Avoiding Miscomprehension:

A Metacognitive Perspective for how Readers Identify and Overcome Comprehension Failure

by

Kole Andreas Norberg

B.A., Indiana University, Bloomington, 2003

M.F.A., Florida Atlantic University, 2009

M.S., University of Pittsburgh, 2020

Submitted to the Graduate Faculty of the

Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2022

UNIVERSITY OF PITTSBURGH

DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Kole Andreas Norberg

It was defended on

March 31, 2022

and approved by

Charles A. Perfetti, Distinguished Professor, Department of Psychology, Director, Senior Scientist, Learning Research and Development Center

Benjamin Rottman, Associate Professor, Department of Psychology, Research Scientist, Learning Research and Development Center

Linda Kucan, Professor, Department of Teaching, Learning, and Leading

Dissertation Director: Scott H. Fraundorf, Associate Professor, Department of Psychology, Research Scientist, Learning Research and Development Center

Avoiding Miscomprehension:

A Metacognitive Perspective for how Readers Identify and Overcome Comprehension Failure

Kole Andreas Norberg, PhD

University of Pittsburgh, 2022

Successful reading comprehension is not a guarantee, even for highly skilled readers. When comprehension fails, the ability of the reader to recognize the failure may be critical to avoiding *mis*comprehension (i.e., false confidence in an inaccurate text representation) and to taking steps to improve comprehension. Generally, people learn the most when they study partially-learned, as opposed to well-learned or completely-unlearned, content (i.e., *Region of Proximal Learning*; Metcalfe & Kornell, 2003). However, this requires the learner both to identify the difficulty level of the material (monitor learning) and to select the proximal material (control). Across two experiments, the current study assessed the interactive effects of monitoring and control in a reading context. Experiment 1a confirmed that readers do make greater gains when reading material of a moderate difficulty level, but Experiments 1b and 1c suggested that poor monitoring is not the reason that readers do not always select this material; rather, readers intentionally selected the hardest material. Although monitoring accuracy was not predictive of reader selections, readers were overconfident in their comprehension across Experiments. Experiment 2 tested the use of an ease-of-process heuristic during comprehension monitoring. Readers were especially *over*confident in their comprehension when a text *seemed* easier to process, in part because they were less likely to attend to difficulties (unfamiliar words) in the text. Texts that "feel" simpler engender shallower processing, which can lead to overconfidence in comprehension. Thus, readers struggle with both comprehension monitoring and metacognitive control, but whereas errors in monitoring appear to be based on incompatibility of the text with the applied heuristic, errors in control may be rooted in the reader’s beliefs about learning.

Keywords: [Reading comprehension, Metacomprehension, Metacognitive Control]

Table of Contents

List of Tables

List of Figures

Preface

I would like to make several personal acknowledgements. First, Nishant Purewal was instrumental in the execution of the experiments reported in this study. His tremendous work in the tedious task of managing participant recruitment was instrumental to the success of my dissertation. I am grateful for his level of professionalism and independence. Second, Catherine Apgar, Lydia Boyer, Rishika Dhanda, and Nishant Purewal all helped with creating and norming stimuli and coding participant data. Their feedback and insights were appreciated and incorporated into the final stimulus set.

I would further like to thank my dissertation committee for their advice and contributions to the construction of this experiment and the selection of individual difference measures.

I would like to thank my partner, Nora Fussner, for her support and patience with the long hours I have dedicated to my education over the past 5 years.

Finally, I would like to thank my advisor, Scott Fraundorf, for his endless support, advice, attention, and patience at all stages of my graduate education, including the many iterations of this and other projects.

The author acknowledges use of MetaMetrics, Inc.'s Lexile® & Quantile® Hub as the source of Lexile and Quantile measures and associated information. MetaMetrics has not endorsed this research.

The author(s) acknowledges that the Lexile measure ranges produced through the Lexile Analyzer® are not certified by MetaMetrics' staff. They are estimates of text complexity. They are not measures of student ability.

1.0 Introduction

Reading comprehension is a multistep process that ideally culminates in the formation of an integrated mental representation of the text (Kintsch, 1988; Van Dijk & Kintsch, 1983). Unfortunately, the process is not always successful. Comprehension failure constitutes an inaccurate or incomplete representation of the text and can occur for any reader. Overcoming comprehension failure likely first requires the reader to recognize that failure has occurred and then to activate appropriate strategies to overcome gaps and inaccuracies in their representation of the text. In the current series of experiments, I develop and test a model that builds on models of metamemory (Nelson & Narens, 1990) by breaking reader responses to comprehension failure into two interactive metacognitive components: *monitoring* of comprehension (i.e., becoming aware of comprehension failure) and *metacognitive control* (i.e., selecting appropriate strategies to improve comprehension).

Neither comprehension monitoring nor metacognitive control are simple. To begin, readers of all skill levels struggle with accurately monitoring their comprehension (Baker, 1989; Carter & Dunning, 2008; Dunning, 2011; Glenberg et al, 1987; Griffin et al., 2008; Kruger & Dunning, 1999; Maki, 1998; Maki et al., 1994; Oakhill et al., 2005; Thiede et al., 2010). Further, selection of appropriate strategies for overcoming comprehension may require a precisely calibrated monitor (Townsend & Heit, 2010). Metcalfe and Kornell (2003) demonstrated that, in some learning domains, the monitor must be sensitive to the relative difficulty of learning materials. They found that learners make the most gains in knowledge when they identify and select study material within their *Region of Proximal Learning* (RPL). Content in a learner's RPL is partially learned, separating it from content that is either well learned or completely unlearned. However, the RPL model was developed for, and tested processes involved in, memorizing comparatively simple materials (e.g., word lists). As I detail later, it is less clear if readers make greater improvements in comprehension if they identify and target text that they are close to comprehending versus far from comprehending. Thus, one goal of the present study is to test the applicability of the RPL in the domain of reading comprehension.

A second goal is to determine if monitoring errors lead readers to select inappropriately difficult texts, and thus if monitoring has an indirect, mediated effect on actual comprehension. Monitoring errors can be indicative of both over- and under-confidence and both could lead readers to select inappropriate learning materials. But overconfidence is particularly common (Commander & Stanwyck, 1997; Garner, 1980; Griffin et al., 2008; Jaeger & Wiley, 2015; Kwon & Linderholm, 2014; Prinz et al., 2020) and of particular concern in this study because this type of error constitutes an *illusion-of-knowing* (i.e., false beliefs about a topic) and can have downstream consequences in the form of diminished learning from later texts (Glenberg et al., 1982; Prinz et al., 2018) or even lead to physical harm (e.g., inaccurately carrying out medical advice; Davis et al., 2006). Thus, in addition to determining how monitoring affects control processes, a final goal of this study is to understand what text-based and individual factors increase the likelihood that an illusion-of-knowing will develop.

Specifically, readers may rely on an ease-of-processing heuristic to assess comprehension. Under an ease-of-processing heuristic, learning is perceived to be successful when processing is perceived to be less effortful (Dunlosky et al., 2002; Maki et al., 1990). Several studies have provided evidence that readers use an ease-of-processing heuristic to evaluate reading comprehension (Finn & Tauber, 2015; Thiede et al., 2010; Rawson & Dunlosky, 2002; Wiley, Griffin, & Thiede, 2005), but use of this heuristic has been linked to inaccurate assessments of learning (Rawson & Dunlosky, 2002; Schommer & Surber, 1986) and to selection of suboptimal study strategies (Kirk-Johnson et al., 2019). An ease-of-processing heuristic may encourage readers to engage in shallow, underspecified reading when a text initially feels easy to process and deeper reading strategies when a text initially feels difficult to process. This strategy may serve the reader well when their inference about the actual difficult of the text is correct. However, when the text contains unexpected complexity, the ease-of-processing heuristic may lead to less attention to the complexities and overconfidence in comprehension.

The current study diverges from past work by focusing on comprehension *monitoring*, rather than comprehension per se. That is, the focus of the work is on creating circumstances which challenge the comprehension processes of even skilled readers. Thus, the two experiments detailed here test (a) the RPL in a reading framework to determine where readers should allocate their time when struggling with comprehension, (b) if errors in comprehension monitoring are at the root of errors in allocation of time to particular regions of a text (i.e., metacognitive control), and (c) if the use of the ease-of-processing heuristic can lead readers to underspecify a text and inflate their confidence in their comprehension. In the remainder of the introduction, I review each of these topics before introducing the design of the present experiments.

1.1 Comprehension Monitoring

Metacognition is the process of assessing one’s cognitive state (Flavell, 1979). It is how learners monitor their learning without the aid of external feedback and how they recognize and select learning strategies that will maximize learning outcomes (Nelson & Narens, 1990; Thiede & Dunlosky, 1999). Nelson and Narens (1990) suggested that when applied to memory, metacognition involves two interactive processes: monitoring and control. During *memory monitoring*, learners assess the state of their memory and the ease with which specific information can be retrieved. Metacognitive *control* in the domain of memory relates to an individual's ability to select memorization strategies which will be most successful. I first discuss the implications of monitoring before returning to a discussion of control in the subsequent section.

Comprehension Monitoring is a specific form of metacognitive monitoring related to text comprehension (Maki & Berry, 1984). Comprehension monitoring is poor among readers of all ages and skill levels (Glover, 1989; Jacobson, 1990; Lin & Zabrucky, 1998; Maki & McGuire, 2002; Zabrucky, 2010) though some researchers have found the deficits to be most pronounced among less-skilled readers (Garner & Taylor, 1982; Kruger & Dunning 1999; Oakhill et al., 2005). Comprehension monitoring accuracy (the match between how well participants believe are performing and how well they actually perform on a reading comprehension assessment) is typically low across reading levels (e.g., r = .24; Prinz et al., 2020), indicating that readers are not typically able to accurately assess how they will perform on specific questions or texts (Dunlosky & Lipko, 2007; Maki, 1998; Thiede et al., 2009).

The high level of errors in comprehension monitoring following reading of texts stands in contrast to memory monitoring following associated learning tasks. In a review of the literature, Wiley et al. (2005) noted that memory monitoring is typically more accurate than comprehension monitoring, and interventions to improve memory monitoring are typically more effective. This difference may arise from the nature of metacognitive monitoring itself, which is believed to rely on cues (e.g., perceived effort) that allow learner to make an inference about their memory or comprehension (Koriat, 1997). Which cues are available or used may differ between memory monitoring and comprehension monitoring. For example, monitoring of long-term memory can be improved by delaying when learners make metacognitive judgments (Dunlosky & Nelson, 1992; Rhodes & Tauber, 2011) because it removes interference from short-term memory; i.e., the item to be recalled is no longer partially active and an inference about the effort required to recall the item is more similar to a future state when the item will need to be recalled again. However, the technique of delaying metacognitive judgments does not typically work for comprehension monitoring (Glenberg & Epstein, 1985). This may be because the types of cues used to assess comprehension are not contaminated by short-term memory; i.e., whether or not a text is comprehended is a separate question from whether or not it will be accurately retrieved at a later time. Further, when assessing comprehension after a delay, readers may not be detecting the accuracy of their comprehension so much as they are assessing their memory for the model they created (regardless of its accuracy). Indeed, improving comprehension monitoring has consistently required not only delaying self-assessment of learning, but other additional interventions that promote assessment of the reader's text representation (Wiley et al., 2005), such as requiring readers to generate key words (Thiede et al., 2003), write summaries (Thiede & Anderson, 2003), or create concept-maps (Thiede et al., 2010). Even among these more effective interventions, it is not clear that comprehension monitoring rather than memory monitoring was what improved. For example, Thiede et al. (2010) found that when asked to make judgments of comprehension readers relied more on memory cues *after* as opposed to *before* a delay. Thus, isolating readers' judgments of their comprehension from their judgments of memory is an important consideration when assessing comprehension monitoring.

Further, long-term memory for details from a text is not always even necessary for successful reading comprehension. Rather, readers are often able to refer back to texts to fill in gaps in their memory. Indeed, many reading comprehension assessments allow readers to look back into the text when answering questions, including the ACT, SAT, GRE, and Nelson-Denny test used in this study (Cummins, 1981; see Ferrer et al., 2017; Ozuru et al., 2007 for a more detailed discussion on text availability during reading comprehension assessment). Thus, in evaluating comprehension monitoring, the current study minimized the reliance on memory monitoring when assessing performance by permitting readers to refer to the text during evaluations of comprehension. In this context, illusions-of-knowing reflect comprehension monitoring failure rather than breakdowns in memory monitoring.

1.1.1 Illusions-of-Knowing

One purpose for measuring comprehension monitoring in this study is to understand its role in the development of illusions-of-knowing (i.e., miscomprehension). In understanding how an illusion-of-knowing might develop, it is helpful to consider how representations of text are developed. Kintsch (1988) proposed that readers engage in three levels of processing: constructing a *surface code*, *textbase*, and *situation model*. Whereas the surface code is tied to the specific lexical and syntactic features of a text (e.g., passive voice), the textbase is a representation of the semantic value of the surface code. For example, sentences (1a) and (1b) have separate surface codes but the same textbase. Although the sentences vary slightly in structure and exact wording, the broader meaning of the two sentences is virtually the same. Construction of the situation model in turn involves integrating concepts from the textbase with prior knowledge, and it is the situation model that the reader generally remembers when recalling the text (van Dijk & Kintsch, 1983).

(1a) A comet is a small chunk of dust and ice that orbits the Sun.

(1b) The Sun is orbited by a cluster of ice and dust called a comet.

Although readers of all levels read incrementally, *mostly* word-by-word (Just & Carpenter, 1980), the pace and level of attention to each word or section of a text varies, sometimes even to the extent that words can get skipped altogether; i.e., the fovea does not focus on the word (Mata et al., 2017; Rayner, 1998). To determine depth and speed of processing, readers may use a "good enough" criteria, stopping when their representation of the text feels complete (Ferreira, 2003; Ferreira et al., 2002; Ferreira & Patson, 2007; Reyna & Brainerd, 1991). As a result, readers may represent the gist of the text, leaving some propositions underspecified within the model or even omitted altogether (Otero & Kintsch, 1992; Sanford & Sturt, 2002; van Dijk & Kintsch, 1983). These omissions can lead to misinterpretations of text, such as interpreting (2) as plausible, concluding from (3) that *the baby was dressed by Anna* (Ferreira, 2003; Ferreira et al., 2002), and responding to (4) with *two* instead of pointing out the faulty premise; i.e., it was Noah, not Moses, who brought animals onto the ark (Erickson & Mattson, 1981).

(2) The dog was bitten by the man.

(3) While Anna dressed the baby played in the crib.

(4) How many animals of each type did Moses bring on the ark?

Cognitive processes can often be divided into two systems in which the first system relies on associative process which are fast and automatic whereas the second system evaluates and analyzes the information from system 1 (Kahneman & Frederick, 2002). The *good enough* account aligns with a dual-process monitoring system in which System 1 processes information quickly, making fast gist-like interpretations, while System 2 monitors the deductions (Kahneman & Frederick, 2002). If the results from System 1 seem reasonable based on available input, System 2 will not notice a problem and an illusion-of-knowing will occur.

1.1.2 Ease-of-Processing Heuristic

A dual-process monitoring system can explain why shallow processes which assemble a gist of the text from the limited number of words processed is accepted by the reader as a complete representation of the text. However, it leaves open the question of what factors affect how deeply the reader will process the text. Two competing hypotheses have been put forward in the literature: that difficult text prompts shallower processing, especially among less-skilled readers (Ferreira et al., 2002; van Dijk & Kintsch, 1983), and that easier texts prompt shallower processing (Mata, 2020; Mata et al., 2017). In both accounts, difficulty is determined by the surface code of the text. However, in the first account, readers engage in greater gist processing and rely more on intuition as a way of compensating for difficulty processing complex syntax (e.g., passive structures [2] and garden-paths [3]; Ferreira et al., 2002). Indeed, this was originally suggested by van Dijk and Kintsch (1983) as an explanation for differences in reading skill. By their account, less-skilled readers have greater difficulty decoding the surface code of a text and so rely more on context cues when forming their textbase. This can lead to error if the context cues are in anyway unreliable.

Although readers misinterpret passive sentences like (2) more often than active sentences like (5) (Ferreira et al., 2002; Ferreira & Stacey, 2000), texts with a more difficult surface code have not consistently resulted in increased illusions-of-knowing. In a study evaluating comprehension monitoring of longer texts, it was the easier texts that resulted in greater overestimations of comprehension (Maki et al., 2005). Further, difficult (versus easy) texts also promoted greater miscomprehension when instructions cued shallow processing (Schommer & Surber,1986). Thus, illusions-of-knowing may be related to shallow, gist processing, but the origin of shallow processing is not necessarily difficulty processing the surface code. In certain circumstances, a more difficult surface code has actually encouraged deeper processing even if it has not improved comprehension.

(5) The man bit the dog. (vs *The dog was bitten by the man.* in [2])

An alternative source of the shallow processing that is associated with illusions-of-knowing stems from the cue-utilization framework for how learners monitor memory (Koriat, 1997). Under this framework, learners use past experience with learning assessments to search for cues that previously predicted their performance. Although primarily used by learners to assess memory and learning, some aspects of the framework could also explain how readers asses the level of processing necessary to read a text and their subsequent confidence in their comprehension. In particular, it is claimed that learners apply an *ease-of-processing heuristic* (Begg et al., 1989; Benjamin et al., 1998; Kornell et al., 2011). Under this heuristic, if the text is perceived as easy to process, it is considered easy to understand, and if it is perceived as hard to process, it is evaluated as hard to understand. To take this a step further, it may also be the case that if it is easy to understand, readers infer that deep processing is unnecessary.

The ease-of-processing cue is likely used because it is broadly accurate (Benjamin et al., 1998). Information that can be retrieved easily in the moment is indeed more likely to be retrieved later. Further, ease of processing is likely a reliable cue for signaling the difficulty level of a text. For example, readers process high-frequency words faster than low-frequency words (making processing feel less effortful), and the high-frequency words are also more likely to result in higher rates of comprehension (Balota et al., 2004; Brysbaert & New, 2009; Dale & Tyler, 1934; Flesch, 1948; Howes & Solomon, 1951; Rayner & Duffy, 1986). Thus, if readers notice that they can read a text quickly, they may reasonably assume that they are finding the text easy to process and conclude it is easy to comprehend.

However, easier processing does not always predict successful comprehension. Indeed, readers who are inaccurate in their metacognitive judgments often self-report using shallow processing cues that are unrelated to improvements in actual comprehension (Thiede et al., 2010), including font-size, interest in the topic, and visual clarity of text (Alter & Oppenheimer, 2009; Alter et al., 2007; Begg et al., 1989; Novemsky et al., 2007; Simmons & Nelson, 2006). Similarly, readers will rate their comprehension higher when they are provided with illustrations and analogies, which may make processing feel easier, despite actual comprehension remaining unaffected (Jaeger & Wiley, 2014, 2015; Serra & Dunlosky, 2010). Thus, ease-of-processing cues can also deceive readers about the actual complexity of the text.

Some evidence for the ease-of-processing heuristic as the source for shallow, underspecified reading comes from the domain of logical reasoning. Mata (2020) suggested that when readers encounter a problem that feels familiar, they do not attend to all parts of the problem and instead fill in the gaps with intuition. Mata offered this as an explanation for what appear to be errors in logical reasoning (e.g., in [6], readers often mistakenly answer $0.10 instead of $0.05). This theory of attentional failure was driven by findings that participants who were less accurate in responding to reasoning problems were also less likely to notice changes in the premise during a change detection task (Mata et al., 2014) and less likely to look at the critical premise (e.g., *more than the ball*) during an eye-tracking study (Mata et al., 2017). Further, when the critical premise was underlined—which may attract attention (e.g., von Restorff effect, Chi et al., 2007; Glynn, 1978; von Restorff, 1933)—participants were more likely to answer the question accurately. That is, when readers were engaged with what otherwise felt like a familiar problem, increasing attention to complex regions of text increased comprehension. But does such attention improve comprehension monitoring? That is, if participants were still unable to accurately answer the questions after the critical premise was underlined, would they be able to recognize that they could not?

(6) A bat and a ball together cost 110 cents. The bat is 100 cents more than the ball.
How much is the ball?

In summary, illusions-of-knowing may occur when readers use shallow processing, resulting in underspecified representations of the text. However, what text features might prompt shallow processing is an open question. In the current study, I tested the features that can produce the illusion-of-knowing by developing a paradigm—discussed in greater detail below—that uses a pseudoword to make comprehension of a target region of text impossible. This allowed me to isolate influences on *monitoring* of comprehension rather than comprehension itself. I then tested the effect of two factors that I hypothesized would influence illusions-of-knowing: (a) the overall ease of the text and (b) the salience (or lack thereof) of the difficult region.

1.2 Metacognitive Control

Although illusions-of-knowing can be frequent, inaccurate comprehension monitoring does not always predict *low* reading comprehension (cf. Begg et al., 1992; Cavanaugh & Perlmutter, 1982; Dunlosky & Connor, 1997; Dunlosky & Hertzog, 1998; Pressley & Schneider, 1997). One reason for this is that inaccurate comprehension monitoring can also occur when readers are underconfident, indicating that they comprehended more than they realized. This has led some researchers to conclude that comprehension monitoring may not be necessary for successful learning from text (e.g., Pressley & Schneider, 1997).

However, Thiede et al. (2003) proposed that a relationship between comprehension monitoring and performance is present but is mediated by metacognitive control. Part of metacognitive control is the regulation of study time and strategies to reduce the discrepancy between a learner's current state and their desired state (Dunlosky & Hertzog, 1998; Nelson & Narens, 1990**.** When the learner believes they have reached their desired state, they stop studying. If a learner is not able to accurately monitor their learning, they may stop studying before mastery has been acquired, or they may "labor-in-vain" (Nelson & Leonesio, 1988) by studying material which they are not ready to master (Dunlosky & Thiede, 2004; Metcalfe & Kornell, 2003, 2005; Thiede & Dunlosky, 1999) or potentially which they have already mastered.

Two dominant models have been proposed to explain what underlies learners' decisions about how to allocate study time. The first is the discrepancy-reduction model (e.g., Dunlosky & Hertzog, 1998; Tullis & Benjamin, 2011), which states that learners will choose to study items that are not yet learned, and critically, they will choose items that are furthest from mastery. These items are expected to maximally reduce the discrepancy between what they know and what they want to know. Multiple studies have demonstrated learners tend to select difficult items and make the greatest gains from restudying them (Dunlosky & Hertzog; 1998; Nelson & Narens, 1990; Son & Metcalfe, 2000; Thiede et al., 2003; Thiede & Dunlosky, 1999). Although most of these studies have used simple materials (e.g., word pairs), Thiede et al. (2003) found that readers who selected the most difficult reading material for restudy had higher metacognitive accuracy and learned more than readers who selected easier reading material.

Critically, studies supporting the discrepancy reduction-model allowed learners unlimited time for study and have often used extreme-groups designs where the materials were either very easy or very hard (Son & Metcalfe, 2000). This may not represent the full spectrum of options available to a learner during self-regulated study. Learning material exists in more gradations of difficulty than hard and easy. Further, learners may have time constraints which preclude devoting the time needed to fully understand the most difficult concepts related to their learning. Thus, another way learners may labor-in-vain is if they devote time to difficult material that they will be unable to learn given the allotted time. Further, periods of struggle which do not result in learning gains may increase frustration and boredom, both of which are likely to cause the learner to stop studying (Metcalfe, 2011; Metcalfe et al., 2020; Morris et al., 2019; Xu & Metcalfe, 2016). In response to these concerns, Metcalfe (2002) proposed the region of proximal learning (RPL) model. This model suggested that rather than selecting the hardest material, the learner will be best served by studying the easiest material not yet learned; that is, the *proximal material*.

The RPL model aligns with broader views of learning. The concept of the zone of proximal development (Piaget & Cook, 1952; Vygotsky, 1934/1987) has been a cornerstone in the learning sciences, and, whether in the domain of motor development or cognition, has shown that learners make the fastest gains when pursuing skills that are partially but not fully developed as compared to skills which have not yet started to develop. Atkinson (1958, 1972, 1974) applied this concept to discrimination learning and created adaptive instructional systems which increase the difficulty of vocabulary learning incrementally as the student improves. And within the realm of learning to read, work in education promotes the use of *instructional level* books based on evidence that, when children read books just above their reading level (i.e., when they comprehend 75-89% of the material), reading skill grows faster than if they read at or far above their reading level (i.e., below 50% comprehension; Betts, 1946; Keene & Zimmerman, 1997; Miller, 2002; Morris et al., 2019; Mounla et al., 2011, but see Shanahan, 2020 for an argument in favor of teaching more complex texts).

The RPL has been supported across multiple studies of learning word pairs in which learners have been observed to select and succeed the most within their region of proximal learning (Metcalfe, 2002; Metcalfe & Kornell, 2003, 2005; Kornell & Metcalfe, 2006; Price & Murray, 2012). However, it has not been applied to learning from reading. This is a critical point because it is not clear that learners (a) have comprehension monitoring skills that are refined enough to identify text in the proximal region (Townsend & Heit, 2010) and (b) would benefit from reading material in their region of proximal learning as contrasted with more difficult material. It is possible that reading hard texts carries benefits that studying hard word lists does not. For example, in reading more complex, difficult texts, learners may be able to glean partial information from the difficult material that affords them a more nuanced understanding when they return to the full text. Recognizing the nuances could then aid them in adjusting their comprehension monitoring as they become aware that the topic is more complex than they would have otherwise realized. This opportunity is not available when studying word lists. Thus, the current study first tests the utility of the RPL within a reading context before exploring whether readers are able to make the nuanced assessments of their comprehension necessary for selecting appropriate supplemental reading material.

1.3 Individual Differences

1.3.1 Reading Skill

Reading skill may add an additional layer of complexity to evaluating metacognitive monitoring and control. As previously mentioned, less-skilled readers are particularly prone to errors in comprehension monitoring, especially overconfidence in their reading comprehension (Garner, 1980; Griffin et al., 2008; Jaeger & Wiley, 2015; Kwon & Linderholm, 2014) and reading ability (Dunlosky & Rawson, 2012; Dunning & Kruger, 1999). Further, studies on reading comprehension have demonstrated that less-skilled readers are especially likely to develop impoverished representations of texts when they have more prior knowledge on the topic (i.e., when the text is more cohesive) as compared to highly skilled readers or less-skilled readers with less prior knowledge (O'Reilly & McNamara, 2007). For less-skilled readers, the text may have seemed easier because they did not need to develop the relationships among the ideas. However, skilled readers may have continued to engage in deeper processing strategies despite finding the text to be easy-to-process. This raises the prospect that less-skilled readers may be particularly prone to using inappropriate cues when assessing text ease or more generally less skilled at monitoring their comprehension (Baker & Beall, 2014; Griffin et al., 2008).

One proposed explanation for why less-skilled readers would also be less skilled at monitoring is that they struggle to simultaneously process text and evaluate their comprehension (Griffin et al., 2008). Some researchers even propose that less-skilled readers do not monitor their comprehension at all (Oakhill et al., 2005; Tighe et al., 2021). Supporting the theory that less-skilled readers struggle to or do not monitor their comprehension during initial reading, their metacognitive accuracy following rereading of a text increases more than for skilled readers (Griffin et al., 2008), and they benefit more from interventions to improve comprehension monitoring (Thiede et al., 2010). However, some studies have found that reading skill is not related to comprehension monitoring at all, with both skilled and less-skilled readers demonstrating equally poor performance (Commander & Stanwyck, 1997).

It is possible that the apparent discrepancy in the direction of metacomprehension errors (over- versus under-confident) between less skilled and skilled readers is an artifact of their overall accuracy differences (Golke & Wittwer, 2017). Skilled readers have high accuracy and therefore less opportunity to demonstrate overconfidence whereas for less-skilled readers the situation is reversed. Thus, ceiling and floor effects may make it difficult to assess the direction of their monitoring errors. Golke and Wittwer (2017) demonstrated that when readers were asked to predict their performance on an upcoming comprehension test, skilled readers did underestimate their comprehension of easy texts, where comprehension was expected to be high, but overestimated their comprehension of hard texts, where comprehension was expected to be low—albeit not to as great an extent as less-skilled readers. Thus, it remains unclear if differences in reading skill are relevant to differences in metacognitive monitoring. One way to circumvent a potential ceiling effect is to assess comprehension monitoring in an environment where comprehension is not possible. In that case, monitoring can be assessed with skilled and less-skilled readers at the same level of comprehension (or lack thereof).

1.3.2 Other Individual Differences

Reading skill is likely not the only individual difference critical to readers' comprehension monitoring ability (see Stolp & Zabrucky, 2009, for a review). Reader beliefs about reading, standards of coherence, and perceptions of reading ability may all play a role in metacognitive monitoring and control. Measures of a reader's perceptions about their own reading skill allows for a wider evaluation of comprehension monitoring ability. Just as the ability to evaluate comprehension of a given text may be important to learning, so too might the ability to assess comprehension potential. In fact, it may be that readers whose perception of their reading skill is high but whose actual reading skill is low are most at risk for illusions-of-knowing (Kwon & Linderholm, 2014). As the current study uses expository texts on topics related to science, I targeted perception of science reading skill rather than their perceptions of reading skill more broadly because readers' perceptions of their reading skill in a science domain, as well as their actual skill, generally differ from that of narrative texts (Berman & Nir-Sagiv, 2007; Kwon & Linderholm, 2014; Singer et al., 1997; Wolfe, 2005).

Readers' beliefs about reading and standards of coherence may also influence both how closely they engage with a text and their reading strategies. *Standards of coherence* include the readers' intrinsic and extrinsic reading goals; i.e., their desire to understand the material and engage with difficult texts (Calloway, 2019). These are separate from external goals related to the nature of the task. All readers in a study may have the same external standard of coherence imposed upon them by the nature of the questions that are asked and the extent of close reading necessary to answer those questions; however, the reader's personal standards may differ (Calloway, 2019; Narvaez et al., 1999; van den Broek et al., 2011). Prior work shows that standards of coherence are lowest among less-skilled readers and that less-skilled readers generally report lower enjoyment and engagement during reading (Calloway, 2019; Guthrie et al., 2004; Crossley et al., 2017). Further, intrinsic reading goals and desired reading difficulty are particularly important characteristic of a reader's standards of coherence (Calloway, 2019). Thus, by measuring readers' standards of coherence, I covary out effects related to engagement with the task (as well as capturing any ways in which standards of coherence themselves influence comprehension monitoring).

1.4 Measuring Comprehension Monitoring

One of the key outcome measures in this work is the accuracy of readers' metacognitive monitoring. Metacognitive monitoring is typically assessed by asking learners to make a judgment of learning (JOL) following learning, either before assessment (prediction) or after assessment (post-diction). Predictive JOLs assess the learner's broad perspective on the state of their comprehension whereas postdictive JOLs assess whether the learner believes they answered a specific question or set of questions correctly (Baker, 1989). Findings from fMRI show that different brain regions are active during pre- versus post-dictions, suggesting that the processes involved in each decision are distinct (Fleming & Dolan, 2012). Further, post-dictions tend to be more accurate than predictions (Busey et al., 2000; Kwon & Linderholm, 2014). Thus, predictions and post-dictions do not have to align (Benjamin, 2003): The reader can be initially confident in their comprehension (high predictive JOL) but then believe they cannot answer the specific question(s) posed (low postdictive JOL). Conversely, the reader may initially be unconfident in their learning (low predictive JOL) but then realize that they learned enough to answer the question(s) correctly (high postdictive JOL).

