and white) pieces of paper, which are flat, indicating equal reflectance across the spectrum. The difference between black, gray, and white is related to the overall amount of light reflected from an object. The black paper in reflects less than 10% of the light that hits it, whereas the white paper reflects more than 80% of the light.

Although most colors in the environment are created by the way objects selectively reflect some wavelengths, the color of things that are transparent, such as liquids, plastics, and glass, is created by selective transmission. Selective transmission means that only some wavelengths pass through the object or substance. For example, cranberry juice selectively transmits long-wavelength light and appears red, whereas limeade selectively transmits medium-wavelength light and appears green.
1. The reflectance curve is a plot of the light reflected off a surface as a function of _____.
   a. spatial frequency
   b. contrast
   c. wavelength
   d. orientation

2. The reflectance curve for a white piece of paper would reflect _____.
   a. mostly short wavelengths, a moderate amount of medium wavelengths, and a little of the long wavelengths
   b. mostly long wavelengths, a small amount of medium wavelengths, and a little of the short wavelengths
   c. a little of short wavelengths, a large amount of medium wavelengths, and a little of the long wavelengths
   d. long, medium and short wavelengths equally

3. The reflectance curve for a purple piece of paper will reflect _____.
   a. short wavelengths
   b. long wavelengths only
   c. all wavelengths equally
   d. long and short wavelengths

4. Which of the following would produce the perception of a chromatically colored object?
   a. An object that selectively transmits some wavelengths
   b. An object that selectively reflects some wavelengths
   c. An object that selectively absorbs some wavelengths
   d. All of the above

**Perceptual Dimensions of Color**

How can we perceive millions of colors when we can describe the visible spectrum in terms of only six or seven colors? The answer is that there are three perceptual dimensions of color, which together can create the large number of colors we can perceive. The three perceptual dimensions of color are: (1) Hue, (2) Saturation, and (3) Value.

The first dimension is hue, which really just means color. Specifically, hue refers to the peak wavelength that reflects off or transmits through an object into our eye. For example, each of the circles below has a red hue. However, the circles vary in the other two dimensions of color – saturation and value.
Saturation is determined by the amount of white that has been added to a particular hue. Specifically, saturation refers to the spread or variance of the wavelengths reflected off or transmitted through an object into our eye. In the circles above, progressively more white has been added to each as you move left-to-right and, as a result, saturation decreases. As hues become desaturated, they can take on a faded or washed-out appearance.

Value refers to the light-to-dark dimension of color. Specifically, value refers to the height or amplitude of the wavelengths reflected off or transmitted through an object into our eye. In the circles above, as you move down, value decreases and the colors become darker.

Another useful way to illustrate the relationship between hue, saturation, and value is to arrange colors systematically within a three-dimensional color space called a color solid. The cylindrical color solid below is called the HSV color solid, because its three dimensions are Hue, Saturation, and Value. Different hues are arranged around the circumference of the cylinder with perceptually similar hues placed next to each other. Notice, in fact, that the order of the hues around the cylinder matches the order of the colors in the visible spectrum. Saturation is depicted by placing more saturated colors toward the outer edge of the cylinder and more desaturated colors toward the center. Value is represented by the cylinder’s height, with lighter colors at the top and darker colors at the bottom. The color solid therefore creates a coordinate system in which our perception of any color can be defined by hue, saturation, and value.
1. Which of the following is not a perceptual dimension of color?
   a. Wavelength
   b. Hue
   c. Saturation
   d. Value

2. Which perceptual dimension of color is determined by the spread or variance of the wavelengths reflected off or transmitted through an object?
   a. Hue
   b. Saturation
   c. Value
   d. None of the above

3. What physical dimension of color is most associated with hue?
   a. Peak reflected or transmitted wavelength
   b. Amplitude
   c. Saturation
   d. Spread or variance of the wavelengths

4. By changing _____, we can create about a million (or more) discriminable colors.
   a. saturation only
   b. value only
   c. hue and saturation, but not value
   d. saturation, value, and hue

5. Adding more white to a color changes the color’s _____.
   a. hue
   b. wavelength
   c. brightness
   d. saturation
6. If a person views three lights that differ only in amplitude, the person would perceive the lights as _____.
   a. differing in brightness  
   b. different colors  
   c. differing in brightness and color  
   d. different shades of the same color

**Specialized Cones**

Now we can return to the relationship between light wavelengths and our receptors. How does the visual system convert wavelengths of light into a perception of color? The process begins with three kinds of specialized receptors, called cones. According to the Trichromatic Theory of Color Vision, developed by Thomas Young in the 1700s and elaborated on by Hermann von Helmholtz in the 1800s, color vision depends on the relative responses of three types of cones. One type is most sensitive to short wavelengths (which we generally see as blue), another to medium wavelengths (seen as green), and another to long wavelengths (seen as red). White light excites all types equally. Every wavelength of light produces its own distinct ratio of responses by the three kinds of cones, which is interpreted and perceived as a specific color by the brain.

