Guerrero, T. A., Griffin, T. D., & Wiley, J. (2022). I think I was wrong: The effect of making experimental predictions on learning about theories from psychology textbook excerpts. *Metacognition & Learning, 17*(2), 337-373. https://doi.org/10.1007/s11409-021-09276-6

I Think I Was Wrong: The Effect of Making Experimental Predictions on Learning about Theories from Psychology Textbook Excerpts

Tricia A. Guerrero, Thomas D. Griffin, & Jennifer Wiley

University of Illinois at Chicago

Acknowledgements The authors wish to thank Marta K. Mielicki and M. Anne Britt for their contributions to this project. We also wish to thank Tom Doonan and Gabriella Lazinek for their assistance in data collection and coding.

Funding The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A160008. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Abstract

Students often struggle with developing understanding from expository science texts. This study explored whether training students to engage in a POE (Predict-Observe-Explain) study strategy might be beneficial when learning from texts that introduce theories by describing experiments and empirical results, a common style in social science textbooks. The main questions tested in this experiment were if training students how to use a POE study strategy while reading textbook excerpts would support better comprehension and comprehension monitoring outcomes when students engaged in future learning attempts for an introductory psychology class. In one condition students were trained to use the POE study strategy, while in a comparison condition students were simply trained to use an explanation study strategy. Analyses suggested that students in the POE strategy training condition may have become preoccupied with whether or not their experimental predictions were correct, prohibiting them from engaging with the POE strategy as intended. Although both POE and explanation strategy training helped students to improve their comprehension monitoring on a new set of texts, students in the explanation condition displayed better comprehension on those new texts than students in the POE condition.

Keywords: Prediction, Explanation, Learning from Text, Metacomprehension, Comprehension

I Think I Was Wrong: The Effect of Making Experimental Predictions on Learning about Theories from Psychology Textbook Excerpts

Although textbook reading assignments are a common part of instruction in many gateway science courses, undergraduate students can struggle with developing understanding from expository science texts. Past work in text comprehension suggests that compared to reading stories or narratives, the processing of expository science texts is more challenging for a variety of reasons (Graesser, 1981; Narvaez, van den Broek, & Ruiz, 1999; Wiley, Griffin, & Thiede, 2005). Differences in the comprehension of narrative and expository genres have generally been discussed in terms of difficulty, as expository texts usually deal with more technical and less familiar subject matter than narrative texts (Graesser, McNamara, & Louwerse, 2003; Lin & Zabrucky, 1998; Weaver, 1990). Starting even at the basic level of vocabulary, expository texts are more likely to use low-frequency words which is one main reason why narrative texts are read faster and more easily (McNamara, Graesser, & Louwerse, 2012).

When reading narrative texts, people possess a great deal of knowledge that they can bring to bear. The information conveyed in stories shares great similarity to experiences in everyday life describing events that occur in space and time, and characters who perform actions in pursuit of goals or due to their motivations or intentions (Graesser, Singer, & Trabasso, 1994). In contrast, expository text is often written to inform students about unfamiliar ideas. By definition, students will have much less relevant prior knowledge about the ideas presented in expository texts that are meant to convey new knowledge in order to help them learn about science topics. Expository texts can describe complex systems, articulate causal mechanisms, or provide justifications for theories. Much of this content is abstract and technical. Because expository texts require more specific background knowledge than narratives to be understood (Graesser & Bertus, 1998; van den Broek, Virtue, Everson, Tzeng, & Sung, 2002), a lack of sufficient topic knowledge is one key contributor to the difficulty that students face when processing expository texts, and learning from expository texts is generally found to be a joint function of prior knowledge and reading skill (Alexander & Judy, 1988; Kintsch, 1994; McNamara, Kintsch, Songer, & Kintsch, 1996; Shapiro, 2004; Voss & Silfies, 1996). Especially because of the absence of topic knowledge, processing expository text can require more effort and rely more on working memory to keep information active as students attempt to generate inferences (Budd, Whitney, & Turley, 1995; Linderholm & van den Broek, 2002; Wiley & Myers, 2003).

Like all text-processing activities, comprehension of expository texts theoretically requires the construction of multiple levels of representation which include the *textbase* and the *situation model* (Kintsch, 1998). The *textbase* is a representation of the propositions explicitly stated in the text. The *situation model* builds on the *textbase* as the reader generates inferences or connections across sentences and recognizes implicit relations among propositions using prior knowledge. The *situation model* is constructed as the reader attempts to create a coherent, integrated representation of what the text is about. It is the level of representation that best represents understanding of a topic and underlies the ability to answer inference and application questions after reading (Kintsch, 1994). Together, the construction of the *textbase* and *situation model* represent two key stages in the process of text comprehension. Yet, given the generally poorer comprehension outcomes that are seen with expository text, it appears many students may fail to engage in the inference generation processes that are required to construct a coherent situation model.

When students do not have much prior knowledge about a topic, then they may find developing even just the *textbase* representation to be challenging. Many approaches have been used to help support students in this respect including instructing students to read a text more than once (Britt & Sommer, 2004; Griffin, Wiley, & Thiede, 2008; Millis, Simon, & tenBroek, 1998; Rawson, Dunlosky, & Thiede, 2000). The logic behind this approach is that during the first reading of a text the students may be devoting more of their attention to low-level processes such as decoding and parsing (Perfetti, 1985), and attempting to represent the meaning of individual sentences (Millis, Magliano, & Todaro, 2006). Other instructional conditions have been designed to more directly support the development of situation models. Encouraging students to be more active and constructive as they study has been one of the more effective ways to improve learning from science text (Chi, 2009; Dunlosky et al., 2013; Kintsch, 1994). In contrast to simply engaging in re-reading (re-exposure to information) or recall (reproduction of information), constructive activities such as generating sketches, questions, or explanations are generally more beneficial for comprehension (Ainsworth & Th Loizou, 2003; Butcher, 2006; Chi, 2000; Davey & McBride, 1986; Hinze, Wiley, & Pellegrino, 2013; King, 1994; McNamara, 2004; Wiley & Voss, 1999; Wiley, 2019).

For similar reasons, properties of the text, such as its cohesion, may become more important for comprehension when students possess relatively little prior knowledge of a topic (McNamara et al., 1996). When more overlapping terms are used across sentences, the text is more cohesive and lower-knowledge readers can process it more easily. Likewise, the presence of more explicit discourse markers such as conjunctions may be needed to guide the processing of expository texts on unfamiliar topics (Singer & O'Connell, 2003; Wiley & Myers,

2003). Yet, many expository texts lack the necessary cohesion and signaling that are needed for students to successfully comprehend them. A related possibility is that students may have different reading goals as they approach narrative and expository text (Linderholm & van den Broek, 2002; Narvaez et al., 1999). Even when markers are present, students may not utilize them effectively unless they are reading with a goal of comprehension. Without a comprehension goal, expository texts may be processed on a descriptive level (Britton & Black, 1985) or in a more item-specific as opposed to global, relational manner (Einstein, McDaniel, Bowers, & Stevens, 1984). It may be that while reading students may attempt to memorize as much as possible, rather than attempting to draw inferences among ideas (Wiley et al., 2005). Finally, students may be challenged by a lack of familiarity with the structure of expository texts. The structure of a text can provide important cues about the processes that one can use to understand it. But, this requires knowledge and recognition of the specific discourse structure on the part of the reader. In Psychology textbooks, a particular social science discourse style is often used which uses references to empirical studies and other forms of evidence in order to support theories. Without explicit training in how to read this particular type of text, it seems likely that students will experience difficulty when tasked with learning from textbook reading assignments in the gateway Introduction to Psychology course.

Expository Text Structures and Sub-Genres of Science Writing

A lack of familiarity with the genres or structures involved in science writing and a lack of instruction about goals for comprehension for different types of science writing and in different disciplines are important reasons why students may struggle in developing understanding (Yore, Bisanz, & Hand, 2003). Stories or narrative texts generally describe events and the goals, thoughts and actions of characters, using a familiar structure that people regularly employ to communicate their own experiences from a very young age. In comparison to the structure of narrative texts, the structure of scientific text tends to be less uniform, less obvious, and less familiar to readers. There are many different rhetorical structures used in expository texts (Cook & Mayer, 1988; Kintsch & Yarbrough, 1982; Lorch, 2015; Meyer & Freedle, 1984) and a variety of sub-genres of science writing in the natural and social sciences (Martin, 1993). Encyclopedia entries or primary school textbooks may use a more informational writing style with the goal of describing or cataloging characteristics of an object or organism. Much of the early work on expository text processing in the 1980s used simple descriptive informational texts as stimuli. The goal when learning from these types of texts may be to read for retention of facts, and the student may not need to engage in much deeper processing for these texts. Other texts, such as history texts, may be written in a linear, chronological

form. These may invite inferences about motives or goals of agents, or triggering events as causes of other events, similar to those that may be made with fictional narrative texts. Expository texts can also be explanatory, with a goal of providing an overview of the steps of a process or causes of a phenomenon. This structure is common in natural science texts. The inferences that are required to understand these texts are primarily causal. Past work in text comprehension suggests that compared to stories, readers are much less likely to draw causal inferences from explanatory science texts than from narrative texts (Graesser, 1981; Millis & Graesser, 1994; Noordman, Vonk, & Kempff, 1992; Singer, Harkness, & Stewart, 1997; Wiley & Myers, 2003). Other kinds of expository science texts that students may encounter include refutation texts that explicitly refute a misconception and explain why a prevailing theory or concept is more correct (Diakadoy et al., 2011; Dole, 2000; Mason et al. 2017), or argumentative texts about controversial issues, where the goal is often to persuade the reader of a position or stance on a topic such as whether watching TV violence causes real violence, whether genetically modified foods are safe, or whether asteroids caused the extinction of the dinosaurs (Iordanou, Kendeou, & Beker, 2016; Mason & Boscolo, 2004; Wolfe, Tanner, & Taylor, 2013).

There is now a substantial body of research on improving comprehension from expository science texts that has focused on improving the likelihood of drawing causal inferences from explanatory texts (Ainsworth & Loizou, 2003; Butcher, 2006; Graesser & Bertus, 1998; Hinze, Wiley, & Pellegrino, 2013; Linderholm & van den Broek, 2002; Millis & Graesser, 1994, Otero, León, & Graesser, 2002; Singer et al., 1997; Singer & O'Connell, 2003). The most commonly researched constructive activities involve having students generate some form of explanation as they study (Chi, 2000). Prompting students to generate explanations can improve their comprehension from text in multiple ways. On one level, during the process of generating explanations many students initially tend to begin by producing paraphrases or summaries of what they read, which can help students to establish a more complete textbase. Building a more complete representation of the textbase can be useful, especially for low-knowledge or low-skilled readers (McNamara, 2017). It can provide them with a basis from which they can begin to construct a deeper understanding of the text (Millis et al., 2006). Additionally, as the process of generating explanations prompts the student to go beyond simply summarizing or paraphrasing. It helps the student to make new connections and to identify implicit relationships between ideas, in other words, to generate inferences. This helps the student to construct a more well-developed situation model or mental model of the phenomenon being described by the text (Chi, 2000; Kintsch, 1986, 1994; Mayer, 1989). Encouraging students to

engage in these activities while learning from explanatory science texts can improve the likelihood that bridging or causal inferences will be constructed. As a result, students achieve better comprehension of explanatory science texts.

In contrast, the ability of students to learn from and comprehend other types of scientific discourse has received much less attention, including scientific journal articles that report the results of empirical research (Bazerman, 1985; Samuels et al. 1988; Yore et al., 2003), and a subset of argumentative texts in which claims or theories are discussed in relation to evidence (Britt, Richter, & Rouet, 2014). These forms seem most similar to the style that is typically used in Psychology textbooks. The lack of familiarity with this disciplinary-specific discourse style is very important to note because the structure of the text, and the specific sorts of inferences that students need to generate to develop an understanding of the content are deeply intertwined. For example, a common treatment of the topic of Fundamental Attribution Error in an introductory psychology textbook will not only attempt to describe the theoretical construct with a definition, but will also provide examples of empirical studies that tested it. While disciplinary experts may appreciate that the important relations to attend to when reading these types of texts are those among theories, predictions, hypotheses, results, and support for theoretical constructs or claims as provided by evidence, few students come to college already familiar with the conventions of these sub-genres (Bazerman, 1985; Berkenkotter & Huckin, 1995; Larson, Britt, & Larson, 2004; Yore et al., 2003). It is unlikely that students will engage in the sort of predictive, hypothetical, and inductive thinking that will allow them to achieve a deeper understanding of the theories and evidence. In contrast to literacy research on reciprocal teaching that has explored benefits from having students ask each other to predict “what will happen next” in a story (Palinscar & Brown, 1984; Rosenshine & Meister, 1994), the predictive inferences that are of interest for these types of expository texts are those that require students to reason forward from the information in the excerpt to make a specific and informed hypothesis about the outcome of an empirical research study. The goal is for students understand which pattern of results would be consistent with and provide support for the theory. Although encouraging students to engage in explanation has been shown to improve the likelihood that some logical and causal inferences will be constructed, it is possible that a more targeted explanation prompts may be needed to better support the specific sorts of inferences that are required to understand Psychology textbook excerpts that describe theories and evidence. Borrowing from research on learning from inquiry activities in science, Carvalho, Manke, and Koedinger (2018) have suggested that encouraging students to engage in a prediction-observation-explanation cycle while reading could be a way to focus

students on the kinds of reasoning and inferences that would support better understanding of theory-evidence relations in these types of texts.