To obtain a measure of metacognitive monitoring, researchers compare the JOLs (either pre- or post-dictive) to actual performance. The comparison typically takes two forms: a measure of sensitivity (i.e., resolution or relative accuracy) and calibration (i.e., absolute accuracy). *Sensitivity* refers to whether the learner was able to accurately assess the comparative probability of getting individual items correct and is one of the primary ways comprehension monitoring skill is assessed (Thiede et al., 2010; Wiley et al., 2005). This measure describes if learners can discriminate between their learning for one item or section of material over another. Such an ability should be relevant to the RPL model, in which the ability to discriminate which material is in the

RPL versus is already learned or too far from being learned is critical. However, sensitivity does not provide a measure of the learner's over- or under-confidence in their performance; rather, this is provided by *calibration*, which measures the discrepancy between a learner's overall confidence compared with their actual performance. When the learner's overall confidence in their learning matches their overall rate of accuracy (score of 0), the learner's metacognitive monitoring is said to be well calibrated. When the score is positive (i.e., calibration increases), the reader is overconfident. When the score is negative (i.e., calibration decreases), the reader is underconfident. This creates some confusion as overconfidence is indicated by an increase in calibration. Thus, in order to facilitate interpretation within the results, I will call this variable *miscalibration*; i.e., overconfidence is indicated by an increase in miscalibration.

Although both sensitivity and miscalibration are measures of metacognitive monitoring, past studies have shown they are not always correlated (Kelemen et al., 2000; Maki et al., 2005; Schraw, 2009). A learner can be sensitive and overconfident or calibrated but not sensitive. Their ability to produce diverging results suggests that a full picture of metacognitive monitoring requires consideration of both the learner's sensitivity and miscalibration for both pre- and post-dictions.

1.5 Current Study

The current study had three broad interactive questions: (a) Does the RPL framework apply to a reading comprehension context? (b) Does readers' use of the ease-of-processing heuristic promote under-specification of texts, leading to a false sense of confidence in comprehension? and

(c) Do errors in comprehension monitoring have downstream consequences for learning by affecting metacognitive control?

Experiments 1a, 1b, and 1c first assessed the importance of metacognitive monitoring ability by testing its effect on comprehension as mediated by metacognitive control processes. Experiment 1a tested the predictions of the RPL framework in the domain of reading comprehension. The RPL model predicts that readers will make the greatest gains in comprehension when they read about moderately difficult aspects of a topic rather than easy or difficult aspects of the topic.

The randomized design of Experiment 1a can establish a causal relationship between reading about moderately difficult aspects of a topic and subsequent improvements in comprehension. However, because readers do not choose material themselves, it cannot provide information about the role of metacognitive control. Thus, Experiment 1b extended Experiment 1a by allowing readers to choose which section they read. Readers who are more sensitive in their comprehension monitoring are expected to select texts which are within their region of proximal learning and selecting these texts is expected to lead to the greatest gains in overall comprehension. In other words, the effect of metacognitive sensitivity on comprehension is expected to be mediated by which supplemental materials readers choose.

One potential concern was that asking readers to evaluate their comprehension prompts them to engage comprehension monitoring that would not have otherwise occurred and that this could change the readers' choices about what subtopics they should learn. Thus, Experiment 1c validated the findings of Experiment 1b by assessing what choices readers made in the absence of questions which prompt them to monitor their learning.

Experiment 1 established the connections between metacognitive monitoring and comprehension. Experiment 2 tested the hypothesis that failures in metacognitive monitoring, specifically illusions-of-knowing, can result from use of the ease-of-processing heuristic and further that this can end up decreasing attention to complex regions of an otherwise easy to read test. Using the same texts from Experiment 1, reading was obscured through inclusion of a pseudoword. Readers who can monitor their comprehension are expected to recognize that they cannot comprehend the text because the pseudoword obscures the meaning of the text. However, under an attention-allocation account, if the reader does not attend to the pseudoword owing to forming a gist representation of the text, they will not recognize that their comprehension is low. This account predicts that when the pseudoword is underlined, it should draw readers' attention to their inability to comprehend the text. Critically, the effect of the underlining manipulation should be a decline in an illusion-of-knowing rather than an increase in actual knowing. That is, readers should still not be able to comprehend the text, but they should be more likely to recognize their low comprehension when the presence of the pseudoword is made more salient.

2.0 Experiment 1a

I tested the effect of studying within the region of proximal learning (RPL) across three experiments. Experiment 1a used random assignment to establish that moderately difficult texts lead to greater gains in learning than reading texts with easier or harder content. Participants read a text broken down into three sections of varying difficulty (Easy, Moderate, and Hard). Critically, they were randomly assigned to receive additional, supplemental material about one of the sections with the difficulty of the supplemental material matched to the difficulty of the section. Receiving the moderately difficult supplemental section was expected to increase accuracy on comprehension questions from a pre- to post-test more than receiving the hard or easy supplements. Thus, Experiment 1a tests the application of the RPL within a reading context. Experiment 1b extended the findings of Experiment 1a by testing whether or not readers choose the supplemental texts most likely to enhance their learning the most. This tested whether metacognitive monitoring accuracy predicted metacognitive control and subsequently gains in comprehension. Finally, Experiment 1c tested whether the choice of text in Experiment 1b was influenced by participants being asked to reflect about their learning (i.e., make JOLs). All Experiments used a within participants design (i.e., all participants received all levels of the manipulation) to test the interaction of Section Difficulty (Easy, Moderate, Difficult) x Supplemental Difficult (Easy, Moderate, Difficult) x Test Phase (pre or post supplemental exposure) on Comprehension Accuracy. Experiment 1a is presented first.

2.1 Method

2.1.1 Participants

Pilot data indicated that the effect size for the critical 3-way interaction (i.e., the effect of Section Difficulty x Supplement Difficulty x Test Phase on response accuracy) was a standardized change in the odds ratio of 0.56. A power analysis for a mixed-effects model using R package *simr* (Green & MacLeod, 2016) indicated that $N = 40$ participants were needed to achieve power of .80. However, as recruitment of readers with greater extremes in reading skill can be difficult, this number was doubled ($N = 80$) to ensure an adequate range of reading levels were represented in the study.[1]

Participants ($N = 86$, $M_{\text{age}} = 26$, 70 female, 1 undisclosed gender) were recruited from the University of Pittsburgh psychology participant pool ($n = 57$) in exchange for partial fulfillment of a course requirement or from the wider urban community ($n = 29$) in exchange for $20. All participants self-reported that they did not hold a bachelor's degree and that their first language was English. Several exclusion protocols were generated *a priori* to ensure participants had given the task their full attention and had knowledge levels which fit within the parameters of the task. The exclusion criteria are discussed in the Results section below. In total, 8 (9%) of participants met the criteria for exclusion. This left 78 participants available for analysis.

[1] The initial goal was to collect 40 highly skilled (Nelson-Denny accuracy >= 80% and completion >= 70%) and 40 low skilled (accuracy < 80% and completion < 70%) readers. Particular focus was given recruiting participants with lower reading skills than the typical undergraduate. However, the reading skills of participants recruited into the study fell along a more continuous distribution with few participants being classified as low skilled ($n = 16$, 21% after exclusions), so the recruitment strategy was modified to ensure a continuous distribution of reading skills ranging from low to high skill (Nelson-Denny Composite Score *Range* = 0 - 32.4 out of 36, *mean* = 18.12, *median* = 19.20, *SD* = 7.97).

2.1.2 Materials

2.1.2.1 Passages

In order for the RPL framework to be tested, reading materials for Session 2 needed to reflect three distinct difficulty levels: Easy (defined as material that was previously known or could be quickly comprehended), Moderately difficult (defined to be partially comprehended on a first pass), and Hard (defined as containing information that a typical undergraduate would not have the knowledge base to comprehend quickly). Although each reader's individual RPL could not be targeted, by ensuring most participants' initial comprehension of the Hard material would be low (at chance) and comprehension of the Moderate material would be partial, the moderately difficult material stood in for a close approximate to the readers RPL. Further, difficulty levels were set within each passage. That is, each passage has an Easy, Moderate, and Hard section. By setting the difficulty within the passage, effects of overarching topic knowledge in the design are controlled.

In order to fit the above criteria, passages were selected and refined over four stages. In stage 1, the experimenter and 3 research assistants reviewed ACT practice passages across numerous ACT prep sites until they identified five, non-overlapping expository passages, each containing at least three subtopics. In stage 2, each of the five passages was revised until the three subtopics had similar word counts and the experimenter was satisfied that each subtopic represented a distinct level of difficulty within the text. To confirm the difficulty rankings of the participants matched those of the experimenter, stage 3 involved a norming study during which participants rank ordered the sections within each passage based on their perceived, relative difficulty (detailed in Appendix A, Norming Study 1). Following norming, the three passages with the most consistent rank order were selected. These passages were on the topics of dinosaur intelligence (title: “Were Dinosaurs Dumb?”[2]), acupuncture treatments (title: “Needles and Nerves”[3]), and comets (title: “A Comment on Comets”[4]).[5]

Stage 4 ensured sections were clearly divided into distinct easy, moderate, and hard difficulty levels. The difficulty of the *hard* sections was increased by combining concepts to form longer sentences and adding distance between related concepts. Similarly, sections perceived as *easy* were further decreased in difficulty by adding supporting details, simplifying vocabulary, and shortening sentences. Revisions were made recursively until the text’s grade level as determined by Lexiles[6] reflected the difficulty levels of the participants; thus, Section Difficulty within passages was determined to be Easy, Moderate, or Hard first through subjective and then through objective measures. The order of section difficulty varied across passages but was consistent across presentations of that passage. See Table 1 for example revisions and Table 2 for descriptive statistics on each passage and section. All passages are provided in Appendix B.

2.1.2.2 Supplemental Text

Participants could also receive one of three supplemental texts about the passage. Each of the three supplemental texts was written to match the difficulty level of one of the sections from the main text and expanded upon the concepts presented in that section. Construction of supplemental texts followed Stages 2-4 as outlined above. Although all questions for a text could

[2] Original ACT prep version no longer available online but from https://www.act.org/ adapted from “Were Dinosaurs Dumb?” by Stephen Jay Gould (© 1980 by Stephen Jay Gould).

[3] Original ACT prep version https://www.powerscore.com/sat/help/content/ACT%202009-2010.pdf adapted from “Needles & Nerves” by Catherine Dold (©1999 by The Walt Disney Company).

[4] Original ACT prep version https://www.crackacc.com/act/reading/test220.html adapted from “A Comment on Comets” by Dr. Anatole C. Thierry (© 2002 by Weak Alliteration Press)

[5] The text versions printed within this document were heavily revised and are not the same as the cited content.

[6] Lexiles establishes a difficulty score for a text based on sentence length and word frequency (see Wright & Stenner, 1998 for more detail) and was independently assessed by the National Center for Education Statistics as a valid measure for assessing text difficulty (White & Clement, 2001).

be answered by the main text, the supplemental texts made the answers to questions for a given section more explicit. Further norming details are in Appendix A, Norming Study 3. Details about the supplemental materials can be found in Table 2 and the texts in Appendix B.

Table 1 Example Revisions

Passage	Revision Direction	Unrevised (selected portions)	Revised (selected portions)
Needles and Nerves	Easier: 1. Sentences shortened. 2. Vocabulary simplified. 3. Clarifying details added. In the revision to the right, this entailed making the relationship between Qi and bodily health more direct.	1. Practiced in China for more than 2,000 years, acupuncture is based on the belief that the body contains energy called Qi (pronounced "chee") that flows throughout the body on pathways called meridians. 2. Practitioners recognize some 1,500 acupoints, most of which have no obvious relationship to their intended targets based on Western models of the body. 3. Acupuncture can also be used to keep Qi in balance and keep the body healthy.	1. Acupuncture has been practiced in China for over 2,000 years. It is based on the belief that the body contains energy called Qi (pronounced "chee"). Qi is energy that flows through the body on pathways called meridians. 2. Over 1,500 acupoints have been found. However, most of the points have no obvious relationship to the parts of the body they are intended to treat. 3. Those who practice acupuncture believe it can keep Qi in balance. By keeping Qi in balance, they believe it can stop the body from getting sick.

A Comment on Comets	Harder: 1. Concepts combined to form longer sentences. 2. Distance added between relevant concepts. In the revision to the right, perihelion is defined 70 words before it is actually needed and “the Sun” has to be inferred as the cause of the evaporation.	1. The Sun's heat causes frozen material to evaporate, shedding gas and dust. A huge cloud of gas and dust builds up around the comet. This cloud is called the coma and is sometimes larger than Earth. 2. When comets are at their closest approach to the Sun, called perihelion, it will be much brighter than if it reaches its closest point to Earth while it is still relatively cold.	1. The Sun's heat causes frozen material to evaporate, and the resulting cloud formation around the nucleus is called the coma and can be larger than Earth. 2. As the comet moves towards its closest point to the Sun, the perihelion, … If a comet reaches its nearest point to Earth after its perihelion, it will be much brighter than if it reaches its nearest point to Earth while it is still relatively cold.

Table 2 Descriptive Statistics for Each Passage

Title	Section (in order of presentation)	Difficulty (in order of presentation)	Lexile Grade Level	Word Count	
				Main	Supplemental
A Comment on Comets	Orbits	Easy	6	194	397
	Light	Difficult	10	212	403
	Research	Moderate	8	195	407
Needles and Nerves	Vision	Moderate	8	223	409
	Origins & Qi	Easy	6	216	405
	Pain Management	Difficult	10	203	408
Were Dinosaurs Dumb?	Brains	Difficult	10	205	407
	Behavior	Moderate	8	206	402
	Changes in Beliefs	Easy	6	209	404

Note. Lexile grade levels were the same for main and supplemental texts.

2.1.2.3 Questions

Section Questions. Three multiple-choice questions were created for each section. As with the passages, the questions were created in stages and normed; details are in Appendix A Norming Study 2. Stage 1 involved the selection of questions from the ACT materials that came with the passage and the construction of new questions. Only inference-based questions were used as these questions are more likely to require readers to access their situation model of the text (Wiley et al., 2005). Testing readers at the level of the situation model, rather than the textbase, was preferred because the situation model represents the reader's deeper understanding of the text (Kintsch, 1994). That is, I sought to test readers on their internalized and integrated representations of the full text (situation model) rather than their ability to discern the meaning of specific sentences (textbase).

All questions initially had five multiple-choice options (one correct and four lures). During stage 2, I initially normed whether questions accurately reflected the text. In this stage of norming, participants answered the questions while they could look back at the passage. Questions which had accuracy levels reflecting the section's perceived level of difficulty were selected and all others discarded. This was done to ensure that no questions were unexpectedly difficult or easy. Further, only the three most-selected multiple-choice responses were used to ensure all lures were plausible alternatives (the correct answer was always among these three responses). Finally, in stage 3, questions were normed to ensure that they could not be answered without reading the text. This ensured that readers would need to rely on their comprehension rather than logical deductions based on prior knowledge.

Full Passage Questions. To help ensure readers were motivated to read the supplemental material for broader comprehension and not just to find the answers to a few questions, participants were told that the post supplemental reading Test Phase would contain three new questions (see 10 in Procedure for exact language). These questions were normed in the same manner as the subtopic questions and were deemed to represent full-passage comprehension either because the question itself required synthesis of information across sections or because the answer choices (correct and lures) were all drawn from different sections. These questions are analyzed separately because full passage questions were only included in the Post-Supplemental Test Phase and pre-to-post improvements in comprehension cannot be measured for these questions.

2.1.2.4 Individual Differences Assessments

Text Vocabulary Knowledge Assessment. My goal was to assess comprehension and metacomprehension at the level of the situation model rather than individual vocabulary words. Thus, all texts were written to contain vocabulary that would be familiar to most readers, including less-skilled readers. Further, to confirm that any observed differences were not due to passage-specific vocabulary knowledge, all participants were assessed on their comprehension of words in the passages as part of Session 1. Three vocabulary words from each section (27 total) which may present a barrier to comprehension were selected for testing. These words were either identified by Lexiles or by the experimenter if Lexiles did not identify three low frequency vocabulary words from a given section. Words which may be unknown to participants, but which were defined within the passage (e.g., *perihelion*) were not included.

Two measures assessed if participants had the requisite vocabulary to read the passages: subjective familiarity and objective knowledge. *Familiarity* was collected to ensure that participants at least believed they knew the definition of each word and was measured on a five-point Likert scale from *1 (No Idea!), 2 (Heard of it), 3 (Kind of know what it means, but I'm not sure), 4 (I could use this in a sentence), 5 (I could use this in a sentence and define it)* (Blachowicz, 1986). *Knowledge* was collected by asking readers to select the answer which most closely related to the target word. Participants respond to a three-alternative forced-choice question consisting of the correct meaning and two lures unrelated to the word (e.g., *agile: a. moves easily, b. tall, c. heavy*). Lures were wholly unrelated to the target word so that selecting the correct option only required the reader to have a basic sense of the word's meaning. To ensure participants did not have time to look up the definitions for the words, a three-minute time restriction was used for the knowledge assessment. All participants were able to complete the assessment in the allotted time. A complete list of vocabulary assessed is in Appendix C.

Any participant whose mean knowledge score was more than 2 standard deviations below the mean was excluded from analysis. See Exclusions in the Results section for more details on participant knowledge, familiarity, and resulting exclusions.

Nelson-Denny. Reading comprehension skill was assessed using the Nelson-Denny test (form E; Cummins, 1981), which contains three subscales: Vocabulary, Comprehension, and Reading Speed.

The vocabulary assessment consisted of 100 multiple-choice items, each containing a prompt which included the target vocabulary word in italics and five options. Participants selected the option which most closely reflected the meaning of the word in the context of the prompt. Participants had 15 minutes to answer all 100 items.

The comprehension assessment consisted of 8 passages and 36 multiple-choice questions (8 questions for the first passage and 4 questions each for the subsequent 7 passages). Each question had a correct target and four lures. Participants could refer to the passages as they answered the questions. Again, participants had 15 minutes to complete the assessment.

Because the Nelson-Denny is a timed test (15 minutes for each portion), participants often do not have time to answer every question. In scoring the Nelson-Denny, reading speed can be accounted for using separate measure of words read per minute (e.g., Cummins, 1981) or by treating the number of attempted questions separately from the total number of questions in the test (e.g., Balass et al., 2010). Using the second procedure a composite score can be generated by collapsing vocabulary and reading scales with speed:

Number correct – (Number Attempted - Number Correct)*.2

Using the composite measures, a participant who attempts 36 questions and answers 27 correct (75%) receives a score of 25.2. A participant who attempts 12 questions with 9 correct (75%) receives a score of 8.40, lower than the participant who was able to attempt all of the questions, but higher than a participant who attempts 36 questions and only gets 9 correct (25%). This participant would receive a 3.6.

Reading comprehension and vocabulary knowledge are often correlated (Calloway, 2019; Landi, 2010; Perfetti & Hart, 2001), and this was also the case in the present study, $r = .59$, $p < .001$. As the relationship between both variables on comprehension were also expected to be similar (and indeed were, $r = .06$ for comprehension and $r = .05$ for vocabulary), the mean of the two standardized scores was used instead to create a single measure of Reading Skill.

Reader-Based Standards of Coherence Questionnaire. A reader's standards of coherence influence performance on measures of comprehension (Calloway, 2019; van den Broek et al., 2011). Standards of coherence can be based on external reading goals imposed by the nature of the task or text (Yeari et al., 2015; Zwaan, 1994). However, they can also be internal and related to enjoyment from reading (Crossley et al., 2017) or to broader interest in learning through reading (Guthrie et al., 2004). Calloway (2019) developed a scale to assess individual differences in standards of coherence, which predicted reading comprehension in her study. The scale includes 31 statements (α = .89[7]) spread across four sub-scales and measured on a Likert scale from strongly disagree (1) to strongly agree (7): intrinsic reading goals (7 statements, 2 reversed coded, α = .91), extrinsic reading goals and learning strategies (8 statements, 3 reverse coded, α = .85), desire for understanding and reading regulation strategies (9 statements, 2 reverse coded, α = .86), and desired reading difficulty (7 statements, 4 reverse coded, α = .87). The mean score across all sub-scales was used as a single measure of Reader-Based Standards of Coherence (RBSC).

Self-Perception of Reading Skill. Metacognitive monitoring may involve a reader's ability to assess their reading skill. Because reading skills can differ across domains and the passages in this study were scientific in nature, a measure of Self-Perception of Reading Skill (SPRS) in a science context was used. The survey includes six questions concerning reading ability in connection to science texts and asks readers to rate their skill on a scale from 1-11 (α = .88; Kwon & Linderholm, 2014). The summed score constituted the readers' overall beliefs about their science reading ability and was included in all models.

Academic Self-Handicapping Scale. Another factor that may affect allocation of reading time and comprehension is self-handicapping. Self-handicapping allows a learner to externalize the cause of their failure (Schwinger et al., 2014). In the context of this study, a participant who believes they are not comprehending the material may choose the hardest content for study or put less thought into their responses because if they later do not know the answers, they will then be

[7] Alpha values reported for the RBSC, SPRS, and ASHS come from the original studies which introduced and validated the measures.

able to protect their self-esteem by assigning blame to the difficulty of the material they read about rather than to their own comprehension. The six-item Academic Self-Handicapping Scale (ASHS, $\alpha = .84$; Urdan et al., 1998) measures tendency to self-handicap in an academic domain. It describes possible student beliefs or behaviors and asks participants to rate how true each statement is for them on a scale from 1 (*not at all true*) and 5 (*definitely true*). The mean score is used across all models.

Judgments of Learning (JOLs). To assess monitoring accuracy, two types of JOLs were collected (Mazzoni & Nelson, 1995): (a) a predictive, text-wise JOL (as in 7) was made after each section, and (b) a postdictive, item-wise JOL (as in 8) after each comprehension question.

(7) How confident are you in your ability to accurately answer multiple choice questions about the text you read on the previous screen?

a. 25% (I think I will just be guessing amongst the choices.)

b. 50% (I think I will be able to answer half of the questions correctly.)

c. 75% (I think I will be able to answer most of the questions correctly.)

d. 100% (I think I will be able to answer all of the questions correctly.)

(8) What do you think is the chance that you answered the previous question correctly?

a. No Chance 0% -- Select only if you ran out of time to answer the question

b. Total Guess (25%)

c. Small Chance (50%)

d. Moderate Chance (75%)

e. High Chance (100%)

Two measures can be calculated using each JOL type:

Sensitivity (as measured by the gamma correlation; Nelson, 1984) refers to whether the learner can accurately assess performance on an individual item; i.e., if learners can discriminate items they have learned well versus not as well. A higher value indicates greater discrimination.

Miscalibration measures the discrepancy between the learner's overall confidence and their overall performance. A positive value reflects overconfidence and a negative value under-confidence.

Both measures of sensitivity and miscalibration were considered in all models. Sensitivity represents comprehension monitoring accuracy (Wiley et al., 2005) and as such, was expected to predict how readers allocated their time in Experiment 1b. Further, its effect on comprehension was expected to be mediated by choice. That is, readers who were more sensitive to their relative comprehension were expected to allocate more time to material in their RPL and thus to make greater improvements in comprehension. It was also possible that sensitivity would have a direct effect on comprehension across Experiments 1a, b, and c. Readers who are more sensitive to which questions they may have gotten incorrect may search for answers to those questions more directly once they are provided with (or choose) supplemental material.

Although sensitivity is more closely aligned with metacognitive accuracy, it cannot explain the direction of any inaccuracies in metacognitive monitoring. As overconfidence is necessary for illusions-of-knowing, measuring miscalibration was also considered important in assessing the role of comprehension monitoring on metacognitive control and comprehension. Without miscalibration, it would be difficult to distinguish whether low sensitivity scores were due to illusions-of-knowing or illusions-of-*not*-knowing.

Finally, it was not clear a priori whether readers predictive, text-wise or postdictive, item-wise sensitivity and miscalibration would be more predictive. Thus, I considered all four measures.

To preview, predictive and postdictive miscalibration were correlated highly and were collapsed (see Analytic Strategy for more detail).

Prior Knowledge. Participants' prior knowledge about the topics may have influenced their comprehension. Several steps were taken to mitigate effects of prior knowledge. Questions were normed to ensure at chance accuracy without the passages. Thus, in norming, participants prior knowledge was not able to improve their accuracy. Further, participants whose accuracy on the difficult material was high (i.e., for whom the materials in the study were not difficult enough to fall into their RPL) were excluded. Finally, participants were asked to self-report prior knowledge about topics related to the passages (see 9): comets ("A Comment on Comets"), dinosaurs ("Were Dinosaurs Dumb?"), and acupuncture, medicine, and neuroscience ("Needles and Nerves").

(9) Which of the below topics do you feel you are more knowledgeable about than the average person? (Select ALL that apply)

a. Dinosaurs

b. Acupuncture

c. Comets

d. Medicine

e. Neuroscience

f. My knowledge about these topics is probably similar to most people

The background knowledge question was intended to be used as a potential exclusion criterion only if self-reported prior knowledge resulted in increased accuracy on questions about the relevant passage. To preview, it did not, thus no participants were excluded based on this measure.

2.1.3 Procedure

All data was collected over the internet. Participants were encouraged to complete the study on a laptop or desktop computer but use of a smart phone or tablet was not prohibited. Participants could take breaks between sections. The second session containing the main experimental manipulation had to be completed within one week of the first session which collected individual difference measures.

2.1.3.1 Session 1

Participants took the Nelson-Denny vocabulary and comprehension tests (maximum of 15 minutes per section), the passage specific vocabulary familiarity (untimed) and knowledge (maximum of 3 minutes) tests, as well as the RBSC, SPRS, and ASHS (all untimed). The measures were presented in a random order with the exception that the Nelson-Denny vocabulary assessment was always immediately followed by the comprehension assessment.

2.1.3.2 Session 2

Following a gap of at least 24 hours but not more than 7 days, participants began Session 2. To increase the difficulty of the task, and in line with the RPL model (Metcalfe, 2002), time limits were placed on both reading and answering questions. During piloting, these time limits were rarely exceeded, and participants of all reading skill levels were able to complete the study without missing responses.

Session 2 included two test phases for each passage, as outlined in Figure 1. During the Pre-Supplemental Test Phase, participants read each passage and answered the questions sequentially, with passage order counterbalanced to ensure even serial placement across participants. To ensure that readers had time to read all sections during their initial exposure and did not skew their reading to favor an early section and thus not get to a later section, each section was displayed on a new screen with a reading time minimum of 30 second and limit of 90 seconds. On a new screen, immediately after participants read each section, participants were asked to provide a predictive, text-wise JOL as previously shown in (7) regarding how well they believed they would do on a subsequent assessment.

After reading all three sections for a passage, comprehension questions were presented one at a time in a random order with the full passage displayed below the question. Participants had one minute to answer each question and could not return to a question once they moved on. After answering each question, participants moved onto a new screen and provided a postdictive, item-wise JOL as previously shown in (8). Participants were not informed as to the accuracy of their response.

Participants then saw instructions for the Post-Supplemental Test Phase as in (10). Critically, the assignment of texts to supplemental-material conditions was manipulated within-subjects so that each participant received the easy supplemental text for one topic, the medium for another, and the hard for a third, with the assignment of topics to conditions counterbalanced across participants. Participants were given two minutes to read the supplemental information because it was slightly longer than the original text. Participants were able to refer back to the supplemental information when answering the comprehension questions in the Post-Supplemental Test Phase.

(10) Great work! You have finished with the first set of 9 questions about this topic. On the next screen, you will be given 2 minutes to continue reading about one of the sub-topics in the passage you just finished. The additional information will be similar to what you just read but will go into greater detail.

> Afterwards, you will be given the same 9 questions as before as well as an additional 3 new questions. Your goal for reading the additional material provided on the next screen is to improve your accuracy and confidence in your answers as much as possible.

After reading supplemental information, participants were asked the same nine questions from the Pre-Supplemental Test Phase in a random order and once again provided postdictive, item-wise JOLs. An additional three full-passage questions, which participants had not previously seen, were also presented. The full passage continued to be available below the question and participants continued to have one minute to answer each question.

At the end of Session 2, participants were asked to provide basic demographic information (age, race, ethnicity, gender, and education), to report any prior knowledge on the topics (see 9), and to state whether or not they had completed the task earnestly. Participants were assured that their response regarding their attention to the task would not affect their compensation.

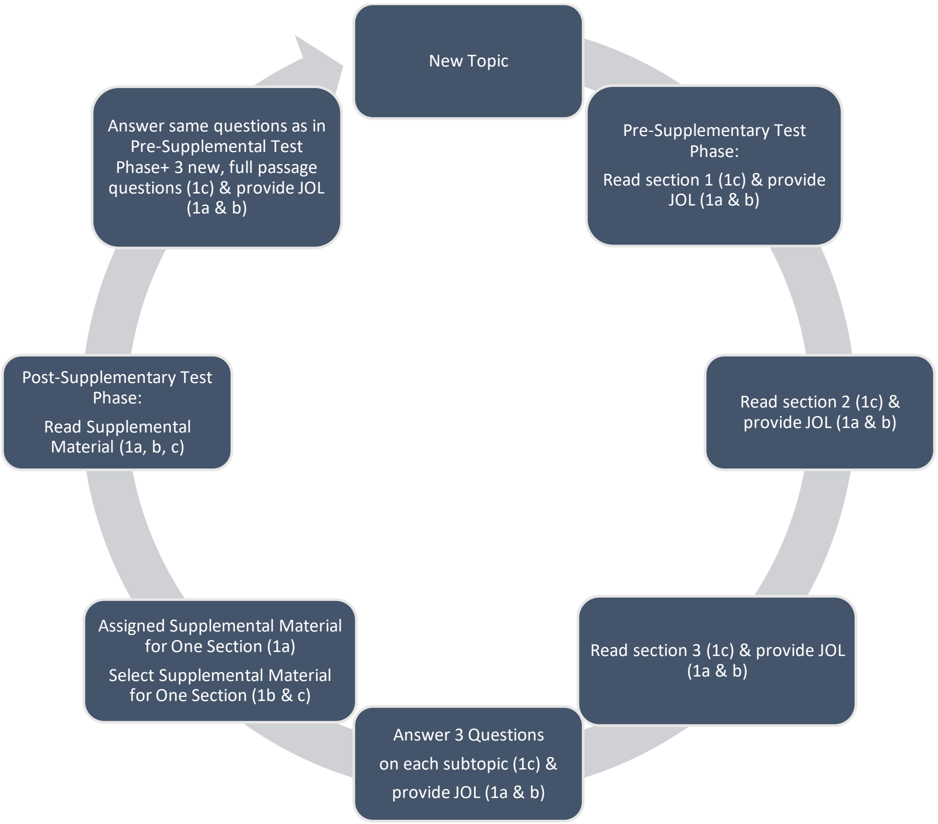

New Topic
Pre-Supplementary Test Phase:
Read section 1 (1c) & provide JOL (1a & b)
Read section 2 (1c) & provide JOL (1a & b)
Read section 3 (1c) & provide JOL (1a & b)
Answer 3 Questions on each subtopic (1c) & provide JOL (1a & b)
Assigned Supplemental Material for One Section (1a)
Select Supplemental Material for One Section (1b & c)
Post-Supplementary Test Phase:
Read Supplemental Material (1a, b, c)
Answer same questions as in Pre-Supplemental Test Phase+ 3 new, full passage questions (1c) & provide JOL (1a & b)

Figure 1 Study Flow

Note. The Pre-Supplementary Test Phase included initial reading, providing predictive, text-wise JOLs, answering questions, and providing postdictive, item-wise JOLs. The Post-Supplementary Test Phase included being assigned (Experiment 1a) or selecting supplemental material (Experiments 1b and 1c), re-answering questions from Pre-Supplementary Test Phase along with 3 new, full passage questions, and providing postdictive, item-wise JOLs. JOLs were only requested in Experiments 1a and 1b. The cycle was repeated once for each of the three topics.

