Enhancing Students' Ability to Correct Misconceptions in Natural Selection

with Refutational Texts and Self-Explanation

by

Micah N. Watanabe

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Arts

Approved November 2020 by the
Graduate Supervisory Committee

Danielle McNamara, Chair
Gene Brewer
Carla Firetto

ARIZONA STATE UNIVERSITY

December 2020

ABSTRACT

This study examined the effects of different constructed response prompts and text types on students' revision of misconceptions, comprehension, and causal reasoning. The participants were randomly assigned to prompt (self-explain, think-aloud) and text type (refutational, non-refutational) in a 2x2, between-subjects design. While reading, the students were prompted to write responses at regular intervals in the text. After reading, students were administered the conceptual inventory of natural selection (CINS), for which a higher score indicates fewer misconceptions of natural selection. Finally, students were given text comprehension questions, and reading skill and prior knowledge measures. Linear mixed effects (LME) models showed that students with better reading skill and more prior knowledge had a higher CINS score and better comprehension compared to less skilled students, but there were no effects of text type or prompt. Linguistic analysis of students' responses demonstrated a relationship of prompt, text, and reading skill on students' causal reasoning. Less skilled students exhibited greater causal reasoning when self-explaining a non-refutational text compared to less skilled students prompted to think-aloud, and less skilled students who read the refutational text. The results of this study demonstrate a relationship between reading skill and misconceptions in natural selections. Furthermore, the linguistic analyses suggest that less skilled students' causal reasoning improves when prompted to self-explain.

TABLE OF CONTENTS

Page

LIST OF TABLES iv

LIST OF FIGURES vi

INTRODUCTION 1

LITERATURE REVIEW 3

Comprehension 3

Misconceptions 4

Refutational Texts 6

Self-Explanation 8

THE CURRENT STUDY 10

METHOD 12

Participants 12

Design 13

Materials 13

Procedure 19

RESULTS 20

Preliminary Analyses 20

Misconceptions 25

Comprehension 27

Cohesion 29

DISCUSSION 43

Page

REFERENCES.. 45

APPENDIX... 52

A: REFUTATIONAL AND NON-REFUTATIONAL TEXTS............................ 52

B: CONSTRUCTED RESPONSE PROMPTS... 55

C: CONCEPTUAL INVENTORY OF NATURAL SELECTION....................... 58

D: TEXT COMPREHENSION QUESTIONS.. 65

E: PRIOR SCIENCE KNOWLEDGE TEST.. 69

F: PRELIMINARY MEDIATION ANALYSIS... 76

G: UNIVERSITY APPROVAL FOR HUMAN SUBJECT RESEARCH............ 78

LIST OF TABLES

Table Page

1. Example Misconceptions in Three Science Domains.. 5

2. Example of Content Word Overlap.. 17

3. Example of Argument Overlap.. 18

4. Example of Verb Overlap.. 19

5. Example of Causal Ratio.. 20

6. Means, Ranges, and Distributions: Pretest, Comprehension Questions, Conceptual Inventory of Natural Selection (CINS), and Individual Difference Measures.............. 21

7. Means and Standard Deviations: Proportion of Correct Answers on the Conceptual Inventory on Natural Selection as a Function of Condition... 25

8. LME Results Predicting Score on the Conceptual Inventory of Natural Selection with Vocabulary and Prior Knowledge.. 25

9. LME Results Predicting Score on The Conceptual Inventory of Natural Selection with Vocabulary, Prior Knowledge, Text Type, and Constructed Response Prompt........... 26

10. Means and Standard Deviations: Average Comprehension Score as a Function of Condition.. 27

11. LME Results Predicting Comprehension Score with Vocabulary and Prior Knowledge.. 27

12. LME Results Predicting Comprehension Score with Vocabulary, Prior Knowledge, Text Type, and Constructed Response Prompt.. 28

Table Page

13. Linear Regression Predicting Content Word Overlap as a Function of Condition with Covariates 30

14. Linear Regression Predicting Content Word Overlap as a Function of Condition and Interactions with Vocabulary.. 31

15. Linear Regression Predicting Content Word Overlap as a Function of Condition and Interactions with Prior Knowledge... 34