Understanding Science as an Inquiry Process and the Predict-Observe-Explain (POE) Learning Cycle

In parallel to the movement toward comprehension activities that promote more constructive processing on the part of the reader in research on learning from expository text, there has also been a movement toward more active, inquiry-based learning activities in research on science education. In contrast to science instruction that focusses on students' acquiring knowledge of discrete facts, recent trends have promoted approaches that focus more on how science is actually conducted (National Research Council [NRC], 2012). If the goal is for students to become educated consumers of science, then students will benefit from understanding the practice of science as an inquiry process (Next Generation Science Standards [NGSS], 2013). These approaches involve having students take a more active role as learners so that they can develop a clearer understanding of the scientific process: how to develop a research question and test it, how theories are supported by results from experiments, as well as how to identify limitations of empirical studies and how results can ultimately lead to new research questions.

One instantiation of such an approach is the Predict-Observe-Explain (POE) learning cycle which has been used in science education as part of hands-on experimentation and inquiry activities (Champagne, Klopfer, & Anderson, 1980; White & Frederiksen, 1998; White & Gunstone, 1992). This process typically consists of presenting students with a problem statement, asking them to make predictions, followed by having them observe phenomena under various experimental conditions, and finally explaining how the results support theories (White & Gunstone, 1992). For example, in one POE study about understanding theories of motion, White and Frederiksen (1998) had students toss a ball to one another. While they were doing this the teacher asked them to generate factors that might be involved in the motion of the ball being tossed, and they were prompted to respond to the following predictive question:

> Imagine that a ball is stopped on a frictionless surface. Suppose that you hit the ball. Then, right after the hit, you hit the ball again in the opposite direction with the same size hit. Would the hit in the opposite direction change the velocity of the ball? If so, describe how it would change and explain why.

The students were then asked to present and explain their predictions, and discuss what might happen with the class. Using computer simulations and real-life experimentation, students then ran experiments to examine the concept of motion. After multiple iterations of the experiments, students developed laws to capture their

observations. In a final step, students were asked to apply their laws of motion to a new hypothetical situation such as how their laws of motion might be used when playing soccer. Students who were taught using this POE learning cycle outperformed students who were taught using a traditional physics curriculum when solving physics problems requiring the application of Newtonian principles to determine movement.

Since the POE cycle was first proposed as an instructional device by Champagne et al. (1980), meta-analyses have shown inquiry activities such as these which require the student to draw conclusions based on evidence and theories to be quite effective in improving learning (Furtak, Seidel, Iverson, & Briggs, 2012). The benefits of POE as part of hands-on experiments (including inquiry activities using both real objects and virtual simulations, de Jong, Linn, & Zacharias, 2013) have been shown across a range of domains in science and engineering (Bolger, Kobiela, Weinberg, & Lehrer, 2012; Chang et al., 2013; Hmelo-Silver, Duncan, & Chinn, 2007; Lehrer & Schauble, 1998; Liew & Treagust, 1995; Triona & Klahr, 2003; White & Frederiksen, 1998). At the same time, research has shown that students cannot be expected to adopt these practices on their own. Despite the prevailing emphasis of the constructivist view of science learning on student-centered learning, it must be acknowledged that students are novices (at best) at the practice of doing and engaging in science learning, and are likely to need guidance and support in order to engage in these activities effectively (Burbules & Linn, 1991; Gil-Pérez et al., 2002; Magoon, 1977).

Applying the POE Strategy to Learning from Text

Even though the POE learning cycle was developed in the context of hands-on learning activities in physical science domains, benefits might also be seen when students are tasked with learning about theories from text in the social sciences because the prediction activity should direct student attention to the theories and evidence in the text, and specifically support the generation of inferences about theory-evidence relations which are the most relevant for understanding. In an initial study that applied this approach with learning from text, Carvalho et al. (2018) had students engage in POE activities when learning about various social psychology topics from textbook excerpts that used descriptions of experimental results to provide evidence for theories. In the POE activity condition, after reading the first part of each text that introduced the theory and the experimental design for a study, students were asked to make a prediction about which of 3 possible outcomes the results of the experiment should be. For example, if an experiment had two conditions, the 3 possible outcomes were that Condition A would do better, Condition B would do better, or there would be no difference. Students were then prompted to explain their

choice. After making and explaining their prediction, they were shown a graph of the results and were asked to explain what they thought the results showed and why. Finally, they read the last part of the text that indicated the conclusions reached by the researchers. In contrast, in the comparison condition students read the same texts and answered the same questions, but the timing and perspective of these questions was different. In the comparison condition, the questions asked students to describe the researchers' predictions (instead of their own), the reasons that the researchers made those predictions, and what the researchers thought the results showed and why. Further, students did not answer these questions until they finished reading all parts of the text, and all of the answers were explicitly mentioned in the text. The main result of this study was that students performed better on tests when the topics were learned through POE activities than in the comparison condition. This result suggests that engaging in POE might be a valuable strategy to teach students to use when they are reading expository texts about theories and evidence. Assessing the potential benefits of training students to use a prediction generation study strategy to improve their future learning on new topics that were assigned after the training was the primary goal of the present work.

To provide clear model of how students should engage in prediction generation activities, in this study the instructions highlighted the theory-evidence structure of the excerpts so that the students could see the relevance and purpose of the POE strategy, what they were being asked to predict (the results of research studies), and what their prediction should be based upon (the logic of the theories in the text and the design of the empirical studies as opposed to their general intuition). Further, it was hoped that by directly connecting the POE strategy to the text structure that this would increase the likelihood that students might focus on theory-evidence relations on their own in the future. If students do not have an epistemic appreciation for the importance of the argumentative structure of these excerpts, then they are unlikely to engage in the sort of predictive, hypothetical, and inductive thinking that will allow them to achieve a deeper understanding of the content.

The Current Study

The current study explored whether training students to generate hypothetical predictions while reading about experiments and theories might improve future learning from expository social science texts. The two main questions tested in this study were whether training students how to use a prediction generation strategy while reading textbook excerpts would support (1) better comprehension and (2) better comprehension monitoring when students engaged in future learning attempts for an introductory psychology class. The effects on future learning

were explored by comparing students who were trained to use a POE study strategy versus an explanation study strategy in terms of their performance on comprehension tests given after they studied a set of 6 new textbook excerpts. The effects on comprehension monitoring accuracy were assessed by examining students' ability to correctly estimate their understanding after studying the new set of topics, as well as how this changed from before to after the study strategy training activity for each of the conditions.

Why a POE Study Strategy Might be Beneficial for Comprehension

The current study explored whether training students to use the underlying theory-evidence structure of the textbook excerpts to generate hypothetical predictions while reading might improve future learning from social science textbook excerpts. Explanation activities have been shown to be one of the more effective ways of improving comprehension from expository texts. However, benefits of explanation may vary due to the type of inferences required for comprehension. Explanation activities may help with bridging and causal inferences, but they may be less effective for encouraging forward inferences or promoting the kind of reasoning that is required to think hypothetically. Thinking hypothetically requires consideration of the relation between theories and evidence. Because the POE study strategy adds a prediction generation phase, it may help to direct the reader's attention to the key relation between theories and evidence. In this way, POE can be thought of as a guided explanation activity.

For social science texts that describe theories and experiments, a prediction strategy could be expected to improve comprehension. The prompts used in POE may help to direct the reader in generating the most appropriate types of inferences for this specific type of text. It should be beneficial for helping students to reflect more deeply on the theories they are learning about, and to better understand relations between hypotheses, designs, and results of studies. It may prompt them to build connections between examples, studies, and theories, and generate the inductive and abductive inferences that one needs to make predictions. Prompting a student to make a prediction before providing an explanation for an outcome may help to make the student more active along the continuum from passive exposure toward more constructive processing. And, if students learn to become more active readers, this should result in better comprehension outcomes during future opportunities for learning from text. To summarize, if engaging in a prediction generation activity as part of reading social science textbook excerpts helps to direct the student's attention to the key relations between theories and evidence, then this activity could also be expected to further improve comprehension beyond the benefits of explanation.

Why a POE Study Strategy Might be Beneficial for Comprehension Monitoring

Generative activities also provide the conditions that have been found to most robustly support better metacomprehension and comprehension monitoring outcomes. Although students' tend to be poor at monitoring their understanding from expository science texts (Griffin, Mielicki, & Wiley, 2019; Maki, 1998b; Thiede, Griffin, Wiley, & Redford, 2009), multiple studies have shown that engaging in generative activities such as sketching while reading, or attempting to explain how and why a scientific phenomenon occurs after reading, can improve students' ability to monitor their understanding from explanatory science texts (Fukaya, 2013; Griffin, Wiley, & Thiede, 2008, 2019; Jaeger & Wiley, 2014; Wiley, 2019). For social science texts that describe theories and experiments, the process of attempting to generate a hypothetical prediction should increase access to cues that are a valid reflection of the quality of one's mental model (Griffin et al., 2008). These cues could then be used to formulate accurate judgments about the status of one's understanding or comprehension (JOCs). Again, if engaging in a prediction activity as part of reading helps to direct the reader's attention to the key relations between theories and evidence that are critical for understanding these textbook excerpts, then this activity could also be expected to improve comprehension monitoring accuracy.

When students are given a set of topics to learn, three distinct measures of monitoring accuracy from JOCs can be computed, capturing unique aspects of judgments that are relevant to self-regulated learning (Griffin et al. 2008; Maki, 1998a). Confidence bias is computed as the average signed difference between the JOCs and the corresponding test performance scores, whereas absolute error is the average absolute difference between JOCs and test performance and captures only the magnitude of judgment errors. Confidence bias is only partially dependent upon the magnitude of judgment errors, because it also captures whether most people in a sample share a similar directional bias in their judgments (Yates, 1990). Since most learners appear to suffer from overconfidence where judgments are higher than actual test performance (Dunlosky & Rawson, 2012), it is useful to know whether an intervention reduces this bias. However, since the proportion of a sample who are overconfident can be reduced without reducing (or even while increasing) the average magnitude of judgement errors, bias is best interpreted in relation to the corresponding absolute error measure. Critiques of these measures as being a poor reflection of monitoring subjective experiences unique to specific learning episodes have still acknowledged that learners' ability to accurately predict their absolute level of performance has pragmatic utility for effective self-regulated learning, such as informing decisions about whether additional or different types of study efforts are needed to achieve some benchmark or goal (Dunlosky & Rawson, 2012; Griffin et al., 2019). The third measure of judgment accuracy,

relative accuracy, is designed to be orthogonal to average absolute levels of either judgments or test performance (Nelson, 1984), and thus may be optimally sensitive to monitoring of experiences that vary from one specific learning episode to the next. Relative accuracy is computed as the intra-individual correlation between JOCs and actual test performance across topics and captures the ability to judge which topics were understood better than others. Whereas confidence bias and absolute error are relevant to whether students decide to persist in studying, better relative accuracy can help students to direct their attention to the topics where restudy is most needed. The current study tested whether these instructional conditions might reduce typical overconfidence while supporting both absolute and relative comprehension monitoring accuracy.

Why a POE Study Strategy Might not be Beneficial for Comprehension or Comprehension Monitoring

Despite these reasons why a POE study strategy should be beneficial for future learning from expository social science texts that describe theories and evidence, there is also the possibility that adding the prediction generation phase could make it less effective than a more general explanation study strategy. While Carvalho et al. (2018) found overall benefits from using a POE learning cycle with expository social science texts, they also found that the effectiveness of the activity varied with the accuracy of the prediction that was made. Students who made incorrect predictions demonstrated worse understanding on the final test. This result raises the possibility that a POE strategy may not be beneficial for comprehension outcomes unless students are able to generate a correct prediction. A further concern is that students could become preoccupied with the accuracy of their own predictions, and this could distract them from mental model construction and derail their comprehension monitoring. Or, students may fail to use the information from the text as a basis for their predictions, and instead rely on their intuition, which would defeat the intended purpose of the activity. The current study provided a test between these alternative hypotheses, and explored whether training students to use a POE study strategy would help to support learning from social science textbook excerpts that describe theories and evidence, or if it might undermine it.

Finally, other work suggests that the quality of the reasons that are produced when students are prompted to explain their reasoning as part of the POE cycle may mediate the benefits of a POE activity. In a hands-on study where students were learning about levers, Bolger et al. (2012) found that when students focused on the relations or connections between parts of the mechanical system to generate their hypotheses, they performed better on a prediction activity than students who focused on more superficial or individual features as a basis for their predictions. Baddock and Bucat (2008) also reported that few of their students were able to articulate the key

relations when prompted to explain the results of a hands-on activity. To further explore this potential mediator, the quality of the written responses given during the study strategy training activity were coded and analyzed in this research.

Method

Participants

Undergraduates (N = 358, 170 females) in Introduction to Psychology completed a series of online homework activities as part of their required course assignments. The Introduction to Psychology course was chosen as a target for this study because it serves as a gateway science course and is generally one of the first science courses taken when students enter college. An additional 158 students did not complete the course or all three activities. Students received course credit for the completion of the activities. The sample reported their average age as 19.71 (SD = 2.57), and racial composition as 23.2% Hispanic, 49.8% White, 27.7% Black, and 9.5% Asian.

To minimize any discussion of the different activities between students, assignment to condition was done at the section level. The 14 sections were taught by 7 different teaching assistants (TA). Each TA taught 2 sections. To minimize any effects of TA on the manipulation, their 2 sections were assigned to separate conditions.

Research Design

An overview of the between-subjects design is shown in Figure 1. During the first week of the semester, prior to any content-based instruction, students completed a baseline assessment of their domain-specific comprehension skill. The following week, students were trained to use either a POE or explanation study strategy to support them in reading psychology texts for understanding. Finally, during the following week, students completed a learning activity to measure the effect of the study strategy training on future comprehension and comprehension monitoring.