2.1.4 Analytic Strategy

All analyses were conducted using R Project for Statistical Computing with the package *lme4* (Bates et al., 2016).

The core model used to evaluate the region of proximal learning model was a logistic mixed effects model[8] in which the dependent measure was Accuracy on the comprehension questions. Planned fixed effects included the primary manipulations of Test Phase (Pre- vs Post-Supplemental), Section Difficulty (Easy, Moderate, Hard), Supplemental Difficulty (Easy, Moderate, Hard), and their interactions. Test Phase was contrast coded (-0.5, 0.5) such that the coefficients indicated the effect of the Post-Supplemental Test Phase relative to Pre-Supplemental Test Phase. As both Section Difficulty and Supplemental Difficulty had an implicit ordering (Easy < Moderate < Hard), I tested the effect across difficulty levels by computing two polynomial contrasts.[9] The *linear* effect tested whether there was a consistent change in accuracy as text difficulty increased. A *quadratic* effect tested for a non-linear relationship; that is, a "sweet spot" of highest or lowest accuracy with a moderately difficult text, as predicted by the RPL model.

Although our primary interest was in the experimental manipulation of Supplemental Difficulty, variables of Reading Skill, Predictive and Postdictive Sensitivity and Miscalibration, SPRS, RBSC, and ASHS were also included as covariates. These continuous variables were standardized. Random intercepts for participants and items were used. Given the complexity of

[8] Model results from logistic models are given in log odds. To facilitate interpretation within the text, log odds were back-transformed to odds for all Experiments. Log odds are still reported in all tables.

[9] An alternative was to use contrast coding which first compared Easy versus Moderate and Hard and then Moderate versus Hard. However, in Experiment 1b, supplemental text choice was tested as a potential mediator making an ordinal comparison more straightforward within the available R statistical packages. Thus, I treated the variable as ordinal rather than nominal throughout Experiments 1a, b, and c. Choice of contrast coding scheme however did not affect the significance of the results. Indeed, contrast coding further supported my interpretations.

fixed-effect structure, a model with random slopes for between-subject and/or between-item variables was not expected to converge. Instead, the final model used the random-slopes structure best supported by the data (Matuschek et al., 2017). In this case, this meant a model with only random intercepts.

One concern with including multiple continuous variables is that they might produce multicollinearity. Although the experimental manipulations were perfectly unconfounded, the individual difference measures had the potential to be highly correlated. Indeed, predictive and postdictive miscalibration had a high degree of correlation, $r = .67$, $p < .001$. Given the high correlation between these two variables, it seemed that the two measures were capturing the same underlying construct and were thus collapsed into one measure of miscalibration (mean of the standardized values). All other correlations were small (see Individual Differences following Experiment 2 Discussion for greater detail). The condition number was used to assess the degree of harmful multicollinearity in the model (Belsley, 1991). A value falling below the 10-30 range is typically accepted as an indication that the model that does not contain harmful multicollinearity (e.g., Brunsdon et al., 2012; Kim, 2019). The condition number for the model in Experiment 1a was 6.59, less than even the most conservative threshold. Thus, the individual difference measures did not cause harmful levels of multicollinearity.

2.2 Results

2.2.1 Exclusions

2.2.1.1 Vocabulary Knowledge

As mentioned in the Materials section, two measures were used to assess whether or not readers had the requisite vocabulary knowledge to comprehend the texts: subjective familiarity and objective knowledge. Scores for both were high, *M* = 4.74 out of 5 (*SD* = .73) and *M* = .97 out of 1 (*SD* = .18), respectively. Thus, knowledge of the specific vocabulary used in the passages should not have limited performance. Nevertheless, 2 (2%) participants had mean knowledge more than 2 standard deviations below the mean and were thus excluded from analysis.

2.2.1.2 Attention to Task

An additional 2 (2%) participants reported not attending to the readings and were thus excluded.

2.2.1.3 Variation in Judgments of Learning (JOL)

A learner is said to be able to differentiate between what they know and do not know if they provide higher JOLs for the items they get correct than for the ones they get incorrect. Thus, calculations of sensitivity require participants to provide varying JOLs across items. Participants (*n* = 3, 3%) who provided the same JOL for all questions were thus excluded from analysis.

2.2.1.4 Task Difficulty

A critical assumption of Experiment 1 was that each passage included sections of an Easy, Moderate, and Hard reading level, respectively. The norming studies reported above demonstrated that this was indeed true for the population of interest (i.e., first language English speakers in Western Pennsylvania without a bachelor's degree). However, it was still possible that participants would find the easy material too hard or the hard material too easy, which would mean that none of the material in the study fell in with their region of proximal learning. Because this possibility could vary on a topic-by-topic basis depending on participants' understanding of an individual topic, exclusions were conducted by passage rather than by participant. Specifically, a passage was deemed to have been *too easy* for a participant if they answered all of the questions correctly during the Pre-Supplemental Test Phase (i.e., no room for growth in the Post-Supplemental Test Phase). This did not occur in Experiment 1a.

To determine if a passage was too easy for a participant, both accuracy and JOLs were considered. It was critical that the hard material *be* hard for the participants and that the moderate material offer room to grow. A passage was also be deemed too easy if (a) the participant answered all of the questions in the hard section correctly *and* (b) indicated that these were not merely "lucky guesses" through greater-than-moderate confidence in their postdictive, item-wise JOLs. Thus, a mean JOL > 4 (greater than 75% chance that the question was answered correctly) indicated that the hard material was either easy or within the readers RPL. This eliminated responses to 1 passage for 3 participants (1% of passages). Similarly, a mean JOL of 4 on the moderate material was acceptable because it indicated that the participant believed they could improve their certainty. Participant passages were excluded if participants answered all of the moderate questions correctly and indicated high confidence in their answers (5 out of 5). This eliminated responses to 1 passage for 5 participants and 2 passages for 1 participant (3%). Thus, participants who did not struggle with the hard section or had no gains to make in JOLs or accuracy for the moderate section were excluded.

A passage was similarly excluded if it was too hard for a participant, defined as scoring at or below chance on the easy section of the passage. 1 passage for 14 participants and 2 passages for 3 participants (8%) met this criterion. This exclusion criteria fully eliminated 1 (1%) participant from Experiment 1a.

The exclusion process left 211 (89%) out of the original 236 passages across participants for analysis ("Needles and Nerves" = 72, "A Comment on Comets" = 62, "Were Dinosaurs Dumb?" = 71). Although the exclusions left an unequal number of observations across passages, a chi-square goodness of fit test indicated the distribution was not significantly changed, χ_2^2 = *.89*, *p* = *.64*. Because exclusions were performed at the level of an entire passage rather than individual sections, these exclusions did not result in an imbalance in exposure to section difficulties. Even after exclusions, the number of participants assigned to each Supplemental Difficulty (Easy = 66, Moderate = 73; Hard = 66 unique occurrences) did not significantly differ, χ_2^2 = .48, *p* = .79.

2.2.1.5 Prior Knowledge

Participants were asked to indicate if they believed they had more knowledge than a typical person on several topics related to the passages. 38 (49%) participants said they had more than typical knowledge about at least one topics related to "Needles and Nerves," 6 (8%) said they had more than typical knowledge about comets, and 7 (9%) said they had more than typical knowledge about dinosaurs. 39 (50%) participants indicated they did not have more than typical knowledge on any of the listed topics. However, participant's perception of their prior knowledge was not reflected in increased accuracy on questions about the relevant passages.[10] Either participants were inaccurate in their estimates or the information in the passages did not build on relevant prior knowledge in such a way that prior knowledge of the topics aided participants. Thus, no exclusions were conducted on the basis of perceived prior knowledge.

2.2.2 Region of Proximal Learning Model

The primary model for Experiment 1a evaluated the effects of Section Difficulty and Supplemental Difficulty on improvements in accuracy across Test Phases. The results from this model are reported in Table 3 and discussed in detail below.

2.2.2.1 Individual Differences

There were no reliable relationships between Academic Self-Handicapping (ASHS), Metacognitive Sensitivity, or Reader-Based Standards of Coherence (RBSC) and comprehension accuracy, *p*s > .10. However, higher reading skill, higher Self-Perceptions of Reading Skill (SPRS), and higher Metacognitive Miscalibration (overconfidence) all predicted more accurate comprehension. Specifically, each standardized unit increase in SPRS predicted a 1.12 times (95% CI:[1.01, 1.23]) increase in the odds of a correct response, *z* = 2.19, *p* = .03. Further, each standardized unit increase in Reading Skill predicted a 1.33 times (95% CI:[1.20, 1.47]) increase

[10] Participants who suggested they had greater knowledge of topics related to the “Needles and Nerves” passage only got 56.46% (se = 1.83%) correct versus 56.11% (se = 1.78%) for those who did not indicate more than typical knowledge, *p* = .90. Participants who suggested they had greater knowledge of topics related to the “A Comment on Comets” passage only got 53.33% (se = 4.89%) correct versus 57.14% (se = 1.43%) for those who did not indicate more than typical knowledge for topics in this passage, *p* = .50. Participants who suggested they had greater knowledge of topics related to the “Were Dinosaurs Dumb?” passage only got 57.94% (se = 4.42%) correct versus 56.41% (se = 1.34%) for those who did not indicate more than typical knowledge for topics in this passage, *p* = .73.

in the odds of a correct response, *z* = 5.31, *p* < .001. Thus, both reading skill and perceptions of reading skill were positively related to accuracy.

Finally, each standardized unit increase in Metacognitive Miscalibration (greater overconfidence) predicted a 0.80 times (95% CI:[0.73, 0.89]) decrease in the odds of a correct response, *z* = -4.39, *p* < .001.

2.2.2.2 Experimental Manipulation

There were multiple effects of the experimental manipulation. First, as would be expected, Section Difficulty had a negative linear effect on the odds of a correct response; across test phases the odds of a correct response decreased by 0.12 times (95% CI:[0.08,0.19]) for each level of increase in Section Difficulty, *z* = 09.37, *p* < .001. There was also a positive quadratic effect illustrating that the difference in accuracy between Easy and Moderate sections was greater than the difference between Moderate and Hard sections, *z* = 2.57, *p* = .01.

There were further effects of Supplemental Difficulty. Although a linear effect of Supplemental Difficulty was not reliable, *p* = .12, there was a negative quadratic effect. Comparison of the means demonstrated that this quadratic effect reflected greater difference between Easy and Moderate supplemental conditions than Moderate and Hard supplemental conditions, *z* = -2.21, *p* = .03.

The presence of a quadratic effect of Supplemental Difficulty indicates that either participants who received the Moderate supplemental text started at a higher baseline accuracy for Pre-Supplemental Test Phase and then remained higher for the Post-Supplemental Test Phase or that the effect of receiving the Moderate supplement text on Post-Supplemental Test performance was large enough to overcome any baseline similarities for the Pres-Supplemental Test Phase. Given that each participant received all of the conditions, the latter seems more likely. Indeed, as can be seen in Figure 2, accuracy during the Pre-Supplemental Test Phase was numerically *lower* for participants receiving the Moderate supplement text than for participants receiving the Hard text.

Further, there were no main effects of Test Phase, *p* = .69. The odds of a correct response did not reliably increase from the Pre-Supplemental to Post-Supplemental Test. Rather increases in accuracy across Test Phases were moderated by Section Difficulty and Supplemental Difficulty.

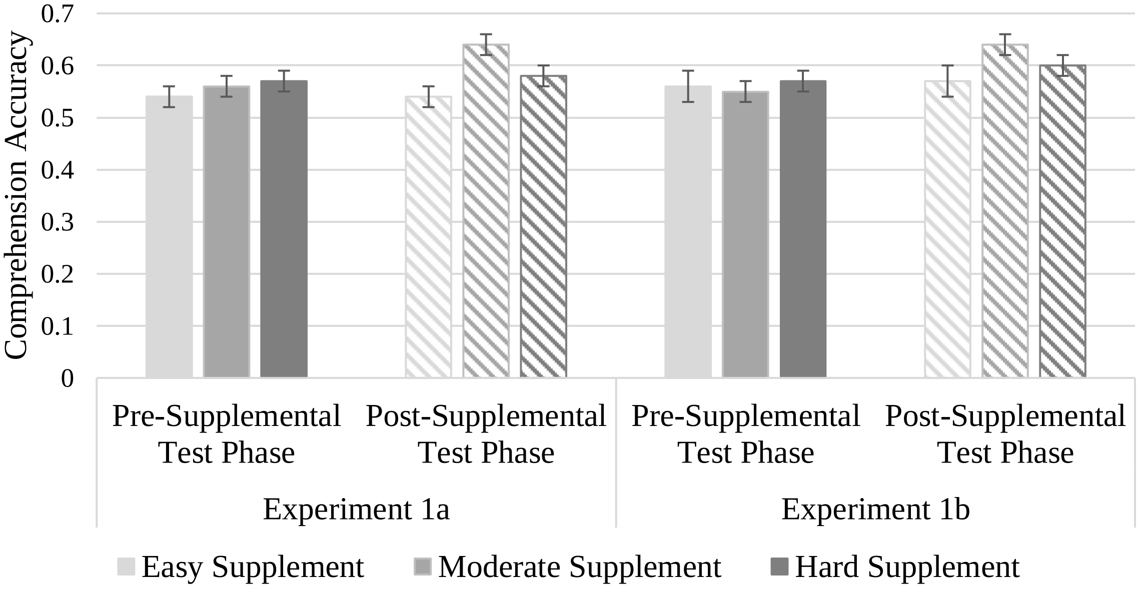


Figure 2 Accuracy Collapsed Across Section Difficulty for Experiment 1a and 1b

Note. Error bars represent standard error of the mean

2.2.2.3 Interactions with Test Phase

The change in accuracy from the Pre- to Post-Supplemental Test Phase as a function of Supplement Difficulty was the core effect of interest. An interactive effect would indicate that the type of supplemental material participants received affected their rate of improvement in comprehension. The interaction between the linear effect of Supplemental Difficulty and Test Phase was not reliable, *p* = .99. Accuracy did not steadily improve or decline across conditions.

Rather, Test Phase interacted with the quadratic effect of Supplemental Difficulty, $z = -1.98$, $p = .05$. Figure 3 shows that the quadratic effect was driven by greater increase from Pre- to Post-Supplemental Test Phase when participants received the Moderate difficulty supplemental material as opposed to either the Easy or Hard supplemental material.

Test Phase also interacted with Section Difficulty. The linear effect of Section Difficulty was less extreme in the Post-Supplemental Test Phase as compared to the Pre-Supplemental Test Phase, $z = 3.18$, $p = .001$. Further, a quadratic trend for Section Difficulty showed once again that this difference was more extreme for Easy versus Moderate rather than Moderate versus Hard Sections, $z = -2.27$, $p = .02$. Together, the two effects demonstrated that readers made greater improvement in their scores from the Pre- to Post-Supplemental Test Phases among questions related to the Moderate and Hard sections rather than Easy sections.

A three-way interaction between Test Phase, Section Difficulty, and Supplemental Difficulty was also tested. All interactions involving the linear terms for Section Difficulty or Supplement Difficulty were nonsignificant, *p*s > .30. However, an interaction between the two quadratic terms and Test Phase existed, $z = 2.18$, $p = .03$. This effect demonstrated that when participants received the Moderate supplemental condition, their accuracy improved from the Pre- to Post-Supplemental Test Phase specifically for the questions related to the Moderate section. Thus, readers improved broadly across Section Difficulties from the Pre- to Post-Supplemental Test Phase when they received the moderate supplemental material, but this improvement was greatest for the moderate section. Figure 3 shows the means broken down by each variable.

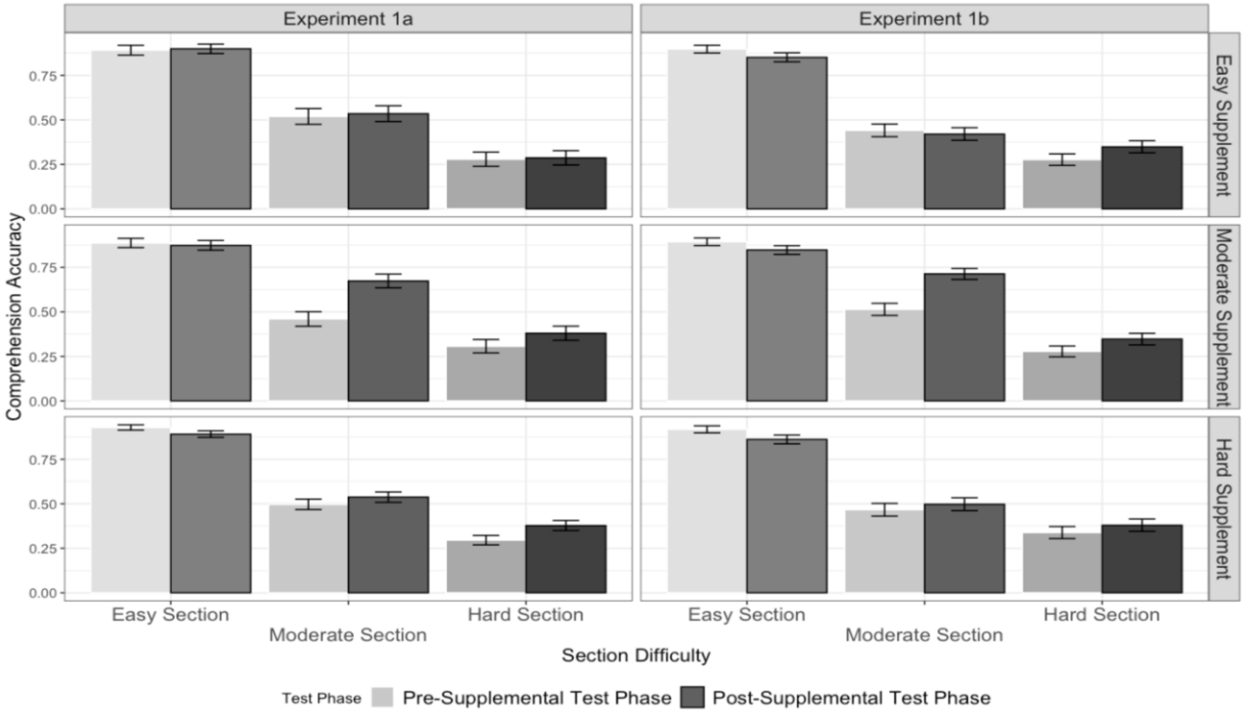


Figure 3 Mean Accuracy for Experiments 1a and 1b

Note. The key interaction of interest shows that the increase in comprehension accuracy from pre- to post-supplemental test phase is greater for the moderate section when the moderate supplemental material is received (center of each Experiment graph) versus for the hard section when the hard supplement is assigned (bottom right of each Experiment graph). subs Error bars represent standard error from the mean.

2.2.2.4 Other Interactions

The two-way interaction between the quadratic effects of Supplemental Difficulty and Section Difficulty was also significant. Overall, the quadratic effect of Section Difficulty had indicated that there was a bigger difference in accuracy between the Easy to Moderate material than between the Moderate to Hard material; however, this was less pronounced when readers received the Moderate supplemental condition, z = 4.50, p < .001, because receiving the supplemental material for the Moderate section increased accuracy on that section. As with the three-way interactions with Test Phase, other two-way interactions that included linear effects of Section Difficulty and/or Supplemental Difficulty were not reliable, *p*s > .62.

Table 3 Logistic Mixed Effects Model Predicting Question Accuracy (Experiment 1a)

Variable	$\hat{\beta}$	***SE***	***Wald-z***	***p***
Intercept	0.51	0.13	3.97	< .001
Main Effects of Manipulation				
Section Difficulty (L)	-2.11	0.23	-9.37	< .001 ***
Section Difficulty (Q)	0.56	0.22	2.57	.01 **
Test Phase	0.04	0.09	0.40	.69
Supplement Difficulty (L)	0.12	0.08	1.55	.12
Supplement Difficulty (Q)	-0.17	0.08	-2.21	.03 *
Main Effects of Individual Differences				
Reading Skill	0.28	0.05	5.06	< .001 ***
ASHS	0.01	0.05	0.21	.83
SPRS	0.11	0.05	2.19	.02 *
RBSC	-0.05	0.05	-1.01	.31
Metacognitive Miscalibration	-0.22	0.05	-4.39	< .001 ***
Predictive Metacognitive Sensitivity	0.07	0.04	1.34	.10
Postdictive Metacognitive Sensitivity	0.06	0.04	1.34	.18
2-Way Interactions Excluding Test Phase				
Supplemental Difficulty (L) x Section Difficulty (L)	0.08	0.15	0.49	.62
Supplemental Difficulty (L) x Section Difficulty (Q)	0.02	0.12	0.19	.85
Supplemental Difficulty (Q) x Section Difficulty (L)	0.01	0.14	0.04	.97
Supplemental Difficulty (L) x Section Difficulty (Q)	0.54	0.12	4.50	< .001***
2-Way Interactions with Test Phase				

Test Phase x Supplement Difficulty (L)	-0.002	0.16	-0.01	.99
Test Phase x Supplement Difficulty (Q)	-0.30	0.15	-1.98	.05 *
Test Phase x Section Difficulty (L)	0.53	0.17	3.18	.001 **
Test Phase x Section Difficulty (Q)	-0.32	0.14	-2.27	.02 *
3-Way Interactions with Test Phase				
Test Phase x Supplemental Difficulty (L) x Section Difficulty (L)	0.02	0.29	0.08	.94
Test Phase x Supplemental Difficulty (L) x Section Difficulty (Q)	-0.24	0.25	-1.03	.94
Test Phase x Supplemental Difficulty (Q) x Section Difficulty (L)	0.08	0.28	0.30	.76
Test Phase x Supplemental Difficulty (Q) x Section Difficulty (Q)	0.53	0.24	2.18	.03 *

Note. L = Linear, Q = Quadratic, ****p* < .001, ***p* < .01, **p* < .05, †*p* < .10

2.2.3 Full Passage Questions Model

As noted in the Method section, participants also answered Full Passage questions which relied on comprehension of multiple sections. It is possible that the difficulty of the supplemental material received would affect participants ability to answer these new questions. These questions were analyzed in a separate model because they were presented only during the Post-Supplemental Test Phase and did not pertain to a specific section, and thus a model including fixed effects of Section Difficulty and Test Phase was inappropriate. Results are listed in Table 4.

There was no significant relation between accuracy on the full passage questions and individual differences in Reading Skill, Self-Handicapping, Reader-Based Standards of Coherence, Metacognitive Sensitivity, or Self-Perceptions of Reading Skill, *p*s > .11.[11] Instead, and as with the primary analysis, there was a negative quadratic effect of Supplemental Difficulty such that accuracy was highest when participants received the Moderate supplemental text, z = -3.88, $p < .001$.

Metacognitive Miscalibration also predicted accuracy. The odds of a correct response decreased by 0.75 times (95% CI:[0.60, 0.95]) for every standardized unit increase in miscalibration, z = -2.44, p = .01. Readers who were more overconfident had lower accuracy on the full passage questions.

[11] Null effects in the model predicting accuracy on the full passage questions may be due to low power. While the primary model had 52 observations per participant, the full passage model only contained 9. As the primary model was sufficient to test the RPL framework, no a priori power analysis was done for the full passage questions model.

Table 4 Logistic Mixed Effects Model Predicting Full Passage Question Accuracy (Experiment 1a)

Variable	$\hat{\beta}$	*SE*	*Wald-z*	*p*
Intercept	0.07	0.30	0.23	.82
Supplement Difficulty (L)	-0.14	0.16	-0.84	.40
Supplement Difficulty (Q)	-0.63	0.16	-3.88	< .001 ***
Reading Skill	0.19	0.12	1.57	.12
ASHS	-0.11	0.11	-1.02	.31
RBSC	0.05	0.12	0.41	.69
SPRS	0.13	0.11	1.09	.27
Metacognitive Miscalibration	-0.28	0.11	-2.44	.01 **
Predictive Metacognitive Sensitivity	-0.07	0.10	-0.68	.49
Postdictive Metacognitive Sensitivity	-0.02	0.10	-0.18	.86

Note. L = Linear, Q = Quadratic, ****p* < .001, ***p* < .01, **p* < .05, †*p* < .10

2.3 Discussion

Experiment 1a tested the Region of Proximal Learning (RPL) model in the context of reading expository texts. The central prediction of the model was confirmed: Pre-to-post increases in comprehension accuracy were greatest for readers receiving Moderately difficult supplemental material rather than Easy or Difficult supplements. The gains in comprehension as a result of receiving the Moderate supplemental material were present broadly, but they were especially pronounced for comprehension of the material related to the moderate section itself. Receiving supplemental details about any of the sections might be expected to increase accuracy *for that* *section*. What is notable about this finding was that (a) the increase was greater for Moderate Sections when receiving the Moderate Supplement than for the Hard Sections when receiving the Hard Supplement, and (b) the improvements in accuracy as a result of receiving the Moderate supplement extended across Sections, even resulting in higher accuracy on full-passage questions. Thus, improvements were not just related to receiving specific details which helped answer a specific set of questions. The reader's model of the text improved also allowing for greater understanding of material not contained in the moderate section. The improvement across subtopics may be related to the interactive nature of learning concepts. Although each concept in the text represented a distinct sub-field of the overall topic, they were all connected by a broader topic. Thus, greater learning in one domain may help readers integrate information from other domains into their larger situation model of the text as a whole.

Notably, there were few significant relationships between individual differences and accuracy. As expected, the model testing the RPL framework also detected higher accuracy odds as reading skill went up. Self-Perception of Reading Skill (SPRS) was also related with higher accuracy odds. It was possible that this relationship occurred because stronger readers perceived themselves to be stronger readers and weaker readers perceive themselves as weaker. However, in Experiment 1a, this was not the case. Rather, the correlation between SPRS and actual reading skill was nearly non-existent, r = -.001, p = .99. Thus, readers who believed they were strong readers tended to perform better regardless of their reading skill. Although Reader Based Standards of Coherence (RBSC) did not predict accuracy in Experiment 1a, it is worth noting that SPRS and RBSC did significantly correlate, r = .46, p < .001. Thus, I speculate that readers who believed they were more skilled had higher accuracy because their standards of coherence during the study were higher.

Finally, increases in Metacognitive Miscalibration (overconfidence) predicted decreases in comprehension. Although the variable was standardized in the model, mean miscalibration was high for participants at all accuracy levels, *M* = 19.32, *SE* = 1.22. Thus, this finding reflects that as overconfidence increased, accuracy decreased. Miscalibration is calculated using accuracy and thus it may appear tautological that it would predict accuracy, but this is not guaranteed. If participants are biased towards over or under confidence by the same degree, a change in accuracy would not occur as a function of overconfidence. That is if one participant has a mean confidence (JOL) of 80% (expects that there is an 80% likelihood of answering questions correctly on average) and a mean accuracy of 70% and another participant has a mean confidence of 70% and a mean accuracy of 60%, both participants are overconfident, but their miscalibration is the same (10%) and thus miscalibration does not in this case predict accuracy. The finding that miscalibration predicts accuracy in this study then indicates that participants with lower mean accuracies actually had similar or higher JOLs than participants who were more accurate. Thus, the finding here is evidence that readers in Experiment 1 were highly overconfident.

Results from Experiment 1a are consistent with prior findings supporting the RPL model in list learning tasks (e.g., Metcalfe, 2002) and extend the model to the domain of reading comprehension. Readers' comprehension generally did not improve when they were given easy supplemental material, presumably because comprehension of the easy sections was already quite high. Reading in greater detail about concepts that are already well understood is unlikely to spark greater insights. Assignment of hard supplemental material did improve accuracy within the harder sections, but critically not to the same extent as moderate supplemental material within the moderate sections.

A key feature of Experiment 1a was that I experimentally manipulated which supplemental material readers received for each text. The use of random assignment allows the inference that studying moderately difficult material in fact *caused* the learning gains. This parallels situations in which an instructor chooses materials *for* students. However, in many other situations, learners must choose materials and their allocation of study time themselves. Experiment 1a provides no information about how readers exercise metacognitive control during supplemental text assignment and whether moderately difficult material would still be optimal under such circumstances. Thus, in Experiment 1b, I instead allow participants to exercise metacognitive control in choosing the supplemental material for themselves so that I can assess the relationship between comprehension monitoring and metacognitive control.

3.0 Experiment 1b

Experiment 1a demonstrated the importance of reading material in the reader’s Region of Proximal Learning, but it did not demonstrate whether or not readers are able to intuit the right reading material for themselves. Kornell & Metcalfe (2006) found that learners who had greater metacognitive sensitivity selected material for study in their Region of Proximal Learning. However, it is not clear if *readers* have enough sensitivity to assess the relative difficulty of texts nor whether they would choose to select material within their RPL. Based on Experiment 1a, readers likely should select the moderately difficult section to improve their learning. Experiment 1b tested whether or not they do. Thus, the goals of Experiment 1b were to determine if participants (a) would choose to further study the material most conducive to generating learning gains, (b) still benefit more if they selected the moderate supplement, and (c) selection of the moderate supplement (metacognitive control) was predicted by metacognitive monitoring (miscalibration and sensitivity).

3.1 Method

3.1.1 Participants

Based on the power analysis for Experiment 1a, a goal of 80 participants was once again set for Experiment 1b.[12] Participants (*N* = 82, M_{age} = 26, 64 female, 3 unspecified) were recruited from the University of Pittsburgh psychology participant pool (*n* = 58) or from the wider urban community (*n* = 24) following the same protocols as in Experiment 1. Five (6.09%) participants were removed following the exclusion criteria, leaving 76 available for analysis.

3.1.2 Materials and Procedure

The materials across both sessions were the same in Experiments 1a and 1b. The primary difference between Experiment 1a and 1b was in how participants received supplemental texts. In Experiment 1b, participants chose the supplemental text they read (as in 11). The subtitles for each section of the topic were provided in the order participants read them and were followed by brief details to remind of them of the what the section discussed (e.g., *Acupuncture and Pain Management – This was the last section you read and contained information about endorphins and nerves*). Participants were asked the following:

[12] Again, the initial recruitment goal to collect equal numbers of 40 high and 40 low skilled readers was adjusted when readers fell along a more continuous distribution (Nelson-Denny Composite Score *Range* = 1.20 - 33.60 out of 36, *mean* = 19.84, *median* = 19.20, *SD* = 7.80).

11. Which of the following sections should you read more about in order to maximize your confidence and chances of answering the greatest number of questions correctly?

Participants in Experiment 1b were also asked queried at the end of the study about the criteria they used in making their selection of reading material. The first question (see 12) broadly asked about participants' criteria.

12. What criteria did you use when selecting which section to read more about? Select all that apply:

a. I selected the section based on how difficult I found the topic to read about.
b. I selected the section based on my confidence in my answers to questions about that section.
c. I selected the section I was most interested in.
d. I selected the section I didn't have enough time to read initially.
e. I selected a section at random.
f. Other ___

Participants who said they selected the section based on the difficulty of the material or their confidence in their answers were asked to elaborate (see 13 and 14).

13. You said you based your selection on how difficult the material was. Please elaborate.

a. I selected the section that I thought was easiest to understand.
b. I selected the section that I thought was moderately difficult to understand.
c. I selected the section that I thought was the hardest to understand.

14. You said you based your selection on how confident you were in your answers. Please elaborate.

   a. I selected the section if I had high confidence in my answers for that section.
   b. I selected the section if I had moderate confidence in my answers for that section.
   c. I selected the section if I had low confidence in my answers for that section.

3.1.3 Analytic Strategy

Planned analysis for the logistic mixed-effect model on Accuracy was the same as in Experiment 1a; however, the variable of Supplemental Difficulty now reflected the participants' choice and thus was renamed Supplemental Choice to distinguish it from Experiment 1a.