![Figure 4.11 Sensitivity of three types of cones to different wavelengths of light. (Based on data of Bowmaker & Dartnall, 1980)](Figure_4.11.png)
1. The trichromatic (Young-Helmholtz) theory of color vision states that color perception is due to ______.
   a. the pattern of activity in four different receptors mechanisms
   b. the activity pattern in the occipital, parietal, and temporal cortical lobes
   c. the pattern of activity in three different receptor mechanisms
   d. processing in layers 1, 2, and 3 in the LGN

2. According to the trichromatic (Young-Helmholtz) theory, what causes perception of green?
   a. The activity of cones is greatest at a point between the fovea and the periphery.
   b. Most of the cones are firing at an intermediate frequency.
   c. The medium-wavelength cones are more active than the other ones.
   d. The velocity of action potentials from the retina is at an intermediate level.

3. According to the trichromatic theory, how do we distinguish red from orange?
   a. The two colors produce different velocities of action potentials.
   b. The two colors produce different ratios of responses by types of cones.
   c. The two colors produce excitation in different areas of the brain.
   d. The two colors produce different durations of response by most of the cones.

**Color Mixing**

A useful way to demonstrate the relationship between wavelengths, receptors, and perceived colors is by mixing different lights together and by mixing different paints together.

**Mixing Lights: Part 1**

Mixing lights is different than mixing paints. To demonstrate, imagine you have a set of flashlights that each produce a different color. If you were to shine a green flashlight on a white wall, you would see a green spot of light because the light reflected back into your eyes is mostly medium wavelength light. When medium wavelength light reaches the eye, it produces a specific pattern of activity in which the S- and L-cones are activated much less than the M-cones. This pattern of electrical activity is sent to the brain, which interprets it as the color green.
The same process occurs for our other flashlights. The blue flashlight produces a spot of blue light on the wall. This occurs because the light is mostly short wavelengths, which excites our S-cones much more than our M- and L-cones, creating a specific ratio of activity that the brain interprets as blue. The red flashlight works the exact same way. Its mostly long wavelengths excite the L-cones much more than the S- and M-cones, causing a pattern of activity interpreted by the brain as the color red.

So, now imagine you take the green and red flashlights out and you position them so that their spots on the wall overlap. What color would that overlapping area be? Yellow, which may not have been your first guess. The overlapping light would contain both medium wavelength (green) light as well as long wavelength (red) light. This produces a specific pattern of activity in which the S-cones are much less excited than the M- and L-cones. This ratio of responding is translated into the color yellow by the brain. Thus, when you perceive the color yellow, it means that the reflected or transmitted light that reaches your eye is a combination of medium and long wavelengths.

1. What color would most likely be perceived if the light reflected into the eye activates the long-cones much more than the short- and medium-cones?
   a. Blue
   b. Green
   c. Red
   d. Yellow
2. What color would most likely be perceived if the light reflected into the eye activates the short-cones much more than the medium- and long-cones?
   a. Blue
   b. Green
   c. Red
   d. Yellow

3. Green and red light are projected on a white screen. What color will the overlapping spot on the screen appear to be?
   a. White
   b. Yellow
   c. Blue
   d. Purple

Mixing Lights: Part 2

Now, what if you were to shine the blue flashlight directly on top of that overlapping, yellow spot? By adding short wavelength light to the mixture, all three types of cones would respond equally, producing a ratio of activity that results in white. Thus, light that contains equal amounts of all wavelengths is perceived as the color white. Sunlight can be thought of as white light because it contains all wavelengths of visible light.

What about the other possible combinations? So far we know that M- and L-cone responding results in the color yellow and S-, M-, and L-cone responding results in the color white. Next, imagine we take out the blue and green flashlights and shine them on the wall to create an overlapping spot of light. This mixture of short and medium wavelengths produces a pattern of activity in our cones (i.e., much more activity in the S- and M-cones than in the L-cones) that is perceived as the color cyan. Similarly, the overlapping spot of light seen when shining the blue and red flashlights is seen as magenta. The combination of short and long wavelengths causes our S- and L-cones to respond much more than our M-cones, and this ratio of responding is interpreted by the brain as the color magenta.
Based on these findings, researchers developed the RGB color model, where red, green, and blue act as primary colors and cyan, magenta, and yellow act as secondary colors.