Materials

The materials for this study included two sets of textbook excerpts and inference test questions. One set was used for the baseline assessment and the study strategy training activity. The second set was used for the learning activity that followed training.

Baseline Assessment Text and Test Set

The baseline assessment tested for domain-specific comprehension skills by asking students to learn from psychology textbook excerpts on 6 topics, and tested their understanding with inference questions for each text. This

baseline assessment given before students engaged in training was intended as a way to control for individual differences and variance in domain-specific comprehension skills of the students who were assigned to each condition. Within the text comprehension literature, performance is generally found to be a joint function of prior knowledge and reading skill (Alexander & Judy, 1988; Kendeou & van den Broek, 2007: Kintsch, 1994; McNamara et al., 1996; Shapiro, 2004; Voss & Silfies, 1996). Given the limitations of conducting the study in a real course context, a separate non-course-related reading comprehension assessment could not be administered. Instead, the baseline comprehension task utilized texts and topics from within the course. By using a student's ability to understand a set of domain-specific texts and answer inference questions about those texts before training, this task captured a student's baseline ability to learn from textbook excerpts in the course which would be a joint product of reading skill and prior topic knowledge. The goal of the baseline assessment was to obtain a covariate that could be used so that any differences between the conditions could be attributed to the manipulations and not pre-existing differences in the ability to learn from texts in this course.

The baseline assessment included psychology textbook excerpts on 6 topics (Placebo Effect, Confirmation Bias, Self-Control, Conformity and Obedience, Fundamental Attribution Error, and Cognitive Dissonance). The average length of the excerpts was approximately 800 words with Flesch Kincaid Grade Levels ranging from 10.5 to 13.5. All of the textbook excerpts were adapted to follow a specific structure. The first paragraph began by presenting a real-life example of the theory or phenomenon followed by a formal definition or description of the concept. Each text then described the results of two empirical research studies that provided support for the theory being described.

Five multiple-choice test inference questions were written to test comprehension of each excerpt. The test questions were designed to measure understanding of the concepts, not just verbatim memory for the material. Answers to these questions were not presented explicitly anywhere in the text and instead required the reader to generate inferences. The inference questions addressed the implicit relationships among ideas in the text, tested for connections among concepts, or asked students to apply their understanding of the concept to a new context. Because these test questions were inference-based, the baseline assessment provided a measure of students' ability to engage in inferencing when reading in this course. For example, the following question from the Cognitive Dissonance text asked students to compare conditions across two studies that had been described in the excerpt:

Which group in the Festinger and Carlsmith (boring experiment) study is most similar to the severe warning group in the Aronson and Carlsmith (kids and toys) study?

A. The control group that did not have to recommend the experiment to other students.

B. The group that was paid $20 to recommend the experiment to other students.

C. The group that was paid $1 to recommend the experiment to other students.

D. The group that was told how fun and exciting the study was by other students.

This relationship between the results of the two empirical studies was not referenced explicitly in the text. To answer this question, the reader must draw upon the individual outcomes of each condition in the studies (which were described in separate paragraphs), then make a comparison of their similarity and their mapping onto the theoretical variables described in the text. Thus, this question required conceptual understanding of the excerpt.

Because this baseline assessment was intended to provide a measure of individual differences in domain-specific comprehension skill, one way of indexing the measurement quality of the assessment was by computing Cronbach's alpha. The internal consistency of the test items was Cronbach's $\alpha = .76$. In addition, norming studies using independent samples showed significant positive correlations of performance on these inference items with self-reported ACT scores, demonstrating convergent validity with an established standardized measure. Descriptive statistics for the 6 textbook excerpts and scores for each set of inference test questions, including these correlations, are reported in the Appendix.

Learning Activity Text and Test Set

The learning activity text and test set included 6 new textbook excerpts that were adapted to follow the same structure as those used in the baseline assessment. They were on 6 new course topics (Classical Conditioning, Operant Conditioning, Observational Learning, False Memories, Twin Studies, and Aphasias). The new excerpts were also approximately 800 words in length with Flesch Kincaid grade levels ranging from 10.1 to 14.2, which did not significantly differ from the set used on the baseline assessment, *t*s < .71.

Five multiple-choice test inference questions were written to test comprehension of each topic. Because the test items were not designed to measure understanding of the same exact concept multiple times, a measure of internal reliability is not an appropriate method of demonstrating measurement quality on these assessments (Taber, 2018). Instead, measurement quality was demonstrated by providing evidence of convergent validity. Norming studies using independent samples showed significant positive correlations between performance on the inference

tests for each individual text after a single reading and scores on a standardized assessment (ACT). Descriptive statistics for the 6 textbook excerpts and scores for each set of inference test questions, including these correlations, are reported in the Appendix.

Procedure

All activities were administered as online homework assignments through the Qualtrics survey platform and were available for students to complete at any time during the assigned week. Students were asked to complete each assignment individually. As shown in Figure 1, during the first week of the semester, prior to any content-based instruction, students completed the baseline assessment which tested domain-specific comprehension skill. The following week, students completed a training activity to learn how to use either a POE or Explanation study strategy to support them in reading psychology texts for understanding. Finally, during the following week, students completed the learning activity to measure the effect of the study strategy training they had received on their future comprehension and comprehension monitoring.

The baseline assessment given during the first week consisted of 4 phases. First, students were asked to read the 6 excerpts. They were told to study the excerpts in the same way they usually study for a class and that they should expect to answer questions about the texts after reading. Second, following the reading phase students were asked to make judgments of their comprehension (JOCs) on a 0-5 scale for each text. They were told, "You will take a multiple-choice test on these texts. How many questions out of 5 do you think you will get correct?". Third, they completed the multiple-choice inference tests containing 5 comprehension questions per text topic in the same order as they were read. Finally, after completing all test questions, students were shown the correct answers to each of the test questions (along with the answers they gave) and were asked to assess whether their answers were correct. This correct-answer feedback was given to ensure that incorrect responses were not carried forward into their learning of the topic for the course (Butler & Roediger, 2008).

During the second week, students completed a study strategy training activity which used the texts from the baseline assessment. Half of the students received an explanation strategy training activity while the other half received a POE strategy training activity.

Explanation Strategy Training Activity. This procedure used in this condition was based upon Griffin et al. (2008). The explanation strategy training activity began with instructions about how to read psychology texts and students were provided with a list of general questions to think about when reading each sentence:

What does this mean?

What new information does this add?

How does this information relate to the title?

How does this information relate to previous sentences or paragraphs?

Does this information provide important insights into the major theme of the text?

Does this sentence or paragraph raise new questions in your mind?

The students were also told that they would be asked to generate explanations as they were reading. After reading the first section of each text students were prompted to write an explanation. The remainder of the text was then presented. At the end of the text, students were again asked to write an explanation. The same procedure was used for each of the 6 excerpts.

POE Strategy Training Activity. As shown in Figure 1, the POE strategy training activity began with a more specific instruction about how to read psychology texts. Instead of a list of general questions to think about while reading, students were directed to consider a common expository structure used in psychology textbooks that first introduces a theory or phenomenon and then describes empirical evidence that supports the theory. Students were told that the goal for this expository structure is to increase students' understanding of the relationship between a theory and the empirical evidence for it. The students were also told that they would be asked to generate predictions of new experimental results as they were reading. The instructions for the POE activity were designed to help support understanding of how to engage in the prediction generation activity. To do this, it drew students' attention to the relationship between the presented experiments and the theory in the text, and showed them how the theory and results from experiments could be used to generate predictions for other experiments. Although the POE cycle closely approximates the general scientific method which all students may be somewhat familiar with from their previous science education, it cannot be assumed that all students would make this connection, or know how to apply principles of the scientific method as they are reading an expository text (Burbules & Linn, 1991; Iordanou et al., 2016). The instruction in the POE condition was designed to help all students to understand how they could use the information from the text to make an experimental or hypothetical prediction, and that they should attempt to engage in this process while reading this type of expository text.

After receiving the initial instructions, students were guided through an example of engaging in the POE study strategy using the textbook excerpt on self-control. After reading the first section of the text (describing the

concept of self-control, the first empirical study, and the description of the experimental design for the second empirical study), students were asked to make a prediction about the outcome of the second empirical study. They were given a list of three possible outcomes to select from. They were then asked to give their reasoning behind selecting that prediction. For just this first example text, the students then received feedback by viewing a good rationale for how the theory and the results of the first empirical study described in the text could be used to inform their prediction about the second. Then after reading the results of the second empirical study, they were prompted to provide a final explanation about how the results of the study provided support for the theory. Again, after writing their own explanation, they were shown another good response as a model response. The goal of providing this model on just the first example text was so that students could see a model of the type of reasoning they should engage in as part of using this POE strategy.

The POE activity then had students practice engaging in this predict-observe-explain study strategy for the remaining 5 texts on their own. Students read the first section of each text, ending with the design of the second empirical study. At this point, the students were asked to predict the outcome of the second study: "Given the concepts and theories discussed in the text, which of the following is the most likely result of this study?" Students selected from three possible outcomes. Similar to Carvalho et al., 2018, the options asked the students to predict which condition from the experimental study would be most likely to show the effect (e.g., Children would be most likely to reduce their opinion of their favorite toy, if they … received no warning/ received a mild warning/ received a harsh warning.). After selecting one of three options from a list, they were asked to "Explain why you think the prediction you made is correct." The remainder of the text was then presented which described the results of the second empirical study. At the end of the text, the student was asked to "Explain how these results support the theory described in the text." They were then told, "If your prediction was incorrect, explain why you think you got it wrong." This process was repeated for each of the remaining texts. After completing the POE activity for each text, students re-read two example test questions that they had seen during the previous week to further illustrate how the POE study strategy would help them to construct the implicit connections that were required to answer test questions.

Across conditions, most students spent slightly less than 75 minutes on the strategy training activity. Given that the goal of this activity was for students to learn how to use the strategy on their own during future reading, and given limited time, students were not tested again on their understanding of the training topics after completing these

study strategy training activities. Instead, the focus was on how this training might impact comprehension and comprehension monitoring on a new set of topics in the course.

During the third week, students engaged in the final learning activity. The main dependent measures for comprehension and comprehension monitoring outcomes were derived from this learning activity where students were asked to study and learn from a new set of textbook excerpts. The instructions for the learning activity prompted students to "Use the strategies and approaches that you learned about during prior online homework assignments to help you study." As in the baseline assessment, students were asked to make JOCs for each topic on the same 0-5 scale following the reading phase. Finally, they completed multiple-choice inference tests containing 5 comprehension questions per text topic in the same order as they were read.

For both the baseline assessment and the learning activity, comprehension test scores and JOCs were converted to proportions out of 5. Three measures of comprehension monitoring were computed from the relation between each students' JOCs and their actual comprehension test performance. Confidence bias was computed by subtracting average test performance from the average JOCs, with higher values indicating overconfidence in comprehension skills. Absolute accuracy was computed by taking the average of the absolute difference between each JOC and performance on each test, with larger values indicating more absolute error. Relative accuracy was computed as the intra-individual Pearson correlation between students' predictions of their test performance and their actual test performance (Griffin et al., 2008). As correlations become more positive and stronger, this indicates a more accurate ability to detect which textbook excerpts were understood better compared to others.

Results

Differences in Outcome Measures due to Study Strategy Training Activity

All analyses were performed using linear mixed-effects models (using lme4 package in R, Bates, Mächler, Bolker, & Walker, 2015). Effect sizes are estimated with Cohen's *d* for significant differences between means.

Comprehension Outcomes

The inference tests that were given as part of the learning activity served as a measure of how the training to use POE or explanation strategies while studying might support better comprehension on future reading assignments. Differences due to the training activity in comprehension outcomes were tested using linear mixed-effects models entering study strategy condition, domain-specific comprehension skill, and their interaction as fixed effects, and including intercepts for TA as a random effect. Domain-specific comprehension skill was included in

the model to ensure that any differences in comprehension test scores on the learning activity represented benefits of the intervention and were not due to pre-existing differences between conditions. Domain-specific comprehension skill was indeed related to future learning, $\beta = .58$, $SE = .04$, $t = 13.53$, $p < .001$. Although performance on the two measures was positively correlated ($r = .57$, $p < .001$), variance inflation factors (< 1.01) indicated that multicollinearity was not an issue.

An effect was seen for study strategy training condition, $\beta = .44$, $SE = .15$, $t = 3.02$, $p = .003$. Contrary to the hypothesis that engaging in POE might support better understanding than explanation alone, as shown in Table 1, better performance was seen in the explanation condition than the POE condition, Cohen's $d = .27$. The interaction between training condition and domain-specific comprehension skill was also significant, $\beta = -.34$, $SE = .15$, $t = -2.37$, $p = .02$. To better understand the interaction, the Johnson-Neyman technique was used. As shown in Figure 2, the interaction was due to students with lower domain-specific comprehension skill performing more poorly in the POE condition than in the explanation condition.

In addition, analyses were performed in an attempt to test for differences between the conditions in time on task. Because this was an unsupervised online study, the timing data needs to be interpreted with caution. Timing measures were derived from page submissions, and about 10% of the sample had very long times that were unlikely to represent actual time on the task (greater than 2 hours on tasks that were meant to take only 1 hour each). A 2-hour cutoff for extreme values was derived from Tukey's (1977) method for detecting outliers in boxplots using the interquartile range (IQR), and removing observations that exceeded 1.5 IQRs. Trimming using this method removed the longest 8% of reading times on the learning activity and longest 15% of times on the study strategy training activity. As shown in Table 1, these estimates suggested that the two study strategy training conditions did not differ in time spent on the training activity or the learning activity, ts < 1.