16. Linear Regression Predicting Verb Overlap as a Function of Condition and Interactions with Vocabulary... 37

17. Linear Regression Predicting Verb Overlap as a Function of Condition and Interactions with Prior Knowledge.. 40

LIST OF FIGURES

Figure Page

1. Scatterplots, Distributions, and Correlations of Vocabulary, Prior Knowledge, and Linguistic Indices 22

2. Scatterplots, Distributions, and Correlations of Vocabulary, Prior Knowledge, and Linguistic Indices 23

3. Content Word Overlap of Responses as a Function of Condition at 1 Standard Deviation Above the Mean of Vocabulary, Holding Prior Knowledge Constant 32

4. Content Word Overlap of Responses as a Function of Condition at the Mean of Vocabulary, Holding Prior Knowledge Constant 32

5. Content Word Overlap of Responses as a Function of Condition at 1 Standard Deviation Below the Mean of Vocabulary, Holding Prior Knowledge Constant 33

6. Content Word Overlap of Responses as a Function of Condition at 1 Standard Deviation Above the Mean of Prior Knowledge, Holding Vocabulary Constant 34

7. Content Word Overlap of Responses as a Function of Condition at the Mean of Prior Knowledge, Holding Vocabulary Constant 35

8. Content Word Overlap of Responses as a Function of Condition at 1 Standard Deviation Below the Mean of Prior Knowledge, Holding Vocabulary Constant 35

9. Verb Overlap of Responses as a Function of Condition at 1 Standard Deviation Above the Mean of Vocabulary, Holding Prior Knowledge Constant 38

10. Verb Overlap of Responses as a Function of Condition at the Mean of Vocabulary, Holding Prior Knowledge Constant 38

Figure Page

11. Verb Overlap of Responses as a Function of Condition at 1 Standard Deviation Below the Mean of Vocabulary, Holding Prior Knowledge Constant.................................... 39

12. Verb Overlap of Responses as a Function of Condition at 1 Standard Deviation Above the Mean of Prior Knowledge, Holding Vocabulary Constant.................................... 40

13. Verb Overlap of Responses as a Function of Condition at The Mean of Prior Knowledge, Holding Vocabulary Constant.. 41

14. Verb Overlap of Responses as a Function of Condition at 1 Standard Deviation Below the Mean of Prior Knowledge, Holding Vocabulary Constant.................................... 41

Introduction

Students of all ages are constantly processing, and potentially learning, new information in various domains such as science, history, and art. Learning can occur in different mediums, for example reading an article, watching a documentary, or listening to a podcast. Regardless of medium, this learning process relies on the integration of new information with prior knowledge. Sometimes this process of learning and integration goes awry, and misconceptions are formed.

Inaccurate understanding and knowledge or *misconceptions* are ubiquitous across all domains. Some misconceptions can be low-stakes, for instance believing seasons are due to the Earth's orbit when in fact seasons are caused by the Earth's axial tilt. However, other misconceptions can have large scale implications. For example, misconceptions about climate change have caused governmental policies with immense financial and human costs (Sterman, 2008). Misconceptions typically fit well with related knowledge and are difficult to identify. Furthermore, misconceptions hinder learning by preventing learners from making inferences to connect inaccurate knowledge and new information. Indeed, misconceptions are an ongoing issue in education, government, and science. Thus, it is important to understand the processes by which misconceptions are formed, and effective means to refute them.

The processes behind conceptual change rely on a complex framework including, but not limited to, epistemology, knowledge activation and inhibition, and causal reasoning. Refutational texts are one method that researchers have used to counter misconceptions (see Allen, McCrudden, & McNamara, 2015; van den Broek & Kendeou, 2008; Watanabe, McCarthy, & McNamara, 2018). Refutational texts target specific misconceptions by stating the misconception, providing the correct concept, and giving evidence on why the correct concept is true. Studies have shown that students hold fewer misconceptions after reading a refutational text compared to a non-refutational text (van den Broek & Kendeou, 2008). In addition, students who read refutational texts use more language indicative of causal reasoning compared to students who read a similar, non-refutational text (Kendeou et al., 2011; van den Broek & Kendeou, 2008). However, because refutational texts are often challenging science texts (see Allen, McNamara & McCrudden, 2015; Kendeou et al., 2011; Kendeou, Braasch, & Braten, 2016), the effectiveness of refutational texts may also depend on the degree to which students can comprehend the text.