![RGB Color Model](https://commons.wikimedia.org/wiki/File:AdditiveColor.svg)

1. Blue and red light are projected on a white screen. What color will the overlapping spot on the screen appear to be?
   a. White
   b. Yellow
   c. Cyan
   d. Magenta

2. Blue and green light are projected on a white screen. What color will the overlapping spot on the screen appear to be?
   a. White
   b. Yellow
   c. Cyan
   d. Magenta

3. Cyan light is projected on a white screen. What pattern of cone-activity would the light reflected off the screen into the eye most likely produce?
   a. Much greater activity in the short- and long-cones than in the medium-cones
   b. *Much greater activity in the short- and medium-cones than in the long-cones*
   c. Much greater activity in the medium- and long-cones than in the short-cones
   d. Fairly equal activity across all three types of cones

4. Magenta light is projected on a white screen. What pattern of cone-activity would the light reflected off the screen into the eye most likely produce?
   a. *Much greater activity in the short- and long-cones than in the medium-cones*
   b. Much greater activity in the short- and medium-cones than in the long-cones
   c. Much greater activity in the medium- and long-cones than in the short-cones
   d. Fairly equal activity across all three types of cones
5. Yellow and blue light are projected on a white screen. What color will the overlapping spot on the screen appear to be?
   a. White
   b. Black
   c. Cyan
   d. Magenta

6. Green and magenta light are projected on a white screen. What color will the overlapping spot on the screen appear to be?
   a. White
   b. Black
   c. Cyan
   d. Yellow

7. Which of the following sets of colored lights would produce the perception of white in the overlapping spot?
   a. Yellow, Red, and Green
   b. Blue, Green, and Red
   c. Cyan, Blue, Green
   d. Blue, Magenta, Red

8. Which of the following is not a primary color in the RGB color model?
   a. Red
   b. Yellow
   c. Blue
   d. Green

Mixing Paints: Part 1

But wait… When you mixed green and red paint together as a child, you didn’t end up with yellow. That’s because mixing paints works differently. Remember that the color we perceive depends on the wavelengths reflected into our eye. Paints, like most other objects, reflect some wavelengths but absorb others. Any wavelengths absorbed by an object are not reflected into the eye. This selective reflection produces the perception of a chromatic color. Thinking back to the reflectance curves discussed earlier, a tomato would need to reflect mostly long wavelengths for us to perceive it as red. By that logic, it must also absorb short and medium wavelengths so that they don’t reach the eye. So, red paint reflects long wavelengths and absorbs short and medium wavelengths. As we said before, when mostly long wavelengths reach the eye,
it causes the L-cones to respond much more than the S- and M-cones, resulting in the perception of the color red.

So, what about cyan paint? We now know that magenta is perceived from a specific pattern of activity in which the S- and M-cones are activated much more than the L-cones. That would mean cyan paint selectively reflects short and medium wavelengths and absorbs long wavelengths. The same logic applies to magenta paint, which selectively reflects short and long wavelengths and absorbs medium wavelengths, and to yellow paint, which selectively reflects medium and long wavelengths and absorbs short wavelengths.

1. Magenta paint must selectively reflect which of the following light-wavelengths?
   a. Short-wavelengths only
   b. Short- and medium-wavelengths only
   c. Short- and long-wavelengths only
   d. Medium- and long-wavelengths only

2. Yellow paint must selectively reflect which of the following light-wavelengths?
   a. Medium-wavelengths only
   b. Short- and medium-wavelengths only
   c. Short- and long-wavelengths only
   d. Medium- and long-wavelengths only

3. Cyan paint must selectively absorb which of the following light-wavelengths?
   a. Medium-wavelengths only
   b. Short- and medium-wavelengths only
   c. Medium- and long-wavelengths only
   d. Long-wavelengths only
4. Which of the following light-wavelengths, if any, do red and blue paint both selectively absorb?
   a. Short-wavelengths only
   b. *Medium-wavelengths only*
   c. Long-wavelengths only
   d. Medium- and long-wavelengths only

5. Which of the following light-wavelengths, if any, do yellow and green paint both selectively reflect?
   a. Short-wavelengths only
   b. *Medium-wavelengths only*
   c. Long-wavelengths only
   d. Medium- and long-wavelengths only

**Mixing Paints: Part 2**

Keeping this in mind, what color would we see if we mixed magenta and yellow paint? Theoretically, if it was perfectly mixed such that the mixture absorbed the same wavelengths that each individual paint absorbed, we would see red because only long wavelength light is not absorbed. Likewise, mixing cyan (absorbs long wavelengths) and magenta (absorbs medium wavelengths) gives you blue, and mixing cyan with yellow, gives you green.