Comprehension Monitoring Outcomes

Measures of comprehension monitoring skills and accuracy were obtained both from the baseline assessment and from the learning activity. The effects on comprehension monitoring accuracy were assessed by examining students' ability to correctly estimate their performance (both absolute and relative levels) after studying the new set of topics, as well as how this changed from before to after the study strategy training activity for each of the conditions. Differences due to the training activity in JOC magnitude, absolute error, confidence bias, and relative metacomprehension accuracy were tested using linear mixed-effects models entering study strategy

condition, time of assessment, and their interaction as fixed effects, and including intercepts for subject and TA as random effects. Descriptive statistics are shown in Table 2.

First, the model for JOCs indicated an effect for time of assessment, β = -.09, *SE* = .03, *t* = -3.47, *p* < .001. Students became more conservative on the learning activity (*M* = .642, *SD* = .134) than they had been on the baseline assessment (*M* = .667, *SD* = .134), Cohen's *d* = .19. No differences were seen due to the particular strategy training activity that students engaged in, β = -.15, *SE* = .09, *t* = -1.59, *p* = .11. There was also no interaction, β = .11, *SE* = .09, *t* = 1.29, *p* = .20.

Second, absolute error was calculated as the average absolute difference between JOC magnitude and test performance for each topic. The model for absolute error indicated an effect for time of assessment, β = -.10, *SE* = .03, *t* = -2.87, *p* = .004. All students showed less error in their JOCs for the learning activity (*M* = .255, *SD* = .093) than they had in the baseline assessment (*M* = .274, *SD* = .100), Cohen's *d* = .20. No differences were seen due to the particular strategy training activity that students engaged in, β = .05, *SE* = .11, *t* = 0.49, *p* = .63, nor was there an interaction, β = -.06, *SE* = .11, *t* = -0.52, *p* = .60.

Third, confidence bias was calculated as the signed difference between the JOC magnitude and test performance for each topic and then averaged. The effect for time of assessment was weaker than that for absolute error, β = -.05, *SE* = .03, *t* = -1.87, *p* = .06. However, similar to the pattern seen for absolute error, all students tended to become less overconfident in their performance on the learning activity (*M* = .114, *SD* = .179) than during the baseline assessment (*M* = .132, *SD* = .180), Cohen's *d* = .10. Again, the decrease in overconfidence was not specific to a particular strategy training activity, β = .09, *SE* = .10, *t* = 0.92, *p* = .36, nor was there an interaction, β = -.12, *SE* = .09, *t* = -1.37, *p* = .17.

Finally, relative accuracy was computed as the intra-individual Pearson correlation between students' predictions of their test performance and their actual test performance. No differences were seen due to strategy training condition, β = -.14, *SE* = .12, *t* = -1.21, *p* = .23, nor was there an interaction, β = .15, *SE* = .12, *t* = 1.26, *p* = .21. However, there was an effect for time of assessment, β = .10, *SE* = .04, *t* = 2.74, *p* = .007. All students showed better relative metacomprehension accuracy for the learning activity (*M* = .136, *SD* = .441) than for the baseline assessment (*M* = .046, *SD* = .451), Cohen's *d* = .20. The effect size for the increase in relative accuracy from baseline to the learning activity in the explanation training condition was Cohen's *d* = .30, while the effect size in the prediction training condition was only Cohen's *d* = .11.

Overall, the results from the comprehension monitoring analyses indicated that although all students tended to be overconfident, students became more conservative following both POE and explanation study strategy. Similarly, although all students showed poor relative accuracy (near zero) at baseline, relative accuracy was better following both POE and explanation study strategy training. The results from both the comprehension and the comprehension monitoring analyses failed to demonstrate any advantage due to the POE study strategy training. In the case of the comprehension outcomes, the explanation training condition was clearly more effective at supporting learning from the new set of texts.

Exploratory Analyses

Several exploratory analyses attempted to discern why the POE study strategy training was not effective by more closely considering the behavior of the students during the training activities. These analyses were performed by coding the language used in the open-ended responses that students gave as they engaged in the strategy training activities. Because the responses that students gave after reading the outcome of the second study (at the end of each text) were more directly comparable across the two strategy training conditions, the nature of those open-ended responses is considered first. These responses were coded for two dimensions, the quality of the response in terms of understanding the theory and evidence, and then secondly, whether the response contained an evaluative comment. Additionally, the accuracy of the prediction made in the initial response is considered. Finally, a mediation analysis was conducted to understand the relationship between these dimensions of the responses.

Quality of Response Given After Second Study Result

Each open-ended response, given after the text provided the results of the second empirical study, was coded for whether the response contained evidence of understanding the result or theory. Quality was coded using a three-level rubric adapted from Guerrero and Wiley (2019), McNamara, Boonthum, Levinstein, and Millis (2007), and Hinze et al. (2013). A score of 0 was given to responses that were devoid of meaningful content (i.e., gibberish, incorrect, or irrelevant). A score of 1 was the modal response and represented what most students did on most texts. These responses provided a summary or paraphrase of the results of the study or the theory that repeated ideas from the text. A score of 2 was given to responses that went beyond correctly describing either the theory or a result by making a connection between the two, making a connection between the two studies discussed in the text, or by adding an elaboration such as a conditional, hypothetical, or new example. Each student wrote one response for each of the 5 texts resulting in 1,790 total responses. (Because the self-control text was used to demonstrate example

responses, students did not generate responses for this text.) Two independent coders both coded all 1,790 responses. Raters initially coded 16% of the responses to ensure that the rubric could be used reliably, and then each coded all of the remaining responses. Interrater agreement resulted in a high degree of reliability, Cohen's $\kappa = .75$. An overview of the rubric, including example responses and frequencies of each category, is shown in Table 3.

A mixed-effects model using ordinal logistic regression (using the ordinal package in R, Christensen, 2019) with study strategy training condition entered as a fixed factor, and including intercepts for subjects and texts as random factors, indicated that students in the explanation condition had greater frequencies of high-quality explanations in the hand-coded scores than students in the POE condition, $B = .51$, $SE = .10$, $z = 5.04$, $p < .001$, even though the latter group were specifically prompted to explain the relation of the results of the experiments to the theory.

As a second way of coding the content of the open-ended responses, the semantic overlap between each actual student response and an "ideal" student response constructed using language that appeared in the best hand-coded student responses in both conditions was computed using latent semantic analysis (LSA, Landauer, Foltz, & Laham, 1998). Prior research that has used LSA to assess student responses has found that comparisons to idealized student responses are better predictors of comprehension than either comparisons to expert responses or to the original text (Guerrero & Wiley, 2019; León et al., 2006; Wiley et al., 2017). The main difference between the idealized student response and an expert response is not content, but the use of more colloquial language. When experts write responses, they tend to use more academic language that students are less likely to use. When idealized responses are constructed from peer examples, they are written in language that is more typical for students which provides a better basis for the LSA comparison (Ventura et al., 2004). The other critical feature of "ideal" responses is that they select out the parts of the text that are most important in building a situation model, which means that high similarity suggests that the reader has focused on the more-important parts of a text. High similarity to the original text could mean that a reader is focusing on less-important parts of the text, and in some cases has been negatively related to comprehension (Wiley et al., 2017). For these reasons, the idealized student response was used as the comparison for LSA.

After editing the student responses to correct for misspellings, abbreviations, and to expand contractions, each student's response was compared to the idealized response for each topic to obtain coefficients representing the degree of semantic overlap, with numbers closer to 1 representing a greater degree of semantic overlap. Semantic

overlap was computed using a one-to-many, document-to-document analysis using the general reading up to first year college LSA space with maximal factors included. As shown in Table 3, the LSA scores increased in parallel with the hand-coded quality scores. The correlation between the hand-coded quality scores and the LSA scores was positive and significant, Spearman *rho* = .43, $p < .001$.

As shown in Table 4, students wrote higher quality responses in the explanation condition than in the POE condition on each topic as measured by both the hand-coded quality score and by LSA scores (with the exception of the placebo effect text). A linear mixed-effects model with condition entered as a fixed factor, and including intercepts for subjects and texts as random factors, indicated that responses in the explanation condition had higher LSA scores than responses in the POE condition, $\beta = .07$, $SE = .03$, $t = 2.03$, $p = .04$. These results are contrary to any expectation that the POE study strategy would help students to make more beneficial connections while studying.

As shown in Table 1, responses written in the explanation condition were also longer on average than those written in the POE condition, $t(356) = 5.15$, $p < .001$, Cohen's $d = .55$. As shown in Table 5, response length correlated with both hand-coded quality scores and LSA scores, consistent with prior work that has found that the length of a response is often related to its quality (Crossley, Allen, Kyle, & McNamara, 2014; Guerrero & Wiley, 2019; Kobrin, Deng, & Shaw, 2007; MacArthur, Jennings, & Philippakos, 2019; Wiley et al., 2017).

Evaluative Comments Given After Second Study Result

Prior analyses showed that despite the fact that those in the POE condition were asked to make a connection between the theory and evidence, those students largely failed to consider *why* the actual outcomes supported the theory and therefore were not engaging in the task as intended. Instead, it seemed that many students tended to focus more on evaluative comments indicating whether or not their prediction was correct. If students were preoccupied with the accuracy of their own predictions, this could have distracted them from mental model construction and derailed their comprehension. Thus, the evaluative comments were coded as a second dimension of the open-ended responses given after the second study results. These evaluative comments included any reference to whether the results of the final study were as expected or not, or if the prediction that was previously made was correct or incorrect (e.g., I think my prediction was incorrect, I was wrong, I did not expect that to happen). Interrater agreement between two independent coders who coded all 1,790 responses resulted in a high degree of reliability, Cohen's $\kappa = .93$. A higher frequency of students made evaluative comments in the POE strategy training

condition (27.5%), however, there were a few students in the explanation condition who spontaneously made an evaluative statement (0.5%). A linear mixed-effects model using binomial logistic regression with condition entered as a fixed factor, and including intercepts for subjects and texts as random factors, indicated that evaluative comments were more likely to occur in the POE condition, $B = -2.40$, $SE = .27$, $z = 8.99$, $p < .001$.

As shown in Table 4, the correlations below the diagonal indicate these evaluative comments were also negatively related to response quality (both hand-coded scores and LSA scores). This suggests that readers were focusing on whether or not their predictions were correct at the expense of engaging in the construction of theory-evidence connections and relations.

Within just the POE training condition, the individual predictions that students made could also be scored for accuracy. As shown in Table 4, the correlations above the diagonal indicate that having made a correct prediction was also related to response quality, while students who made an incorrect prediction were more likely to generate an evaluative comment. A linear mixed-effects model with evaluative comments entered as a fixed factor, and including intercepts for subjects and texts as random factors, showed a weak negative relation between evaluative comments and response quality, $\beta = -.03$, SE = .03, $t = -1.01$, $p = .31$, similar to what is shown in the correlation table. However, when prediction accuracy was added as a second fixed factor, prediction accuracy was strongly related to response quality, $\beta = .16$, SE = .04, $t = 4.64$, $p < .001$, and evaluation was not, $\beta = .02$, SE = .03, $t = 0.66$, $p = .51$. This highlights how it was students who made incorrect predictions who were the ones least likely to engage in the full POE cycle as intended, and offers suggestions as to why the POE strategy training was less beneficial for future learning than explanation alone. Students were not only harmed by making an incorrect prediction, but engagement in the subsequent explanation process was fundamentally changed by making a prediction first. The quality of the explanations decreased, and students were more likely to restate results instead of making connections between results and the theory. It is possible that prediction activities were causing students to focus on their own performance, instead of comprehension of the text. Hence, in the final explanation of the text, some students may have been derailed by the previously-made inaccurate prediction.

Quality of Prediction Justifications

In the POE condition, the quality of responses given at the first prompt during the strategy training activity was also explored. After making a selection from the 3 possible outcomes, students were asked to explain why they made that prediction. An ideal justification for the prediction would be to draw a connection between the results of

the first empirical study described in the text and the theory to support the prediction for the second study. Unfortunately, this was a rarity. Similar to the results of Baddock and Bucat (2008), only 2 students referenced the results of the first empirical study when justifying their prediction.

Connection between Activity Quality and Learning from Future Texts

The purpose of the strategy training activity was to teach and provide practice with the POE and explanation study strategies. The main question of interest was whether students would be able to utilize these strategies effectively to support future learning. As shown in Table 5, the quality scores for the responses provided during the strategy training activity were positively related to students' ability to comprehend on the final set of texts that they later studied during the learning activity. Even when using partial correlations to take into account domain-specific comprehension skill as a control variable, future learning was still negatively related to the likelihood of making evaluative comments during the training activity, $r = -.09$, $p = .01$, and tended to be positively related to whether students made accurate predictions in the POE condition, $r = .06$, $p = .06$. Even though the average length of responses provided during the strategy training activity was correlated with future learning, when partial correlations were computed to take the hand-coded quality of the responses into account, then the relation between length and test scores on the learning activity was reduced to $r = .02$, $p = .36$. In contrast, the relation between learning and hand-coded quality scores still remained even after the average length of responses were taken into account, $r = .22$, $p < .001$. Thus, quality of engagement in the activity was more important than the length of the responses.