In Experiment 1b, I additionally tested (a) whether participants favored supplemental material of a particular difficulty, (b) if individual differences, particularly those related to metacognitive sensitivity, could predict Supplemental Choice, and (c) whether or not Supplemental Choice mediated the effect of Metacognitive Sensitivity on Accuracy. To determine if participants preferred a certain type of supplemental material, a chi-square goodness-of-fit test was used to determine if the frequencies with which participants selected a section difficulty level significantly differed from chance. This test was performed with all three texts, and then follow-up tests were used to probe each pairwise combination of two texts.

Supplemental Choice was then regressed on the individual-difference measures; because this was an ordinal outcome, regression was conducted using the R package *ordinal* using the logit link (Christensen, 2019).[13]

Lastly, testing a mediation with an ordinal mediator and multiple hierarchical levels (with the dependent variable and mediator on two different levels) is not easily accommodated with current statistical packages. Thus, I applied Baron & Kenny's (1986) test for mediation. Step 2 of this process tests whether the treatment (Metacognitive Sensitivity) predicts the mediator (Supplemental Choice). If the treatment does not predict the mediator, a mediation is not present. To preview, no measures of individual difference significantly predicted Supplemental Choice, and thus there was no significant evidence for mediation.

The same checks for multicollinearity were applied to Experiment 1b as in Experiment 1a. Predictive and postdictive miscalibration again had a high degree of correlation, r = .70, p < .001, and were collapsed as in Experiment 1a. All other correlations were small. The condition number was again below the most conservative threshold, 8.21. Thus, the models in Experiment 1b do not contain harmful levels of multicollinearity.

[13] Use of a logit link function meant that effects were given in log odds of selecting one supplemental text versus another. As with the logistic models, log odds are reported in the tables but back-transformed within the text.

3.2 Results

3.2.1 Exclusions

3.2.1.1 Vocabulary Knowledge

As in Experiment 1a, Familiarity and Knowledge of the tested vocabulary was high, $M_{\text{familiarity}}$ = 4.65 out of 5 (*SD* = .78) and M_{accuracy} = .97 out of 1 (*SD* = .17), respectively. These results suggest that specific vocabulary used in the passages should not have limited performance. 1 (1%) participant had mean knowledge more than 2 standard deviations below the mean and was thus excluded from analysis.

3.2.1.2 Attention to Task

An additional 2 (2%) participants reported not attending to the readings and were thus excluded.

3.2.1.3 Variation in Judgments of Learning (JOL)

Participants (*n* = 1, 1%) who provided the same JOL for all questions were excluded from analysis.

3.2.1.4 Task Difficulty

As in Experiment 1a, the difficulty of the task for each participant was assessed, and exclusions were made if the task was either too hard or too easy. 1 participant answered all of the questions correctly for 1 passage and thus the passage was excluded (0.43% of passages). Following the exclusion criteria outlined for Experiment 1a, 1 additional (0.43%) passage was excluded for one participant because the hard section was too easy, 7 (3%; 1 for each of 7 participants) passages were excluded because the moderate section was too easy, and 23 (10%; 2 for each of 3 participants and 1 for each of 17 participants) passages were excluded due to at chance accuracy on the easy section. This exclusion procedure also eliminated 1 (1%) participant from Experiment 1b.

The exclusion process left 208 (88%) out of the original 234 passages for analysis ("Needles and Nerves" = 73, "A Comment on Comets" = 67, "Were Dinosaurs Dumb?" = 68). As in Experiment 1a, a chi-square goodness-of-fit test indicated the distribution was not significantly changed, χ^2_2 = .30, p = .86. Because supplemental material was freely chosen by participants, I would not necessarily expect each difficulty level to be chosen with equal frequency, and indeed they were not (Easy = 53, Moderate = 63; Hard = 118 unique occurrences); critically, however this distribution was not significantly altered (Easy = 44, Moderate = 54; Hard = 110 unique occurrences) by the exclusions, χ^2_2 = .53, p = .77.

3.2.1.5 Prior Knowledge

40 (53%) participants said they had more than typical knowledge about topics related to "Needles and Nerves," 13 (17%) said they had more than typical knowledge about comets, and 15 (20%) said they had more than typical knowledge about dinosaurs. 32 (42%) participants indicated they did not have more than typical knowledge on any of the listed topics. Once again, participants' perception of their prior knowledge was not reflected in increased accuracy on questions about the relevant passages.[14] As in Experiment 1a, either participants were inaccurate in their estimates or

[14] Participants who suggested they had greater knowledge of topics related to the "Needles and Nerves" passage only got 55.39% (*SE* = 1.76%) correct versus 60.27% (*SE*= 1.81%) for those who did not indicate more than typical knowledge, *p* = .09. Although this effect is marginal, it suggests that perceived prior knowledge may have interfered

the information in the passages did not build on relevant prior knowledge in such a way that it aided participants with prior knowledge of the topics.

3.2.2 Region of Proximal Learning Model

Experiment 1b evaluated the effects of Section Difficulty and Supplemental Choice on improvements in accuracy from Pre- to Post-Supplemental Test Phase (see Figure 3). The effects of this model are reported in Table 5 and discussed in detail below.

3.2.2.1 Individual Differences

The relationships between accuracy Academic Self-Handicapping (ASHS), Self-Perceptions of Reading Skill (SPRS), and both measures of Sensitivity were not reliable, *p*s > .21. Reader-Based Standards of Coherence (RBSC) marginally predicted accuracy such that for every standardized unit of increase the odds of a correct response improved by 1.10 times (95% CI:[1.00, 1.21]), *z* = 1.88, *p* =.06. Once again, Reading Skill and Metacognitive Miscalibration reliably predicted comprehension. For each standardized unit of increase in reading skill, the odds of a correct response increased by 1.32 times (95% CI:[1.17, 1.50]), *z* = 4.48, *p* < .001. Standardized increases in Metacognitive Miscalibration (greater overconfidence) predict a 0.80 times (95% CI:[0.72, 0.89]) decrease in the odds of a correct response, *z* = -4.08, *p* < .001.

with comprehension rather than aided it. Participants who suggested they had greater knowledge of topics related to the "A Comment on Comets" passage only got 58.01% (*SE* = 3.25%) correct versus 61.82% (*SE* = 1.42%) for those who did not indicate more than typical knowledge for topics in this passage, *p* = .39. Participants who suggested they had greater knowledge of topics related to the "Were Dinosaurs Dumb?" passage only got 60.54% (*SE* = 2.86%) correct versus 55.20% (*SE* = 1.54%) for those who did not indicate more than typical knowledge for topics in this passage, *p* = .13.

3.2.2.2 Experimental Manipulation

Experiment 1b replicated multiple effects of the manipulations of Section Difficulty. These included the negative linear effect of Section Difficulty on the odds of a correct response; the odds of a correct response decreased by 0.11 times (95% CI:[0.07, 0.16]) for each level of increase in Section Difficulty, z = -10.39, p < .001. There was also a positive quadratic effect illustrating that the difference in the odds of a correct response between Easy and Moderate sections were greater than the difference between Moderate and Hard sections, *z* = 2.18, *p* = .03.

However, other main effects from Experiment 1a did not persist. There was no main effect of Supplemental Choice, *p*s > .42, but there was a marginal effect of Test Phase. The odds of correct response marginally increased by 1.19 times (95% CI:[0.99,1.44]) from the Pre- to Post-Supplemental Test Phase, *z* = 1.83, *p* = .07.

3.2.2.3 Interactions with Test Phase

Replicating Experiment 1a, there was a quadratic trend in the Text Phase x Supplemental Choice interaction, *z* = -1.99, *p* = .05. Accuracy increased more from the Pre- to Post-Supplemental Test Phase when participants selected the *Moderate* supplemental material rather than the *Easy* or *Hard* material. This finding replicates the conclusion from Experiment 1a that the RPL also applies to a reading framework and further indicates that the benefits of the RPL externalize to a context in which readers can choose how to allocate their reading time; i.e., a context more with greater external validity in that being allowed to allocate reading time is more similar to self-regulated reading.

Test Phase also interacted with Section Difficulty. A linear trend demonstrated that accuracy from the Pre- to Post-Supplemental Test Phase increased more as Section Difficulty increased, *z* = -1.80, *p* = .07. A further quadratic trend showed that this increase in accuracy was greatest for the Moderate Section, *z* = 2.23, *p* = .03.

As in Experiment 1a, the only reliable three-way interaction included the two quadratic terms for Section Difficulty and Supplemental Choice along with Test Phase, *z* = 2.01, *p* = .04. Thus, even when given a choice of materials, readers improved broadly across Section Difficulty from Pre- to Post-Supplemental Test Phase when they read the Moderate supplemental material and the improvement was greatest for the Moderate section.

3.2.2.4 Other Interactions

There were no interactions which did not involve differences between Test Phases, all *p*s > .11.

Table 5 Logistic Mixed Effects Model Predicting Section Specific Question Accuracy (Experiment 1b)

Variable	$\hat{\beta}$	***SE***	***Wald-z***	***p***
Intercept	0.56	0.12	4.53	< .001
Main Effects of Manipulation				
Section Difficulty (L)	-2.24	0.22	-10.39	< .001 ***
Section Difficulty (Q)	0.45	0.21	2.18	.03 **
Test Phase	-0.18	0.10	-1.83	.07 †
Supplemental Choice (L)	0.07	0.09	0.79	.43
Supplemental Choice (Q)	0.01	0.09	0.09	.93
Main Effects of Individual Differences				
Reading Skill	0.28	0.06	4.48	< .001 ***
ASHS	-0.04	0.05	-0.87	.38
SPRS	0.06	0.05	1.24	.21
RBSC	0.09	0.05	1.88	.06 †
Predictive Metacognitive Sensitivity	-0.04	0.05	-0.71	.48
Postdictive Metacognitive Sensitivity	0.03	0.06	0.54	.59
Metacognitive Miscalibration	-0.23	0.06	-4.08	< .001 ***
2-Way Interactions with Excluding Test Phase				
Supplemental Choice (L) x Section Difficulty (L)	0.005	0.16	0.03	.98
Supplemental Choice (L) x Section Difficulty (Q)	0.19	0.13	1.46	.15
Supplemental Choice (Q) x Section Difficulty (L)	-0.25	0.16	-1.58	.11
Supplemental Choice (Q) x Section Difficulty (Q)	0.20	0.14	1.47	.14
2-Way Interactions with Test Phase				

Test Phase x Supplement Choice (L)	-0.04	0.16	-0.24	.81
Test Phase x Supplement Choice (Q)	-0.34	0.17	-1.99	.05 *
Test Phase x Section Difficulty (L)	0.33	0.18	-1.80	.07 †
Test Phase x Section Difficulty (Q)	-0.34	0.15	-2.23	.03 *
3-Way Interactions with Test Phase				
Test Phase x Supplemental Choice (L) x Section Difficulty (L)	0.46	0.31	1.49	.14
Test Phase x Supplemental Choice (L) x Section Difficulty (Q)	-0.16	0.26	-0.62	.54
Test Phase x Supplemental Choice (Q) x Section Difficulty (L)	-0.08	0.32	-0.24	.81
Test Phase x Supplemental Choice (Q) x Section Difficulty (Q)	0.55	0.27	2.01	.04 *

Note. L = Linear, Q = Quadratic, ****p* < .001, ***p* < .01, **p* < .05, †*p* < .10

3.2.3 Full Passage Questions Model

Accuracy on the Full Passage questions was again evaluated using the same model from above but excluding Test Phase and Section Difficulty. Results are listed in Table 6.

Reading Skill did not predict accuracy on the full-passage questions, *p* = .23. However, Self-Perceptions of Reading Skill (SPRS) did marginally predict accuracy. Increases in SPRS were marginally associated with a 1.23 times increase (95% CI:[1.00, 1.52]) in the odds of a correct response, *z* = 1.94, *p* = .05. This was also true for Metacognitive Miscalibration, which was again associated with a 0.69 times (95% CI:[0.55, 0.86]) decrease in the odds of a correct response, *z* = -3.24, *p* < .001.

Critically, these results were accompanied by a quadratic effect of Supplemental Choice, *z* = -2.14, *p* = .03 such that accuracy on the full-passage questions was highest when participants received the Moderate versus Hard or Easy supplemental text.

There were no other reliable effects in the model.

Table 6 Logistic Mixed Effects Model Predicting Full Passage Question Accuracy (Experiment 1b)

Variable	$\hat{\beta}$	*SE*	*Wald-z*	*p*
Intercept	0.11	0.34	0.34	.73
Supplement Difficulty (L)	0.26	0.19	1.55	.13
Supplement Difficulty (Q)	-0.37	0.17	-2.14	.03 *
Reading Skill	0.15	0.13	1.20	.23
ASHS	-0.06	0.10	-0.60	.55
RBSC	0.002	0.10	0.02	.98
SPRS	0.21	0.11	1.94	.05 .
Predictive Metacognitive Sensitivity	0.01	0.10	0.10	.92
Postdictive Metacognitive Sensitivity	-0.05	0.11	-0.40	.69
Metacognitive Miscalibration	-0.40	0.11	-3.51	< .001 ***

Note. L = Linear, Q = Quadratic, ****p* < .001, ***p* < .01, **p* < .05, †*p* < .10

3.2.4 Supplemental Choice Model

The prior two models tested the RPL in reading framework when participants could self-allocate their reading time. However, a critical goal of Experiment 1b was also to consider what individual difference factors, and specifically Metacognitive Sensitivity, predicted participant choice in supplemental material.

Participants were more likely to choose supplemental material for the Hard section (53% of the time) than Easy (21% of the time) or Moderate (26% of time) section, χ^2_2 = 36.50, p < .001. Indeed, Hard material was selected more often than Easy and Moderate material combined. Thus, there was a significant difference in participant choices between the Hard and Easy sections, χ^2_1 =

28.29, $p < .001$ and Hard and Moderate, $\chi_1^2 = 19.12$, $p < .001$ but not between Easy and Moderate, $\chi_1^2 = 1.02$, $p = .31$.

Participants appear to have selected the Hard section more often *because* it was most difficult. In explaining their criteria for allocating their time to specific texts, 74% (56) of participants stated that they considered the difficulty of the corresponding section when making their choice. Of these participants, 70% (39) said that the selected the material they perceived to be the most difficult. 11% (6) selected the material they found to be easiest, and 16% (9) selected what they believed to be moderate difficulty when making their choice.

Other reasons participants cited for making their choice included *Personal interest* (26% of participants), *Needing more time to read* (9% of participants), and *At random* (7% of participants). Participants were allowed to select more than one option, so some participants considered multiple factors in making their choice.

One possibility for the finding that participants intentionally selected the hard material was that they believed their performance on the moderate section was at ceiling. However, JOLs for the moderate section among participants who reported selecting hard material indicated that participants were only moderately confident, $M_{\text{JOL}} = 3.21$, $SE = .05$. Further, participants who selected the harder material were no more overconfident across all sections than participants who selected the easy material; $M_{\text{miscalibration}} = 20.14$, $SE = 0.34$ for participants who selected harder material versus $M_{\text{miscalibration}} = 19.97$, $SE = 0.46$ for participants who selected easier material. I discuss this pattern of effects further in the Experiment 1b Discussion and General Discussion.

I also tested whether individual differences in participants' reading skill, metacognitive skill, or metacognitive beliefs predicted their choice in supplemental material. As text choice involved three options, ordinal regression was performed with Easy set as the lowest value and

Hard as the highest (see Table 7). None of the variables considered in the model reliably predicted Supplement Difficulty Choice.

Table 7 Model Predicting Supplemental Choice in Experiment 1b

Variable	$\hat{\beta}$	*SE*	*Wald-z*	*p*
Intercept (Easy Supplement | Moderate Supplement)	-1.34	0.23	-5.91	
Intercept (Moderate Supplement | Hard Supplement)	-0.004	0.18	-0.02	
Predictive Metacognitive Sensitivity	0.08	0.19	0.43	.67
Postdictive Metacognitive Sensitivity	0.16	0.21	0.78	.44
Metacognitive Miscalibration	0.21	0.21	1.00	.32
Reading Skill	0.33	0.23	1.41	.16
SPRS	-0.08	0.19	-0.41	.68
RBSC	-0.19	0.18	-1.05	.29
ASHS	0.18	0.19	0.90	.32

Note. ****p* < .001, ***p* < .01, **p* < .05, †*p* < .10

3.2.5 Mediation

Because no individual differences predicted Supplemental Choice, no mediation was present according to Baron and Kenny (1986)’s test.

3.3 Discussion

Experiment 1b replicated and extended the core findings from Experiment 1a: self-selecting moderately difficult reading material predicted higher improvements in accuracy across two test phases and higher accuracy on questions assessing full passage comprehension.

Did the choice of moderately difficult material *cause* the improvements in comprehension? Taken alone, Experiment 1b would not permit a causal inference; it would be possible that participants who were likely to show the most increase in accuracy across test phases also happened to self-select into the moderate supplement condition. That is, these participants may have had the same performance if they had self-selected into easy or hard condition. However, Experiment 1b used the same materials and an extremely similar procedure as Experiment 1a, in which random assignment *did* allow the causal claim that moderately difficult material led to the greatest learning gains. Taken together, then, an inference of causality seems warranted for Experiment 1b: Choosing the moderate supplement likely did cause readers to learn more.

Experiment 1b also demonstrated that readers were most prone to select the hardest materials. Further, participants indicated that selecting the hardest materials was their intention.

Unlike Thiede et al. (2003), the decision to select the hardest materials did not appear to be related to readers' inability to identify which material was in their region of proximal learning. Instead, the majority of readers stated that they intended to select the material that was most difficult; they were aware that the hardest sections were the hardest and selected them anyway.

Why did many readers adopt a strategy of targeting the hardest material? There are at least two possible explanations. One is that readers tried to identify material in their RPL, but because they were generally overconfident in their comprehension, they believed that the hardest material was in fact the material that was in their RPL. However, this did not appear to be the case as their JOLs for the moderate section indicated they believed they had room to improve. The other possibility, then, is that readers had adopted a strategy—that, in this case, was not the optimal choice for their learning —of studying the hardest material. Their suboptimal allocation of reading time appeared to be related to the heuristic they used to guide their metacognitive *control* processes rather than an error in comprehension *monitoring.*

4.0 Experiment 1c

Experiment 1a and 1b demonstrated that participants made greater learning gains when they studied moderately difficult material as compared to harder or easier material on the same topic. Despite this, in Experiment 1b, participants selected the Hard sections at a greater frequency than Easy or Moderate sections. Thus, their selection in study material did not align with the type of material which promoted the greatest learning gains.

However, there is one potential confound. To measure comprehension monitoring, I had to ask participants in Experiments 1a and 1b to also made judgments of learning (JOLs) after reading each section and after answering each question. It is possible that rating their perceived comprehension may have introduced metacognitive monitoring into participants' processing that would not have otherwise occurred. Indeed, some research has indicated that readers may not monitor their comprehension at all (Oakhill et al., 2005; Tighe et al., 2021). Learners are not normally asked by an external source to consider how confident they are in their ability to answer a question or comprehend a text. Thus, asking participants to supply this information may have engaged them in atypical metacognitive processes that influenced their later choices of which material to study. Further, the mere act of making JOLs may itself improve learning (Myers et al., 2020; Soderstrom et al., 2014; Witherby & Tauber, 2017).

To rule out this possibility, Experiment 1c replicated Experiment 1b except that participants did not provide JOLs. Experiment 1c thus also acts as a third replication of Experiment 1a and 1b in testing whether reading moderately difficult material is most conducive to learning while also testing whether the prior effects were an artifact of asking participants to make JOLs.

4.1 Method

4.1.1 Participants

Students (N = 50, 31 female, M_{age} = 20) participated in partial fulfillment of a course requirement. All participants self-reported that they attended to the task, had no learning disabilities, and were 1st language speakers of English. Two participants were excluded per criteria detailed under Exclusion in the Results section, leaving 48 for analysis.

4.1.2 Materials and Procedure

The materials and procedure were identical to Session 2 of Experiment 1b except that participants no longer provided JOLs. Individual differences measures (Session 1 of Experiments 1a and 1b) were not collected in Experiment 1c.

4.1.3 Analytic Strategy

I fit a logistic mixed effects model with Test Phase, Section Difficulty, and Supplemental Choice as fixed effects and Items and Participants as random effects similar to Experiments 1a and 1b. The model in Experiment 1c only differed in the absence of the individual-difference variables.

To detect differences in the choice of Supplemental Difficulty, a chi-squared goodness of fit test was once again conducted.

4.2 Results

4.2.1 Exclusions

No participants reported being inattentive to the task. In Experiments 1a and 1b, exclusion criteria involved assessments taken during Session 1 as well as interactions between participant confidence and accuracy. As JOLs were not collected in Experiment 1c, only exclusions due to the task being too difficult were used. Participant responses related to specific passages were excluded if accuracy on the Easy section was at chance during the Pre-Supplemental Test Phase. This criterion resulted in the exclusion of 2 (4%) participants, 1 passage from a further 10 participants and 2 passages from 6 participants (15% of passages).

4.2.2 Region of Proximal Learning Model

Experiment 1c replicated the results of Experiments 1a and 1b. Both a linear and quadratic effect of Section Difficulty existed such that the odds of a correct response declined by 0.15 times (95% CI:[0.12, 0.20]) as difficulty increased, $z = -12.72$, $p < .001$, but the decline in accuracy diminished from Moderate to Hard, $z = 4.07$, $p < .001$.

Critically, the interaction between Test Phase and the quadratic effect of Supplemental Difficulty Choice was significant, $z = -2.88$, $p = .004$. There was a greater increase in accuracy from Pre- to Post-Supplemental Test Phase across all Section Difficulties when readers selected the Moderate supplemental text. However, the three-way interaction which included Section Difficulty was not reliable, $p = .33$. Thus, the benefits of selecting the Moderate supplemental text were spread over accuracy for all questions and did not differentially benefit questions about the Moderate section itself.

Finally, linear and quadratic interactions between Section Difficulty and Test Phase suggest that learning gains from Pre- to Post-Supplemental Test Phase were greater for the two harder sections (Moderate and Hard) than for the Easy section, *p*s < .03.

4.2.3 Full Passage Questions Model

A second model with accuracy for full-passage questions was also conducted as in Experiments 1a and 1b. Here, Supplemental Choice was the only fixed effect. Unlike in the prior two experiments, Supplemental Choice did not reliable predict the odds of a correct response for full passage questions, *p*s > .61. However, the full passage model contained 3 observations per participant versus 18 for the RPL model (9 in the Post-Supplemental Test Phase). Thus, it may not have been powered to detect effects on accuracy.

4.2.4 Supplemental Choice Model

Participants chose to read the Hard section more frequently (51% of the time) than both the Moderate (22% of the time) and Easy (27% of the time) sections, χ^2_2 = 17.23, p < .001. This held true in pairwise comparisons between the Hard and Moderate sections, χ^2_1 = 13.76, p < .001, and Hard and Easy sections, χ^2_1 = 8.85, p = .003. However, there was no reliable differences in the frequency of selecting the Moderate versus Easy sections, χ^2_1 = 0.60, p < .44.

As in Experiment 1b, participants' choice of the Hard section aligned with their self-reported reasons for their choice. 64% (31) of participants reported selecting a section based on its perceived difficulty and 67% (21) of these participants reported selecting the section with the greatest perceived difficulty. Meanwhile, only 22% (7) reported selecting the section they perceived as moderately difficult and 9% (3) reported selecting the section they perceived to be the easiest. The remaining participants reported using some other criteria to inform their text select (*Personal Interest* = 29%, *Needing more time to read* = 6%, and *At random* = 8%).

4.3 Discussion

Experiment 1c demonstrates the findings of Experiment 1a and 1b persist when readers are not asked to reflect on their confidence levels. Further, as in Experiment 1b, readers preferred harder texts despite their learning being most supported by texts of moderate difficulty. This rules out two possible confounds: that the relations observed in prior experiments arose because the JOL prompts engaging additional monitoring that would not otherwise occur or because the JOLs themselves contributed additional learning. Rather, Experiment 1c demonstrates that readers allocate their reading time similarly regardless of whether or not they provide overt JOLs.

5.0 Experiment 2

Experiment 1 demonstrated that readers learn the most when they choose moderately difficult material to focus on. However, readers rarely made this choice, instead opting to choose the most difficult material. Although readers were able to correctly identify which material was most difficult, they may have been overconfident in their ability to comprehend that material with further exposure. Indeed, participants throughout Experiment 1 were overconfident in their comprehension. It is not clear what heuristics readers use to monitor their comprehension or why they sometimes fail to recognize their low comprehension (i.e., become overconfident). Experiment 2 tested the hypothesis that one way readers evaluate their comprehension is through the use of an ease-of-processing heuristic. Specifically, when a text *feels* easy to read, readers assess it as comprehended, even when it may not be. This may lead to shallower processing in which portions of the text are underspecified, resulting in complex sections of text going unattended. An alternate view, however, that processing becomes shallower when the surface code of a text is more complex, leaving few resources available for forming a semantic representation of the text (e.g., Ferreira et al., 2002; van Dijk and Kintsch, 1983) and potentially leading to illusions-of-knowing. If taxing readers working memory leads to greater gist processing, then gist processing should be more likely to occur for harder texts and simpler texts should not be able to increase a reader's illusions-of-knowing over harder text or even decrease the likelihood.

In order to test the circumstances that lead to illusions-of-knowing in a laboratory setting, I designed a paradigm in which comprehension is improbable. Specifically, I embedded pseudowords within the easy and hard sections of the texts used in Experiment 1. Because the pseudoword is inherently meaningless, any perception of comprehension represents an illusion-of-knowing. When reading easier sections, if participants rely on their overall ease of processing to judge their comprehension, they may miss the comprehension barrier caused by the pseudoword and become especially overconfident in their comprehension as compared to when they are reading harder sections. Further, if illusions-of-knowing are due to readers not attending to the pseudoword in the easy section, then calling attention to the pseudoword should reduce overconfidence. Following, Mata's (2020) successful manipulation of attention to a complex region of text via underlining it, attention in Experiment 2 was drawn to the pseudoword via underlining it.

This resulted in a 2 (Section Difficulty – Easy vs Hard) x 3 (Pseudoword Conditions – Absent, Merely Present, Underlined) within-subjects design. I assessed participants' processing of the pseudoword in two ways. First, I assessed readers' confidence in their ability to answer questions about the text, as in Experiment 1; the critical questions referred to the sections including the pseudoword. Then, as a second test of reader attention to the pseudoword, readers were also later asked to indicate how likely it was that they had seen the pseudoword in the earlier text.

5.1 Method

5.1.1 Participants

Pilot data estimated that the critical 2-way interaction (i.e., the effect of Section Difficulty x Pseudoword Condition on Metacognitive Miscalibration) has a standardized effect size of 0.48. A power analysis for a mixed-effect model using R package *simr* (Green & MacLeod, 2016) indicated that N = 60 participants were needed to detect this effect with power of .80. As in

Experiment 1 this number was doubled (*N* = 120) to ensure readers from a variety of skill levels were recruited.[15]

Participants (*N* = 153, 101 female, 7 undisclosed gender) were recruited in the same manner as Experiment 1 (*n* = 105 from the University of Pittsburgh and *n* = 48 from the surrounding urban community). 15 participants were excluded based on a priori exclusion criteria (see Exclusions in Results for more details). This left 138 participants available for analysis.

5.1.2 Materials

5.1.2.1 Passages

To simplify the experimental design, only the Easy and Hard Sections from Experiment 1 were included in the experimental manipulations; the Moderate section was retained as filler in the study to distract from the experimental manipulation.[16] The easy and hard sections were modified as detailed in Appendix A Norming Study 4. Briefly, a new sentence was embedded within the easy and hard sections of each passage as in 15. A word or words in this sentence was identified as the critical word, which was replaced with a pseudoword in the pseudoword conditions. Norming (again detailed in Appendix A) confirmed that replacing the word with the pseudoword

[15] As in Experiment 1, the initial goal was to collect 60 skilled and 60 less skilled readers. However, the reading skills of participants recruited into the study fell along a more continuous distribution, so the recruitment strategy was modified to ensure a continuous range of reading skills would be included in the models (Nelson-Denny Composite Score *Range* = 0 - 34.80 out of 36, *mean* = 19.32, *median* = 19.20, *SD* = 7.59).

[16] Including the pseudoword manipulation within the moderate section would have required at least two pseudowords to be present in most texts and thus increased the likelihood that participants would have recognized the manipulation. Further, as the moderate section was by definition a mix of easier to comprehend and harder to comprehend material, it would have been difficult to determine how reader's ease-of-processing heuristics were influencing their attention to the pseudoword. Thus, the moderate section was retained for each text to include sections which never received a pseudoword and the same questions as in Experiment 1 were used to probe comprehension of the moderate section, but results from this section were not analyzed.

successfully obscured the meaning of the sentence, as required by the experimental design; i.e., when the pseudoword was present accuracy for questions related to the new sentence fell to chance.

(15) Indeed, their extensive ear canals/trooks/<u>trooks</u> likely coordinated rapid eye movements and quick reflexes.

Across topics, all participants received all versions of the pseudoword manipulation (Absent, Merely Present, and Emphasized) for the Easy and Hard Sections. Because each section addressed a different subtopic, a separate sentence was constructed for the easy and hard sections. Thus, a given text could have a pseudoword in one of the sections, both, or none.

5.1.2.2 Questions

Three new comprehension questions were created for each of the Easy and Hard Sections (6 per passage). New questions targeted comprehension of the sentence containing the pseudoword manipulation as in 16.

(16) Which of the following likely helped researchers to realize that dinosaurs were agile creatures?

a. A finding of an intact, fossilized ear, complete with inner ear bones

b. A finding of an intact, fossilized brain with extensive motor cortex

c. A finding of an intact, fossilized spine showing extensive neck muscles

The comprehension questions from Experiment 1 were retained for used as filler questions in Experiment 2. These served purposes: (a) to obscure the experimental manipulation and (b) to assess participants' overall level of comprehension so that I could exclude participants who did not attend to the texts at all.

5.1.2.3 End-of-task Memory for Pseudowords

The final task in the study required participants to estimate how likely it was that they had seen each of a series of pseudowords and obscure real words during the experiment (Likert scale 1-5: Extremely Unlikely – Extremely Likely). The task included the 6 pseudowords used in the experiment, as well as 10 real words and 4 novel pseudowords (not used in the study) as lures.

5.1.3 Procedure

As in Experiment 1, the experiment was conducted over two sessions. Session 1 included collecting individual-difference measures. Session 2 also broadly followed the procedure in Experiment 1 except the use of the supplemental material and by extension post-supplemental questions was dropped. At the end of Experiment 2, participants took the end-of-task memory test and answered demographic questions.

5.1.4 Analytic Strategy

In Experiment 2, I focused my analysis on the item-wise judgments of confidence since the text-wise judgments of confidence necessarily included readers' beliefs about their comprehension of the material not including the new, pseudowords sentence. Thus, it would not be possible to ensure their judgments were taking the new sentence into account. Further, Postdictive Miscalibration was calculated only for the questions related to the pseudoword manipulation, not for filler questions.

Three mixed-effects models were planned, differing only in their dependent variable. All models contained fixed effects of Section Difficulty (contrast coded), Pseudoword condition (contrast coded), and the individual differences (standardized) of Reading Skill, Self-Perceptions of Reading Skill (SPRS), Reader-Based Standards of Coherence (RBSC), and Academic Self-Handicapping (ASHS). All fixed-effects were tested for multicollinearity. The condition number was 4.10; thus, none of the models contain harmful levels of multicollinearity. Random intercepts of Participant and Items were also used.

The first model served as a manipulation check to confirm that the two conditions with a pseudoword lowered comprehension accuracy.[17] This check ensures that any differences in metacognitive judgments reflect differences in overconfidence rather than veridical differences in accuracy in one condition.