Unlike with lights, mixing many paints together doesn’t produce white. Instead, the mixture will turn dark brown and eventually black. Because the mixture absorbs any and all wavelengths absorbed by each individual paint, little is reflected into the eye and we perceive an absence of color – black.
Researchers have also developed a color model for paint mixtures, the CMYK model, where cyan, magenta, and yellow are the primary colors and blue, green, and red are now the secondary colors (K = black). If these sound familiar, it’s because the RBG model is used with media that transmit light (e.g., televisions), while the CMYK model is used to produce printed colors (e.g., inks and dyes).

1. If you were to mix yellow and cyan paint together, the mixture would be seen as which of the following colors?
   a. White
   b. Blue
   c. Green
   d. Magenta

2. If you were to mix yellow and magenta paint together, the mixture would selectively reflect which of the following light-wavelengths?
   a. Short-wavelengths only
   b. Medium-wavelengths only
   c. Long-wavelengths only
   d. Medium- and long-wavelengths only

3. If you were to mix cyan and magenta paint together, the mixture would selectively reflect which of the following light-wavelengths?
   a. Short-wavelengths only
   b. Medium-wavelengths only
   c. Long-wavelengths only
   d. Short- and medium-wavelengths only
4. Which of the following is not a primary color in the CMYK color model?
   a. Cyan
   b. Magenta
   c. Yellow
   d. Green

**Objects Under Different Lights**

As previously mentioned, sunlight can be thought of as white light because it contains all visible light wavelengths. But what happens when an object is placed under a different light? Normally, our brain can figure out the “normal” color of the object by comparing the cone activity produced by that object to the activity produced by other nearby objects. For example, if you were to replace your light bulb with a red bulb that only produces long wavelength light, everything would have a red tint to it. Your brain, which knows the “normal” colors of objects, would notice each object is now causing a pattern of cone activity that is much more L-cone dominated. So if every object produces a ratio of activity that is dominated by the L-cones, your brain can understand that it is probably not that the objects themselves are all red, rather it is more likely that the usual white light of the sun has been replaced by a light producing a different proportion of wavelengths.

Thus, under normal conditions, our brains can achieve color constancy by comparing multiple objects in the environment as well as using prior knowledge of an object’s color. However, what if we were to isolate an object we have not seen before and illuminate it with one of our colored flashlights? For example, what if we have an object that “normally” appears magenta but we isolate it and illuminate it using only our yellow flashlight? Remember that magenta objects absorb medium wavelengths and that yellow light is a combination of medium and long wavelengths. So, under the yellow light which has no short wavelengths, the object would continue to absorb medium wavelengths but now reflect only long wavelengths. This would produce a pattern of cone activity that the brain interprets as the color red.
Although color mixing is a helpful way to understand the relationship between wavelengths, receptors, and colors, it requires a specific type of viewer. What if the person mixing and viewing these colors did not have the typical three cone types? What if they only had two of the three types, or even just one? Would they perceive the same colors we perceive when looking at the same objects and lights?

1. Theoretically, an object that appears magenta under the white light of the sun would appear to be which of the following colors under a yellow light?
   a. Magenta
   b. Red
   c. Yellow
   d. White

2. Theoretically, an object that appears yellow under the white light of the sun would absorb which of the following light-wavelengths when under a magenta light?
   a. Short-wavelengths only
   b. Medium-wavelengths only
   c. Short- and medium-wavelengths only
   d. Long-wavelengths only

3. Theoretically, an object that appears yellow under the white light of the sun would appear to be which of the following colors under a cyan light?
   a. Blue
   b. Green
   c. Yellow
   d. White

**Color Blindness**

When people typically use the term color blindness, they are usually referring to a color deficiency called dichromatism, in which one of the three types of cones are missing or deficient. Dichromats are not truly colorblind – they do not see in black and white. On the other hand, someone with only one of the three types of cones would only perceive different shades of black and white. This is called monochromatism. Color perception depends on the ratio of signals sent by the different types of cones. Since a single type of cone cannot produce any ratio information on its own, no color is perceived. Thus, someone with only one receptor mechanism would not
be able to perceive different colors but a person with two types would be able to perceive some colors. Someone with all three types of cones is called a trichromat.

**Monochromatism**

Monochromatism is a rare form of color blindness that is usually hereditary and occurs in only about 10 people out of 1 million. Monochromats’ vision has the characteristics of rod vision in both dim and bright lights. Monochromats see only in shades of lightness (white, gray, and black); they can therefore be called color blind (as opposed to dichromats, who see some chromatic colors and therefore are called color deficient).

**Dichromatism**

People with dichromatism are missing one of the three cone pigments and hence experience some colors. However, they cannot distinguish as many colors as can trichromats. One way to diagnose color deficiency is by a color vision test that uses stimuli called Ishihara plates. Two example plates are shown below. People with normal color vision should see the number “74” in the left plate and the number “6” in the right plate. However, people with a form of color deficiency may not be able to see one of, or even both, of the numbers.