Connection between Domain-specific Comprehension Skill and Learning from Future Texts

The initial analyses of learning outcomes revealed not just that students trained with the POE study strategy performed worse than students who were trained with the explanation study strategy, but also that it was the less-skilled comprehenders who were most negatively affected by the training condition. A sequential mediation model showed the relationship between domain-specific comprehension skill (performance on inference questions on the baseline assessment) and future comprehension (performance on the inference questions after the learning activity) was partially mediated by the accuracy of the prediction and the presence of evaluative comments. As Figure 3 illustrates, the standardized regression coefficient between domain-specific comprehension skill and the accuracy of the prediction was statistically significant, as was the standardized regression coefficient between accuracy of the prediction and evaluative comments, and between evaluative comments and test score on the learning activity. The

significance of the indirect effect was computed for each of 10,000 bootstrapped samples, and the 95% confidence interval was computed by determining the indirect effects at the 2.5th and 97.5th percentiles. The bootstrapped standardized effect was .01, and the 95% confidence interval ranges from .001, .01. Thus, the indirect effect was small, yet statistically significant.

Even though the final prompt provided to those in the POE condition may have included a distraction to discuss if their prediction was incorrect, the evaluative comments did not fully account for the poor performance on the learning activity. When evaluative comments were added to the original model of comprehension outcomes, the interaction between condition and domain-specific comprehension skill remained, $\beta = .32$, $SE = .15$, $t = 2.07$, $p = .04$.

In summary, these exploratory analyses on the quality of responses written during the strategy training activity help to clarify why the POE instructional condition might not have been effective. A closer examination indicated that a possible reason why making predictions hurt performance was that students may have focused primarily on whether they made incorrect predictions, and failed to engage in the subsequent explanation task to complete the POE learning cycle as intended. Additionally, less-skilled comprehenders may have been most negatively affected by this manipulation because they were more likely to make incorrect predictions and to focus on them.

Discussion

This study examined whether training students to use a POE study strategy would benefit performance on future learning from social science textbook excerpts. Engaging in a Predict-Observe-Explain cycle has been shown to be a successful approach to learning with hands-on activities because it is thought to draw students' attention toward the theory-evidence relationships (Champagne et al., 1980; White & Frederiksen, 1998; White & Gunstone, 1992). It was hypothesized that if engaging in a prediction activity as part of reading helps to direct attention to the key relation between theories and evidence, then this activity could be expected to benefit comprehension as well as comprehension monitoring during subsequent learning attempts from social science textbook excerpts. The results of this study showed that both POE and explanation study strategy training activities improved comprehension monitoring on multiple measures. However, in contrast to the hypothesis that training students on a POE study strategy would be superior, students in the explanation condition showed better comprehension for the new topics.

Effects of Strategy Training on Future Learning

The effect size of the study strategy training manipulation on comprehension of new textbook excerpts was modest, with test scores in the explanation condition being 4 percentage points higher than the POE condition. That difference corresponds to one-third of a standard deviation and represents almost a half letter grade difference in a classroom context. In addition, the POE study strategy was especially problematic for the lowest-skilled comprehenders whose test scores were more than 10 percentage points (a full letter grade) lower than in the explanation condition.

Exploratory analyses also revealed that the lowest-skilled comprehenders were more likely to make inaccurate predictions, and then subsequently focus on the inaccuracy of their predictions during the final stage of the POE learning cycle. These findings are similar to those of Carvalho et al. (2018) who found that students who made incorrect predictions also performed more poorly on their tests. In the present study, reflection on the accuracy of the predictions seemed to interfere with constructing an integrated model of the theory and evidence. There are at a number of possible alternative accounts for how incorrect predictions might disrupt learning. By one account, it could have been that students perseverated about the accuracy of their response. The preoccupation with being incorrect derailed their focus and interfered with their ability to comprehend the text. It is also possible that the interference stemmed from the maintenance of the incorrect prediction and the inclusion of an incorrect relationship into the mental model of the phenomenon. However, because the sequential mediation model did not show a significant relationship directly between prediction accuracy and future learning, and because the indirect effect was only significant when it included the evaluative response, the former explanation appears more likely. Additionally, it is possible that students in the POE condition engaged in these same ineffective behaviors during the learning activity, and failed to engage in appropriate explanation behaviors, which could account for the poorer comprehension outcomes on the new topics. It is also possible that incorrect predictions could have discouraged students and derailed motivation to engage in the task moving forward. Future work will need to explore how making incorrect predictions may have an effect on motivation.

More generally, the exploratory analyses showed that students did not engage in the POE study strategy training activities as intended. Students in the explanation condition tended to write both longer and higher quality responses. When these two features were examined simultaneously, the quality of the response given during the training task was found to be the key predictor for later learning. Similar to past research that has coded the quality of responses that are generated during study activities (Hinze et al., 2013), there was a positive relationship between

response quality during the study strategy training activities and future learning on new textbook excerpts in both conditions. This suggests that the benefits seen in the explanation condition were because the important connections were more likely to be generated by students. The mechanism by which explanation improved the quality of the situation model was by prompting the generation of the key theory-evidence relations that were left implicit in the text.

Implications for Research on Metacomprehension

Engaging in study strategy training also led to modest effects on metacomprehension measures. Students became more conservative in their estimates of how much they understood from reading each textbook excerpt after study strategy training, which resulted in less absolute error and improved absolute accuracy. While students were generally overconfident in their estimates of their understanding, the amount of overconfidence tended to decrease from estimates on the initial set of textbook excerpts to estimates on the later set of excerpts that followed study strategy training. The general tendency for students to be overconfident is consistent with the literature showing that overconfidence is the norm among students from middle school to college (Kent State, 2007; Maki, 1998a). Overconfidence can be problematic as it may cause students to terminate study earlier than they should. Dunlosky and Rawson (2012) found that those who were most overconfident did not study enough to achieve a mastery level which resulted in lower final test scores. Theoretically, underconfidence could also have harmful effects on learning when there are time constraints and if students spend time studying topics they have already mastered at the cost of those that they have not. However, reports of negative effects of underconfidence are much less common, perhaps because students generally do not achieve full mastery of materials -- which is the only condition under which additional study would not be useful (Metcalfe & Finn, 2013).

In addition, engaging in study strategy training activities also led to improvements in relative metacomprehension accuracy. Again, no differences were seen due to study strategy training condition, and the effect size was again modest with post-training relative accuracy still being quite low (r = .14 averaged across training type). However, given that baseline accuracy was near zero, the increase in the explanation condition still corresponded to one-third of a standard deviation. The near zero relative accuracy in the baseline measures is striking, because it is much lower than has been typically reported in college-aged samples. Several reviews of average levels of metacomprehension accuracy (without any instructions or activities) report positive relations between judgments and performance at around .27 (Dunlosky & Lipko, 2007; Griffin et al., 2019; Lin & Zabrucky,

1998; Maki, 1998b; Thiede et al., 2009). One possibility for why this study obtained such low levels of baseline relative accuracy could be due to the similarity in the topics of the texts. It is common in studies on metacomprehension for students to read sets of texts on diverse topics where each topic might be quite distinct from the next. However, because the current study was conducted with actual class materials, the texts were on related topics from a single textbook. The results of this study suggest an emerging pattern when combined with a few other reports of very low metacomprehension accuracy. For example, poor relative accuracy (with means not different from zero) have been seen when students read related texts from a single domain, such as excerpts from a textbook on psychological research methods (Wiley et al., 2016), and for text sets that are either all about music or all about physics (Glenberg & Epstein, 1987). Negative average relations between predictions and performance have even been reported when students judge comprehension of individual sentences within a Psychology text on brain structure (Ozuru, Kurby, & McNamara, 2012). When excerpts or sentences seem highly related or just fall within the same domain, it may increase the difficulty of judging understanding of each one separately. Put another way, relative accuracy requires discriminating between different texts, so the less objectively discriminable the texts are, the greater the obstacle to making accurate relative judgments.

It is also possible that the low baseline levels of relative accuracy could be due to the lack of familiarity with how to learn from these types of texts, or the difficulty of learning from texts written at a collegiate level. Lower accuracy has been observed for texts written at a collegiate versus 12th grade level of difficulty (Weaver, Bryant, & Burns, 1995). Further, differences have been seen due to text genre or structure (Weaver & Bryant, 1995). Relative accuracy is lower when tests require causal or bridging inferences rather than memory for details (Griffin et al., 2019), so it is likely that relative accuracy would be especially hindered when comprehension requires unfamiliar types of inferences such as those that integrate theories with the empirical results of particular experiments. It may be that at the transition to college, students are not prepared to read textbook excerpts at this level of difficulty which may combine several expository structures, include unfamiliar subgenres of scientific writing, and require specific disciplinary literacy skills that they have not yet been taught. Therefore, they cannot accurately predict their understanding. It is also possible that most existing work that has shown higher levels of relative accuracy in college samples has used texts written at a below-grade level. Nevertheless, despite the baseline levels of relative accuracy being quite low, training in explanation activities as well as prediction activities appeared to improve skills in accurate monitoring.

The decrease in absolute accuracy and the trend toward a decrease in overconfidence that was seen from before to after the study strategy training activities was not simply a result of the oft-cited *underconfidence with practice effect* (UWP; for a review, see Koriat, Sheffer, & Ma'ayan, 2002), because these decreases were the result of the magnitude of JOCs actually reducing with training. In the standard UWP effect, judgments of learning (JOLs) do not decrease with practice or become more conservative. Rather, the standard effect of practice on the magnitude of judgments is just the opposite. Koriat et al. (2002, p. 152) found that across 11 experiments from multiple publications "both JOL and recall increased strongly with [practice], and the function was steeper for recall than for JOL." In other words, the UWP effect does not reflect more conservative judgments or reduced subjective confidence, but emerges as a byproduct of increases in learning with practice outpacing increases in confidence with practice. Scheck and Nelson (2005) showed that when extremely difficult test items are used and when initial test performance is at floor, the UWP effect takes the form of initial overconfidence reducing to a more accurate unbiased confidence score with practice. But even then, there was no reduction in judgment magnitude, just an increase in test performance that rose to converge with judgments. Thus, the current results provide an uncommon example of a change in accuracy that reflects a reduction in subjective confidence, and suggest that the study strategy training activities are reversing the typical effects of practice on judgment magnitude by causing judgments to decrease and become more conservative.

Furthermore, relative accuracy is statistically orthogonal to average absolute levels of either judgments or test performance, and thus to overconfidence or underconfidence. Hence, the observed improvements in relative accuracy cannot be explained by these factors. Rather, better relative accuracy can only be achieved if students are better able to judge what they understand best from least, and this can help students to direct their attention to the topics where restudy is most needed. While relative accuracy is important for the monitoring process, improvements in absolute accuracy are also important as this makes students better able to judge when they should persist or terminate study. Engaging in these study strategy training activities appears to be helping students to gain a greater awareness of their understanding in both respects. On a new set of texts, they were able to both adjust their expected test performance downward to better match the general difficulty of the task, and to better attend to the cues that reflected their actual understanding of particular texts.

Though not large, these increases in relative accuracy are promising given the low levels that students are starting at. Even modest improvements may have implications for learning in a course context. As a mechanism,

metacomprehension impacts whether and how additional study and learning is engaged in prior to summative learning assessments. In this experiment, the textbook excerpts were studied in only a single session. When students have greater opportunity for iterative self-regulated study over the course of a semester, as they do in authentic learning contexts, then the benefits from improved metacomprehension accuracy may be magnified. Because metacomprehension accuracy can have a positive impact during many phases of learning, even a small improvement could result in notable changes in long-term learning gains.

Limitations and Future Directions

The motivation for the intervention tested here was to develop a generative study activity that supported students in understanding the process of how to make a hypothetical prediction. Even if students have some general knowledge of the scientific method, they are unlikely to be familiar with this subgenre of scientific writing that links theories to evidence, and the types of constructive processing that are required to comprehend it. In contrast to other prediction activities that simply ask students to make an intuitive guess of "what might happen next", in this intervention students needed to reason forward from the theoretical information provided in the text to make predictions about which result would provide support for a theory. Thus, to highlight how the students should do this, the instruction needed to discuss the structure of the text in order to provide clear information about the basis that students should use to make predictions. It can be seen as a limitation of this study that both the generation prompt and instruction about the structure of the text were varied simultaneously. Although it seems unlikely given the lack of benefits from this combined POE instruction, it is possible that independent manipulations could show an effect, and this could be tested in future work.

Another limitation of this work was the relatively modest effects that were seen due to either study strategy training. There is a need for future work that can uncover the factors and conditions that might lead to more robust benefits when learning from authentic texts in course contexts. Past work has used more extensive explanation interventions that have included more information about the goals for study as being reading for comprehension or understanding, more clearly setting up expectations about nature of the test questions as being inference-based, and providing more extensive exposure to example test questions (Griffin et al., 2019; Thiede, Wiley, & Griffin, 2011; Wiley et al., 2016). It may be that the simplified explanation instructions that were provided in this study were too vague for students to obtain the full benefits from explanation. Also, the similarity among excerpts could have been an added obstacle to improving relative monitoring accuracy. Further, as suggested in the introduction, the effects of

an explanation generation activity may work best for improving comprehension and comprehension monitoring from explanatory texts. This was the motivation for developing a new sort of generative activity that would better match the theory-evidence structure of these textbook excerpts.