One technique to improve students' comprehension and causal reasoning while reading is prompting students to self-explain. Self-explanation is the practice of explaining the text to oneself while reading. This includes paraphrasing, monitoring comprehension, and asking questions. Skilled readers self-explain naturally, and less skilled readers prompted to self-explain demonstrate gains in comprehension and increases in causal connections while reading (McNamara, 2004). Therefore, self-explanation prompts have the potential to enhance students' comprehension of refutational texts and the effectiveness of the texts.

The current study assessed the degree to which self-explanation, in combination with refutational or non-refutational texts, improves students' ability to correct their misconceptions about natural selection. Students received different types of reading prompts while reading either a refutational or non-refutational text. It was hypothesized that students who self-explain would better comprehend the texts, leading to fewer misconceptions.

Literature Review

Comprehension

Comprehension is the processing of new information to extract meaning (McNamara & Magliano, 2009) which relies on a number of lower and higher-order cognitive processes (Balota, Flores d'Arcais, & Rayner, 1990). In addition, prior knowledge of the topic is one of the strongest predictors of comprehension (McNamara & Kintsch, 1996). Text comprehension is generally examined in the context of theoretical models of how readers create and access mental representations or mental models of text (Cote & Goldman, 1999; Johnson-Laird, 1983; McNamara & Magliano, 2009; van den Broek et al., 2002). The majority of text comprehension models include at least two levels in mental representations of a text: a surface structure, which refers to the explicit information in the text, and a deeper structure, which refers to the underlying meaning and concepts in a text (McNamara & Magliano, 2009). According to the construction-integration model (CI model; Kinstch, 1988; 1998), the mental representation includes three levels: (1) The *surface code*, which refers to the exact words and syntax used in the text; (2) the *textbase*, which refers to the explicit meaning of the words and sentences in the text; and (3) the *situation model*, which refers to the deeper structure of the text. The situation model is reflected by the degree to which the reader integrates the information in the text with prior knowledge by generating inferences to fill in the gaps in the text (Kintsch, 1988, 1998). The stability of the situation model depends on the ability of the reader to generate these inferences. A stable situation model is necessary for the reader to apply the new information to future learning. Since comprehension relies on connecting prior knowledge to the new information through inferencing, comprehension can be compromised when the reader has misconceptions about the topic.

Misconceptions

Misconceptions are inaccuracies within mental representations that arise from incorrect knowledge (van den Broek & Kendeou, 2008). Misconceptions can develop when attempts to generate inferences are made without adequate prior knowledge of a topic. While attempting to connect the new information with insufficient prior knowledge, the learner generates an inference that is intuitive and fits well with prior knowledge. Table 1 contains a list of some common misconceptions. For example, many students learn the Earth's orbit around the sun is an elliptical and make the inference that seasons are caused by the earth's orbit; when the earth is close to the sun it is summer, and when it is far from the sun it is winter. However, this is an inaccurate inference: seasons are caused by the axial tilt of the Earth.

Table 1.
Example misconceptions in three science domains

Field	Misconception	Correct Concept
Astronomy	Seasons are caused by Earth's orbit around the sun.	Seasons are caused by the axial tilt of the Earth.
Astronomy	The Great Wall of China is the only man-made object visible from space	The Great Wall of China is not visible from space, however city lights are easily visible from space.
Biology	Species adapt to their environment in order to survive.	Random mutations in genetic code produce beneficial matches between organism & environment which then propagate
Biology	Plants get their food from the soil.	Interactions between photosynthesis and respiration create sugars the plant uses for energy.
Physics	The Coriolis effect influences the rotation of draining water (clockwise or counter-clockwise) depending on the hemisphere.	The Coriolis effect is far weaker than any minor rotation that is present when the water begins to drain.
Physics	Lightning never strikes the same place twice.	Lightning is more likely to strike similar areas based on conductivity.