Once we have determined that a person’s vision is color deficient, we are still left with the question: What colors does a person with color deficiency see? It is often suggested that we can answer this question by pointing to objects of various colors and asking a color deficient person what they see. This method does not really tell us what the person perceives, however, because a color deficient person may say “red” when we point to a strawberry simply because they have learned that people call strawberries “red.” It is quite likely that the color deficient person’s experience of “red” is very different from the experience of the person without color
deficiency. For all we know, they may be having an experience similar to what a person without deficient color vision would call “yellow”.

To determine what a dichromat perceives, we need to locate a unilateral dichromat – a person with trichromatic vision in one eye and dichromatic vision in the other. Both of the unilateral dichromat’s eyes are connected to the same brain, so this person can look at a color with their dichromatic eye and then determine which color it corresponds to in their trichromatic eye. Although unilateral dichromats are extremely rare, the few who have been tested have helped us determine the nature of a dichromat’s color experience.

1. A monochromat experiences _____.
   a. black, white, and grays
   b. black, grays, and greens
   c. shades of yellow instead of red and green
   d. different shades of blue

2. Which of the following would not be able to perceive color?
   a. Monochromats
   b. Dichromats
   c. Trichromats
   d. Monochromats and Dichromats

3. A unilateral dichromat _____.
   a. has trichromatic vision in one eye and dichromatic vision in the other eye
   b. can only see black, white, and grays
   c. is missing two of the three cone types
   d. is more common in the U.S. than protanopes

There are three major forms of dichromatism: protanopia, deuteranopia, and tritanopia. The two most common kinds, protanopia and deuteranopia, are recessive traits inherited through a gene located on the X chromosome. Males (XY) have only one X chromosome, so a defect in the visual pigment gene on this chromosome causes color deficiency. Females (XX), on the other hand, with their two X chromosomes, are less likely to become color deficient because only one normal gene is required for normal color vision. These forms of color vision are therefore called sex-linked because women can carry the gene for color deficiency without being color deficient themselves. Thus, many more males than females are dichromats.
Protanopia affects 1% of males and 0.02% of females. A protanope is missing the long-wavelength pigment. As a result, a protanope perceives short-wavelength light as blue, and as the wavelength is increased, the blue becomes less and less saturated until, at 492 nm, the protanope perceives gray. The wavelength at which the protanope perceives gray is called the neutral point. At wavelengths above the neutral point, the protanope perceives yellow, which becomes less intense at the long wavelength end of the spectrum.

Deuteranopia affects about 1% of males and 0.01% of females. A deuteranope is missing the medium-wavelength pigment. A deuteranope perceives blue at short wavelengths, sees yellow at long wavelengths, and has a neutral point at about 498 nm.

If long-wavelength light is perceived as red and medium-wavelength light is perceived as green, why don’t protanopes only have issues perceiving reds and deuteranopes only have issues perceiving greens? Why do they both perceive yellow as the wavelength increases? The answer comes back to the importance of the ratio in firing between the multiple cone types. Remember that yellow is perceived when a mixture of medium- and long-wavelength light is reflected into the eye. This means both the medium-cones and the long-cones are involved in the perception of yellow. The brain determines the perceived color (green, red, yellow, etc.) by comparing the signals sent from the medium-cones to the signals sent from the long-cones. When one of those
cone types is missing, neither green nor red are perceived. Instead, medium- and long-wavelength light appear yellowish to both protanopes and deuteranopes.

Tritanopia is very rare, affecting only about 0.002% of males and 0.001% of females. A tritanope is missing the short-wavelength pigment. Again, it may seem strange that they still perceive blues without short-cones. This is because the ratio of signals coming from the remaining two cone types allows for the perception of certain shades of blue. However, without any input from the short-cones, tritanopes are unable to perceive yellow.

1. Which of the following statements is true about dichromatism?
   a. Males are more likely to be dichromats than females.
   b. Experience, not genetics, is the major cause of dichromacy.
   c. There are six major forms of dichromacy.
   d. There are nine major forms of dichromacy.

2. Why is color vision deficiency more common in men than in women?
   a. It depends on a dominant gene on the Y chromosome.
   b. It depends on a recessive gene on the Y chromosome.
   c. It depends on a dominant gene on the X chromosome.
   d. It depends on a recessive gene on the X chromosome.