Although training students to use a POE study strategy did not improve comprehension on new textbook excerpts in this study, some alterations to this activity could potentially improve its effectiveness. Overall, the quality of the responses suffered when students engaged in making predictions. They were less likely to address the connections between the theory and evidence in the text than those who engaged in explanation only. Similar to Gunstone and White (1981), students showed particular difficulty explaining why they chose their prediction. They tended to use circular justifications by just restating results or general definitions of the topic instead of making connections between the results and theory to motivate their predictions. It may be that students need more guidance and scaffolding during the POE process to engage in more thoughtful predictions (White & Frederiksen, 1998). The POE study strategy training could place more emphasis on theory-evidence relations by simply clarifying prompts to explicitly state that when making a prediction it is important to think about the results of the prior empirical studies and how they support the theory. Or, for struggling readers, prompting students to engage in summarization of the theory and first empirical study immediately before prompting them to make a prediction and to justify it could help to make the relevant textbase information available to them. This would be consistent with other work showing how some readers benefit from support at the textbase level before they can move onto developing an integrated situation model (McNamara, 2017; Millis et al., 2006).

While improving the students' justifications of their predictions may increase attentiveness to the relationship between theory and evidence at the first step of the POE cycle, the low quality of responses given in the final step suggests students also need additional support to complete the activity as intended. At the end of each excerpt students were asked to explain how the results provided evidence for the theory as well as to reflect on their predictions. Unfortunately, it seemed that many students perseverated on the accuracy of their predictions. When viewed in light of the results of the exploratory analyses, including the additional part of the prompt that directed the readers to consider the accuracy of their predictions may have done more harm than good. Yet, at the same time prior work emphasizes the role of reflection, and reconciling any discrepancies between the predictions that were made and what was actually observed, as key factors underlying the benefits of prediction (White & Frederiksen, 1998; White & Gunstone, 1992). It is possible that students may need extended practice, or more extensive

examples or modelling of the full POE process, in order to develop the reasoning skills that will allow them to benefit more fully from making predictions based in theory (Chang et al, 2013; Lehrer & Schauble, 1998; White & Frederiksen, 1998). It is also possible that providing online feedback about the quality of the predictions, justifications, and explanations could help students to be more likely to engage in the POE activity as intended, just as such feedback has proven useful in supporting higher quality summaries and explanations from expository texts (McNamara, O'Reilly, Best, & Ozuru, 2006; VanLehn, 2011; Wade-Stein & Kintsch, 2004). As students become more familiar with the POE approach during reading and are able to see its utility in improving their learning about theories and evidence, the effects would be expected to become more robust and have practical relevance for academic performance.

The exploratory analyses also revealed that it was the poorer comprehenders who were harmed by the POE activities. A question that might be asked is whether there are particular individual differences that may influence the utility of these activities for improving comprehension outcomes. If future work is interested in determining the specific and unique roles that factors such as prior knowledge, working memory capacity, or different aspects of reading ability might play, then it would be important to measure them separately with standardized measures following the lead of other work that has explored which individual differences interact with manipulations intended to support understanding from expository texts (Budd, Whitney, & Turley, 1995; Cromley, Snyder-Hogan, & Luciw-Dubas, 2010; Linderholm & van den Broek, 2002; Kendeou & van den Broek, 2007; McNamara et al., 1996; Sanchez & Wiley, 2006; Voss & Silfies, 1996).

The goal behind the development of this intervention was to help students to appreciate the need to focus on theory-evidence relations when they are assigned readings in Introduction to Psychology. Most work on expository text comprehension has been focused on other types of expository structures, most notably informational and explanatory texts. There is not yet a body of work in the text comprehension literature on learning about theories from text, and this marks an initial foray into learning more about this structure and the difficulties it presents. Once a condition is developed that is effective for improving learning in this domain-specific context (psychology textbook excerpts), it will be interesting to explore whether text from other disciplines that follow a similar structure (in which theories are presented with supporting evidence) might benefit from a similar strategy training.

In summary, this study tested whether training students to generate experimental predictions as they read social science texts might be beneficial for helping students to reflect more deeply on the theories they are learning

about, to prompt them to better understand relations between hypotheses, designs, and results of studies, and to be in a better position to accurately monitor their own learning about these theories. While it is still possible that a prediction study strategy could be better than an explanation-only strategy if students actually engaged in each stage of the POE process as intended, the results of this study suggested that training students to use an explanation strategy was more effective for comprehension.

References

Ainsworth, S., & Th Loizou, A. (2003). The effects of self-explaining when learning with text or diagrams. *Cognitive Science*, *27*, 669-681. https://doi.org/10.1207/s15516709cog2704_5

Alexander, P. A., & Judy, J. E. (1988). The interaction of domain-specific and strategic knowledge in academic performance. *Review of Educational Research*, *58*(4), 375-404. https://doi.org/10.3102/00346543058004375

Baddock, M., & Bucat, R. (2008). Effectiveness of a classroom chemistry demonstration using the cognitive conflict strategy. *International Journal of Science Education*, *30*(8), 1115-1128. https://doi.org/10.1080/09500690701528824

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2014). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67,* 1-48. https://doi.org/10.18637/jss.v067.i01

Bazerman, C. (1985). Physicists reading physics: Schema-laden purposes and purpose-laden schema. *Written Communication*, *2*(1), 3-23. https://doi.org/10.1177/0741088385002001001

Berkenkotter, C., & Huckin, T. N. (1995). *Genre knowledge in disciplinary communication: Cognition/culture/power*. Routledge.

Bolger, M. S., Kobiela, M., Weinberg, P. J., & Lehrer, R. (2012). Children's mechanistic reasoning. *Cognition and Instruction, 30,* 170-206. https://doi.org/10.1080/07370008.2012.661815

Britt, M. A., Richter, T., & Rouet, J.-F. (2014). Scientific literacy: The role of goal-directed reading and evaluation in understanding scientific information. *Educational Psychologist*, *49*(2), 104-122. https://doi.org/10.1080/00461520.2014.916217

Britt, M. A., & Sommer, J. (2004). Facilitating textual integration with macro-structure focusing tasks. *Reading Psychology*, *25*(4), 313-339. https://doi.org/10.1080/02702710490522658

Britton, B. K., & Black, J. B. (1985). Understanding expository text: From structure to process and world knowledge. In B. K. Britton & J. B. Black (Eds.), *Understanding expository text: A theoretical and practical handbook for analyzing explanatory texts* (pp. 1-9). Erlbaum.

Budd, D., Whitney, P., & Turley, K. J. (1995). Individual differences in working memory strategies for reading expository text. *Memory & Cognition*, *23*(6), 735-748. https://doi.org/10.3758/BF03200926

Burbules, N. C., Linn, M. C. (1991). Science education and philosophy of science: Congruence or contradiction? *International Journal of Science Education, 13*(3) 227-241. https://doi.org/10.1080/0950069910130302

Butcher, K. R. (2006). Learning from text with diagrams: Promoting mental model development and inference generation. *Journal of Educational Psychology, 98*, 182-197. https://doi.org/10.1037/0022-0663.98.1.182

Butler, A. C., & Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition*, *36*(3), 604-616. https://doi.org/10.3758/MC.36.3.604

Carvalho, P. F., Manke, K. J., & Koedinger, K. R. (2018). Not all active learning is equal: Predicting and explaining improves transfer relative to answering practice questions. In T. T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *CogSci 2018 proceedings* (pp. 1458-1463). https://cogsci.mindmodeling.org/2018/papers/0284/0284.pdf

Champagne, A. B., Klopfer, L. E., & Anderson, J. H. (1980). Factors influencing the learning of classical mechanics. *American Journal of Physics, 48,* 1074-1079. https://doi.org/10.1119/1.12290

Chang, J. L., Chen, C. C., Tsai, C. H., Chen, Y. C., Chou, M. H., & Chang, L. C. (2013). Probing and fostering students' reasoning abilities with a cyclic predict-observe-explain strategy. In M. H. Chiu, H. L. Tuan, H. K. Wu, J. W. Lin, & C. C. Chou (Eds.), *Chemistry education and sustainability in the global age* (pp. 49-57). Springer.

Chi, M. T. H. (2000). Self-explaining expository texts: The dual processes of generating inferences and repairing mental models. In R. Glaser (Ed.), *Advances in instructional psychology* (pp. 161-238). Lawrence Erlbaum.

Chi, M. T. H. (2009). Active-constructive-interactive: A conceptual framework for differentiating learning activities. *Topics in Cognitive Science*, *1*, 73-105. https://doi.org/10.1111/j.1756-8765.2008.01005.x

Christensen, R. H. B. (2019). ordinal - Regression Models for Ordinal Data. R package version 2019.12-10. https://CRAN.R-project.org/package=ordinal

Cook, L. K., & Mayer, R. E. (1988). Teaching readers about the structure of scientific text. *Journal of Educational Psychology, 80*(4), 448-456. https://doi.org/10.1037/0022-0663.80.4.448

Cromley, J. G., Snyder-Hogan, L. E., & Luciw-Dubas, U. A. (2010). Reading comprehension of scientific text: A domain-specific test of the direct and inferential mediation model of reading comprehension. *Journal of Educational Psychology*, *102*(3), 687-700. https://doi.org/10.1037/a0019452

Crossley, S. A., Allen, L. K., Kyle, K., & McNamara, D. S. (2014). Analyzing discourse processing using a simple natural language processing tool. *Discourse Processes, 51*(5-6), 511-534.

Davey, B., & McBride, S. (1986). Effects of question-generation training on reading comprehension. *Journal of Educational Psychology*, *78*, 256-262. https://doi.org/10.1037/0022-0663.78.4.256

de Jong, T., Linn, M. C., & Zacharias, Z. C. (2013). Physical and virtual laboratories in science and engineering education. *Science, 340*, 305-308. https://doi.org/10.1126/science.1230579

Diakidoy, I.-A. N., Mouskounti, T., & Ioannides, C. (2011). Comprehension and learning from refutation and expository texts. *Reading Research Quarterly*, *46*(1), 22-38. https://doi.org/10.1598/RRQ.46.1.2

Dole, J. A. (2000). Readers, texts and conceptual change learning. *Reading & Writing Quarterly*, *16*(2), 99-118. https://doi.org/10.1080/105735600277980

Dunlosky, J., & Lipko, A. R. (2007). Metacomprehension: A brief history and how to improve its accuracy. *Current Directions in Psychological Science*, *16*, 228-232. https://doi.org/10.1111/j.1467-8721.2007.00509.x

Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention. *Learning and Instruction*, *22*(4), 271-280. https://doi.org/10.1016/j.learninstruc.2011.08.003

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest*, *14*, 4-58. https://doi.org/10.1177/1529100612453266

Einstein, G. O., McDaniel, M. A., Bowers, C. A., & Stevens, D. T. (1984). Memory for prose: The influence of relational and proposition-specific processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(1), 133-143. https://doi.org/10.1037/0278-7393.10.1.133

Fukaya, T. (2013). Explanation generation, not explanation expectancy, improves metacomprehension accuracy. *Metacognition and Learning, 8*, 1-18. https://doi.org/10.1007/s11409-012-9093-0

Furtak, E. M., Seidel, T., Iverson, H., & Briggs, D. C. (2012). Experimental and quasi-experimental studies of inquiry-based science teaching: A meta-analysis. *Review of Educational Research*, *82*(3), 300-329. https://doi.org/10.3102/0034654312457206

Gil-Pérez, D., Guisasola, J., Moreno, A., Cachapuz, A., De Carvalho, A. M. P., Torregrosa, J. M., ... & Dumas-Carré, A. (2002). Defending constructivism in science education. *Science & Education*, *11*(6), 557-571. https://doi.org/10.1023/A:1019639319987

Glenberg, A. M., & Epstein, W. (1987). Inexpert calibration of comprehension. Memory & Cognition, 15, 84-93. https://doi.org/10.3758/BF03197714

Graesser, A. C. (1981). *Prose comprehension beyond the word.* Springer.

Graesser, A. C., & Bertus, E. L. (1998). The construction of causal inferences while reading expository texts on science and technology. *Scientific Studies of Reading*, *2*, 247-269. https://doi.org/10.1207/s1532799xssr0203_4

Graesser, A. C., McNamara, D. S., & Louwerse, M. M. (2003). What do readers need to learn in order to process coherence relations in narrative and expository text? In C. Snow & A. Sweet (Eds.), *Rethinking reading comprehension* (pp. 82-98). Guilford Press.

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review, 101*(3), 371-395. https://doi.org/10.1037/0033-295X.101.3.371

Griffin, T. D., Mielicki, M. K., & Wiley, J. (2019). Improving students' metacomprehension accuracy. In J. Dunlosky & K. Rawson (Eds.), *Cambridge handbook of cognition and education* (pp. 619-646). Cambridge University Press.

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,* 93-103. https://doi.org/10.3758/MC.36.1.93

Griffin, T. D., Wiley, J., & Thiede, K. W. (2019). The effects of comprehension-test expectancies on metacomprehension accuracy. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 45*, 1066-1092. http://dx.doi.org/10.1037/xlm0000634

Gunstone, R. F., & White, R. T. (1981). Understanding of gravity. *Science Education, 65,* 291-299. https://doi.org/10.1002/sce.3730650308

Guerrero, T. A., & Wiley, J. (2019). Using “idealized peers” for automated evaluation of student understanding in an introductory psychology course. In S. Isotani, E. Millan, A. Ogan, P. Hastings, B. McLaren, & R.

Luckin (Eds.), *Artificial Intelligence in Education - 20th international conference, AIED 2019* (pp. 133-143). Springer. https://doi.org/10.1007/978-3-030-23204-7_12

Hinze, S. R., Wiley, J., & Pellegrino, J. W. (2013). The importance of constructive comprehension processes in learning from tests. *Journal of Memory and Language*, *69*, 151-164. https://doi.org/10.1016/j.jml.2013.03.002

Hmelo-Silver, C. E., Duncan, R. G., & Chinn, C. A. (2007). Scaffolding and achievement in problem-based and inquiry learning: A response to Kirschner, Sweller, and Clark (2006). *Educational Psychologist*, *42*, 99-107. https://doi.org/10.1080/00461520701263368

Iordanou, K., Kendeou, P., & Beker, K. (2016). Argumentative reasoning. In J. A. Greene, W. A. Sandoval, & I. Braten (Eds.), *Handbook of epistemic cognition* (pp. 51-65). Routledge.