The second model tested the core hypothesis that overconfidence (Postdictive Miscalibration) would increase when a pseudoword was present versus absent but that this increase would be moderated (increase less) when the pseudoword was Underlined versus Merely present.

Finally, the third model tested whether participants attended to the presence of a pseudoword in the passage. The dependent variable was the participant rating of how likely it was that they saw each pseudoword.

[17] Accuracy models in Experiment 2 reflect the pseudoword manipulation. However, a model of accuracy was conducted on filler questions and outcomes were consistent with those of Experiment 1. Further, the pseudoword manipulation did not significantly alter the accuracy for filler questions, *p*s > .54. Thus, the effects of Pseudoword Presence were confined to questions relating to the new manipulation.

5.2 Results

5.2.1 Exclusions

5.2.1.1 Vocabulary Knowledge

Scores for both familiarity and knowledge remained high in Experiment 2, *m* = 4.67 out of 5 (*SD* = .80) and *m* = .96 out of 1 (*SD* = .20), respectively. However, 3 (2%) participants had mean knowledge more than 2 standard deviations below the mean and were thus excluded from analysis.

5.2.1.2 Attention to Task

An additional 5 (3%) participants reported not attending to the readings and were thus excluded.

5.2.1.3 Variation in Judgments of Learning (JOL)

Participants (*n* = 4, 3%) who provided the same JOL for all questions related to the pseudoword manipulation were excluded from analysis.

5.2.1.4 Task Difficulty

As in Experiment 1, passages which were deemed too difficult or too easy for a given participant were excluded from analysis. So that exclusions could be made independent of the outcomes of interest, only filler questions were used in making this determination. 8 (2%) passages from 8 participants were excluded due to high accuracy on the Hard Section. 51 (12%) passages were excluded from 34 participants due to the Easy Section being too hard for the participant. The exclusion procedure eliminated 5 (3%) participants from Experiment 2.

The exclusion process left 364 (86%) out of the original 423 passages for analysis ("Needles and Nerves" = 127, 74 real word, 74 Pseudo, 69 Pseudo Underlined, "A Comment on Comets" = 116, 67 real word, 65 Pseudo, 69 Pseudo Underlined, "Were Dinosaurs Dumb?" = 121, 69 real word, 66 Pseudo, 72, Pseudo Underlined). Although inclusion of each passage was no longer exactly equal, a chi-square goodness of fit test indicated the distribution was not significantly changed, χ^2_2 = *.50, p = .78* and χ^2_8 = 1.5, *p* = .99 for each level of the condition in each passage.

5.2.1.5 Prior Knowledge

Participants were again asked to indicate if they believed they had more knowledge than a typical person on several topics related to the passages. 50 (34%) participants said they had more than typical knowledge about at least one of the topics related to "Needles and Nerves," 6 (4%) said they had more than typical knowledge about comets, and 16 (11%) said they had more than typical knowledge about dinosaurs. 75 (51%) participants indicated they did not have more than typical knowledge on any of the listed topics. Once again, however, participant's perception of their prior knowledge was not reflected in increased accuracy on questions about the relevant passages, so I did not use this an exclusion criteria.[18]

[18] Participants who suggested they had greater knowledge of topics related to the "Needles and Nerves" passage only got 48.30% (*se* = 1.84%) correct versus 48.63% (*se* = 1.46%) for those who did not indicate more than typical knowledge, *p* = .89. Participants who suggested they had greater knowledge of topics related to the "A Comment on Comets" passage only got 57.33% (*se* = 5.75%) correct versus 49.07% (*se* = 1.22%) for those who did not indicate more than typical knowledge for topics in this passage, *p* = .20. Participants who suggested they had greater knowledge of topics related to the "Were Dinosaurs Dumb?" passage only got 46.19% (*se* = 3.45%) correct versus 46.42% (*SE* = 1.25%) for those who did not indicate more than typical knowledge for topics in this passage, *p* = .95.

5.2.2 Accuracy Model

There was no reliable difference in accuracy on the critical questions based on whether they were embedded in the easy or hard section, *p* = .88. However, the presence of a pseudoword reduced the odds of a correct response by 0.22 times (95% CI:[0.18, 0.27]) compared to when the real word was used, *z* = -14.66, *p* < .001. Further, there were no reliable differences in accuracy based on whether or not the pseudoword was underlined, *p* = .61. Thus, the primary factor in whether or not a reader could answer the question correctly was whether or not the real word had been replaced with a pseudoword. Underlining the pseudoword did not reliably change accuracy (see Figure 4 and Table 8).

Reading skill did not reliably predict accuracy, *p* = .31, but there was a significant effect Self-Perceived of Reading Skill (SPRS) such that for each 1-standard-deviation increase in SPRS, the odds of a correct response increased by 1.14 times (95% CI:[1.02, 1.27]), *z* = 2.39, *p* = .02.

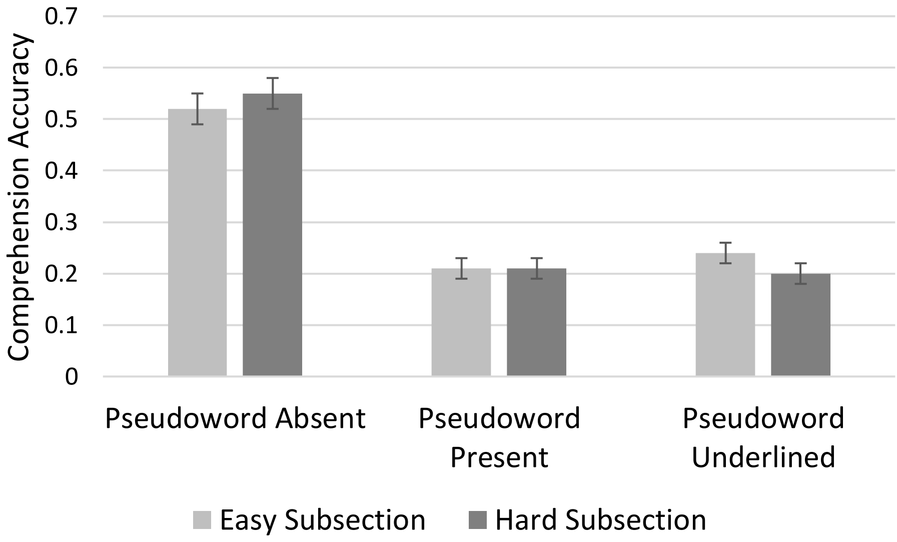


Figure 4 Mean Accuracy by Condition

Note. Error bars represent standard error of the mean.

Table 8 Logistic Mixed Effects Model Predicting Accuracy (Experiment 2)

Variable	$\hat{\beta}$	***SE***	***Wald-z***	***p***
Intercept	-0.87	0.14	-6.16	< .001
Effects of Manipulation				
Section Difficulty (Easy versus Hard)	-0.04	0.28	-0.15	.88
Pseudoword Presence	-1.51	0.10	-14.66	< .001 ***
Pseudoword Underlined	-0.07	0.13	-0.51	.60
Supplemental Difficulty x Pseudoword Presence	-0.28	0.20	-1.37	.17
Supplemental Difficulty x Pseudoword Underlined	0.29	0.26	1.12	.26
Effects of Individual Differences				
Reading Skill	-0.07	0.07	-1.02	.31
ASHS	-0.04	0.05	-0.80	.43
SPRS	0.10	0.05	2.39	.02 *
RBSC	0.01	0.05	0.11	.92

Note. ***p < .001, **p < .01, *p < .05, †p < .10

5.2.3 Calibration Model

As in Experiment 1, participants were broadly overconfident ($M_{\text{miscalibration}}$ = 23.52). Thus, all increases in Miscalibration were effects of increased overconfidence rather than reduced under confidence. Overconfidence was 0.75 (95% CI:[0.68, 0.82]) standardized units higher when a pseudoword was used versus a real word, $t(2065) = 20.95$, $p < .001$. Further, when the pseudoword was not underlined, overconfidence was 0.27 (95% CI:[0.20, 0.37]) standardized units higher than when it was underlined, $t(2067) = 6.78$, $p < .001$. However, our critical interest was in whether the Pseudoword x Section Difficulty interaction predicted readers' overconfidence in their comprehension. The interaction demonstrated that this discrepancy in overconfidence was indeed more prevalent for the Easy versus Hard sections, $t(2075) = -2.00$, $p = .05$. In fact, in pairwise comparisons, overconfidence was 0.24 (95% CI:[0.04, 0.43]) standardized units higher when a pseudoword was present in the Easy versus Hard condition, $t(27.30) = 2.31$, $p = .03$, but no reliable differences in overconfidence were found between the Easy and Hard conditions when the pseudoword was absent (real word condition), $p = .10$, or underlined, $p = .53$. Thus, readers' overconfidence only differed based on Section Difficulty (Easy vs Hard) in the Pseudoword Merely Present condition. See Table 9 for a list of effects and Figure 5 for a comparison between JOLs and Miscalibration across conditions.

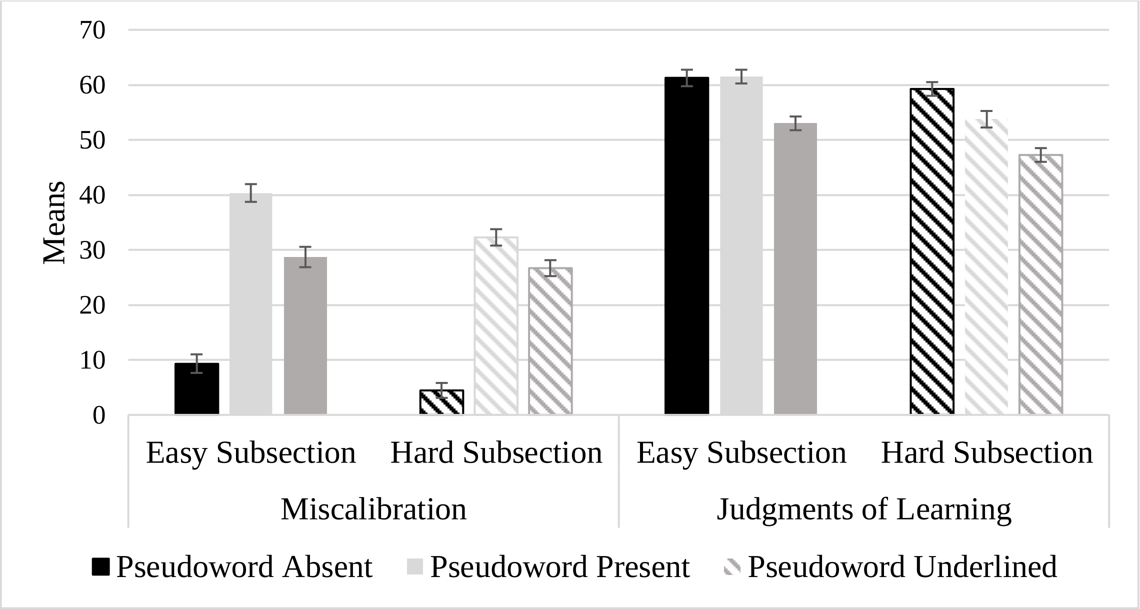


Figure 5 Mean Miscalibration and JOLs by Condition

Note. Means on the y-axis reflect mean Miscalibration (left) or Mean Judgments of Learning (right). From left to right, conditions within each subsection reflect Pseudoword Absent, Pseudoword Present, Pseudoword Underlined. JOLs are multiplied by 25 to place them on the same scale (0-100) as Miscalibration. JOLs are provided for reference. No statistical tests were conducted on raw JOLs. Error bars represent standard error of the mean.

Table 9 Linear Mixed Effects Model Predicting Overconfidence (Experiment 2)

Variable	$\hat{\beta}$	***SE***	***t***	***p***
Intercept	-0.001	0.06	-0.02	.99
Effects of Manipulation				
Section Difficulty (Easy versus Hard)	-0.016	0.09	-1.77	.10
Pseudoword Presence	0.76	0.04	20.95	< .001 ***
Pseudoword Underlined	0.29	0.04	6.78	< .001 ***
Section Difficulty x Pseudoword Presence	0.09	0.15	0.30	.76
Section Difficulty x Pseudoword Underlined	-0.17	0.08	-2.00	.05 *
Effects of Individual Differences				
Reading Skill	0.09	0.05	1.80	.07 †
ASHS	0.01	0.04	0.27	.79
SPRS	0.01	0.04	0.16	.88
RBSC	0.0002	0.04	0.004	> .99

Note. ****p* < .001, ***p* < .01, **p* < .05, †*p* < .10

5.2.4 End-of-Task Memory Model

The final model tests how likely readers were to recognize the pseudoword if they saw it later (see Table 10 for all effects). Greater likelihood of recognition may be a sign of deeper processing (e.g., Craik & Tulving, 1975). Participants were broadly able to remember the pseudowords later. The participants reported likelihood of seeing the pseudoword rose by 0.63 (95% CI:[0.50, 0.76]) standardized units if they had actually encountered the word within the texts, *t*(681.87) = 9.74, *p* < .001. Critically, this difference was magnified when the pseudoword was underlined versus when it was merely present, *t*(681.87)= 4.98, *p* < .001. Finally, although the two-way interaction between section difficulty and pseudoword condition (merely present vs. underlined) was not reliable, *p* = .12, in pairwise comparisons, participants reported likelihood of seeing the pseudoword decreased by -0.33 (95% CI:[-0.58, -0.09]) standardized units when it was merely present in the Easy sections versus Hard sections, *t*(14.40) = -2.67, *p* = .02. As in the miscalibration model, a similar effect was not reliable in the real word or the pseudoword underlined conditions, *p*s > .44. See Figure 6 for a visualization of effects.

Greater reading skill also predicted the likelihood of recognizing the word, *t*(133) = 2.25, *p* = .03, suggesting that readers with a higher reading skill may have been more likely to attend to the words across conditions.

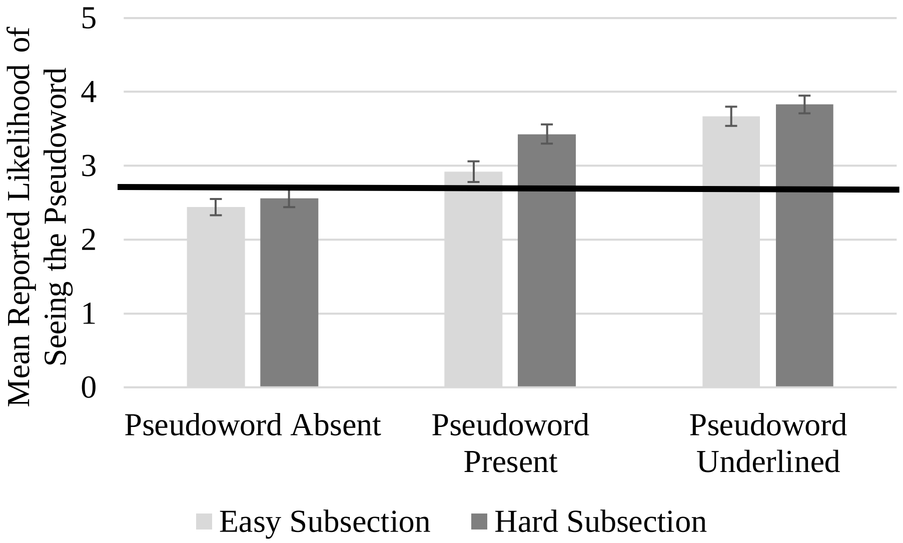


Figure 6 Mean Likelihood of Seeing Words

Note. Error bards represent standard error of the mean. Mean rating for the likelihood of seeing words which were never used in the passages are provided for reference (black bar, *M* = 2.22, *SE* = .03).

Table 10 Linear Mixed Effects Model Predicting Memory for Pseudowords (Experiment 2)

Variable	$\hat{\beta}$	*SE*	*t*	*p*
Intercept	-0.003	0.05	-0.06	.95
Main Effects of Manipulation				
Section Difficulty (Easy versus Hard)	-0.17	0.09	-1.88	.13
Pseudoword Presence	0.63	0.06	9.74	< .001 ***
Pseudoword Underlined	0.37	0.07	4.98	< .001 ***
Supplemental Difficulty x Pseudoword Presence	-0.14	0.13	-1.09	.28
Supplemental Difficulty x Pseudoword Underlined	0.23	0.15	1.57	.12
Main Effects of Individual Differences				
Reading Skill	0.12	0.05	2.25	.03 *
ASHS	-0.01	0.04	-0.25	.80
SPRS	0.04	0.04	0.88	.38
RBSC	-0.04	0.04	-0.82	.41

Note. ****p* < .001, ***p* < .01, **p* < .05, †*p* < .10

5.3 Discussion

Experiment 2 tested whether or not readers use an ease-of-processing heuristic to assess their comprehension. The results support this conclusion. When a pseudoword made comprehension impossible, overconfidence increased more for Easier versus Harder texts despite accuracy being similar across pseudoword conditions.

Further, Experiment 2 provides a reason *why* easier texts may sometimes lead to overconfidence. Readers were most overconfident when the pseudoword replaced the real word and no extra attention was drawn to the word. When attention was drawn to the complex word, readers were less overconfident. Further, when the pseudoword was underlined, readers were more likely to remember seeing the word later. Thus, as readers processed the easier text, they may have used an ease-of-processing heuristic to determine that the text only required superficial processing. This superficial processing may have increased the likelihood that readers would not attend to the pseudoword, which should have alerted the reader that parts of the Easy Section could not be understood.

6.0 Individual Differences

Multiple individual differences were tested for in each Experiment with mixed results. Metacognitive Miscalibration and Reading Skill consistently predicted accuracy in primary models and Academic Self-Handicapping (ASHS) consistently failed to predict outcomes. However, Self-Perceptions of Reading Skill (SPRS) and Reader-Based Standards of Coherence (RBSC) only *sometimes* predicted accuracy. Further, the significance of their relationships with accuracy varied even within the same experiment. For Experiment 1b, RBSC marginally predicted accuracy in the RPL model and SPRS marginally predicted accuracy in the full passage questions model.

Although the focus of the study is on the experimental manipulation, it is also worthwhile to consider the relationships among the individual differences. In this section, the correlations among these measures are described and explored.

6.1 Method

6.1.1 Participants

All participants (*N* = 383) who completed Session 1, including those who did not complete Session 2, were included.

6.2 Results

A correlation matrix was created to determine the relationship among individual differences. Measures included the ASHS, SPRS, RBSC, and both Nelson-Denny scores (vocabulary and comprehension) as well as measures of Miscalibration and Sensitivity for participants who completed Session 2 of the study. For participants in Experiments 1a and 1b, miscalibration and sensitivity were calculated based on their performance across pre and post supplemental question sets. For Experiment 2, they were calculated only using questions from Experiment 1 (i.e., not questions related to the pseudoword manipulations). Figure 7 provides details on the correlations and their changes across reading skill levels.

6.2.1 Reading & Vocabulary Skill

ASHS, SPRS, and RBSC all correlated with both measures of reading skill. More skilled readers were less likely to self-handicap (r = -.19 , p < .001), more likely to perceive themselves as skilled readers in a science domain (r = .22 , p < .001), and more likely to have higher standards of coherence (r = .22 , p < .001). A further marginal correlation existed between Reading Skill and Predictive Sensitivity, r = .11, p = .09. However, reading skill did not reliably correlate with any measure of predictive or postdictive miscalibration or postdictive sensitivity, *p*s > .26.

The correlation between SPRS and Reading Skill suggests that readers are broadly aware of their reading skill; i.e., stronger readers perceived themselves as stronger readers. However, this also raise two open questions. (a) Why was self-handicapping tendency negatively correlated with reading skill? It may be that less skilled readers were more likely to self-handicap to protect themselves from facing failure in comprehension, or it may be that participants who were prone to self-handicapping scored lower on measures of reading skill because they were self-handicapping rather than because their reading skills were poor. (b) Why are more skilled readers more likely to have higher standards of coherence? As with self-handicapping, the direction of this relationship is unclear. It may be that higher standards of coherence are part of what makes reader skilled or it may be that it just helped them perform better on the reading test. It may also be that skilled readers have higher standards of coherence because they know they are capable of it. This last explanation may coincide with the relationship between reading skill and self-handicapping. Less skilled readers have lower standards of coherence because they are self-handicapping; though this interpretation is purely speculative.

It is clear that the relationship among reading skill, self-handicapping, and standards of coherence has the potential to add insight into how reading skill is discussed and measured. However, as far as I am aware, self-handicapping has not been considered in a reading context before (see Schwinger et al., 2014 for a meta-analysis). In future research, greater use of self-handicapping metrics may help distinguish truly less skilled readers from readers who performed poorly.

6.2.2 Academic Self-Handicap Scale

ASHS outcomes only correlated with reading skill. There were no other significant correlations with the remaining individual differences, *p*s > .15.

6.2.3 Self-Perception of Reading Skill

In addition to the correlations with reading skill measures, readers who perceived their reading of science texts as stronger also tended to have higher standards of coherence ($r = .26$, $p < .001$) as well as higher predictive ($r = .20$, $p < .001$) and postdictive ($r = .14$, $p = .02$) miscalibration (overconfidence). Believing oneself to be a skilled reader may result in a greater likelihood of believing a text has been comprehended even if, in reality, comprehension is lacking. Although no causal conclusions can be drawn from this relationship, it is notable that when the correlation between RSPS and miscalibration was weaker among the highest skilled readers. In other words, the correlation between miscalibration an RSPS mostly existed for readers whose reading test scores did not align with their perceptions of their skills. Thus, it is possible that miscalibration in perception of reading skill led to overconfidence in actual comprehension of the text, though I emphasize that this is speculative.

6.2.4 Reader-Based Standards of Coherence

As might be expected given the correlation with SPRS, RBSC also correlated with predictive miscalibration (greater overconfidence, $r = .14$, $p = .02$) though, notably, not postdictive miscalibration, $p = .15$. Again, this correlation with overconfidence was primarily true for less skilled readers; indeed, the correlation among higher skilled readers is negligible, $r = .01$. In other words, less-skilled readers who expressed greater focus on ensuring their own comprehension also seemed to have an inflated sense of their comprehension immediately after reading a passage. However, this inflated sense of comprehension did not persist after the reader encountered the comprehension questions perhaps indicating that they were able to adjust their JOLs and recognize their specific information was missing from their text representation. As a result, standards of coherence appeared to have little relationship with the readers' ability to assess whether or not they answered questions correctly.

6.2.5 Predictive and Postdictive Miscalibration

As noted elsewhere, my two measures of metacognitive miscalibration are highly correlated ($r = .73$, $p < .001$) , suggesting that readers are consistent in their predictions about their ability to answer questions following reading a passage and their confidence in their accuracy once they actually answer the question. Additional correlations also existed between Postdictive Sensitivity and Predictive Miscalibration, $r = -.14$, $p = .02$: As readers grew more overconfident in their initial beliefs about their comprehension of a text, they also grew less able to differentiate between questions they answered correctly versus incorrectly. To some degree, this may reflect measurement limitations insofar as readers who consistently give high confidence ratings had fewer points on the JOL scale to discriminate among well-learned versus poorly-learned materials.

6.2.6 Predictive and Postdictive Sensitivity

Readers' ability to predict their performance on one text relative to another before encountering questions about the passages did not reliably correlate with their ability to assess how likely they were to answer a given question correctly relative to a separate question. Indeed, the correlation was close to zero, $r = .01$, $p = .86$. Readers were generally better at postdictive sensitivity ($M = .46$ gamma correlation, $SE = .01$) than predictive sensitivity ($M = .18$ gamma correlation, $SE = .03$). Intuitively, it makes sense that recognizing relative difficulty in answering questions would be easier than assessing the likelihood of being able to answer questions before seeing the questions, but it is not clear why there is not at least some correlation. If some readers can *predict* their relative ability to answer questions about a set of passages, why can't they also postdict their relative probability having answered a set of questions correctly? This finding potentially supports theories that predictive and postdictive judgments involve separate processes (Benjamin, 2003; Kelemen et al., 2000; Maki et al., 2005; Schraw, 2009). If readers use cues to establish their JOLs (e.g., Korait, 1997) it may be that quite different are available to people for predictions vs. post-dictions. However, this is complicated by the fact that there was a significant correlation (though of only moderate magnitude) between predictive and postdictive JOLs across all experiments (r = .33, t(1744) = 14.67 for Experiment 1a; r = .34, t(1843) = 15.73 for Experiment 1b; r = .32, t(5203) = 24.00 for Experiment 2). Thus, it may be that there are some commonalities between pre and postdictive JOLs (e.g., tendency to be over -or under-confident), but the cues that readers use to assess relative accuracy vary across pre- and post-dictions.

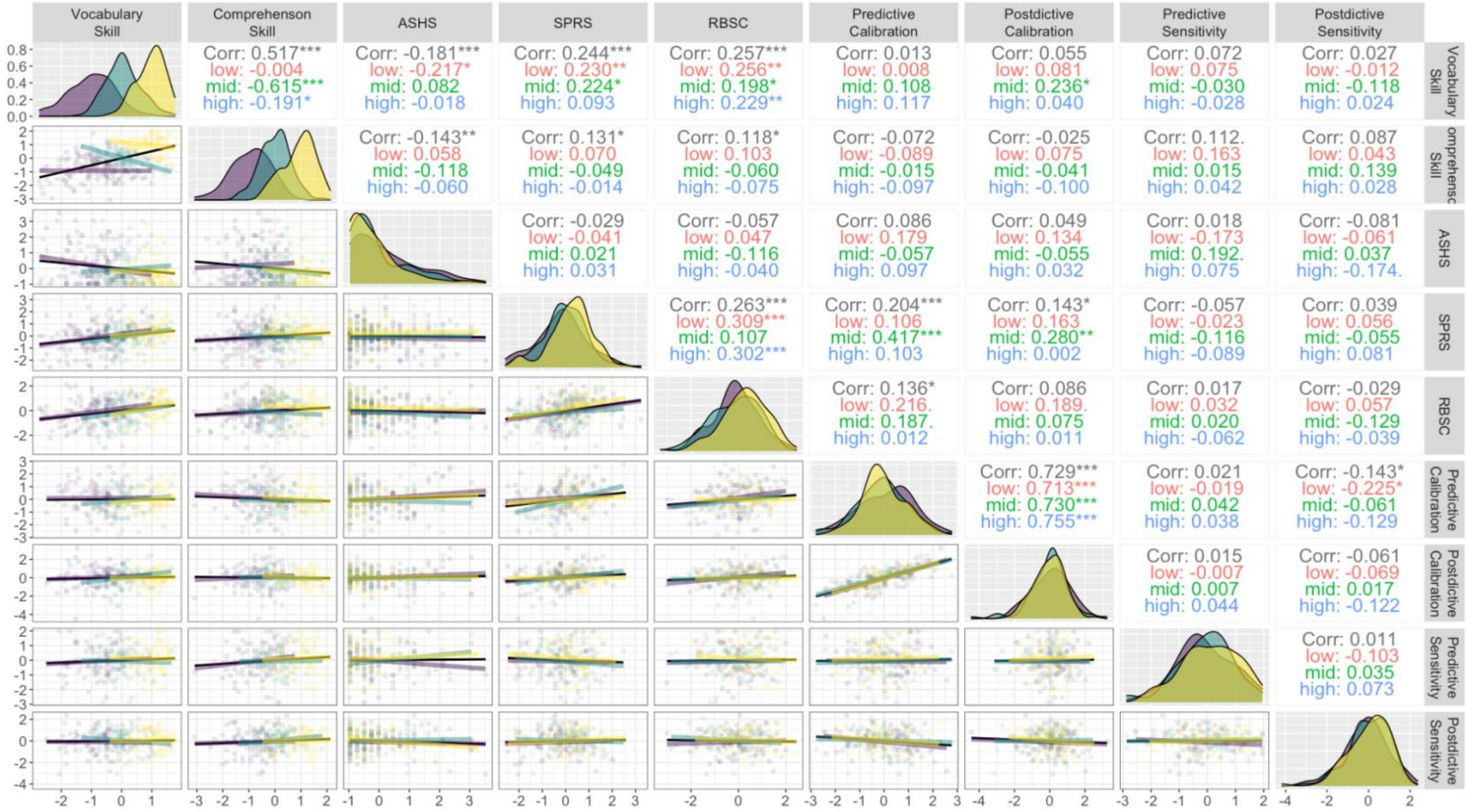


Figure 7 Correlations Among Individual Differences

Note. Three quantiles were created to separate reading skill (the combined vocabulary and reading comprehension score on the Nelson-Denny) into distinct levels (low = purple, moderate = green, and high = yellow). Black lines represent overall trend. All values are standardized (y-axis = columns, x-axis = rows). Diagonal contains the distribution of scores for each reading skill level. ***p < .001, **p < .01, *p < .05, . p < .10

7.0 General Discussion

The goal of the above set of experiments was to determine which factors may influence comprehension monitoring and how those factors in turn influence metacognitive control. Three core findings stood out: (a) In Experiment 1a, I validated the Region of Proximal Learning (RPL) model within a reading framework. When readers were assigned to read additional information about moderately difficult sections of a text, they made more gains in comprehension than when they were assigned to allocate additional time to easy or hard sections of the text. (b) Experiment 1b demonstrated the RPL held even when readers chose which sections of a text they would read in greater detail. However, a majority of readers intentionally chose sections with harder material despite being aware (i.e., not overconfident) that they were still lacking comprehension of the moderately difficult material. Thus, Experiment 1b implied that readers' selection of suboptimal reading material is an error in metacognitive control rather than an error in comprehension monitoring. (c) Participants in Experiment 1 were also broadly overconfident in their comprehension. In Experiment 2, I found that one source of this overconfidence in comprehension may be the heuristics readers use to determine the required depth of processing. Readers engaged in shallower processing (i.e., had greater overconfidence in their comprehension and were less likely to recognize the pseudoword later) when texts *felt* easier to process and thus missed the presence of pseudowords which obscured comprehension. Overconfidence was higher for easier versus harder texts but drawing attention to (underlining) the pseudoword reduced overconfidence.

7.1 Region of Proximal Learning

Metcalfe (2002) introduced the Region of Proximal Learning framework as an alternative to studies suggesting that learners will make the greatest gains when they study material that is farthest from their desired state of knowledge (i.e., the hardest available material). Multiple studies have since found that if given the choice between studying almost mastered material and completely unmastered material, learners will make more gains if they study the almost mastered material (Dunlosky & Ariel, 2011; Kornell & Metcalfe, 2006; Metcalfe, 2011; Metcalfe & Kornell, 2003, 2005; Metcalfe et al., 2020). However, the majority of the data supporting the RPL comes from list-learning tasks. Prior to the introduction of RPL model, Son and Metcalfe (2000) found that readers' allocation of reading time varied depending on the constraints of the tasks. They found that readers spent more time reading easier texts when short time limits were placed on study and more time reading harder texts when longer time limits were used. However, only two text levels (easy and hard) were used, and comprehension of the texts was not assessed. Although this study introduced an early challenge to prior findings that readers benefitted from engaging with the most difficult material available, it was designed to support the full RPL model (i.e., measuring accuracy and selection of material in a context which provided material at three distinct difficulty levels). Thus, Experiment 1 is the first time the RPL framework has been tested in full in a reading comprehension context, and its applicability was confirmed.

Despite similarities in learning and comprehension gains across paradigms (list learning and reading), several key differences also showed up. Although in list learning tasks, learners choose to engage with moderately difficult material, readers in the present study chose, on average, the hardest available materials. Further, prior work has found that comprehension monitoring has an indirect effect on comprehension by first predicting the allocation of reading time to not-yet-mastered material (Thiede et al., 2003). However, Thiede et al. (2003) only considered two levels: mastered and not-yet-mastered. It was not clear then if given more levels of difficulty, comprehension monitoring would still have the similar effects. Indeed, in the present study, choice of material did not mediate the effect of comprehension monitoring on accuracy. Better comprehension monitoring did not reliably affect what additional texts readers chose. Readers who were sensitive to the relative difficulty of the different sections of text were just as likely to select the most difficult texts as readers who were less able to discriminate the relative difficulty of the material.