3. Which types of dichromatism are most similar in terms of the colors perceived?
   a. Protanopia and Deuteranopia
   b. Protanopia and Tritanopia
   c. Deuteranopia and Tritanopia
   d. Trichromatic and Tritanopia
Color Perception Topic Test

1. Why can humans only perceive a specific range of light?
   a. We only have two eyes
   b. Our pupils only let in certain wavelengths
   c. We only have two types of visual receptors
   d. Our receptors only respond to certain wavelengths

2. What eye structure provides about 80% of the eye’s focusing power?
   a. The iris
   b. The pupil
   c. The cornea
   d. The lens

3. Which part of the retina, if any, has the greatest proportion of cones, relative to rods?
   a. The periphery
   b. The fovea
   c. The area surrounding the blind spot
   d. All parts have equal proportions

4. Which of the following is true about the visual receptors?
   a. Cones are more sensitive to light than rods
   b. Cones have better detailed vision/acuity than rods
   c. There are many more cones than rods in the retina
   d. All of the above are true

5. Perception of the brightness of a color is affected mainly by the _____ of light waves.
   a. wavelength
   b. amplitude
   c. hue
   d. saturation

6. Which perceptual dimension of color is determined by the peak wavelength that is reflected off or transmitted through an object?
   a. Hue
   b. Saturation
   c. Value
   d. None of the above
7. Which of the following would lead to the perception of an achromatically colored object?
   a. An object that selectively transmits some wavelengths
   b. An object that equally reflects all wavelengths
   c. An object that selectively reflects some wavelengths
   d. An object that selectively absorbs some wavelengths

8. The trichromatic (Young-Helmholtz) theory emphasizes which of these points?
   a. The brain compares responses of one retinal area to that of another to infer colors.
   b. Certain brain cells increase response for some colors and decrease it for others.
   c. Three types of cones react differently depending on the wavelength of light.
   d. Red-green color deficiency is more common in men than it is in women.

9. What color would most likely be perceived if the light reflected into the eye activates the medium-cones much more than the short- and long-cones?
   a. Blue
   b. Green
   c. Red
   d. Yellow

10. Blue light is projected on a white screen. What pattern of cone-activity would the light reflected off the screen into the eye most likely produce?
    a. Much greater activity in the short-cones than in the medium- and long-cones
    b. Much greater activity in the medium-cones than in the short- and long-cones
    c. Much greater activity in the long-cones than in the short- and medium-cones
    d. Fairly equal activity across all three types of cones

11. If you project a red, a green, and a blue light onto a white wall, the overlapping spot of light would be perceived as what color?
    a. Black
    b. Magenta
    c. White
    d. Brown

12. Which of the following wavelengths do red and green paint both selectively absorb?
    a. Short-wavelengths only
    b. Medium-wavelengths only
    c. Long-wavelengths only
    d. Medium- and long-wavelengths only
13. Theoretically, an object that appears yellow under the white light of the sun would absorb which of the following light-wavelengths when under a cyan light?
   a. Short-wavelength only
   b. Medium-wavelengths only
   c. Short- and medium-wavelengths only
   d. Medium- and long-wavelengths only

14. Which of the following would be able to perceive color?
   a. Dichromats
   b. Trichromats
   c. Unilateral Dichromats
   d. All of the above

15. Which two types of dichromatism are the most perceptually similar?
   a. Protanopia and Deutanopia
   b. Protanopia and Tritanopia
   c. Deutanopia and Tritanopia
   d. Trichromatic and Tritanopia
APPENDIX D: SAMPLE LEARNING PHASE TOPIC TEXT PASSAGE

1. Color Perception

The ability to see a tree, or any other object, depends on visible light being reflected from that object into the eye. Receptors in the retina transform the light into electrical signals that are sent to the brain and produce our perception of the tree.

1.1 **Light.** What we call visible light is part of the electromagnetic spectrum, a continuum of energy produced by electric charges and radiated as waves. The spectrum’s energy can be described by its wavelength (i.e., the distance between wave peaks). What makes light visible? Our receptors – which respond only to wavelengths from 400-700 nanometers (nm). With different receptors, we would see a different range of light, like the many bird and insect species that perceive ultraviolet light.

![Image of the electromagnetic spectrum](linked_image)

**Figure 4.1** Visible light is a small part of the electromagnetic spectrum. We see these wavelengths because we have receptors that respond to them.


Our visual system converts wavelengths of light into specific perceptions of color. The process begins with three kinds of specialized cone receptors. One type is most sensitive to short wavelengths (seen as blue), another to medium wavelengths (seen as green), and another to long wavelengths (seen as red). When light activates all three types equally, as sunlight does, we see the color white. Every wavelength produces its own unique ratio of responses from the short-, medium-, and long-cones, which is translated by the brain into a specific color.
One easy way to learn about color perception is through color mixing.