Inaccurate inferencing creates misconceptions that do not accurately reflect the scientific reality (Guzzetti et al., 1993; van den Broek & Kendeou, 2008). Misconceptions differ from inaccurate knowledge by implying inaccurate causal relationships. For example, a student may erroneously think that hydrogen has two protons when it only has one; a misattribution of characteristics or qualities of an object. This misattribution is different compared to the misconception of natural selection. Greater than 75% of people think species adapt to their environment, when in fact the random mutations in genetic codes produce beneficial matches between organisms and environment, encouraging reproduction of that code (Coley & Tanner, 2015).

Misconceptions can pose significant problems in comprehension by interfering with future attempts to learn new information (Feltovich, Coulson, & Spiro, 2001). This is because misconceptions often develop at a young age, are difficult to identify, and are resistant to change (van den Broek & Kendeou, 2008). As such, investigating the processes involved in effectively correcting misconceptions (conceptual change) has been considered an important avenue of research (Vosniadou, 2003).

Refutational Texts

One method of reducing readers' misconceptions is through the reading of refutational texts. A refutational text is a text either written or adapted specifically to refute a misconception or set of misconceptions in a topic area, typically in science (Sinatra & Broughton, 2011). Refutational texts are defined by three characteristics: (a) They contain a definition of a common misconception, (b) there are explicit statements addressing the inaccuracies in these beliefs, and (c) these statements are followed by explanations and evidence of the correct view of the concept (Guzzetti et al., 1993).

Refutational texts may be more effective than other types of texts in correcting misconceptions due to co-activation (Kendeou et al., 2014). In refutational texts, both the misconception and the refutation are activated at the same time, increasing the likelihood the reader identifies and confronts the error (Kendeou & van den Broek, 2007; van den Broek & Kendeou, 2008). This *co-activation hypothesis* is supported from past work demonstrating that readers spend more time reading a refutational text compared to a control text. Longer reading times indicate the reader detects the contradiction and is working to resolve it (Kendeou & van den Broek, 2007).

Reading a refutational text has been shown to improve learning of a topic compared to reading a text that simply repeats the correct information (Braasch, Goldman, & Wiley, 2013) and decrease the number of misconceptions compared to reading an expository text on the same topic (Broughton, Sinatra, & Reynolds, 2010). Furthermore, there is evidence that reading a refutational text increases the frequency at which readers use language indicative of conceptual change (i.e., “Because of this…”, “This makes me believe…”) in think-aloud responses to the text compared to participants reading a control text (Kendeou et al., 2011; van den Broek & Kendeou, 2008).

Though some studies report positive effects of refutational texts on retention of science knowledge (Kendeou et al., 2011; Kendeou & van den Broek, 2007), other studies report null results (Ariasi & Mason, 2011; Mason & Gava, 2007). One explanation for these inconsistent findings is that students may be struggling to comprehend the refutational texts. These texts are often about complex scientific phenomena (Kendeou & van den Broek, 2007; Mason & Gava, 2007) and readers often struggle to understand texts of this nature (O’Reilly & McNamara, 2007).

One alternative hypothesis stems from the *knowledge activation theory* (McNamara & O’Reilly, 2009; McNamara, O’Reilly, & de Vega, 2007). Accordingly, if a reader possesses more prior knowledge on a topic, the reader can more quickly resolve ambiguous words, phrases, and ideas while reading the text. For example, Bransford and Johnson (1972) presented ambiguous texts to readers with or without a title. When reading a text with a title, readers recalled approximately twice as much compared to reading a text without a title. The presence of a title affords the reader the opportunity to activate related prior knowledge while reading.

In addition, readers with more prior knowledge on the topic generate more inferences between the text and their prior knowledge which contributes to the coherence and stability of the reader's mental representation of the text (McNamara & Kintsch, 1996). Thus, the effectiveness of refutational texts may be limited by the degree to which students comprehend the text. While high-knowledge or skilled readers may be able to generate inferences and develop a robust situation model of a refutational text, low-knowledge or less skilled students may struggle to decode the surface level information. As a result, less successful readers may be unable to activate the relevant knowledge, which prevents co-activation of the information and hinders conceptual change processes.