The lack of relationship between comprehension monitoring and metacognitive control, though, does not imply that comprehension monitoring is irrelevant for metacognitive control. Indeed, at the end of the study, readers reported intentionally selecting the hardest material precisely because they were able to recognize its difficulty. This finding leaves an open question: Why did readers select the hardest material? It was possible that readers believed that the hardest material was in their RPL. That is, readers may have believed that their comprehension on the moderate material was at ceiling and could not be improved. Thus, while they recognized that the moderate material was easier than the hard material, they did not recognized the moderately difficult material as being in their RPL. Despite readers' general overconfidence in their comprehension, this conclusion seems unlikely. Readers were less overconfident in their judgments of learning for the moderately difficult versus hard texts. Given that mean accuracy on the moderately difficult materials was decidedly below ceiling (*M* = .49, *SD* = .02), this indicates that readers knew they had only partial comprehension of the moderate materials.

Another explanation for the finding that readers tend to allocate extra time to harder sections is that readers erroneously thought selecting the harder material was better for improving their performance on the reading comprehension measures than selecting more proximal material. Thus, comprehension monitoring may be important to metacognitive control as it can help readers accurately differentiate the difficulty of materials, but in this study, comprehension monitoring errors were not the issue. Rather the error was in metacognitive control. 74% of participants reported using information on difficulty as opposed to personal interest or some other criteria when selecting which section to read about in additional detail. Thus, readers in this study were on the right track in that they identified the relative difficulty of the sections of text and made a choice in what material to read based on their assessment. However, their metacognitive beliefs about learning may have steered them away from the optimal choice for learning.

One potential caveat to these results was that readers were required to make judgments of learning (JOL) both after reading each section of a text and after answering a question. This introduced an atypical procedure to the reading process. Readers do not typically stop to make a JOL as they read. Indeed, some evidence suggests that some readers may not monitor their comprehension at all (Oakhill et al., 2005; Tighe et al., 2021). Thus, the process of assessing their comprehension may have interfered with decisions about which section to read in greater detail. Experiment 1c ruled this out as a possibility. When JOLs were removed from the experimental procedure, readers were still more likely to allocate additional time to the hardest sections and still more likely to perform better if they did allocate time to the moderately difficult section. Of course, Experiment 1c should not be taken to imply that readers typically overtly monitor their comprehension while reading, only that asking readers to make JOLs did not appear to influence their decisions or criteria.

7.2 Illusions-of-Knowing

Readers displayed overconfidence in their comprehension across all Experiments. Their tendency towards overconfidence is consistent with prior findings (Garner, 1980; Griffin et al., 2008; Jaeger & Wiley, 2015; Kwon & Linderholm, 2014). Experiment 2 illustrated that the tendency towards overconfidence may be rooted in gist processing and low attention to sections of text that would signal to the reader that lower confidence is warranted. A critical question for Experiment 2 was what prompts readers to engage in greater gist processing. Although some researchers have suggested that illusions-of-knowing arise from processing sentences with complex surface features (e.g., syntax; Ferreira et al., 2002; van Dijk & Kintsch, 1983), others have suggested the gist processing is more likely to occur when texts are *easier* to read (Mata, 2020). The current study supports the latter prediction. Readers were less aware of gaps in their representations of simpler texts than those of harder texts. That is, readers' use of the ease-of-processing heuristic may lead to greater gist processing and subsequently to a failure to attend to complex regions of a text (in this case of the studies presented here, the pseudoword itself). Experiment 2 provided multiple lines of support for this rationale. First, when readers encountered a pseudoword in the text, it should have cued low confidence in their comprehension of at least that sentence. Thus, when answering questions that required knowing the meaning of the pseudoword, confidence should have been low. Instead, readers expressed confidence in the accuracy of statements that were not reflected in the text. Critically, this sense of overconfidence was higher for the easy versus the harder text segment. When readers processed a sentence containing a pseudoword in an easier section of text, they were more likely to believe they comprehended the sentence than when they processed a sentence with a pseudoword in a hard section of text. This suggests that readers' sense of comprehension is inflated, not necessarily for difficult material per se, but for difficult material embedded within easy material.

Second, at the end of the experiment, participants who were exposed to a pseudoword were more likely to recognize it if it had been in the hard versus easy section, suggesting that the words in the hard section had been attended more closely. Finally, if the pseudowords were underlined, readers' overconfidence in their comprehension dropped. When attention was drawn to the complex section of an otherwise easy text, readers lowered their confidence in their comprehension of the text. Thus, readers' miscalibration of their comprehension was off because they did not attend to the complex region of text. Once this region was attended, their miscalibration became more accurate.

7.3 Broader Implications

Readers' use of an ease-of-processing heuristic to determine when gist processing can be engaged may explain a seemingly paradoxical result from prior studies: The comprehension of less-skilled readers who are *more* knowledgeable about a topic declines when the relationships among the ideas in a text are made more explicit via increased cohesion (O'Reilly & McNamara, 2007; Ozuru et al., 2008). Although the work by McNamara and colleagues (2007, 2008) did not assess comprehension monitoring, a separate study found that comprehension monitoring for more-cohesive texts was less accurate than that of less-cohesive texts (Rawson & Dunlosky, 2002). When a reader is more knowledgeable about a text, it may make the text *feel* easier to process and thus lead to greater gist processing and under specification.

Critically, in the prior studies on cohesion discussed above, the effect of cohesion on comprehension was smaller for skilled readers. Skilled readers with higher levels of prior knowledge on a topic still performed well on a comprehension assessment. McNamara and colleagues (2007, 2008) explained this by proposing that skilled readers continue to engage in deeper processing even when prior knowledge is high and the text feels easier to process. The results of the present study challenge this conclusion. Reading skill positively, though marginally, predicted overconfidence in Experiment 2, where comprehension was improbable for readers of all skill levels. Thus, the explanation for why skilled readers with prior knowledge were not as negatively affected by cohesive texts as less-skilled readers may be more complex than skilled readers avoiding shallow processing.

Finally, the tendency for texts that feel easy to read to promote illusions-of-knowing when the texts are in fact less easy than they feel may be an important consideration when writing for public consumption. It may be necessary for writers to consider if they are making complex concepts appear simpler than they actually are. It is possible that simply underlining the complex concepts could reduce overconfidence, but additional work will be necessary to verify this in a context in which the complexity of the content is manipulated rather that the reader's ability to recognize the vocabulary word.

7.4 Individual Differences

All of the effects reported above existed over and above individual differences related to reading skill, standards of coherence, perceptions of reading ability, and self-handicapping tendencies. Despite the study containing a wide range of reading skills, there were few significant relations between the individual differences and comprehension or comprehension monitoring. Although, as mentioned above, reading skill did marginally predict miscalibration in Experiment 2, it did not predict miscalibration or sensitivity more broadly (i.e., when comprehension was not rendered improbable).

Self-perception of reading skill, on the other hand, did show a positive correlation with overconfidence. Readers who perceived themselves to be skilled readers were also more overconfident than readers who perceived themselves as less skilled, although it is not possible to determine causal directions from these data alone. It would be worthwhile to consider if overconfidence in reading ability also influences overconfidence in comprehension. Do readers rely on their beliefs about their reading skill when evaluating their comprehension? Or do readers who consistently feel confident in their comprehension come to believe they are skilled readers? Or is there an alternative mechanism responsible for both generating readers' overconfidence in their reading skill and their overconfidence in their comprehension. The present study was not designed to answer these questions, but the relationships among the independent variables highlight an important area for future research.

Greater overconfidence was also predicted by higher standards of coherence. This finding is somewhat at odds with the conclusion that illusions-of-knowing are associated with shallower processing. However, this correlation was only with *predictive* miscalibration, not postdictive. High predictive miscalibration does not necessarily represent an illusion-of-knowing. Predictive miscalibration, as operationalized in this study, involved an assessment of future ability to answer questions about a topic. If a reader provides high predictive JOLs and later scores poorly on a comprehension assessment, it may be that they had an illusion-of-knowing or it may be that they were tested on the material they believed they understood. That is, they may have underestimated the level of detail they would need to comprehend to do well on the test. Instead, the illusion-of-knowing in this study is more reliably related to postdictive miscalibration in which the reader selects an incorrect answer and expresses confidence in its accuracy. In the case of the postdictive JOL, the reader is expressing confidence that they comprehended a specific fact from the text. If they are incorrect, they have an illusion-of-knowing that fact.

Predictive miscalibration may be more related to standards of coherence because readers with higher standards of coherence may be more likely to believe that they understood all they needed from a text because they attended to it closely and tried to comprehend it. They may in turn correct that belief when they reach an actual question and realize that despite having attended closely to the text, they are not able to answer the question. Supporting this speculation is the correlation between perceptions of reading skill and standards of coherence. Readers who tend to try harder to comprehend a text also perceive themselves as more skilled readers and thus, at least initially, expect that they have comprehend a text. However, it is only (potentially false) beliefs about reading skill which are related to the stronger measure of illusions-of-knowing (postdictive miscalibration).

The complex correlations among measures of reading skill, perceptions of reading skill, standards of coherence, and miscalibration suggest that greater attention should be given to the relationship among these variables as they may yield new insights regarding the individual factors that can promote illusions-of-knowing and how those illusions are constructed. Further, this study found support for differences between pre- and post-dictions withing the measure of sensitivity but not the measure of miscalibration. Given that some studies only use one of these measures (e.g., Mata, 2020; Wiley et al., 2005), further research into how each measure differs may help provide precision to interpretations of results.

7.5 Limitations

One potential limitation of the conclusions is in their generality across genres and topic domains. In the present work, I used only a limited set of texts so that I could carefully norm and control them. Passages were of similar length and were normed to ensure perceptions of text difficulty from the target population matched an objective difficulty measures (grade level in Lexiles). Further, all three texts were from the science domain, which suggest it is reasonable to expect the results to replicate across a wide array of science materials. However, it is not immediately clear if the results would replicate within other reading domains (e.g., narrative or humanities texts). Self-perception of reading skill was positively correlated with miscalibration, but perceptions of reading ability for science texts and actual science reading ability tend to be lower than for other text (Berman & Nir-Sagiv, 2007; Singer et al., 1997; Wolfe, 2005). Thus, in a genre where beliefs about reading skill are higher, there may be more differentiation in miscalibration and sensitivity. As a result, different effects may emerge. Replications of these findings in other domains will be an important step in understanding their wider applicability.

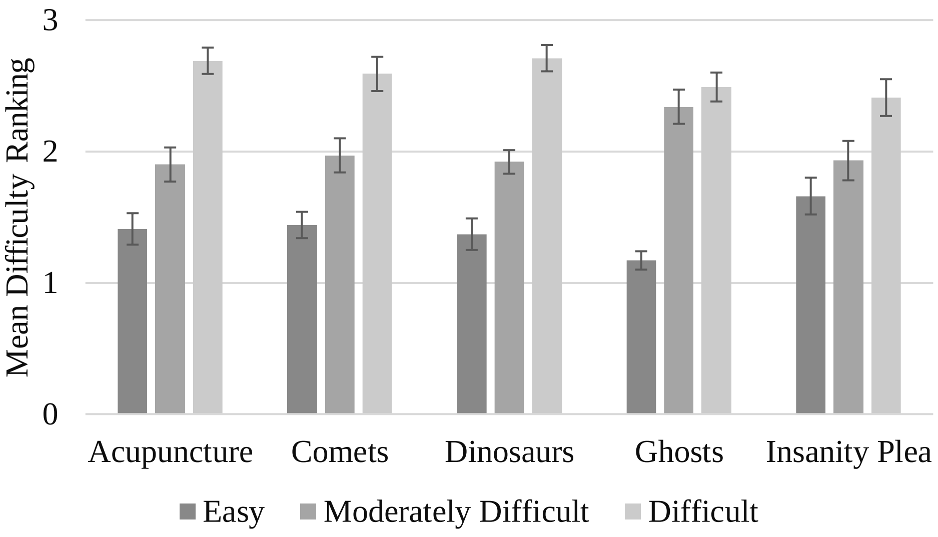
7.6 Conclusion

Although accurate reading comprehension is critical to learning and making informed decisions, all readers will experience moments of low comprehension. In these moments, it is critical that readers both recognize their low comprehension, thus avoiding illusions-of-knowing, and take appropriate steps to improve their comprehension. However, readers appear unaware of which strategies will improve their comprehension: Most appear to subscribe to a discrepancy-reduction belief rather than targeting their region of proximal learning, despite evidence from these very materials favoring proximal material as ideal for learning. Further, readers tend to be overconfident in their comprehension, which might decrease the likelihood that they recognize their comprehension needs revision. This overconfidence may stem from underspecified representations of the text, particularly when texts feel easy to process. Thus, although difficult texts may provoke greater comprehension failure, simpler texts embedded with complex information may actually lead to greater illusions-of-knowing.

Appendix A Norming Studies

Appendix A.1 Norming Study 1 (Experiment 1)

Five texts were developed to have distinct sections of varying difficulty. The first norming study was conducted to validate the difficulty levels. Participants (*N* = 29 undergraduates recruited for partial credit for a course requirement) read five passages and ranked them based on their perceived difficulty. Participants also answered a subset of 45 questions (9 per passage) from a possible 105 comprehension questions about the passages. In a regression model, participants' rankings were indeed predicted by the intended difficulty of the sections, *p*s < .001 (see Appendix A Figure 1). The three passages with the greatest distinction between difficulty levels within their sections were selected for use in the study. For those passages, questions which were appropriate to the perceived difficult level of the section were selected to be used in Norming Study 2. If a section was perceived by participant as easy, but accuracy for a question about content in that section was at chance, the question was discarded. This was to ensure that the perceived difficulty matched the actual difficulty.

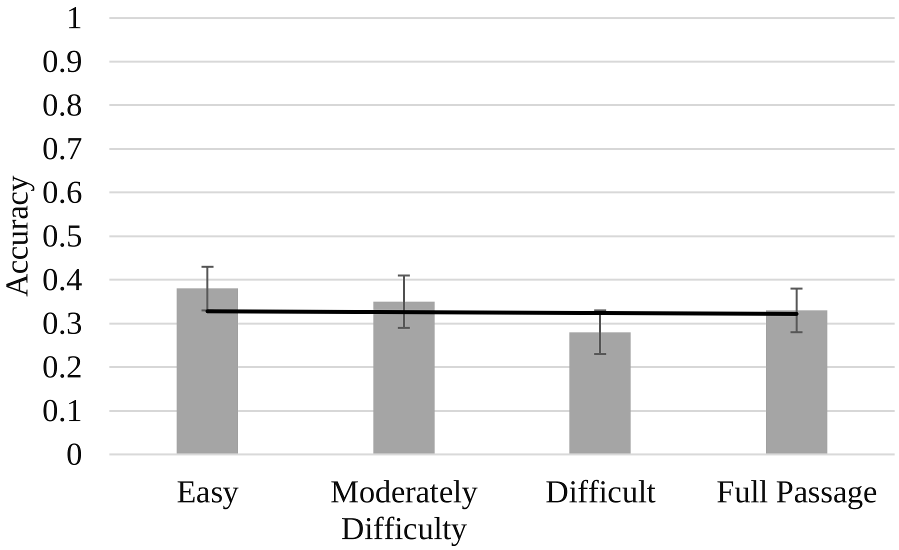


Appendix Figure 1 Mean Participant Rankings from Norming Study 1

Note. Mean rankings for the original five passages as matched with experimenter intended difficulty. Errors bars represent standard error from the mean.

Appendix A.2 Norming Study 2 (Experiment 1)

The second norming study was conducted to determine if questions could be answered without reading the passages. Amazon Mechanical Turk (MTurk) workers (*N* = 20) were paid $5 for the ~10 min task. They were asked to answer questions selected from Norming Study 1 without reading the text. Questions started as 5-alternative forced-choice items with the intention of removing lures that were never selected (i.e., lures participants deemed obviously incorrect). Questions which could be answered at greater than chance levels were revised and resubmitted to a new set of MTurk workers until all moderate and difficult questions had near chance accuracy rating across participants (see Appendix A Figure 2).

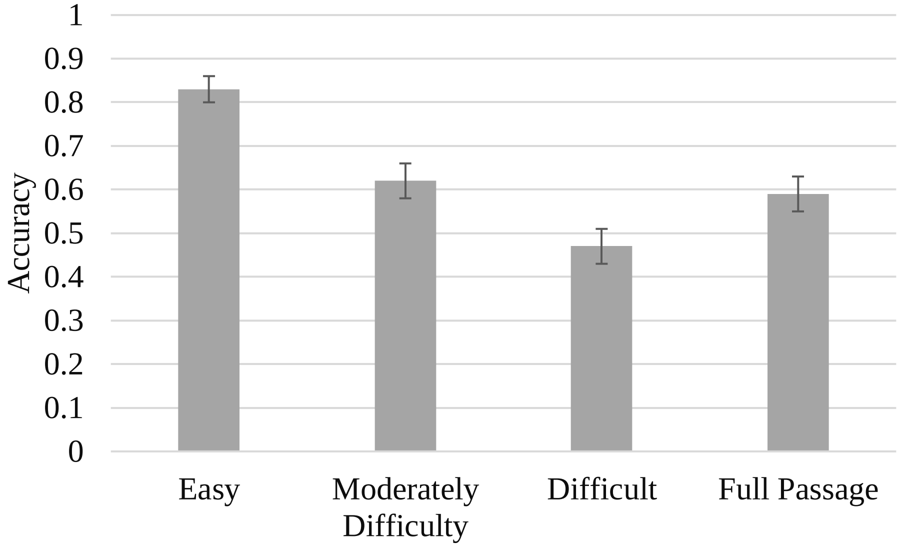


Appendix Figure 2 Question Accuracy Norming Study 2

Note. Accuracy on questions when participants did not have access to the passage. The black bar represents chance accuracy (33%). Errors bars represent std. error of the mean.

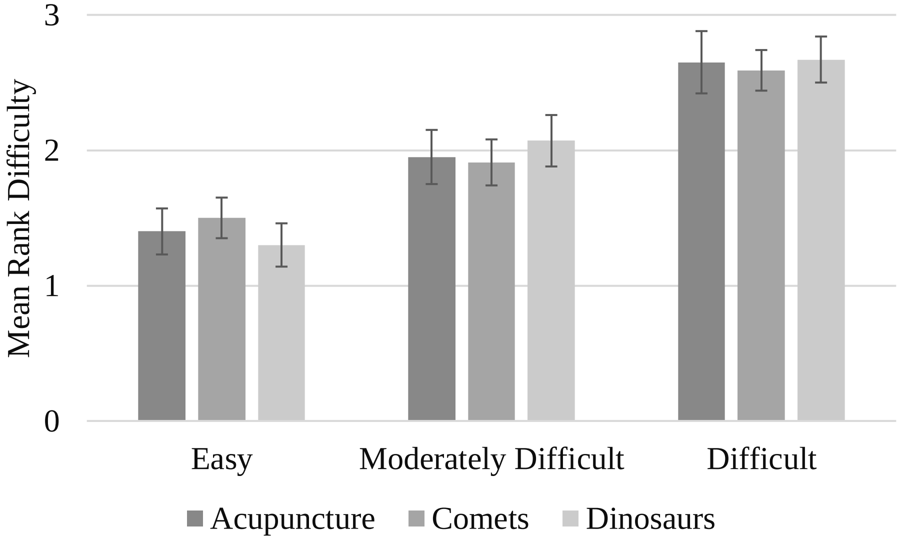
Appendix A.3 Norming Study 3 (Experiment 1)

The third norming study was conducted to determine if the supplemental texts were also of the same relative perceived difficulty. The same procedure as Norming Study 1 was followed, but with the supplemental texts instead of the main texts. Participants (*N* = 14 undergraduates recruited for credit) read three passages (containing the 3 supplemental sections) and ranked them based on their perceived difficulty (see Appendix A Figures 3 and 4). These participants also answered the questions chosen following Norming Study 2. Perceived difficulty and participant accuracy on the questions matched the intended difficulty of the experimenter with participant rank-order and accuracy showing clear distinctions among the sections, *p*s < .001.



Appendix Figure 3 Question Accuracy Norming Study 3

Note. Accuracy when participants had access only to the supplemental reading materials. Errors bars represent standard error from the mean.



Appendix Figure 4 Mean Supplemental Text Rankings from Norming Study 3

Note. Mean participant rankings for supplemental material as matched with experimenter intended difficulty. Errors bars represent standard error from the mean.

Appendix A.4 Norming Study 4 (Experiment 2)

Norming for Experiment 2 proceeded in three phases. The overarching goal of the norming process was to generate (a) six pseudowords which participants would believe were obscure real words, (b) six sentences to be embedded in the Easy and Hard sections of the three passages, (c) three questions related to each new sentence (18 total) for which (d) the ability to answer hinged on comprehension of a single word in the new sentence which could be replaced with a pseudoword, completely obscure its meaning.

Appendix A.4.1 Phase 1

In the first phase, 45 pseudowords and 45 obscure (e.g., *velleity*) or lesser known (e.g., *obsequious*) real words were generated by the experimenter and three research assistants. The list of words was shown in randomized order to people familiar with the experimenter (N = 12). They were then asked to indicate the likelihood that each word represented a real English word on a 5-point Likert scale (*Extremely unlikely* [1] – *Extremely likely* [5]). 16 pseudowords had a mean rating of 3 (*Neither likely nor unlikely*) or greater, suggesting the participants believed they could be real words. Indeed, a mixed effects model confirmed there was no reliable difference between the ratings of these 16 pseudowords and the real words, $t(50.09) = 0.26$, $p = .80$, indicating they were highly plausible pseudowords.

Appendix A.4.2 Phase 2

Passages in Experiment 1 were revised to add a new sentence whose meaning could not be fully understood when a real word(s) was substituted for one of the pseudowords. In some sections, additional modifications were made to the surrounding sentences to make this possible. Out of the 16 pseudowords identified in Phase I, 6 were chosen to be used in these new sentences based on how their tense fit with the generated sentences. See Appendix A Table 1 for selected pseudowords and the corresponding sentences.

Three multiple choice questions were then created, each with three possible responses (1 correct and 2 lures). All questions were constructed so that they could only be answered when the sentence with the real word was present. To confirm this was the case, new participants (*N* = 25, recruited from the University of Pittsburgh in exchange for partial fulfillment of a course requirement) were shown the full passages from Experiment 1 with the new sentences added and answered the three new comprehension questions along with all of the questions from Experiment 1. A mixed-effects model demonstrated that, for the three new questions, the odds of a correct response were significantly lower when the pseudoword versus the real word was present, z = -2.16, p = .0, (M_{accuracy} = 0.38, SE = .04 with the real word vs. M = 0.33, SE = .04 with the pseudoword). Indeed, mean accuracy for the pseudoword condition was at chance (chance = $0.\overline{33}$). Further, the presence of the pseudoword in the new sentence did not reliably affect accuracy for the original set of questions asked in Experiment 1, which did not concern the pseudoword, z = .01, p > .99. This suggests that addition of the pseudoword did not interfere with the participants' ability to construct a representation for the broader section.

To further confirm that participants who saw the pseudowords believed they were real words, the same participants were shown 10 of the obscure, real words from Phase 1 and 10 pseudowords: the six used in the passage and four of the other the top 16 words from Phase 1. Participants who saw a pseudoword in the passages were more likely to rate it as a real word than participants who did not see it ($M_{\text{rating}} = 3.36$ out of 5, $SE = .15$ vs $M = 2.99$, $SE = .15$), $t(147.21) = 1.96$, $p = .05$. Further, if a participant saw the pseudoword in the passage, they were more likely to believe it was a real word than *actual* real words not shown in the passage, ($M_{\text{rating}} = 2.62$, SE=.08 for the obscure, real words), $t(14.60) = 2.40$, $p = .03$. Thus, the pseudowords were deemed to be successfully perceived as real words. Participants may have more prone to believe a pseudoword was real because they saw it in the context of a scientific passage as part of a university conducted experiment (e.g., *the False Fame Effect,* Jacoby et al., 1989; Jacoby et al., 1989). Regardless of the reason, a core criteria of the pseudoword was that it not be immediately recognized by participants as a fake word introduced for an experimental manipulation. This check verifies that the criteria was met.

Finally, these participants were also shown the new sentences containing the pseudowords and asked to define the pseudoword. Ability to define the pseudoword, even with the sentence present to provide context, was low. Two raters scored all definitions to see if they were synonymous with meaning of the real word, *Cohen's kappa* = .96. For the single item of disagreement, the Experimenter's value was used. Only eight (9%, SE = 3%) pseudowords were given a definition close to the real word that it replaced. Further, seven of these eight successful definitions came from participants who had previously seen the relevant sentence in the real-word condition and thus could rely on their memory for the real word in providing the definition. Only one participant who had originally see the pseudoword in the passage was able to determine one definition in context.

Although Phase 2 made it clear that participants believed the pseudowords were real words that obscured the meaning of the sentence, one remaining issue was that the comprehension questions were still quite difficult overall. Although participants' accuracy was higher with the real word than with the pseudoword (and by extension chance), it was still low, creating a concern that some questions were unlikely to be answered correctly even with the real word in the passage. Phase 3 of norming addressed this remaining issue.

Appendix A.4.3 Phase 3

In Phase 3, revisions were made to the questions to make the correct answers more explicit (including using the real word that would be substituted for the pseudoword in the answer choice). The sentences participants saw and the ones ultimately used in Experiment 2 are displayed in Appendix A Table 1.

Then, in a second round of norming, participants (N = 31 recruited in the same manner as in Phase 2) read only the new sentences which contained either the real word or the pseudo words and answered the relevant questions. For participants who saw the real word, accuracy was again significantly higher than for participants who saw the pseudoword, z = -6.27, p < .001. Further, accuracy was now well above chance when the real word was present in the sentence (M = 0.62, SE = .03). However, accuracy was near chance (M = 0.36, SE = .03) when the pseudoword was present. The result, then, was a set of materials that could be readily comprehended when the real word was present but not when that real word was replaced with a pseudoword.

Appendix B outlines the changes to the passages and the questions used.

Appendix Table 1 Pseudowords and Sentences

Pseudoword	Passage	Placement	Sentence
salants	Needles and Nerves	Easy Section	Further, considerations for **individual differences** have to be made when selecting the correct point used to treat a specific condition.
cernous	Needles and Nerves	Hard Section	During acupuncture endorphins increase **relaxation** which primes the body for healing.
cepulized	A Comment on Comets	Easy Section	The solar system, along with the Kuiper Belt, is constantly **surrounded** by the rocks in the Oort Cloud.
ocelants	A Comment on Comets	Hard Section	The radiation pressure stems from when the dust particles reflect photons of light, which have **momentum** but no weight.
trooks	Were Dinosaurs Dumb?	Easy Section	Indeed, their extensive **ear canals** likely coordinated rapid eye movements and quick reflexes.
bassals	Were Dinosaurs Dumb?	Hard Section	Further, as part of a cooling process, dinosaurs' brains had **holes** that don't exist in most other animals, including modern reptiles.

Note. Bolded words in the sentences were replaced with pseudowords.

Appendix B Stimuli (Passages and Questions)

Appendix B.1 Were Dinosaurs Dumb?

Dinosaur Brains (Difficult)

Beliefs about dinosaur intelligence have changed over the years primarily because of a change in understanding of stupidity and its correlation with size. Brain mass (weight) relative to body size, known as the encephalization quotient (EQ), correlates with intelligence. A higher ratio generally means greater intelligence. To put this in perspective, the *Brontosaurus* is 12x larger than a human, but its skull is 4x larger. [Further, as part of a cooling process, dinosaurs' brains had holes/bassals that don't exist in most other animals, including modern reptiles.][19]

Part of the discrepancy may be that there wasn't evolutionary growth in brain size among dinosaurs over time, unlike in mammalian and bird groups. Instead, dinosaurs evolved a "second brain" (a bundle of neurons in their tails) to help speed up processing. Also, only the correlation of brain size with body size among similar animals (all reptiles, all mammals, for example) is reliable because brain size increases less than body size and at different rates among different types of animals, so when calculating EQ, an adjustment for brain to body growth rates has to be made based on animal type. Further, we must conclude that large animals require relatively less brain to

[19] Sentences contained in [] were only present in Experiment 2.

do as well as smaller animals. The current view doesn’t claim that dinosaurs are highly intelligent, only that they had the right brains for their bodies.

Dinosaur Behavior (Moderate)

Beyond brain size, behavior is another way of determining dinosaur intelligence. If dinosaurs were intellectually capable, we should find evidence of behavior that demands social and mental coordination. Indeed, we do. Multiple trackways have been uncovered, with evidence for more than twenty animals and multiple species traveling together. At the Davenport Ranch sauropod trackway, small footprints lie in the center and larger ones at the edge. As a further indication of herd life, upwards of thirty juveniles have been found next to a single adult dinosaur as well. That is among plant-eaters, but similar signs are present among meat-eaters, too. A group of Velociraptors were found in a quicksand pit next to an Iguanodon.

Further, few reptiles today are involved in the lives of their young, although the crocodile and pythons are notable exceptions. A finding of dinosaur bones next to unhatched eggs was once believed to be evidence of one dinosaur eating another's eggs. Multiple similar findings have changed that belief. Care for young may have been particularly important among dinosaurs. The Tyrannosaurus Rex hatched from an egg the size of a pigeon. It would take some time to grow from that small size to the 40-foot-long and eight-ton beast of the adult.

Changes in Beliefs about Dinosaurs (Easy)

The discovery of dinosaurs in the 1800s provided, or so it appeared, proof that big bodies meant less smarts. With their pea brains and giant bodies, dinosaurs became a symbol of stupidity. Their extinction seemed only to confirm their bad design. For some time after their discovery, dinosaurs were thought of as slow and clumsy. For example, a typical image from the past involves a giant dinosaur called a *Brontosaurus* wading in a murky pond because he cannot hold up his own weight on land.

Modern scientists do not agree with this image. Modern imaginings of dinosaur bodies show them as strong, fast, and agile. Today, most paleontologists view dinosaurs as active and capable animals. [Indeed, their extensive ear canals/trooks likely coordinated rapid eye movements and quick reflexes.] Now, the *Brontosaurus* is imagined running on land. They even believe pairs of males could wrap their long necks around each other in combat (much like the neck wrestling of giraffes).

But the best example of dinosaur ability may well be the fact most often used against them—their extinction. What's remarkable about dinosaurs is not that they became extinct, but that they lasted on Earth for so long. Dinosaurs dominated the Earth for 160 million years before they become extinct. Meanwhile, humans have only been around ~300,000 years.

Appendix B.1.1 Changes in Beliefs about Dinosaurs Additional Information

Dinosaurs were the main animals on Earth for more than 160 million years. Some of them were the largest animals that ever walked on land. The last dinosaurs went extinct, or died out, about 65 million years ago. Dinosaur bones were around for a long time before people knew what they were. In fact, people thought one dinosaur bone belonged to a giant human. The term dinosaur was used for the first time in 1842 and literally translates to "terrible lizard." The study of dinosaurs is less than 200 years old, and early beliefs about dinosaurs have now been proven wrong.

One of the biggest mistakes scientists made was assuming that dinosaurs went extinct because they were too dumb and too big to survive. In fact, scientists thought that the *Brontosaurus*, a dinosaur with a really long neck, could not have held up its own neck on land. As a result, they drew pictures of the animal in water and believed that the water must have helped support the animal's neck. In reality, the *Brontosaurus* had extra tissue to connect the muscles in the neck to the bone and could support its neck on land. They may have used their necks similar to a giraffe, but scientists are not certain.