1.2 Mixing Lights. Imagine you have a set of flashlights that each produce a different color. If you were to shine a green flashlight on a white wall, you would see a green spot of light because the light reflected back into your eyes is mostly medium wavelength light. Now imagine you take a second flashlight that produces red light and you position them so that the spots overlap on the wall. What color would that overlapping area be? Yellow, which may not have been your first guess. The overlapping light would reflect both medium (green) and long (red) wavelengths. When both the medium- and long-cones respond equally, they produce a ratio of activity that is translated by the brain into yellow.

Now, what if you were to shine a blue flashlight directly on top of that yellow spot? By adding short-wavelength light, all three types of cones would respond equally, producing a ratio of activity that results in white.

What about the other possible combinations? A mixture of short and medium wavelengths produces a ratio of activity that results in cyan, while a mixture of short and long wavelengths produces a pattern of responding that results in magenta.
Based on these findings, researchers developed the RGB color model, where red, green, and blue act as primary colors and cyan, magenta, and yellow act as secondary colors.

1.3 Mixing Paints. But wait… When you mixed green and red paint as a child, you didn’t get yellow. That’s because mixing paints works differently. Remember that the color we perceive depends on the wavelengths reflected into our eye. Paints, like most other objects, reflect some wavelengths but absorb others. Any wavelengths absorbed are not reflected into the eye. That means blue paint, for example, must absorb medium and long wavelengths but reflect short wavelengths. Similarly, magenta paint (produced by short and long-cone activity) absorbs medium wavelengths, while yellow paint (produced by medium- and long-cone activity) absorbs short wavelengths.

So, what color would we see if we mixed magenta and yellow paint? Theoretically, if it was perfectly mixed, we would see red because only long-wavelength light is not absorbed. Likewise, mixing cyan (absorbs long wavelengths) and magenta paint gives you blue, and mixing cyan with yellow, gives you green.
Unlike lights, mixing many paints together doesn’t produce white. Instead, it will turn dark brown and eventually black. Because the mixture absorbs all wavelengths absorbed by each individual paint, little is reflected into the eye and we perceive an absence of color – black.

Researchers have also developed a color model for paint mixtures, the CMYK model, where cyan, magenta, and yellow are the primary colors and blue, green, and red are now the secondary colors (K = black). These models may sound familiar – the RBG model is used with media that transmit light (e.g., televisions), while the CMYK model is used with printed colors (e.g., inks and dyes).

Finally, what if we combined what we know about colored lights and objects? Imagine an object that appears magenta under the sun’s white light. Theoretically, if we were to isolate and illuminate the object with only our yellow flashlight, it would appear to be a different color. Remember that magenta objects absorb medium wavelengths and that yellow light is a combination of medium and long wavelengths. So, under yellow light which has no short wavelengths, the object would still absorb medium wavelengths but now reflect only long wavelengths, and would be seen as red.
APPENDIX E: CRITICAL RESULTS USING ALL-OR-NOTHING SCALE

The critical final test results using the all-or-nothing scale are reported below. Each answer could earn either 0 points or 1 point. Two raters independently scored the first 20% of responses and then compared scores to determine interrater reliability. Reliability between raters was very high ($r = .956$, CI 95% [.944, .968]; $ICC = .977$, CI 95% [.976, .979]) and one rater scored the remaining responses.

All Question Types: 2 (prior knowledge) x 2 (learning condition) ANOVA

- Main effect of prior knowledge, $F(1, 126) = 115.74$, $MS_e = .01$, $p < .001$, $\eta^2_p = .48$.
- Main effect of learning condition, $F(1, 126) = 20.66$, $MS_e = .04$, $p < .001$, $\eta^2_p = .14$.
- No prior knowledge by learning condition interaction, $F(1, 126) = 0.10$, $p = .75$.
  - Data are 5.56 times more likely under a model with main effects only than under a model with main effects and an interaction term ($BF_{01} = 5.556$).

2 (prior knowledge) x 2 (learning condition) x 3 (question type) ANOVA

- Main effect of prior knowledge, $F(1, 126) = 115.74$, $MS_e = .03$, $p < .001$, $\eta^2_p = .48$.
- Main effect of learning condition, $F(1, 126) = 20.66$, $MS_e = .11$, $p < .001$, $\eta^2_p = .14$.
- Main effect of question type, $F(2, 252) = 72.35$, $MS_e = .02$, $p < .001$, $\eta^2_p = .37$.
- No prior knowledge by learning condition interaction, $F(1, 126) = 0.10$, $p = .75$.
  - Data are 8.79 times more likely under a model with main effects only than under a model with main effects and a prior knowledge x learning condition interaction term ($BF_{01} = 8.788$).
  - Data are 7.94 times more likely under a model with main effects, a learning condition x question type interaction term, and a prior knowledge x question type interaction term than under a model with main effects, a learning condition x question type interaction term, a prior knowledge x question type interaction term, and a prior knowledge x learning condition interaction term ($BF_{01} = 7.939$).
  - Trending (but non-significant) prior knowledge x question type interaction, $F(2, 252) = 68.76$, $MS_e = .02$, $p < .001$, $\eta^2_p = .35$. 
• Learning condition x question type interaction, $F(2, 252) = 68.76, MS_e = .02, p < .001, \eta^2_p = .35$.
• No prior knowledge x learning condition x question type interaction, $F(1.57, 197.89) = 0.22, p = .75$.