Self-explanation

Self-explanation is used by teachers and researchers to encourage activation of prior knowledge and the generation of inferences. Self-explanation is the act of explaining the text to oneself while reading (McNamara, 2004). Providing self-explanation reading training to students has been demonstrated to enhance students' comprehension of difficult science texts. This effect has been observed for students with different levels of prior knowledge (McNamara, 2015; O'Reilly, Best, & McNamara, 2004; Ozuru et al., 2010).

Rereading a text or being prompted to think-aloud while reading does not significantly improve comprehension of a text (Millis, Golding, & Barker, 1995). In comparison, self-explanation in combination with reading strategy instruction and training alters the comprehension and learning process while reading (Allen, McCrudden, & McNamara, 2015). Previous research has shown self-explanation primarily enhances the construction of causal connections between events, rather than referential connections among concepts (Legare & Lombrozo, 2014) and promotes more coherent mental representations of text (Allen, Snow, & McNamara, 2015; McNamara, 2004).

Allen, McCrudden, and McNamara (2015) examined the effects of prompting students to self-explain a refutational text. Students who were prompted to self-explain scored higher on a conceptual test than those who were prompted to think-aloud or reread. Linguistic analyses indicated that while reading, students prompted to self-explain engaged in more causal reasoning compared to those prompted to think-aloud. This was observed through measures of causal cohesion. In general, cohesion of students' self-explanations is a strong predictor of ability to comprehend texts (Allen, Snow, & McNamara, 2015). These findings indicate that conceptual change does not solely rely on the type of texts presented to students – it also depends on the comprehension processes students employ while reading the text.

The collective body of research suggests self-explanation enhances conceptual change processes in two ways (a) improving students' ability to comprehend the text and (b) improving students' ability to create causal connections between events in the text. While refutational texts have been shown to increase the causal language students use in think-aloud responses (Kendeou et al., 2011; van den Broek & Kendeou, 2008), the effectiveness of a refutational text depends on the comprehension processes the reader employs (Allen, McCrudden, & McNamara, 2015). Therefore, self-explanation may enhance the conceptual change process even when students are reading a non-refutational text. In addition, self-explanations are more easily transferable between domains than refutational texts because most refutational texts must be crafted individually for specific misconceptions.

The Current Study

The current study examined the extent to which prompting students to self-explain a non-refutational or refutational text affected their understanding of natural selection in comparison to thinking-aloud while reading a non-refutational text.

Comprehension

H1a: ***Prior Knowledge.*** In line with the literature on prior knowledge and comprehension, it was hypothesized that students with high prior knowledge would have higher scores on a comprehension test compared to students with low prior knowledge.

H1b: ***Reading Skill.*** In line with the literature on reading skill and comprehension, it was hypothesized that skilled readers would have higher scores on a comprehension test compared to less skilled readers.

H1c: Based on the body of research on self-explanation and comprehension, it can be hypothesized that participants prompted to self-explain while reading would have higher scores on a comprehension test compared to those prompted to think-aloud while reading.

H2a: Consistent with previous research on self-explanation and comprehension, an interaction between prior knowledge and constructed response prompt was expected such that the effect of prior knowledge would be attenuated when students self-explained as compared to when students engaged in think-aloud.

H2b: Consistent with previous research on self-explanation and comprehension, an interaction between reading skill and constructed response prompt was expected such that the effect of reading skill would be attenuated when students self-explained as compared to when students engaged in think-aloud.

H3: According to the co-activation hypothesis, an interaction between text type (refutational, non-refutational) and constructed response prompt (self-explain, think-aloud) was not expected to predict differences in comprehension score.

Linguistic Analyses of Responses

H4: Based on the body of research on referential and causal cohesion in constructed responses, an effect of either text type (refutational, non-refutational) and constructed response prompt (self-explain, think-aloud) on referential cohesion in constructed responses was not expected.