We also now know that dinosaurs did not go extinct because they were too dumb or too big. This should have been obvious to scientists early on because we knew dinosaurs as a group survived for over 165 million years. That is far longer than most groups of animals survive. For example, humans belong to a group of animals called primates, which include apes and monkeys. Primates have only existed for 50 million years. That's far less than dinosaurs. There is no doubt that humans and other primates are among the most intelligent animals to ever live. There's also no doubt that primates, and many other mammals, are smarter than dinosaurs. But dinosaurs must have been doing something right to have lived for so long.

Dinosaurs went extinct because a large asteroid hit the Earth. The asteroid instantly killed off a lot of animals and plants. With so few animals left, there was not enough food for dinosaurs to eat, and they quickly went extinct. There was nothing dinosaurs could have done to stop it. Some scientists even believe that if an asteroid had not hit the Earth, dinosaurs would still be alive today.

Appendix B.1.2 Dinosaur Brains Additional Information

In the 1800s, scientists started to notice that the size of an animal's brain in proportion to its body size correlated with its intelligence. At the same time, scientists realized that dinosaur skulls were small relative to their body. The theory went that the larger the brain is relative to the body, the more brain mass available for complex cognitive tasks. The fact that an animal weighing over 5 tons could have a brain of no more than 2.8 oz led to the idea that dinosaurs were unintelligent.

The ratio between brain size and body mass is a generally reliable predictor of intelligence, as long as you know how to apply it. The encephalization quotient (EQ) formula varies but is usually $Ew(brain) = 0.12w(body)^{2/3}$. From this, we get a mean EQ for mammals around 1, with meat-eaters, marine mammals, and primates above 1, and insects and plant-eaters below. The power sign corrects for the brain growing at 2/3 the rate of the body in mammals. For example, mice have a brain/body size ratio similar to humans (1:40), while elephants have comparatively small brain/body size (1:560), though elephants are obviously intelligent. What is likely happening is that the brain can only be so small and still function. In some ways the cost for each additional neuron in terms of overall brain volume gets smaller as the number of neurons go up. The end result is that larger animals don't need as much brain mass to support greater intelligence and as a result their brain to body ratios reduce.

The formula for EQ is based on data from mammals, so it should be applied to other animals with caution. Reptile brain to body growth is less understood, but a power of ¼ might be more appropriate. Differences in brain to body growth rates may reflect differences in evolutionary selection pressures. Only in mammals and birds has evolution favored large brains. In mammals and birds, brain size relative to body size increased over the course of evolution. In dinosaurs, evolution made the brain more efficient, though not more intelligent, by adding additional neurons to the base of the spine which could speed up processing related to movement and reflexes. This was highly adaptive but didn't require more intelligence and didn't require a larger brain. Ultimately, dinosaurs had small brains but given corrections for EQ this does not mean they had to be unintelligent, at least compared to other reptiles.

Appendix B.1.3 Dinosaur Behavior Additional Information

In 1923 scientists discovered eggs believed to belonged to a dinosaur called a Protoceratops. Next to the eggs were the bones of another dinosaur, an Oviraptor, that they believed died while trying to eat the eggs. Years later, they discovered more of these same eggs, but inside one was the body of an Oviraptor. This meant that the Oviraptor was the parent and was not eating the eggs.

The new discovery shouldn't have been surprising. Exposed eggs require an adult to keep them warm. Dinosaurs at a minimum would have had to stay with their eggs until they hatched. Supporting this, the eggs were arranged in a wide circle so a large dinosaur could keep them warm without crushing them. But the evidence for dinosaur parenting does not stop there. Young dinosaurs, but not babies, have been found in nests next to left-over food presumably brought by adults. The reason dinosaurs cared for their young was likely related to their size. Bigger eggs need thicker shells for support, but if the shell is too thick, oxygen can't get in. As a result, dinosaurs had to be born proportionally much smaller than many other species and would need protection to reach their adult size.

Dinosaurs also had a social network outside their immediate family. Trackways, which are just dinosaur footprints, show that that they sometimes traveled in herds. But even more important are the Davenport Ranch Trackways which show a small herd of two adults and twenty-one juveniles of the same species traveling together. This would have required attention to the position of the other adults at a minimum. Further evidence that herd behavior was important to dinosaurs comes from a nest containing more than 30 eggs. Dinosaurs did not produce this many eggs at once, so the number suggests multiple mothers were taking turns keeping them warm.

There is also some evidence of social coordination among meat-eating dinosaurs. In Utah, researchers found several Velociraptors and an Iguanodon trapped together in quicksand. It is possible that the Velociraptors were each trapped separately, but a trackway in China shows an Iguanodon running from multiple Velociraptors supporting the pack-hunting theory.

Altogether, the evidence suggests that despite their small brains, dinosaurs could engage in complex social coordination. This does not mean they were as smart as primates or even dogs, but it does mean they weren't dumb and were likely smarter on average than modern reptiles.

Appendix B.1.4 Questions about Changes in Beliefs about Dinosaurs - Easy

1. How has the perception of dinosaurs changed?
    a. It was once believed that they were slow and not likely to survive, but we now believe they could move quickly and easily.
    b. It was once believed that they only existed for 300,000 years, but now we know they lasted 160 million years.
    c. It was once believed that dinosaurs could swim, but now we know they were too big to swim.
2. The author says, "Now, the *Brontosaurus* is imagined running on land " to make the point that:
    a. Scientists' understanding of the *Brontosaurus* has changed within the last generation.
    b. The *Brontosaurus* evolved from living in the water to living on land.
    c. The *Brontosaurus* eventually learned to hold up its weight on land.

3. Which of the following statements would the author most likely agree with regarding dinosaur extinction?
    a. They lasted far longer than most animals before going extinct.
    b. Their extinction proves their bad design.
    c. They were too big to survive in the ice age that came after the asteroid hit.

Appendix B.1.5 Questions about Dinosaur Behavior - Moderate

1. Why are the Davenport Ranch Sauropod Tracks evidence of dinosaur intelligence?
    a. They demonstrate coordinated efforts among the herd.
    b. They demonstrate that some dinosaurs lived in large herds.
    c. They demonstrate that dinosaurs had cross-species social organization.
2. What can be assumed about the relationship between parental care and dinosaurs?
    a. Dinosaurs looked after their young until the young were large enough to survive on their own.
    b. Dinosaurs looked after their young until the young had learned enough to survive on their own.
    c. Dinosaurs may have watched their eggs, but they likely did not "raise" their young.
3. Which of the following is implied about dinosaurs?
    a. Dinosaurs in herds may have cared for each other’s young.
    b. Raising the large numbers of young produced by a single dinosaur would have taken a lot of mental coordination.
    c. Dinosaur parental behavior was unique among reptiles.

Appendix B.1.6 Questions about Dinosaur Brains – Hard

1. Which of the following is true about large animals?
    a. Larger animals typically have lower brain to body ratios than smaller animals.
    b. EQ should not be used to compare intelligence between small and large animals.
    c. Larger animals do not need as much intelligence as smaller animals to survive.
2. Which of the following best states the relationship of brain size to body size?
    a. The brain grows at two-thirds the rate of the body.
    b. Brain size is not related to body size.
    c. If an animal has a bigger body, they will have a smaller EQ.
3. Which of the following is a potential problem for judging dinosaur intelligence based on EQ?
    a. The ratio of brain size to brain mass works within animal types (e.g., mammals) but not across animal types because of variation in brain to body growth rates.
    b. The ratio of brain size to brain mass is less relevant among animal types which have not have not experienced an evolutionary increase in brain size over time.
    c. The ratio of brain size to brain mass is off among dinosaurs because they had "second brain" at the base of their spine that reduced the need for a large brain.

Appendix B.1.7 Questions about Dinosaur Full Passage

1. Which of the following is probably a current belief about the *Brontosaurus*?
    a. Its small brain to body ratio was likely related to its large body size.

    b. Its small brain meant that Brontosaurus herds probably did not have complex social behavior.
    c. Its size meant it did not need as much intelligence as smaller dinosaurs to survive.

2. Which of the following questions can be answered best by the passage?
    a. Why might dinosaurs have needed more intelligence than many modern reptiles?
    b. What social behaviors were characteristic of the *Brontosaurus*?
    c. Why did dinosaurs travel in herds when no modern reptiles do?
3. Which of the following best demonstrates the author’s beliefs about dinosaurs?
    a. Interpretation of dinosaur fossils and behavior was influenced by false beliefs about brain size.
    b. Recent ideas about dinosaur behavior suggest that they were highly intelligent animals.
    c. New discoveries about dinosaurs have changed the way scientists view the relationship between brain and body size.

Appendix B.1.8 Pseudoword Questions for Dinosaur Easy Section

1. Which of the following likely helped researchers to realize that dinosaurs were agile creatures?
    a. A finding of an intact, fossilized ear, complete with inner ear bones
    b. A finding of an intact, fossilized brain with extensive motor cortex
    c. A finding of an intact, fossilized spine showing extensive neck muscles
2. It’s likely that which of the following gave dinosaurs superior ability to coordinate their neck and eyes?

a. Their ear canals gave them a superior sense of balance.

b. They had long, flexible necks with excellent peripheral vision.

c. They had a second brain in their tail allowing faster responses.

3. Dinosaurs have which of the following?

   a. Elongated ear canals

   b. Increased brain area devoted to motor function (movement)

   c. Large visual cortexes

Appendix B.1.9 Pseudoword Questions for Dinosaur Hard Section

1. Why might measuring dinosaur EQ based on patterns of other reptiles be misleading?

   a. Dinosaur brains had holes in their brains making it difficult to determine how much brain matter they actually had.

   b. Large blood vessels for cooling the brain existed in dinosaur brains and as result, dinosaurs likely had less brain matter than their skulls suggest.

   c. The size of brain areas used for complex cognitive functions is not clear from measuring skull size.

2. What property of dinosaur brains magnified their difference compared to humans?

   a. Dinosaurs had holes in their brains.

   b. Dinosaur had a larger gap between their brain and skull.

   c. Dinosaurs had larger areas of brain matter dedicated to keeping their brain and body cool rather than to cognitive capacity.

3. Which of the following is a reason why scientists might not want to use skull size to determine dinosaur intelligence?

a. Unlike reptiles, dinosaur brain mass cannot be determined from skull size.

b. EQ does not correlate with intelligence in reptiles.

c. Dinosaurs were unique in that they had larger frontal cortexes than modern reptiles.

Appendix B.2 Needles & Nerves

Acupuncture and Vision - Moderate

Acupuncture is the practice of inserting tiny, hair-thin needles into the skin at specific points to treat pain and illness. Doctors and acupuncturists give millions of treatments each year in the U.S., usually for pain control. But studies show that acupuncture is also extremely useful for the type of nausea caused by chemotherapy and pregnancy. It can even reverse effects of eye degeneration which typically cannot be helped by Western medicine. Acupuncturists believe eye degeneration is caused by problems with Qi flowing through the spleen, liver, and kidney. However, the area they apply the needles to treat the problem is in the outside of the foot.

To understand if a point in the foot could affect the eyes, physicist Zang-Hee Cho strapped volunteers into an fMRI (functional magnetic resonance imaging) machine to get a photograph of their brain activity. Cho flashed a light in front of the volunteers' eyes so the fMRI image would show him what regions of their brain were involved in vision. Then, Cho had an acupuncturist stimulate the side of the foot. The very same areas of the brain lit up on the fMRI. To remove the possibility of a placebo effect, Cho also stimulated a nonacupoint in the big toe. This time, there was no response in the areas of the brain related to vision.

Acupuncture Origins and Qi - Easy

Acupuncture has been practiced in China for over 2,000 years. It is based on the belief that the body contains energy called Qi (pronounced "chee"). Qi is energy that flows through the body on pathways called meridians. When you're healthy, the energy flows freely, but during illness, the energy may be weak or blocked. The goal of acupuncture is to improve the energy flow. In fact, those receiving acupuncture sometimes report feeling a small, slightly painful pinch, followed by a tug in the body. They believe the tug is related to the movement of energy.

According to acupuncturists, the flow of Qi through the meridians is greater in certain areas—these are the acupuncture points. Over 1,500 acupoints have been found. However, most of the points have no obvious relationship to the parts of the body they are intended to treat. For example, a point on the second toe is used to treat headaches. [Further, considerations for individual differences/salants have to be made when selecting the correct point used to treat a specific condition.]

Acupuncture is also used to promote general health. Those who practice acupuncture believe it can keep Qi in balance. By keeping Qi in balance, they believe it can stop the body from getting sick. Acupuncture is popular in the United States, but the explanation for how acupuncture actually works has long been a mystery for most Western doctors.

Acupoints and Pain Management - Hard

Although a medical reason for all of acupunctures benefits has not been found, scientists agree about how it reduces pain. The points at which acupuncture needles are inserted are likely the spots where nerves are gathered together. According to neuroscientist Bruce Pomeranz, many studies have shown that acupuncture stimulates nerves in the muscles. Researchers believe the stimulated nerves send signals up the spinal cord to the pituitary gland which produces and stores chemicals called endorphins. With a strong enough signal, the pituitary gland will begin releasing endorphins.

Endorphins are a well-understood chemical primarily involved in blocking pain signals from reaching the brain. Pain is a chemical message which travels from the source of a nerve through multiple cells on its way to the brain. Endorphins bind to opiate receptors which triggers the release of additional chemicals that block the reception of chemicals created by distressed nerves. Because of this, endorphins also trigger a positive feeling throughout the body and are responsible for the feeling of a "runner's high.” [During acupuncture endorphins increase relaxation/cernous, which primes the body for healing.] However, unlike a runner’s high, the brain keeps releasing endorphins up to 24 hours after acupuncture. This can improve blood flow, reduce inflammation, and allow the body to heal more rapidly.

Appendix B.2.1 Acupuncture Origins Additional Information

Acupuncture is an ancient Chinese form of healing. It involves a patient, the person receiving the acupuncture, and an acupuncturist, the person giving the acupuncture. The patient lies on a table, and the acupuncturist sticks special needles into points on the body. The needles are made of metal and are about as thick as a human hair. They normally go less than 0.5 inches into the skin.

When a needle is pushed into the skin, the patient may feel a slight pinch or tug and then a tingling sensation that spreads out from where the needle pierced the skin. The pinch can be a little painful, but the tug is believed to be the feeling of Qi moving through the body. Qi is like breath. According to acupuncturists, all of the parts of the body are connected by lines called meridians. The meridian lines are like a giant web that links different parts of the body together. Every organ has its own meridians that connect a specific area of the body to the larger web. Qi moves from one area to the body to another by traveling along the meridians.

Sometimes these lines cross. These are the acupuncture points. When energy in the body flows easily, we don't feel Qi, and we are balanced and healthy. But when energy gets blocked at the acupoints, it causes pain and disease. Acupuncture gets Qi unstuck so that energy can flow through the body again. This helps the body heal and stay healthy.

Acupuncture requires exact placement of needles at spots on the body called acupoints. Placing the needles requires in-depth knowledge of the body. There are over a thousand possible points where the needles can be stuck, each with a different effect. One of the interesting things about acupuncture is that acupoints do not have a clear relationship to the parts of the body they affect. For example, putting a needle into an acupoint on the wrist does not help with wrist injuries. Instead, it helps with heart problems. It is not clear to modern doctors why this works.

Doctors today do not believe that acupuncture is related to Qi. But they do believe that acupuncture can help the body heal and reduce pain. They do not know how it works, but they do see that it works in their patients. As a result, it is not unusual for doctors to tell their patients to try acupuncture.

Appendix B.2.2 Acupuncture and Pain Management Additional Information

Acupuncture triggers the release of endorphins, which are “feel-good” chemicals that stop the brain from feeling pain. Endorphins are also released through exercise. In fact, both acupuncture and exercise cause the same series of events. The nerves within the muscles or skin receive a negative sensation, cells in the skin or muscles release a chemical called adenosine which travels to the hypothalamus (a portion of the brain which essentially routes incoming signals), and the hypothalamus then produces a separate chemical which is sent to the pituitary gland where endorphins are stored. When the pituitary gland receives the message from the hypothalamus, it releases the endorphins.

Endorphins reduce pain in a similar way to pain killers. In fact, pain killers are essentially man-made endorphins. They both bind to opiate cells involved in making the body feel good. When they bind to these cells, even more chemicals are released, some of which go back to the brain and reduce stress, and some of which compete to bind at the same receptors as the adenosine, effectively preventing it from doing its job. When adenosine or similar chemicals are not able to reach the brain as easily, pain is reduced. Of course, when these pain signals stop reaching the pituitary gland, it stops releasing endorphins and the "feel good" feelings go away.

Acupuncture works because, unlike with exercise, the adenosine is still present as endorphin levels reduce, thus prompting the pituitary gland to release endorphins over a longer timeframe. Exercise produces adenosine through temporary stress on the muscles and joints whereas acupuncture needles cause minor damage to tissue under the skin. In fact, a critical component of making acupuncture effective is twisting the needle, which increases the tissue damage. That means that adenosine will continue to be produced until the repair is complete, which can take 24 hours.

The longer period of increased endorphins following acupuncture has a number of long-term health benefits. In the short term, inflammation helps the body heal by increasing blood flow, but if it stays too high for too long it can actually cause more problems. Inflammation is caused by an increasing immune cells. At first, they help the body heal by attacking viruses and bacteria. But if they stay active too long, they will actually start attacking the body. Endorphins can attach themselves to immune cells and turn them off. This reduces inflammation and allows the body to finish healing and return to its normal state.

Appendix B.2.3 Acupuncture and Vision Additional Information

Macular degeneration is the most common cause of severe vision loss in people over age 50. The disease causes a breakdown of cells in the central part of the eye, called the macula, and results in blurred vision. Eventually, it can cause a blind spot to form in the person’s central vision. According to acupuncturists, macular degeneration is caused when the body's *yin* is reduced in the kidney and liver. They believe that the spleen makes the *yin* during digestion. When the body's ability to turn food into energy decreases, the body’s ability to produce *yin* decreases. This could happen if Qi was blocked in the spleen. As a result, acupuncturists increase *yin* in the liver and kidneys by unblocking Qi in the spleen. This in turn reverses macular degeneration.

Meridians in the spleen, cross with meridians from other areas of the body in the foot. Acupuncturists therefore unblock Qi in the spleen using acupoints on the side of the foot. Do Western doctors believe this actually works? One possibility is that it works like a placebo effect. That just means that it works because people believe it will work. However, placebo effects do not usually last for very long. Researchers tested whether or not acupuncture was related to placebo effects by seeing if stimulating vision-related acupoints on the foot would produce activity in vision-related areas of the brain. They used fMRI, which shows if a specific area of the brain is being used when the picture is taken. The researchers used the acupoint on the side of the foot that is believed to be related to reversing macular degeneration and they also used a point on the big toe which is not believed to be related to vision. Acupuncture at the vision-related acupoint caused more increases in activity within the area of the brain responsible for vision than at the non-vision-related acupoint.

However, macular degeneration is related to changes in the eye, not the brain. The results of the fMRI study show that acupuncture can produce activity in seemingly unrelated brain regions, but it does not explain how this improves macular degeneration. The same is true for other disorders. Acupuncture treats nausea from pregnancy, also called morning sickness, but interestingly there is no evidence that it can treat nausea from the flu. Instead, it helps with the flu by boosting the body's immune system. Right now, the only explanations for how acupuncture works lie outside of Western Medicine.

Appendix B.2.4 Questions about Acupuncture Origins and Qi - Easy

1. What might you feel if you get acupuncture?
   a. A feeling as if energy is moving within the body
   b. Nothing. Acupuncture is painless.
   c. Instant relief from pain
2. How do Western doctors view acupuncture?
   a. As having potential benefits, although how it works is still unclear
   b. As a type of alternative medicine that is not based on science
   c. As a good example of the placebo effect
3. During illness, what can happen to the body's Qi?
   a. Qi gets blocked
   b. Qi gets drained
   c. Qi breaks down

Appendix B.2.5 Questions about Acupuncture and Vision – Moderate

1. Why do acupuncturists believe acupuncture improves degenerative eye disease?
    a. It improves the functioning of the spleen.
    b. It releases endorphins, which reduce eye inflammation.
    c. It helps to unblock Qi within the eye.
2. Acupuncture may help with all of the following EXCEPT:
    a. Nausea from the flu
    b. Morning sickness
    c. Blurred vision
3. Why do acupuncturists use the acupoints on the outside of the foot to treat degenerative eye disease?
    a. They are connected via meridians to the spleen.
    b. They are connected via meridians to the eye.
    c. They are connected via meridians to areas of the brain involved in vision.

Appendix B.2.6 Questions about Acupoints and Pain Management – Hard

1. Why might acupuncture therapy continue to reduce pain even weeks treatment?
    a. Endorphins can reduce inflammation and give the body time to heal.
    b. It causes the body to start consistently releasing more endorphins, which block pain signals from being sent to the brain.
    c. It stimulates endorphins in the muscles which promote relaxation and healing.
2. Why might multiple needles be needed during acupuncture?

a. Because prolonged release of endorphins requires a buildup of signals from the body

b. Because many areas of the body need to be pierced with a needle for treatment to work

c. Because the area being treated is large and requires a greater release of endorphins

3. How might acupuncture and runner's highs be similar?
   a. They both stimulate the same areas of the brain.
   b. Western medicine is not able to explain their health benefits.
   c. They both provide a short-term rush of endorphins.

Appendix B.2.7 Questions about Acupuncture Full Passage

1. The passage suggests that acupuncture research:
   a. has found evidence that acupuncture reduces inflammation
   b. has demonstrated that acupuncture increases blood flow
   c. has not found evidence that acupoints on one area of the body are connected to other, seemingly unrelated, areas of the body
2. Which of the following best explains the author’s perspective about how acupuncture works?
   a. There are multiple ways acupuncture may work and most of them are not well understood.
   b. Most of the health benefits from acupuncture are related to increased endorphin levels in the body.

   c. Each acupoint has the ability to stimulate a seemingly unrelated area of the brain which promotes healing of the affected area of the body.

3. The article supports all of the following points about acupuncture EXCEPT:
   a. Western medicine has been ignoring the benefits of acupuncture treatment for too long.
   b. Pressure placed at acupoints can cause activity in surprising areas of the brain.
   c. Acupuncture can help a lot of modern problems, including general pain and nausea.

Appendix B.2.8 Pseudoword Questions for Acupuncture Easy Section

1. Which of the following makes identifying exact acupuncture points difficult?
   a. The exact point varies based on individual differences.
   b. The relationship between specific acupoints and the number of needles needed to treat specific conditions has not been determined.
   c. The distance between the source of the problem and the closest meridian affects the exact acupoint.
2. If acupuncture fails, what reason might an acupuncturist give for the failure?
   a. Individual differences can mean that a point which works on one person will not work on another.
   b. Additional points needed to be targeted due to the extent of blockage at the meridian.
   c. The selected point was too far the targeted meridian.
3. Which of the following statements is true?

a. The location of an acupoint used to treat a specific disease can vary from individual to individual.

b. The location of an acupoint may need to be adjusted based on external factors, such as the person's body position while laying down.

c. The location of an acupoint may depend on the number of needles that need to be used during the session.

Appendix B.2.9 Pseudoword Questions for Acupuncture Hard Section

1. Which of the following is a possible explanation for why endorphins promote healing?
    a. Endorphins relax the body.
    b. Endorphins increase dopamine levels in the brain, which decreases the body's stress response.
    c. Endorphins quickly increase white blood cell counts, which are critical to the immune system.
2. What might a person experience the night following an acupuncture session?
    a. They might sleep deeply due feeling more relaxed than usual.
    b. They might fall asleep later due to increases in dopamine levels which promote wakefulness and immune system function.
    c. They might fall asleep later due to the energetic high they are experiencing.
3. Which of the following would be considered immediate effects of an acupuncture session?
    a. Decline in blood pressure associated with relaxation.
    b. An increase in energy, similar to what someone might feel after a workout

c. An increase in multiple chemicals, like melatonin and dopamine, which promote healing

Appendix B.3 A Comment on Comets

Comet Orbits - Easy

A comet is a small chunk of dust and ice that orbits (travels around) the Sun in an irregular but mostly oval shape. It is sometimes described as a "dirty snowball." The main part of a comet is called the nucleus. The nucleus is usually a few miles wide and has many holes in the surface which give it a spongy appearance, but it is not actually soft.

The most famous comet is called Halley's Comet. It can be seen from Earth without a telescope about every 76 years. Comets come from two areas at the farthest edges of the solar system. These areas are called the Kuiper Belt and Oort Cloud. {Comets that can only be seen from Earth every several hundred years are from the Oort Cloud. Any comet that passes by Earth more frequently comes from the Kuiper Belt.}[20] [The solar system, along with the Kuiper Belt, is constantly surrounded/cepulized by the rocks in the Oort Cloud.]

There are billions of comets in the solar system. Most never come close to the Earth. The comets that are seen from Earth have been pushed out of their normal orbits by the gravity of passing stars from other solar systems. The change in orbit can put a comet on a path closer to the Sun and Earth.

Comet Light - Hard

[20] Sentences in the {} were removed from the passage for Experiment 2 to decrease the probability that readers would infer the meaning of the sentence with the pseudoword.

Most of the time, a comet only has a dark nucleus. The bright portions, called the coma and tail, are temporary and depend on the distance from the Sun and Earth. The Sun's heat causes frozen material to evaporate, and the resulting cloud formation around the nucleus is called the coma and can be larger than Earth. As the comet moves towards its closest point to the Sun, the perihelion, the momentum of solar photons creates radiation pressure as it meets dust in coma. The speed of each dust particle as it meets the radiation pressure varies according to its size which creates a tail of dust. [The radiation pressure stems from when the dust particles reflect photons of light, which have momentum/ocelants but no weight.] Gas particles break away because the magnetic field of plasma of the outward bound solar winds attracts magnetized ions in the gas.

If a comet reaches its nearest point to Earth after its perihelion, it will be much brighter than if it reaches its nearest point to Earth while it is still relatively cold. However, the tails, which are sometimes longer than the Earth's distance to the Sun, and coma last only while the comet is fairly close to the Sun. After each pass, the nucleus of the comet is smaller and will eventually evaporate.

Comet Research - Moderate

After the explosion that created our solar system about four billion years ago, some of the materials that were pushed farthest from the Sun froze together. Comets are believed to be made-up of these materials. Because comets spend most of their time in the outer reaches of space, they have remained relatively unchanged and are thought of as a "fossil record" of the solar system. Comets may even carry the secret to life. Water and some organic materials may have been brought to Earth by comets hitting our planet during its earliest days.

Scientists are unlocking these answers by studying comets directly rather than through a telescope. For example, a collection of tiny dust particles left behind by a comet led to the discovery of a previously unknown mineral. Even more recently, the Rosetta probe caught up with a comet beyond the asteroid belt after a ten-year flight. It sent back data from water vapor surrounding the comet that was fundamentally different from water on Earth. The probe also found organic compounds that could be the building blocks for DNA. Unfortunately, the solar battery died two days after landing in a crater, and no additional data was collected.

Appendix B.3.1 Comet Orbits Additional Information

Our Sun was formed 4.5 billion years ago through an explosion. The material from the explosion included gas, water, and dust. The explosion kicked many of these materials out far away from the Sun. In fact, they ended up so far away from the Sun that they froze. When some of the gas and dust froze together, it created comets. That means that comets are made of frozen gas and dust. The frozen ball of gas and dust is called a nucleus. You may have heard of a nucleus in biology or physics. They form the center of cells and atoms. Just like with cells and atoms, the nucleus of a comet forms the center of a comet.

The combination of frozen gas and dust gives comets an unusual appearance. They look like a sponge because there are many holes in the comet. Of course, the frozen material that makes up a comet does not feel like a sponge. It is not soft. Instead, the material is hard, like a rock. In fact, it’s hard enough that probes sent by NASA have been able to successfully land on comets.

Comets orbit the Sun. An orbit is just an object’s path as it moves around the Sun. The Earth orbits the Sun too. Because comets are so far away, it takes them a long time to orbit around the Sun, and the farther away they are, the longer it takes. Objects that are close to the Sun, like Earth and Mars, orbit the Sun more quickly than objects in the Kuiper Belt and Oort Cloud. And objects in the Kuiper Belt are closer to the Sun than objects in the Oort Cloud. Comets in the Oort

Cloud and Kuiper Belt are so far way that we cannot see them from Earth, even with a high-powered telescope. In order for comets to be visible from Earth, something has to force them out of their normal orbit. And then the comet's new orbit needs to bring it close to the Earth. Only then, will we have a chance of seeing the comet.

The chance of a comet's orbit bringing it close to the Earth is very small. But, when it does happen, we can sometimes see them with our naked-eye. That means that a telescope is not needed to see near-Earth comets as long as the comet is lit up.

Appendix B.3.2 Comet Light Additional Information

In the outer Solar System, comets remain frozen and inactive and are extremely difficult or impossible to see from Earth due to their small size. Statistical detections of inactive comet nuclei in the Kuiper belt have been reported from observations by the Hubble Space Telescope but these detections have been questioned. As a comet approaches the inner Solar System, solar radiation causes the volatile materials within the comet to vaporize and stream out of the nucleus, carrying dust away with them. Some of the dust is left behind as the ice changes. It forms a dark, protective crust on the surface of the nucleus and slows the melting. In some places the protective layer is thinner, and jets of gas break through. The gas and dust form a cloud around the nucleus called a coma.

Two distinct tails develop from the coma — the plasma (gas) tail and the dust tail, and each form their own distinct tail, pointing in slightly different directions. The different shapes and angles of the tails are caused by the way different particles are affected by the Sun. The thinner, longer plasma tail forms a straight line extending from the comet. The particles in this ion tail are electrically charged and are pushed away from the Sun by solar wind. The solar wind is made-up of a constant flow of gas and particles (mostly protons and electrons) that stream outward at 220 miles per second.

The shorter dust tail is curved slightly. The larger particles in the dust tail do not have an electric charge and are not affected by the solar wind. Dust-size particles that escape from the comet experience a much weaker push from the Sun caused by the pressure of sunlight itself (called radiation pressure), rather than by the charged particles of the solar wind. Radiation pressure is the mechanical pressure exerted upon any surface due to the exchange of momentum between the object and the electromagnetic field. While the dust tail also points generally away from the Sun, it has a slight curve back in the direction the comet came from.

Comet tails get longer and more impressive as the comet gets closer to our Sun. As the comet approaches our Sun, it gets hotter and material is released more rapidly, producing a larger tail. Scientists estimate that a comet loses between 0.1 and 1 percent of its mass each time it orbits our Sun.

Appendix B.3.3 Comet Research Additional Information

Comets were created during an explosion that created the Sun and marked the beginning of our solar system. Because comets spend most, if not all, of their time in the outer solar system, far from the Sun, they are frozen, and, because frozen material does not change, comets have not changed much since the beginning of the solar system. As a result, learning about comets means learning about the origins of the solar system and everything within it.

One question scientists are trying to answer is how water got to Earth. When the Earth was forming, it was so hot that most of its water evaporated. Once the Earth cooled down, there was virtually no water left, meaning that the water that makes up the oceans had to have come from somewhere else. One theory is that the water came from the frozen ice on comets that hit the Earth.

All water contains H_2O: two parts hydrogen and one part oxygen. But hydrogen comes in two types: regular hydrogen and deuterium. Deuterium is just hydrogen with an added neutron. Earth's water contains far more regular hydrogen than deuterium. If Earth's water came from comets, then the water in comets should also contain more regular hydrogen.