Jaeger, A. J., & Wiley, J. (2014). Do illustrations help or harm metacomprehension accuracy? *Learning and Instruction, 34*, 58-73. https://doi.org/10.1016/j.learninstruc.2014.08.002

Kendeou, P., & van den Broek, P. (2007). The effects of prior knowledge and text structure on comprehension processes during reading of scientific texts. *Memory & Cognition,* 35**,** 1567-1577. https://doi.org/10.3758/BF03193491

Kent State University. (2007, November 27). Middle school students 'extremely overconfident' in their own learning. *ScienceDaily*. Retrieved from www.sciencedaily.com/releases/2007/11/071126162526.htm

King, A. (1994). Guiding knowledge construction in the classroom: Effects of teaching children how to question and how to explain. *American Educational Research Journal*, *31*, 338-368. https://doi.org/10.3102/00028312031002338

Kintsch, W. (1986). Learning from text. *Cognition and Instruction*, *3*, 87-108. https://doi.org/10.1207/s1532690xci0302_1

Kintsch, W. (1994). Text comprehension, memory, and learning. *American Psychologist*, *49*, 294-303. https://doi.org/10.1037/0003-066X.49.4.294

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge University Press.

Kintsch, W., & Yarbrough, J. C. (1982). Role of rhetorical structure in text comprehension. *Journal of Educational Psychology*, *74*(6), 828-834. https://doi.org/10.1037/0022-0663.74.6.828

Kobrin, J. L., Deng, H., & Shaw, E. J. (2007). Does quantity equal quality? The relationship between length of response and scores on the SAT essay. *Journal of Applied Testing Technology, 8*(1), 1-15.

Koriat, A., Sheffer, L., & Ma'ayan, H. (2002). Comparing objective and subjective learning curves: Judgments of learning exhibit increased underconfidence with practice. *Journal of Experimental Psychology: General, 131*(2), 147-162. http://doi.org/10.1037/0096-3445.131.2.147

Larson, M., Britt, M. A., & Larson, A. A. (2004). Disfluencies in comprehending argumentative texts. *Reading Psychology*, *25*(3), 205-224. https://doi.org/10.1080/02702710490489908

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*, 259-284. https://doi.org/10.1080/01638539809545028

Lehrer R., & Schauble, L. (1998). Reasoning about structure and function: Children's conceptions of gears. *Journal of Research in Science Teaching, 35*, 3-25. https://doi.org/10.1002/(SICI)1098-2736(199801)35:1<3::AID-TEA2>3.0.CO;2-X

León, J. A., Olmos, R., Escudero, I., Cañas, J. J., & Salmerón, L. (2006). Assessing short summaries with human judgments procedure and latent semantic analysis in narrative and expository texts. *Behavior Research Methods*, *38*(4), 616-627. https://doi.org/10.3758/BF03193894

Liew, C. W., & Treagust, D. F. (1995). A Predict-Observe-Explain teaching sequence for learning about students' understanding of heat and expansion liquids. *Australian Science Teachers Journal, 41,* 68-71.

Lin, L. M., & Zabrucky, K. M. (1998). Calibration of comprehension: Research and implications for education and instruction. *Contemporary Educational Psychology*, *23*, 345-391. https://doi.org/10.1006/ceps.1998.0972

Linderholm, T., & van den Broek, P. (2002). The effects of reading purpose and working memory capacity on the processing of expository text. *Journal of Educational Psychology*, *94*(4), 778-784. https://doi.org/10.1037/0022-0663.94.4.778

Lorch Jr., R. F. (2015). What about expository texts? In E. J. O'Brien, A. E. Cook, & R. F. Lorch Jr. (Eds.), *Inferences during reading* (pp. 348-361). Cambridge University Press.

MacArthur, C. A., Jennings, A., & Philippakos, Z. A. (2019). Which linguistic features predict quality of argumentative writing for college basic writers, and how do those features change with instruction? *Reading and Writing, 32(*6), 1553-1574. https://doi.org/10.1007/s11145-018-9853-6

Magoon, A. J. (1977). Constructivist approaches in educational research. *Review of Educational Research*, *47*(4), 651-693. https://doi.org/10.3102/00346543047004651

Maki, R. H. (1998a). Metacomprehension of text: Influence of absolute confidence level on bias and accuracy. In D. L. Medin (Ed.), *The psychology of learning and motivation: Advances in research and theory, Vol. 38* (p. 223-248). Academic Press.

Maki, R. H. (1998b). Test predictions over text material. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in Educational Theory and Practice* (pp. 117-145). Erlbaum.

Martin, J. R. (1993). Literacy in science: Learning to handle text as technology. In M. A. K. Halliday, & J. R. Martin (Eds.), *Writing science: Literacy and Discursive Power* (pp. 166-202). Routledge.

Mason, L., Baldi, R., Di Ronco, S., Scrimin, S., Danielson, R. W., & Sinatra, G. M. (2017). Textual and graphical refutations: Effects on conceptual change learning. *Contemporary Educational Psychology, 49,* 275-288. https://doi.org/10.1016/j.cedpsych.2017.03.007

Mason, L. & Boscolo, P. (2004). Role of epistemological understanding and interest in interpreting a controversy and in topic-specific belief change. *Contemporary Educational Psychology, 29,* 103-128. https://doi.org/10.1016/j.cedpsych.2004.01.001

Mayer, R. E. (1989). Models for understanding. *Review of Educational Research*, *59*, 43-64. https://doi.org/10.3102/00346543059001043

McNamara, D. S. (2004). SERT: Self-explanation reading training. *Discourse Processes*, *38*, 1-30. https://doi.org/10.1207/s15326950dp3801_1

McNamara, D. S. (2017). Self-explanation and reading strategy training (SERT) improves low-knowledge students' science course performance. *Discourse Processes*, 1-14. https://doi.org/10.1080/0163853X.2015.1101328

McNamara, D. S., Boonthum, C., Levinstein, I. B., & Millis, K. (2007). Evaluating self-explanations in iSTART: Comparing word-based and LSA algorithms. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 227-241). Lawrence Erlbaum.

McNamara, D. S., Graesser, A. C., & Louwerse, M. M. (2012). Sources of text difficulty: Across genres and grades. In J. P. Sabatini, E. R. Albro, & T. O'Reilly (Eds.), *Measuring up: Advances in how to assess reading ability* (pp. 89-116). Rowman & Littlefield Publishers, Inc.

McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and instruction*, *14*(1), 1-43. https://doi.org/10.1207/s1532690xci1401_1

McNamara, D. S., O'Reilly, T. P., Best, R. M., & Ozuru, Y. (2006). Improving adolescent students' reading comprehension with iSTART. *Journal of Educational Computing Research*, *34*, 147-171. https://doi.org/10.2190/1RU5-HDTJ-A5C8-JVWE

Metcalfe, J., & Finn, B. (2013). Metacognition and control of study choice in children. *Metacognition and Learning 8*(1), 19-46. https://doi.org/10.1007/s11409-013-9094-7

Meyer, B. J. F., & Freedle, R. O. (1984). Effects of discourse types on recall. *American Educational Research Journal, 21*(1),121-143. https://doi.org/10.3102/00028312021001121

Millis, K. K., & Graesser, A. C. (1994). The time-course of constructing knowledge-based inferences for scientific texts. *Journal of Memory and Language*, *33*, 583-599. https://doi.org/10.1006/jmla.1994.1028

Millis, K., Magliano, J., & Todaro, S. (2006). Measuring discourse-level processes with verbal protocols and latent semantic analysis. *Scientific Studies of Reading*, *10*(3), 225-240. https://doi.org/10.1207/s1532799xssr1003_2

Millis, K. K., Simon, S., & tenBroek, N. S. (1998). Resource allocation during the rereading of scientific texts. *Memory & Cognition*, *26*(2), 232-246. https://doi.org/10.3758/BF03201136

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*, 488-496. https://doi.org/10.1037/0022-0663.91.3.488

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. The National Academies Press.

Nelson, T. O. (1984). A comparison of current measures of feeling-of-knowing accuracy. *Psychological Bulletin, 95*, 109-133. http://doi.org/10.1037/ 0033-2909.95.1.109

NGSS Lead States. (2013). *Next generation science standards: For states, by states.* The National Academies Press.

Noordman, L. G., Vonk, W., & Kempff, H. J. (1992). Causal inferences during the reading of expository texts. *Journal of Memory and Language*, *31*, 573-590. https://doi.org/10.1016/0749-596X(92)90029-W

Otero, J., León, J., & Graesser, A. C. (Eds.). (2002). *The psychology of science text comprehension*. Routledge.

Ozuru, Y., Kurby, C. A., & McNamara, D. S. (2012). The effect of metacomprehension judgment task on comprehension monitoring and metacognitive accuracy. *Metacognition and Learning, 7*(2), 113-131. https://doi.org/10.1007/s11409-012-9087-y

Palinscar, A. S., & Brown, A. L. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and Instruction*, *1*(2), 117-175. https://doi.org/10.1207/s1532690xci0102_1

Perfetti, C. A. (1985). *Reading ability*. Oxford University Press.

Rawson, K. A., Dunlosky, J., & Thiede, K. W. (2000). The rereading effect: Metacomprehension accuracy improves across reading trials. *Memory & Cognition*, *28*(6), 1004-1010. https://doi.org/10.3758/BF03209348

Rosenshine, B., & Meister, C. (1994). Reciprocal teaching: A review of the research. *Review of Educational Research*, *64*(4), 479-530. https://doi.org/10.3102/00346543064004479

Samuels, S.J., Tennyson, R., Sax, L., Mulcahy, P., Schermer, N., & Hajovy, H. (1988). Adults' use of text structure in the recall of a scientific journal article. *Journal of Educational Research, 81*(3), 171-174. https://doi.org/10.1080/00220671.1988.10885818

Sanchez, C., & Wiley, J. (2006). An examination of the seductive details effect in terms of working memory capacity. *Memory & Cognition, 34*, 344-355. https://doi.org/10.3758/BF03193412

Scheck, P., & Nelson, T. O. (2005). Lack of pervasiveness of the underconfidence-with-practice effect: Boundary conditions and an explanation via anchoring. *Journal of Experimental Psychology: General, 134*(1), 124-128. http://doi.org/10.1037/0096-3445.134.1.124

Shapiro, A. M. (2004). How including prior knowledge as a subject variable may change outcomes of learning research. *American Educational Research Journal*, *41*(1), 159-189. https://doi.org/10.3102/00028312041001159

Singer, M., Harkness, D., & Stewart, S. T. (1997). Constructing inferences in expository text comprehension. *Discourse Processes, 24*(2-3), 199-228. https://doi.org/10.1080/01638539709545013

Singer, M., & O'Connell, G. (2003). Robust inference processes in expository text comprehension. *European Journal of Cognitive Psychology*, *15*(4), 607-631. https://doi.org/10.1080/095414400340000079

Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education, 48,* 1273–1296. https://doi.org/10.1007/s11165-016-9602-2

Thiede, K. W., Griffin, T. D., Wiley, J., & Redford, J. (2009). Metacognitive monitoring during and after reading. In D. J. Hacker, J. Dunlosky & A. C. Graesser (Eds), *Handbook of metacognition in education* (pp. 85-106). Routledge.

Thiede, K. W., Wiley, J., & Griffin, T. D. (2011). Test expectancy affects metacomprehension accuracy. *British Journal of Educational Psychology*, *81*(2), 264-273. https://doi.org/10.1348/135910710X510494

Triona, L. M., & Klahr, D. (2003). Point and click or grab and heft: Comparing the influence of physical and virtual instructional materials on elementary school students' ability to design experiments. *Cognition and Instruction, 21,* 149-173. https://doi.org/10.1207/S1532690XCI2102_02

Tukey, J. W. (1977). *Exploratory data analysis*. Pearson.

van den Broek, P., Virtue, S., Everson, M. G., Tzeng, Y., & Sung, Y. C. (2002). Comprehension and memory of science texts: Inferential processes and the construction of a mental representation. In J. Otero, J. A. Leon, & A. C. Graesser (Eds.), *The psychology of science text comprehension* (pp. 131-154). Routledge.

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, *46*, 197-221. https://doi.org/10.1080/00461520.2011.611369

Ventura, M. J., Franchescetti, D. R., Pennumatsa, P., Graesser, A. C., Hu, G. J., & Cai, Z. (2004). Combining computational models of short essay grading for conceptual physics problems. In J. C. Lester, R. M. Vicari, & F. Paraguacu (Eds.), Proceedings of the Intelligent Tutoring Systems Conference (pp. 423-431). Springer.

Voss, J. F., & Silfies, L. N. (1996). Learning from history text: The interaction of knowledge and comprehension skill with text structure. *Cognition and Instruction*, *14*(1), 45-68. https://doi.org/10.1207/s1532690xci1401_2

Wade-Stein, D., & Kintsch, E. (2004). Summary Street: Interactive computer support for writing. *Cognition and instruction*, *22*, 333-362. https://doi.org/10.1207/s1532690xci2203_3

Weaver, C. A. (1990). Constraining factors in calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(2), 214-222. https://doi.org/10.1037/0278-7393.16.2.214

Weaver, C. A., & Bryant, D. S. (1995). Monitoring of comprehension: The role of text difficulty in metamemory for narrative and expository text. *Memory & Cognition*, *23*(1), 12-22. https://doi.org/10.3758/BF03210553

Weaver, C. A., Bryant, D. S., & Burns, K. D. (1995). Comprehension monitoring: Extensions of the Kintsch and van Dijk model. In C. A. Weaver, S. Mannes, & C. Fletcher (Eds.), *Discourse comprehension: Essays in honor of Walter Kintsch* (pp. 177-193). Erlbaum.