H5a: In line with the co-activation hypothesis, it was hypothesized that students who read a refutational text would demonstrate greater causal cohesion in constructed responses and have higher scores on a conceptual knowledge test compared to students who read a non-refutational text.

H5b: According to the knowledge activation hypothesis, it was hypothesized that students who were prompted to self-explain while reading would demonstrate greater causal cohesion in constructed responses and have higher scores on a conceptual knowledge test compared to students who were prompted to think-aloud while reading. This finding would replicate the study conducted by Allen, McCrudden, and McNamara, (2015).

H6a: According to the co-activation hypothesis, an interaction between text type (refutational, non-refutational) and constructed response prompt (self-explain, think-aloud) was not expected to predict differences in causal cohesion of students' constructed responses.

H6b: Consistent with the knowledge activation hypothesis, the effect of text type (refutational, non-refutational) on the causal cohesion in the students' constructed responses would depend on constructed response prompt (self-explain, think-aloud). This interaction between text type and constructed response prompt would occur such that the effect of text type would be attenuated when students self-explained as compared to when students engaged in think-aloud.

Method

Participants

A power analysis for linear regression was conducted using the program G*Power 3. In a similar study comparing the effects of constructed response prompt on misconceptions, there was a medium effect size (Cohen's $d = 0.55$) such that prompting self-explanation reduced students' misconceptions compared to prompting think-aloud (Allen, McCrudden, & McNamara, 2015). In two similar experiments on refutational texts, Kendeou et al. (2014) demonstrated effect sizes of $d = 1.08$ and $d = 0.44$. Therefore, in the power analysis for the current study, the parameters entered were a small-medium effect size ($d = 0.4$), and 5 predictors (Vocabulary and prior knowledge as covariates, and main effects and interactions of constructed response prompt and text type), with $\alpha = .05$. The results of the power analysis showed a total sample of 215 was adequate for the planned analyses. That sample size was rounded up to 240 to ensure adequate power and equal group sizes.

Participants (*n* = 240) were recruited from introductory psychology courses and offered course credit for participating. Participants reported a number of ethnic backgrounds with the majority being White (62%), Hispanic (21%), Asian (13%), and African-American (4%). Approximately 60% of the participants were women, and approximately 65% were freshmen.

Design

Students were randomly assigned to one of four conditions in a 2 (Constructed Response Task: self-explanation, think-aloud) x 2 (Text: refutational text, non-refutational text) between-subjects design.

Materials

Texts. The non-refutational text (707 words, 8 paragraphs, Flesch-Kincaid Grade Level: 11) is an excerpt from How the Mind Works (Pinker, 1997), which describes the concept of natural selection. The author uses the example of the eye to explain how the world can appear to be the product of intelligent design but does not have a designer.

The text was adapted by Allen, McCrudden, and McNamara (2015) to include explicit references to, and direct refutations of, alternative concepts of natural selection. The adapted version has similar length and reading difficulty (716 words, 8 paragraphs, Flesch-Kincaid Grade Level: 10.5) as the original text. Appendix A contains both the original non-refutational and adapted refutational versions of the text. The text was selected because of the prevalence of misconceptions regarding natural selection and the existence of a validated scale (The CINS) which measures misconceptions in natural selection.

Constructed response prompt. Prior to reading the text, participants were instructed to either self-explain or think-aloud during reading using the same constructed response instructions from Allen, McCrudden, and McNamara (2015).

The self-explanation instructions were:

You can see that an explanation doesn't just restate the passage. It explains what the passage means. You can use anything you know about the sentence to explain it. After this definition, students were provided with a sample passage and example self-explanation.

The think-aloud instructions were:

You can see that your think-aloud can be about any thoughts you had while reading.

Participants were provided with a sample passage and example think-aloud. During reading, students were prompted to either self-explain or think-aloud for 14 target sentences. See Appendix B for the full self-explanation and think-aloud prompts.