Retention Questions Only: 2 (prior knowledge) x 2 (learning condition) ANOVA
• Main effect of prior knowledge, $F(1, 126) = 39.57, MS_e = .02, p < .001, \eta^2_p = .24$.
• Main effect of learning condition, $F(1, 126) = 87.29, MS_e = .05, p < .001, \eta^2_p = .41$.
• No prior knowledge by learning condition interaction, $F(1, 126) = 0.13, p = .72$.
  ○ Data are 5.29 times more likely under a model with main effects only than under a model with main effects and an interaction term ($BF_{01} = 5.285$).

Near Transfer Questions Only: 2 (prior knowledge) x 2 (learning condition) ANOVA
• Main effect of prior knowledge, $F(1, 126) = 37.05, MS_e = .05, p < .001, \eta^2_p = .23$.
• Main effect of learning condition, $F(1, 126) = 10.70, MS_e = .05, p = .001, \eta^2_p = .08$.
• No prior knowledge by learning condition interaction, $F(1, 126) = 0.04, p = .85$.
  ○ Data are 5.54 times more likely under a model with main effects only than under a model with main effects and an interaction term ($BF_{01} = 5.537$).

Far Transfer Questions Only: 2 (prior knowledge) x 2 (learning condition) ANOVA
• Main effect of prior knowledge, $F(1, 126) = 82.19, MS_e = .02, p < .001, \eta^2_p = .40$.
• No main effect of learning condition, $F(1, 126) = 0.63, p = .43$.
• No prior knowledge by learning condition interaction, $F(1, 126) = 0.57, p = .45$.
  ○ Data are 4.44 times more likely under a model with main effects only than under a model with main effects and an interaction term ($BF_{01} = 4.443$).
# Table E.1

*Testing phase final test proportion correct using all-or-nothing scale, split by question type.*

<table>
<thead>
<tr>
<th></th>
<th>Restudy Mean (SD)</th>
<th>Retrieval Mean (SD)</th>
<th>Testing Effect (Retrieval - Restudy)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Trained Topics: Lower Element Interactivity</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Retention Questions</td>
<td>.48 (.19)</td>
<td>.73 (.17)</td>
<td>+.25</td>
</tr>
<tr>
<td>Near Transfer Questions</td>
<td>.52 (.23)</td>
<td>.61 (.22)</td>
<td>+.09</td>
</tr>
<tr>
<td>Far Transfer Questions</td>
<td>.50 (.18)</td>
<td>.49 (.20)</td>
<td>-.01</td>
</tr>
<tr>
<td><strong>Untrained Topics: Higher Element Interactivity</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Retention Questions</td>
<td>.38 (.15)</td>
<td>.64 (.19)</td>
<td>+.26</td>
</tr>
<tr>
<td>Near Transfer Questions</td>
<td>.36 (.20)</td>
<td>.44 (.21)</td>
<td>+.08</td>
</tr>
<tr>
<td>Far Transfer Questions</td>
<td>.36 (.17)</td>
<td>.33 (.17)</td>
<td>-.03</td>
</tr>
</tbody>
</table>
REFERENCES


Dinsmore, D. L., & Alexander, P. A. (2012). A critical discussion of deep and surface processing: What it means, how it is measured, the role of context, and model...


JASP Team (2020). JASP (Version 0.14.1) [Computer software].


Lechuga, M. T., Ortega-Tudela, J. M., & Gómez-Ariza, C. J. (2015). Further evidence that concept mapping is not better than repeated retrieval as a tool for learning from texts. Learning and Instruction, 40, 61-68. https://doi.org/10.1016/j.learninstruc.2015.08.002

education: Infusing psychological science into the curriculum (pp. 32-44). Society for the Teaching of Psychology.


Pennebaker, J. W., Gosling, S. D., & Ferrell, J. D. (2013). Daily online testing in large classes: Boosting college performance while reducing achievement gaps. *PloS One, 8*(11), e79774. [https://doi.org/10.1371/journal.pone.0079774](https://doi.org/10.1371/journal.pone.0079774)


https://dx.doi.org/10.23668/psycharchives.844