White, B. Y., & Frederiksen J. R. (1998). Inquiry, modeling, and metacognition: Making science accessible to all students. *Cognition and Instruction, 16,* 3-118. https://doi.org/10.1207/s1532690xci1601_2

White, R., & Gunstone, R. (1992). Prediction-observation-explanation. In R. White & R. Gunstone (Eds.), *Probing understanding* (pp. 44-64). The Falmer Press.

Wiley, J. (2019). Picture this! Effects of photographs, diagrams, animations, and sketching on learning and beliefs about learning from a geoscience text. *Applied Cognitive Psychology, 33*, 9-19. https://doi.org/10.1002/acp.3495

Wiley, J., Griffin, T. D., Jaeger, A. J., Jarosz, A. F., Cushen, P.J., & Thiede, K. W. (2016). Improving metacomprehension accuracy in an undergraduate course context. *Journal of Experimental Psychology: Applied, 22,* 393-405. http://dx.doi.org/10.1037/xap0000096

Wiley, J., Griffin, T. D., & Thiede, K. W. (2005). Putting the comprehension in metacomprehension. *The Journal of General Psychology*, *132*(4), 408-428. https://doi.org/10.3200/GENP.132.4.408-428

Wiley, J., Hastings, P., Blaum, D., Jaeger, A. J., Hughes, S., Wallace, P., Griffin, T. D., & Britt, M. A. (2017). Different approaches to assessing the quality of explanations following a multiple-document inquiry activity in science. *International Journal of Artificial Intelligence in Education, 27,* 758-790. https://doi.org/10.1007/s40593-017-0138-z

Wiley, J., & Myers, J. L. (2003). Availability and accessibility of information and causal inferences from scientific text. *Discourse Processes*, *36*, 109-129. https://doi.org/10.1207/S15326950DP3602_2

Wiley, J., & Voss, J. F. (1999). Constructing arguments from multiple sources: Tasks that promote understanding and not just memory for text. *Journal of Educational Psychology, 91*, 301-311. https://doi.org/10.1037/0022-0663.91.2.301

Wolfe, M. B., Tanner, S. M., & Taylor, A. R. (2013). Processing and representation of arguments in one-sided texts about disputed topics. *Discourse Processes, 50*, 457-497. https://doi.org/10.1080/0163853X.2013.828480

Yates, J. F. (1990). *Judgment and decision making*. Prentice-Hall.

Yore, L., Bisanz, G. L., & Hand, B. M. (2003). Examining the literacy component of science literacy: 25 years of language arts and science research. *International Journal of Science Education*, *25*(6), 689-725. https://doi.org/10.1080/09500690305018

Table 1

Means (and Standard Deviations) of Test Scores, Time on Task Measures, and Response Length by Study Strategy Training Condition

	Explanation $n = 173$		POE $n = 185$	
Test scores on learning activity	.545	(.127)	.511	(.127)
Time spent on study strategy training (min)	67.525	(27.872)	63.419	(29.905)
Time spent reading during learning activity (min)	31.915	(26.186)	32.002	(26.632)
Time spent on tests during learning activity (min)	17.213	(8.572)	17.475	(8.237)
Response length on training activity (words)	50.143	(34.099)	35.888	(15.427)

Table 2

Means (and Standard Deviations) of JOCs and Comprehension Monitoring Accuracy Measures by Study Strategy Training Condition and Time of Assessment

	Baseline Assessment		Learning Activity	
	Explanation	POE	Explanation	POE
JOC	.655 (.140)	.678 (.129)	.640 (.140)	.643 (.127)
Absolute Error	.276 (.105)	.272 (.096)	.254 (.091)	.256 (.096)
Confidence Bias	.134 (.183)	.131 (.178)	.102 (.184)	.126 (.174)
Relative Accuracy	.024 (.444)	.068 (.457)	.156 (.433)	.117 (.449)

Table 3

Example Responses Given During Study Strategy Training for the Cognitive Dissonance Text for Each Hand-coded Quality Scoring Category, with Frequencies by Category and Condition, and Average LSA Scores for Responses in Each Category

Score	Explanation	POE	Example Responses	Mean LSA Score
0	18.1%	21.2%	They support the theory because the children who received a mild warning were less likely to change their viewpoints because they did not receive a harsh or lack of warning. When you keep on warning about the toy to the children they are more likely to listen because they still learning what is ok and not ok. If you do not tell them anything they do not know any better. I got it wrong because I thought if a kid got a severe warning about playing with a certain toy they wouldn't play with it at all and pick a new favorite toy. Same thoughts going along with the end of the passage that I felt about it.	.26
1	58.5%	74.5%	There's an inconsistency between one's belief and their actions. Because the child received a mild warning, rather than no warning at all or a severe warning, their favoritism of their originally favorite toy was greatly reduced.	.39
2	21.1%	4.3%	Cognitive dissonance and the need to reduce it are greatest when the person cannot easily attribute their behavior to some external influence, such as being paid or forced by someone else. The mild warning given by the experiment represents an outside force acting on the children causing them to either change their beliefs or behaviors. As stated it is hard to change behavior, thus the children lowered their belief that a certain toy was their favorite. Similar to those in the control group and the 20 dollar group in the other experiment, the mild warning is similar to the 1 dollar group who was not well compensated and resulted in exhibiting dissonance.	.48

Note. Mean LSA (Latent Semantic Analysis) score is average semantic overlap of student responses with ideal response.

Table 4

Means (and Standard Deviations) of LSA and Hand-coded Quality Scores for Responses Given During Study Strategy Training by Condition

	LSA Score		Hand-Coded Quality Score	
	Explanation	POE	Explanation	POE
Placebo Effect	.39 (.16)	.45 (.11)	1.11 (.69)	.94 (.51)
Confirmation Bias	.36 (.13)	.29 (.11)	1.13 (.57)	.86 (.36)
Conformity & Obedience	.43 (.16)	.39 (.13)	1.00 (.56)	.89 (.47)
Fundamental Attribution Error	.42 (.14)	.39 (.11)	.85 (.67)	.69 (.50)
Cognitive Dissonance	.34 (.13)	.32 (.09)	1.07 (.64)	.77 (.49)
Overall	.39 (.15)	.37 (.13)	1.03 (.63)	.83 (.48)

Note. LSA (Latent Semantic Analysis) score is semantic overlap of student responses with ideal response.

Table 5

Simple Correlations Among Characteristics of Responses Given During Study Strategy Training and Test Scores on Learning Activity (Spearman's rho)

	Hand-Coded Quality Score	LSA Score	Evaluative Comments	Length of Responses	Prediction Accuracy
Hand-Coded Quality Score	-	.32***	-.03	.38***	.17***
LSA Score	.43***	-	-.16***	.42***	.13***
Evaluative Comments	-.08**	-.14***	-	.14***	-.36***
Length of Responses	.54***	.55***	.03	-	-.01
Test Scores on Learning Activity	.27***	.13***	-.11***	.22***	.17***

Note. POE condition only (n = 925) above the diagonal; Full sample below the diagonal (n = 1790).

*** $p < .001$.
** $p < .01$.
* $p < .05$.

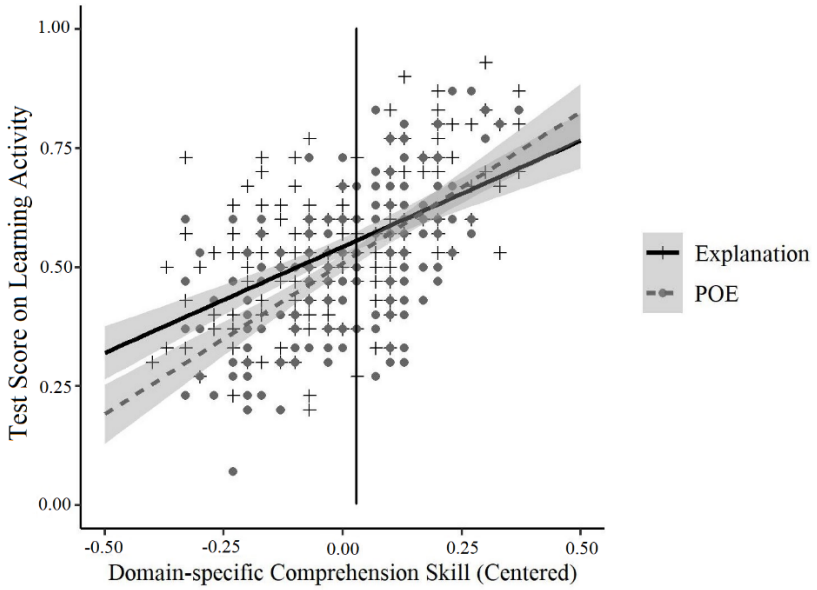
Figure 1

Procedure for Explanation and POE Study Strategy Training Conditions

Explanation Condition	POE Condition
Baseline Assessment (Week 1)	
1. Read 6 textbook excerpts 2. Make JOCs for each textbook excerpt 3. Complete inference test for each textbook excerpt	
Study Strategy Training Activity (Week 2)	
1. Goals for reading textbook excerpts Explanation support: General questions to keep in mind while reading excerpts	1. Goals for reading textbook excerpts Prediction and explanation support: How to use excerpt to make and justify experimental predictions
2. Read first section of Text 1 Write explanation	2. Read first section of Text 1 Make prediction Write explanation See example response as feedback
3. Read remainder of Text 1 Write explanation	3. Read remainder of Text 1 Write explanation See example response as feedback
4. Repeat for Texts 2-6 (no feedback)	4. Repeat for Texts 2-6 (no feedback)
Learning Activity (Week 3)	
1. Read 6 new textbook excerpts 2. Make JOCs for each textbook excerpt 3. Complete inference test for each textbook excerpt	

Figure 2

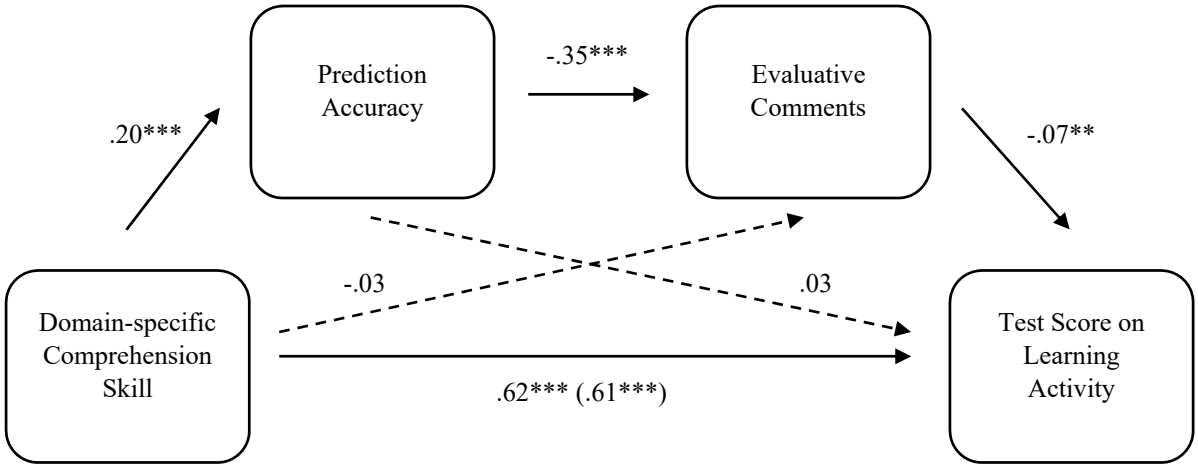
Test Score on Learning Activity by Study Strategy Training Condition as Predicted by Domain-specific Comprehension Skill



Note. Johnson-Neyman technique indicated that conditions differ to the left of the vertical line (.04).

Figure 3

Sequential Mediation Model



Note. Standard regression coefficient for the relationship between domain-specific comprehension skill and inference test score on learning activity as mediated by prediction accuracy and inclusion of evaluative comments. The standardized regression coefficient between domain-specific comprehension skill and inference test score on the learning activity, controlling for prediction accuracy and evaluative comments, is in parentheses.

*** $p < .001$.
** $p < .01$.

Appendix

Table A1

Text and Test Characteristics

	Word Count	Flesch Kincaid Grade Level	Average Test Score *M (SD)*	Test-ACT Correlation (*n* = 170)
Baseline Assessment				
Placebo Effect	784	13.5	.54 (.24)	.26
Confirmation Bias	739	10.5	.48 (.27)	.29
Self-Control	879	11.6	.69 (.31)	.35
Conformity & Obedience	675	12.4	.60 (.24)	.39
Fundamental Attribution Error	767	12.9	.47 (.25)	.36
Cognitive Dissonance	863	12.5	.44 (.24)	.25
Learning Activity				
Classical Conditioning	836	11.0	.63 (.28)	.33
Operant Conditioning	812	12.3	.60 (.24)	.22
Observational Learning	657	11.7	.43 (.24)	.31
False Memories	868	10.1	.57 (.25)	.25*
Twin Studies	762	14.2	.49 (.27)	.39
Aphasias	648	11.0	.44 (.23)	.19

Note. Correlations between inference test scores and ACT (standardized test scores) from separate norming studies.

*$n = 67$