Conceptual Inventory of Natural Selection. The Conceptual Inventory of Natural Selection (CINS; Anderson et al., 2002) is a 20-item multiple-choice test that evaluates both accurate ideas and common misconceptions related to natural selection. Higher scores indicate a more accurate concept of natural selection (and fewer misconceptions). This assessment has been used in studies of natural selection concepts, including those using the excerpted Pinker text (Allen, McCrudden, & McNamara, 2015;

Watanabe et al., 2018) and is considered to be a reliable measure (Gregory, 2009; Nehm & Shonfield, 2008: α =~0.65).

Two expert raters independently compared the CINS and target text and identified five questions on the CINS that contained concepts that could not be learned from the information in the text. These five questions were given to students in a pre-survey to assess baseline misconceptions on natural selection to assess a priori misconceptions about natural selection. All 20 items were given to students after they read the refutational text. See Appendix C for the full CINS.

Text Comprehension Questions. A set of 14 short-answer comprehension questions were created and piloted for use in this study (see Appendix D). The questions required students to inference between two sentences while reading and could be answered from the material in both the refutational and non-refutational texts. A pilot study was conducted to test the validity of the comprehension questions. Five of the items were selected to be included in the current study because they correlated with measures vocabulary and prior science knowledge, and because student scores on the five items were normally distributed. A sample question from the five-item comprehension assessment is below:

Why do replicators accumulate changes for the better?

The five selected questions were used to assess students' text comprehension of the texts. The students' responses were scored independently by two expert raters (κ = 0.79) in 0.5 increments from 0 to 1 based on the completeness of the response.

Gates-MacGinitie Vocabulary Test. The vocabulary test from the Gates–MacGinitie Vocabulary Test served as a proxy for reading skill (GMVT; MacGinitie & MacGinitie, 1989). It is a standardized test with strong psychometric properties (α= .85-.92; Phillips et al., 2002). The test consists of 45 multiple-choice questions in which a word is presented in the context of a sentence and students must select the word or phrase most synonymous with the target word.

Science prior knowledge. Participants were given a 20-item science knowledge assessment (See Appendix E) used in several studies on reading comprehension and validated with over 4,000 high school students. The test contains multiple-choice questions on biology, chemistry, physics, and mathematics, and yields a single proportion score. Previous studies using this assessment have demonstrated the test has good reliability, with a range from α = .71-.74 (see O'Reilly, Best, & McNamara, 2004; O'Reilly & McNamara, 2007). Inclusion of the prior knowledge test adds the concern of priming prior knowledge during reading. Low-knowledge students are particularly disadvantaged when prior knowledge tests are provided prior to reading (McNamara & Kintsch, 1996). As such, presenting additional cues in prior questions can influence readers' comprehension of science texts. Thus, the prior knowledge test was presented at the end of the study to prevent priming effects for students (O'Reilly & McNamara, 2007).

Cohesion measures. The students' constructed responses were aggregated into a single text file and analyzed using Coh-Metrix (McNamara & Graesser, 2012), and the

Tool for the Automatic Analysis of Cohesion (TAACO: Crossley, Kyle, & Dascalu, 2018) to derive cohesion indices.

Two measures of referential cohesion were used from TAACO: content word overlap (*adjacent_overlap_cw_sent*) and argument overlap (*adjacent_overlap_argument_sent*). Table 2 shows an example of content word overlap. Content word overlap is calculated by summing all of the content word lemma that occur in the first sentence and the second sentence, then summing all the content word lemma that occur in the second sentence and the third sentence, and continuing for all adjacent sentences beginning with the first sentence in the text through the second-to-last sentence. The total sum of all the content word lemma that overlap in adjacent sentences is divided by the total number of content word lemma in the text except for content word lemma in the last sentence.

Table 2.

Example of content word overlap

Sentence	Overlapping lemmas	Total lemmas
1. The ability to move is a very special ability.	1 (is)	4 (ability, is, very, special)
2. There is such an abundance of things that cannot move around us.	0	5 (is, such, abundance, things, move)
3. We have the ability to get from place a to place b on our own because of evolution.		
Content word sums	1	9
Content word overlap	1 / 9 = 0.11	

Table 3 shows an example of argument overlap. Argument overlap is calculated in the same manner as content word overlap, but using noun and pronoun lemmas instead of content word lemma.