Answering this question was one of the reasons scientists launched the Rosetta probe to land on a comet, analyze its materials, and send back the results. Rosetta's measurements revealed far more deuterium in the water on the comet than exists in Earth's water. This makes it highly unlikely that Earth's water came from a comet.

Nevertheless, the Rosetta probe found other materials on the comet which are critical to life. The most important was glycine, a building block of DNA. Unfortunately, the Rosetta probe's solar died faster than expected. When the probe landed on the comet, it unexpectedly bounced and ended up in the shadow of cliff. It was out of view of the Sun, and its batteries could not recharge. The probe fell silent when its solar batteries ran out of power.

Thankfully, we don't have to wait for the next probe to study comets. By plotting the orbit of a comet, NASA can pinpoint the date when its dust will enter the Earth's atmosphere. During one such occasion, NASA collected dust in the stratosphere and found a brand new type of mineral. It was a type of manganese silicide which has been named "Brownleeite" after the researcher who found it.

Appendix B.3.4 Questions about Comet Orbits - Easy

1. What is a comet made of?

a. Dust and ice

b. Cast off materials from when meteors hit planets

c. Organic materials that do not exist on Earth

2. Which of the following best describes the nucleus of a comet?

    a. It’s full of holes

    b. It's a dense block of solid rock

    c. It’s soft, like a sponge

3. How likely is it that a comet pushed out of its orbit will come close enough to Earth to be seen by astronomers?

    a. Unlikely

    b. Likely

    c. A comet cannot be pushed out of its orbit

Appendix B.3.5 Questions about Comet Research - Moderate

1. Why did the Rosetta probe only collect data for 2 days after it landed on the comet?

    a) The probe happened to land in a crater, blocking it from view of the sun.

    b) The probe had used a lot of battery power to reach the comet and had little left after its arrival.

    c) The probe landed harder than expected, causing it’s the battery to malfunction.

2. Why are comets considered a “fossil record”?

    a) The materials that make up the nucleus are unchanged since the origin of the solar system.

b) During their orbits, they pick up material from many regions of the solar system and can document its evolution.

c) Material from comets helped create many of the planets, so they hold the key to understanding planet origins.

3. Which of the following were NOT found by the Rosetta probe?

a) A previously undiscovered mineral

b) Organic compounds similar to parts of DNA

c) Water vapor

Appendix B.3.6 Questions about Comet Light - Hard

1. What might cause a comet near the Earth to be less visible to astronomers?

a) If it approaches Earth before it has moved closest to the Sun

b) If it approaches Earth after it has moved closest to the Sun

c) If its tails have begun to separate as they are attracted by magnetic fields in the Sun

2. Which of the following are true about the tail of a comet?

a) The tails are created by radiation pressure blowing dust off the coma and solar winds ionizing and attracting the gasses.

b) The tail is created when material is, in essence, blown off the coma by solar winds and magnetized ions force separation of the dust and gas particles.

c) The tails become smaller during each orbit around the sun as the nucleus loses more and more material.

3. What direction is a comet tail pointed as the comet travels around the Sun?

a) Away from the Sun

b) Away from the Sun as it approaches and toward the Sun as it departs

c) The dust tail points away from the sun while the ionized gas tail is attracted towards the sun

Appendix B.3.7 Questions about Comet Full Passage

1. Information in the passage indicates that seeing a near-Earth comet requires all of the following EXCEPT:
   a. the viewer to have a powerful telescope
   b. solar winds and radiation pressure to blow against the coma
   c. the comet to have been pushed out of its typical orbit
2. The passage mentions astronomers observing all of the following about comets EXCEPT:
   a. comets that give off bright light from their nucleus
   b. orbits that take comets to the edges of the Sun's gravitational influence
   c. comets with sponge-like appearances
3. Scientists are most interested in comets directly because:
   a. They contain chemicals from the origin of the solar system.
   b. The temporary effects of close encounters with solar radiation can reveal how planets and atmospheres were formed.
   c. They come from the edge of the solar system, and so hold clues as to what lies beyond.

Appendix B.3.8 Pseudoword Questions for Comet Easy Section

1. Which of the following is true about the Oort Cloud?
    a. It forms a sphere around the entire solar system, completely surrounding it.
    b. It forms a divider, separating the solar system into an inner and outer half.
    c. Rocks from the Oort Cloud are constantly being hurled into the inner solar system.
2. Why might it be more likely for a comet that passes the Earth to have originated within the Oort Cloud than the Kuiper Belt?
    a. Objects in the Oort Cloud are further from the gravitational pull of the Sun
    b. Objects in the Oort Cloud are closer to the Sun and therefore more likely to pass the Sun once they leave their orbit
    c. Objects from the Oort Cloud are closer to planetary bodies which may pull their orbit into the inner solar system
3. Which of the following is true about the Oort Cloud and Kuiper Belt?
    a. Comets orbiting within the Kuiper Belt do not pass through the Oort Cloud
    b. Comets orbiting within the Oort Cloud pass through the Kuiper belt
    c. Comets orbiting within the Kuiper Belt were originally in the Oort Cloud

Appendix B.3.9 Pseudoword Questions for Comet Hard Section

1. Why is it surprising that photons have no weight?
    a. Because they have momentum
    b. Because they exert a gravitational pull.
    c. Because they generate heat.

2. Which of the following is true for the formation of the dust tail but not for the formation of the gas tail?
    a. Particles in the dust tail are pushed away from the coma due to the momentum of the photons.
    b. The dust tail is pulled by the mass of the photons as they move past the coma.
    c. Particles in the dust tail form an electrical charge as they come in contact with photons from the Sun.
3. Based on the article, which of the following do photons have that may result in radiation pressure?
    a. Momentum
    b. Gravity
    c. Electrical charge

Appendix C Passage-Related Vocabulary Test

Appendix Table 2 Vocabulary Assessed during Session 1

Topic	Text Difficulty	Word	Correct Response	Lure 1	Lure 2
Acupuncture	Easy	flow	move freely and easily	air	fly
Acupuncture	Easy	tug	pull at something	stop something	jump
Acupuncture	Easy	obvious	easy to understand	friendly	smart
Comets	Easy	irregular	oddly shaped	slow	large
Comets	Easy	spongy	has holes	dark	small
Comets	Easy	telescope	item used to make far objects appear closer	item used to see in the dark	item used to see underwater
Dinosaurs	Easy	agile	move easily	tall	heavy
Dinosaurs	Easy	clumsy	uncoordinated	smart	fast
Dinosaurs	Easy	murky	dirty	clean	bright

Acupuncture	Moderate	nausea	feeling sick to your stomach	feeling a headache	feeling a muscle cramp
Acupuncture	Moderate	reverse	move in the opposite direction	repeat an action	jump
Acupuncture	Moderate	placebo	fake medicine	a sickness	an animal
Comets	Moderate	water vapor	gas	liquid	solid
Comets	Moderate	fossil	something old (usually a bone)	a fish	a painting
Comets	Moderate	organic	natural	watery	made in a factory
Dinosaurs	Moderate	trackway	path where an animal has walked	mountain	a type of car race
Dinosaurs	Moderate	quicksand	loose wet sand	a superhero	beach
Dinosaurs	Moderate	coordination	work together	meeting	not similar
Acupuncture	Difficult	nerves	a type of cell in the body	bones	muscle
Acupuncture	Difficult	distressed	feeling anxiety	not afraid	at rest
Acupuncture	Difficult	inflammation	swelling in the body	a big fire	a type of cancer

Comets	Difficult	momentum	moving fast in one direction	staying still	a way to measure time
Comets	Difficult	evaporate	go from water to air	evolution	become hard
Comets	Difficult	magnetic	can be drawn to another object	can start a fire	can create life
Dinosaurs	Difficult	discrepancy	a difference	a signal	an idea
Dinosaurs	Difficult	mammalian	had a "live birth" (not born from an egg)	can fly	is dangerous
Dinosaurs	Difficult	stupidity	not smart	excited	interesting

Bibliography

Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review, 13*(3), 219-235.

Alter, A. L., Oppenheimer, D. M., Epley, N., & Eyre, R. N. (2007). Overcoming intuition: metacognitive difficulty activates analytic reasoning. *Journal of Experimental Psychology: General, 136*(4), 569.

Atkinson, R. C. (1958). A Markov model for discrimination learning. *Psychometrika, 23*(4), 309-322.

Atkinson, R. C. (1972). Optimizing the learning of a second-language vocabulary. *Journal of Experimental Psychology, 96*(1), 124.

Atkinson, R. C. (1974). Teaching children to read using a computer. *American Psychologist*, *29*(3), 169.

Baker, L. (1989). Metacognition, comprehension monitoring, and the adult reader. *Educational Psychology Review, 1*(1), 3-38.

Baker, L., & Beall, L. C. (2014). Metacognitive processes and reading comprehension. In *Handbook of research on reading comprehension* (pp. 397-412). Routledge.

Balass, M., Nelson, J. R., & Perfetti, C. A. (2010). Word learning: An ERP investigation of word experience effects on recognition and word processing. *Contemporary educational psychology*, *35*(2), 126-140.

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. J. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General*, *133*(2), 283.

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology*, *51*(6), 1173.

Bates, D. M., Maechler, M., Bolker, B., & Walker, S. (2016). lme4: Mixed-effects modeling with R; 2010. URL: http://lme4. r-forge. r-project. org/book [8 April 2015].

Begg, I. M., Martin, L. A., & Needham, D. R. (1992). Memory monitoring: How useful is self-knowledge about memory? *European Journal of Cognitive Psychology, 4*(3), 195-218.

Begg, I., Duft, S., Lalonde, P., Melnick, R., & Sanvito, J. (1989). Memory predictions are based on ease of processing. *Journal of Memory and Language, 28*(5), 610-632.

Belsley, D. A. (1991). A guide to using the collinearity diagnostics. *Computer Science in Economics and Management*, *4*(1), 33-50.

Benjamin, A. S. (2003). Predicting and postdicting the effects of word frequency on memory. *Memory & Cognition, 31*(2), 297-305.

Benjamin, A. S., Bjork, R. A., & Hirshman, E. (1998). Predicting the future and reconstructing the past: A Bayesian characterization of the utility of subjective fluency. *Acta Psychologica, 98*(2-3), 267-290.

Berman, R. A., & Nir-Sagiv, B. (2007). Comparing narrative and expository text construction across adolescence: A developmental paradox. *Discourse Processes, 43*(2), 79-120.

Betts, E. (1946). Foundations of reading instruction. New York: American Book Company.

Blachowicz, C. L. (1986). Making connections: Alternatives to the vocabulary notebook. *Journal of Reading, 29*(7), 643-649.

Brunsdon, C., Charlton, M., & Harris, P. (2012). *Living with Collinearity in Local Regression Models.* In: Accuracy 2012 - 10th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, 10th - 13th July, 2012, Florianópolis, SC, Brazil.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977-990.

Busey, T. A., Tunnicliff, J., Loftus, G. R., & Loftus, E. F. (2000). Accounts of the confidence-accuracy relation in recognition memory. *Psychonomic Bulletin & Review, 7*(1), 26-48.

Calloway, R. C. (2019). Why do you read? Toward a more comprehensive model of reading comprehension: The role of standards of coherence, reading goals, and interest (Doctoral dissertation, University of Pittsburgh).

Carter, T. J., & Dunning, D. (2008). Faulty self-assessment: Why evaluating one's own competence is an intrinsically difficult task. *Social and Personality Psychology Compass*, *2*(1), 346-360.

Cavanaugh, J. C., & Perlmutter, M. (1982). Metamemory: A critical examination. *Child development*, 11-28.

Chi, E. H., Gumbrecht, M., & Hong, L. (2007, July). Visual foraging of highlighted text: An eye-tracking study. In *International Conference on Human-Computer Interaction* (pp. 589-598). Springer, Berlin, Heidelberg.

Christensen, R. H. B. (2019). Regression models for ordinal data [R package ordinal version 2019.12-10].

Commander, N. E., & Stanwyck, D. J. (1997). Illusion of knowing in adult readers: Effects of reading skill and passage length. *Contemporary Educational Psychology, 22*(1), 39-52.

Crossley, S. A., Skalicky, S., Dascalu, M., McNamara, D. S., & Kyle, K. (2017). Predicting text comprehension, processing, and familiarity in adult readers: New approaches to readability formulas. *Discourse Processes*, *54*(5-6), 340-359.

Cummins, R. P. (1981). Test review: The Nelson-Denny Reading Test (Forms E and F*). Journal of Reading*, *25*(1), 54-59.

Dale, E., & Tyler, R. W. (1934). A study of the factors influencing the difficulty of reading materials for adults of limited reading ability. *The Library Quarterly, 4*(3), 384-412.

Davis, T. C., Wolf, M. S., Bass, P. F., Middlebrooks, M., Kennen, E., Baker, D. W., ... & Parker, R. M. (2006). Low literacy impairs comprehension of prescription drug warning labels. *Journal of general internal medicine*, *21*(8), 847-851.

Dunlosky, J., & Ariel, R. (2011). Self-regulated learning and the allocation of study time. *Psychology of Learning and Motivation*, *54,* 103-140.

Dunlosky, J., & Connor, L. T. (1997). Age differences in the allocation of study time account for age differences in memory performance. *Memory & Cognition*, *25*(5), 691-700.

Dunlosky, J., & Hertzog, C. (1998). Training programs to improve learning in later adulthood: *Helping older adults educate themselves.* (pp. 263-290). Routledge.

Dunlosky, J., & Lipko, A. R. (2007). Metacomprehension: A brief history and how to improve its accuracy. *Current Directions in Psychological Science*, *16*(4), 228-232.

Dunlosky, J., & Nelson, T. O. (1992). Importance of the kind of cue for judgments of learning (JOL) and the delayed-JOL effect. *Memory & Cognition*, *20*(4), 374-380.

Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention. *Learning and Instruction*, *22*(4), 271-280.

Dunlosky, J., & Thiede, K. W. (2004). Causes and constraints of the shift-to-easier-materials effect in the control of study. *Memory & Cognition*, *32*(5), 779-788.

Dunlosky, J., Rawson, K. A., & Hacker, D. J. (2002). Metacomprehension of science text: Investigating the levels-of-disruption hypothesis. In J. Otero, J. A. León, & A. C. Graesser (Eds.), *The psychology of science text comprehension* (pp. 255–279). Lawrence Erlbaum Associates Publishers.

Erickson, T. D., & Mattson, M. E. (1981). From words to meaning: A semantic illusion. *Journal of Verbal Learning and Verbal Behavior*, *20*(5), 540-551.

Ferreira, F. (2003). The misinterpretation of noncanonical sentences. *Cognitive Psychology, 47*(2), 164-203.

Ferreira, F., Bailey, K. G., & Ferraro, V. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science, 11*(1), 11-15.

Ferreira, F., & Patson, N. D. (2007). The 'good enough' approach to language comprehension. *Language and Linguistics Compass*, *1*(1-2), 71-83.

Ferreira, F., & Stacey, J. (2000). The misinterpretation of passive sentences. *Unpublished manuscript*.

Ferrer, A., Vidal-Abarca, E., Serrano, M. Á., & Gilabert, R. (2017). Impact of text availability and question format on reading comprehension processes. *Contemporary Educational Psychology, 51*, 404-415.

Finn, B., & Tauber, S. K. (2015). When confidence is not a signal of knowing: How students' experiences and beliefs about processing fluency can lead to miscalibrated confidence. *Educational Psychology Review, 27*(4), 567-586.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist*, *34*(10), 906-911.

Fleming, S. M., & Dolan, R. J. (2012). The neural basis of metacognitive ability. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1338-1349.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, *32*(3), 221-233.

Garner, R. (1980). Monitoring of understanding: An investigation of good and poor readers' awareness of induced miscomprehension of text. *Journal of Reading Behavior*, *12*(1), 55-63.

Garner, R., & Taylor, N. (1982). Monitoring of understanding: An investigation of attentional assistance needs at different grade and reading proficiency levels. *Reading Psychology: An International Quarterly, 3*(1), 1-6.

Glenberg, A. M., & Epstein, W. (1985). Calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*(4), 702-718.

Glenberg, A. M., Sanocki, T., Epstein, W., & Morris, C. (1987). Enhancing calibration of comprehension. *Journal of Experimental Psychology: General*, *116*(2), 119-136.

Glenberg, A. M., Wilkinson, A. C., & Epstein, W. (1982). The illusion of knowing: Failure in the self-assessment of comprehension. *Memory & Cognition, 10*(6), 597-602.

Glover, J. A. (1989). The" testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology, 81*(3), 392-399.

Glynn, S. M. (1978). Capturing readers' attention by means of typographical cuing strategies. *Educational Technology*, *18*(11), 7-12.

Golke, S., & Wittwer, J. (2017). High-performing readers underestimate their text comprehension: Artifact or psychological reality?. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Green, P., & MacLeod, C. J. (2016). SIMR: an R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution, 7*(4), 493-498.

Griffin, T. D., Wiley, J., & Thiede, K. W. (2008). Individual differences, rereading, and self-explanation: Concurrent processing and cue validity as constraints on metacomprehension accuracy. *Memory & Cognition, 36*(1), 93-103.

Guthrie, J. T., Wigfield, A., Barbosa, P., Perencevich, K. C., Taboada, A., Davis, M. H., Scafiddi, N. T., & Tonks, S. (2004). Increasing Reading Comprehension and Engagement Through Concept-Oriented Reading Instruction. *Journal of Educational Psychology, 96*(3), 403–423.

Howes, D. H., & Solomon, R. L. (1951). Visual duration threshold as a function of word-probability. *Journal of Experimental Psychology, 41*(6), 401.

Jacobson, J. M. (1990). Congruence of pretest predictions and posttest estimations with grades on short answer and essay tests. *Educational Research Quarterly, 14*(2), 41-47.

Jaeger, A. J., & Wiley, J. (2014). Do illustrations help or harm metacomprehension accuracy?. *Learning and Instruction, 34,* 58-73.

Jaeger, A. J., & Wiley, J. (2015). Reading an analogy can cause the illusion of comprehension. *Discourse Processes, 52*(5-6), 376-405.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. *Heuristics and Biases: The Psychology of Intuitive Judgment, 49,* 81.

Keene, E. O., & Zimmermann, S. (1997). *Mosaic of thought: Teaching comprehension in a reader's workshop.* Heinemann, 361 Hanover Street, Portsmouth, NH 03801-3912.

Kelemen, W. L., Frost, P. J., & Weaver, C. A. (2000). Individual differences in metacognition: Evidence against a general metacognitive ability. *Memory & Cognition, 28*(1), 92-107.

Kim, J. H. (2019). Multicollinearity and misleading statistical results. *Korean journal of anesthesiology*, *72*(6), 558.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review, 95*(2), 163.

Kintsch, W. (1994). Text comprehension, memory, and learning. *American psychologist*, *49*(4), 294.

Kirk-Johnson, A., Galla, B. M., & Fraundorf, S. H. (2019). Perceiving effort as poor learning: The misinterpreted-effort hypothesis of how experienced effort and perceived learning relate to study strategy choice. *Cognitive Psychology, 115*, 101237.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126*(4), 349.

Kornell, N., & Metcalfe, J. (2006). Study efficacy and the region of proximal learning framework. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(3), 609.

Kornell, N., Rhodes, M. G., Castel, A. D., & Tauber, S. K. (2011). The ease-of-processing heuristic and the stability bias: Dissociating memory, memory beliefs, and memory judgments. *Psychological Science*, *22*(6), 787-794.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*(6), 1121.

Kwon, H., & Linderholm, T. (2014). Effects of self-perception of reading skill on absolute accuracy of metacomprehension judgements. *Current Psychology, 33*(1), 73-88.

Landi, N. (2010). An examination of the relationship between reading comprehension, higher-level and lower-level reading sub-skills in adults. *Reading and Writing*, *23*(6), 701–717.

Lin, L. M., & Zabrucky, K. M. (1998). Calibration of comprehension: Research and implications for education and instruction. *Contemporary Educational Psychology, 23*(4), 345-391.

Maki, R. H. (1998). Test predictions over text material. In *Metacognition in educational theory and practice* (pp. 131-158). Routledge.

Maki, R. H., & Berry, S. L. (1984). Metacomprehension of text material. Journal of *Experimental Psychology: Learning, Memory, and Cognition*, *10*(4), 663.

Maki, R. H., & McGuire, M. J. (2002). Metacognition for text: Findings and implications for education. In T. J. Perfect & B. L. Schwartz (Eds.), *Applied metacognition* (pp. 39–67). Cambridge University Press.

Maki, R. H., Foley, J. M., Kajer, W. K., Thompson, R. C., & Willert, M. G. (1990). Increased processing enhances calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(4), 609.

Maki, R. H., Jonas, D., & Kallod, M. (1994). The relationship between comprehension and metacomprehension ability. *Psychonomic Bulletin & Review, 1*(1), 126-129.

Maki, R. H., Shields, M., Wheeler, A. E., & Zacchilli, T. L. (2005). Individual Differences in Absolute and Relative Metacomprehension Accuracy. *Journal of Educational psychology, 97*(4), 723.

Mata, A. (2020). An easy fix for reasoning errors: Attention capturers improve reasoning performance. *Quarterly Journal of Experimental Psychology, 73*(10), 1695-1702.

Mata, A., Ferreira, M. B., Voss, A., & Kollei, T. (2017). Seeing the conflict: An attentional account of reasoning errors. *Psychonomic Bulletin & Review, 24*(6), 1980-1986.

Mata, A., Schubert, A. L., & Ferreira, M. B. (2014). The role of language comprehension in reasoning: How "good-enough" representations induce biases. *Cognition, 133*(2), 457-463.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language, 94*(1), 305-315.

Mazzoni, G., & Nelson, T. O. (1995). Judgments of learning are affected by the kind of encoding in ways that cannot be attributed to the level of recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(5), 1263.

Metcalfe, J. (2002). Is study time allocated selectively to a region of proximal learning?. *Journal of Experimental Psychology: General, 131*(3), 349.

Metcalfe, J. (2011). Desirable difficulties and studying in the region of proximal learning. *Successful remembering and successful forgetting: A Festschrift in honor of Robert A. Bjork*, 259-276.

Metcalfe, J., & Kornell, N. (2003). The dynamics of learning and allocation of study time to a region of proximal learning. *Journal of Experimental Psychology: General, 132*(4), 530.

Metcalfe, J., & Kornell, N. (2005). A region of proximal learning model of study time allocation. *Journal of Memory and Language, 52*(4), 463-477.

Metcalfe, J., Schwartz, B. L., & Eich, T. S. (2020). Epistemic curiosity and the region of proximal learning. *Current Opinion in Behavioral Sciences, 35*(1), 40-47.

Miller, D. (2002). *Reading with meaning: Teaching comprehension in the primary grades.* Stenhouse Publishers.

Morris, D., Trathen, W., Gill, T., Perney, J., Schlagal, R., Ward, D., & Frye, E. M. (2019). Reading Instructional Level from a Print-Processing Perspective. *Reading & Writing Quarterly*, *35*(6), 556-571.

Mounla, G., Bahous, R., & Nabhani, M. (2011). The Reading Matrix© 2011. *Reading, 11*(3), 279-291.

Myers, S. J., Rhodes, M. G., & Hausman, H. E. (2020). Judgments of learning (JOLs) selectively improve memory depending on the type of test. *Memory & Cognition, 48*(5), 745-758.

Narvaez, D., Van Den Broek, P., & Ruiz, A. B. (1999). The influence of reading purpose on inference generation and comprehension in reading. *Journal of Educational Psychology, 91*(3), 488.

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin, 95*(1), 109.

Nelson, T. O., & Leonesio, R. J. (1988). Allocation of self-paced study time and the "labor-in-vain effect.” *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*(4), 676.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In *Psychology of learning and motivation* (Vol. 26, pp. 125-173). Academic Press.

Novemsky, N., Dhar, R., Schwarz, N., & Simonson, I. (2007). Preference fluency in choice. *Journal of Marketing Research, 44*(3), 347-356.

O’Reilly, T., & McNamara, D. S. (2007). The impact of science knowledge, reading skill, and reading strategy knowledge on more traditional “high-stakes” measures of high school students’ science achievement. *American Educational Research Journal, 44*(1), 161-196.

Oakhill, J., Hartt, J., & Samols, D. (2005). Levels of comprehension monitoring and working memory in good and poor comprehenders. *Reading and Writing, 18*(7), 657-686.

Otero, J., & Kintsch, W. (1992). Failures to detect contradictions in a text: What readers believe versus what they read. *Psychological Science, 3*(4), 229-236.

Ozuru, Y., Dempsey, K., & McNamara, D. S. (2009). Prior knowledge, reading skill, and text cohesion in the comprehension of science texts. *Learning and Instruction, 19*(3), 228-242.

Perfetti, C. A., & Hart, L. (2001). The lexical basis of comprehension skill. In D. S. Gorfein (Ed.), *On the consequences of meaning selection: Perspectives on resolving lexical ambiguity* (pp. 67–86). American Psychological Association.

Piaget, J., & Cook, M. T. (1952). *The origins of intelligence in children.* (M. Cook, Trans.). W W Norton & Co.

Pressley, M., & Schneider, W. (1997). *Introduction to memory: Development during childhood and adolescence.* Psychology Press.

Price, J., & Murray, R. G. (2012). The region of proximal learning heuristic and adult age differences in self-regulated learning. *Psychology and Aging, 27*(4), 1120.

Prinz, A., Golke, S., & Wittwer, J. (2018). The double curse of misconceptions: Misconceptions impair not only text comprehension but also metacomprehension in the domain of statistics. *Instructional Science, 46*(5), 723-765.

Prinz, A., Golke, S., & Wittwer, J. (2020). How accurately can learners discriminate their comprehension of texts? A comprehensive meta-analysis on relative metacomprehension accuracy and influencing factors. *Educational Research Review, 31*(2020), 100358.

Rawson, K. A., & Dunlosky, J. (2002). Are performance predictions for text based on ease of processing?. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(1), 69.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological bulletin, 124*(3), 372.

Rayner, K., & Duffy, S. A. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition, 14*(3), 191-201.

Reyna, V. F., & Brainerd, C. J. (1991). Fuzzy‐trace theory and framing effects in choice: Gist extraction, truncation, and conversion. *Journal of Behavioral Decision Making, 4*(4), 249-262.

Rhodes, M. G., & Tauber, S. K. (2011). The influence of delaying judgments of learning on metacognitive accuracy: a meta-analytic review. *Psychological bulletin, 137*(1), 131.

Sanford, A. J., & Sturt, P. (2002). Depth of processing in language comprehension: Not noticing the evidence. *Trends in Cognitive Sciences, 6*(9), 382-386.

Schommer, M., & Surber, J. R. (1986). Comprehension-monitoring failure in skilled adult readers. *Journal of Educational Psychology, 78*(5), 353.

Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. *and Learning, 4*(1), 33-45.

Schwinger, M., Wirthwein, L., Lemmer, G., & Steinmayr, R. (2014). Academic self-handicapping and achievement: A meta-analysis. *Journal of Educational Psychology, 106*(3), 744.

Metacognition

Serra, M. J., & Dunlosky, J. (2010). Metacomprehension judgements reflect the belief that diagrams improve learning from text. *Memory, 18*(7), 698-711.

Shanahan, T. (2020). Limiting Children to Books They Can Already Read: Why It Reduces Their Opportunity to Learn. *American Educator*, *44*(2), 13.

Simmons, J. P., & Nelson, L. D. (2006). Intuitive confidence: choosing between intuitive and nonintuitive alternatives. *Journal of Experimental Psychology: General, 135(3),* 409.

Singer, M., Harkness, D., & Stewart, S. T. (1997). Constructing inferences in expository text comprehension. *Discourse Processes, 24*(2-3), 199-228.

Soderstrom, N. C., & Rhodes, M. G. (2014). Metacognitive illusions can be reduced by monitoring recollection during study. *Journal of Cognitive Psychology*, *26*(1), 118-126.

Son, L. K., & Metcalfe, J. (2000). Metacognitive and control strategies in study-time allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(1), 204.

Stolp, S., & Zabrucky, K. M. (2009). Contributions of metacognitive and self-regulated learning theories to investigations of calibration of comprehension. *International Electronic Journal of Elementary Education, 2*(1), 7-31.

Thiede, K. W., & Anderson, M. C. (2003). Summarizing can improve metacomprehension accuracy. *Contemporary Educational Psychology, 28*(2), 129-160.

Thiede, K. W., & Dunlosky, J. (1999). Toward a general model of self-regulated study: An analysis of selection of items for study and self-paced study time. *Journal of experimental psychology: Learning, Memory, and Cognition, 25*(4), 1024.

Thiede, K. W., Anderson, M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology, 95*(1), 66.

Thiede, K. W., Griffin, T. D., Wiley, J., & Anderson, M. C. (2010). Poor metacomprehension accuracy as a result of inappropriate cue use. *Discourse Processes, 47*(4), 331-362.

Thiede, K. W., Griffin, T. D., Wiley, J., & Redford, J. S. (2009). Metacognitive monitoring during and after reading. In *Handbook of metacognition in education* (pp. 97-118). Routledge.

Tighe, E. L., Kaldes, G., Talwar, A., Crossley, S. A., Greenberg, D., & Skalicky, S. (2021) Adults with low academic skills: Do struggling adult readers monitor their reading? Understanding the role of online and offline comprehension monitoring processes during reading [Conference presentation]. Society for the Scientific Study of Reading 2020, Virtual.

Townsend, C., & Heit, E. (2010). Metacognitive Judgments of Improvement are Uncorrelated with Learning Rate. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 32, No. 32).

Tullis, J. G., & Benjamin, A. S. (2011). On the effectiveness of self-paced learning. *Journal of Memory and Language, 64*(2), 109-118.

Urdan, T., Midgley, C., & Anderman, E. M. (1998). The role of classroom goal structure in students' use of self-handicapping strategies. *American Educational Research Journal, 35*(1), 101-122.

van den Broek, P., Bohn-Gettler, C. M., Kendeou, P., Carlson, S., & White, M. J. (2011). *When a reader meets a text: The role of standards of coherence in reading comprehension.* In M. T. McCrudden, J. P. Magliano, & G. Schraw (Eds.), *Text relevance and learning from text* (pp. 123–139). IAP Information Age Publishing.

van Dijk, T. A., and Kintsch, W. (1983) *Strategies of discourse comprehension*. New York: Academic Press.

Von Restorff, H. (1933). About the effect of area formations in the trace field. *Psychological Research* , *18* (1), 299-342.

Vygotsky, L. S. (1987). In RW Rieber & AS Carton (Eds) *The collected works of LS Vygotsky: Vol. 1. Problems of general psychology.* Plenum Press.

Wiley, J., Griffin, T. D., & Thiede, K. W. (2005). Putting the comprehension in metacomprehension. *The Journal of General Psychology, 132*(4), 408-428.

Witherby, A. E., & Tauber, S. K. (2017). The influence of judgments of learning on long-term learning and short-term performance. *Journal of Applied Research in Memory and Cognition*, *6*(4), 496-503.

Wolfe, M. B. (2005). Memory for narrative and expository text: independent influences of semantic associations and text organization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(2), 359.

Xu, J., & Metcalfe, J. (2016). Studying in the region of proximal learning reduces mind wandering. *Memory & Cognition, 44*(5), 681-695.

Yeari, M., van den Broek, P. & Oudega, M. (2015). Processing and memory of central versus peripheral information as a function of reading goals: Evidence from eye movements. *Reading and Writing*, *28*(1), 1071–1097.

Zabrucky, K. M. (2010). Knowing what we know and do not know: Educational and real world implications. *Procedia-Social and Behavioral Sciences, 2*(2), 1266-1269.

Zwaan, R. A. (1994). Effect of genre expectations on text comprehension. *Journal of experimental psychology: learning, memory, and cognition*, *20*(4), 920.