Table 3.

Example of argument overlap

Sentence	Overlapping lemmas	Total lemmas
1. The ability to move is a very special ability.	0	1(ability)
2. There is such an abundance of things that cannot move around us.	1 (us/our/we)	3 (abundance, things, us)
3. We have the ability to get from place a to place b on our own because of evolution.		
Argument sums	1	4
Argument overlap	1 / 4 = 0.25	

One measure of causal cohesion was used from TAACO, and one measure of causal cohesion from Coh-Metrix. TAACO was used to derive verb overlap (*adjacent_overlap_verb_sent).* Table 4 shows an example of verb overlap. Verb overlap is calculated the same as content word overlap, but using verb lemmas instead of content word lemmas.

Table 4.

Example of verb overlap`

Sentence	Overlapping lemmas	Total lemmas
1. The ability to move is a very special ability.	1 (is)	1 (is)
2. There is such an abundance of things that cannot move around us.	0	1 (is, move)
3. We have the ability to get from place a to place b on our own because of evolution.		
Verb sums	1	3
Verb overlap	1 / 3 = 0.33	

Table 5 shows an example of causal ratio. Coh-Metrix was used to derive causal ratio (*coh_causr*). Causal ratio is measured by summing the average ratio of causal particles to causal verbs in each sentence with a causal verb, and dividing that sum by the number of sentences with causal verbs + 1.

Table 5.

Example of causal ratio

Sentence	Causal particles	Causal verbs	Ratio
1. Living things move around because of evolution.	1 (because)	1 (move)	1
2. For example, birds fly through the sky.	0	1 (fly)	0
3. However, non-living things are always the same.	0	0	N/A
Causal ratio sum		1	
Sentences with a causal verb + 1		3	
Causal ratio		0.33	

Procedure

The pretest was administered to students in a larger survey which included questions from other researchers. The pretest was given approximately two months prior to the study, and only a subset of participants completed the pretest. The entire sample of students were recruited through the university's online recruiting portal and given course credit for participation. The participants were told that the study focused on science learning and involved a series of reading, writing, and memory tasks. Participants were randomly assigned to prompt (self-explain, think-aloud) and text type (refutational, non-refutational) in a 2x2, between-subjects design. The students were given instructions and an example for their assigned reading prompt (self-explain, think-aloud). Students then read their assigned text (refutational, non-refutational) and were prompted to write 14 responses at regular intervals in the text. After reading, students were administered the conceptual inventory of natural selection, text comprehension questions, demographics, and reading skill and prior knowledge measures.

Results

Preliminary Analyses

Table 6 provides the grand means, standard deviations, range, skew, and kurtosis for pretest, conceptual inventory of natural selection (CINS), text comprehension, prior knowledge, and vocabulary scores.

Scores on the CINS and text comprehension were normally distributed. Pretest scores had a moderate negative skew. The vocabulary scores had a moderate negative skew, and the prior knowledge scores had a high negative skew, indicating that the sample of students had high reading skill and prior knowledge.

Table 6.

Means, Ranges, and Distributions: Pretest, comprehension questions, conceptual inventory of natural selection (CINS), and individual difference measures

Measure	Mean (*SD*)	Range	Skew	Kurtosis
Pretest	0.52 (0.29)	0.0 - 1.0	0.69	-0.85
CINS	0.47 (0.17)	0.06 - 0.93	0.28	-0.06
Text Comprehension	0.30 (0.18)	0.0 - 0.8	0.34	-0.15
Prior Knowledge	0.78 (0.19)	0.11 - 1.00	-1.10*	0.91
Vocabulary	0.73 (0.16)	0.27 - 1.00	-0.53	-0.37

Figure 1 presents the distributions, scatterplots, and correlations of the dependent variables and individual difference. Figure 2 presents the distributions, scatterplots, and correlations of the individual differences and linguistic indices.

For the survey measures, the pretest, CINS, prior knowledge, and vocabulary all had moderate correlations with each other; prior knowledge and vocabulary had a high correlation. The comprehension questions had low correlations with the other measures.

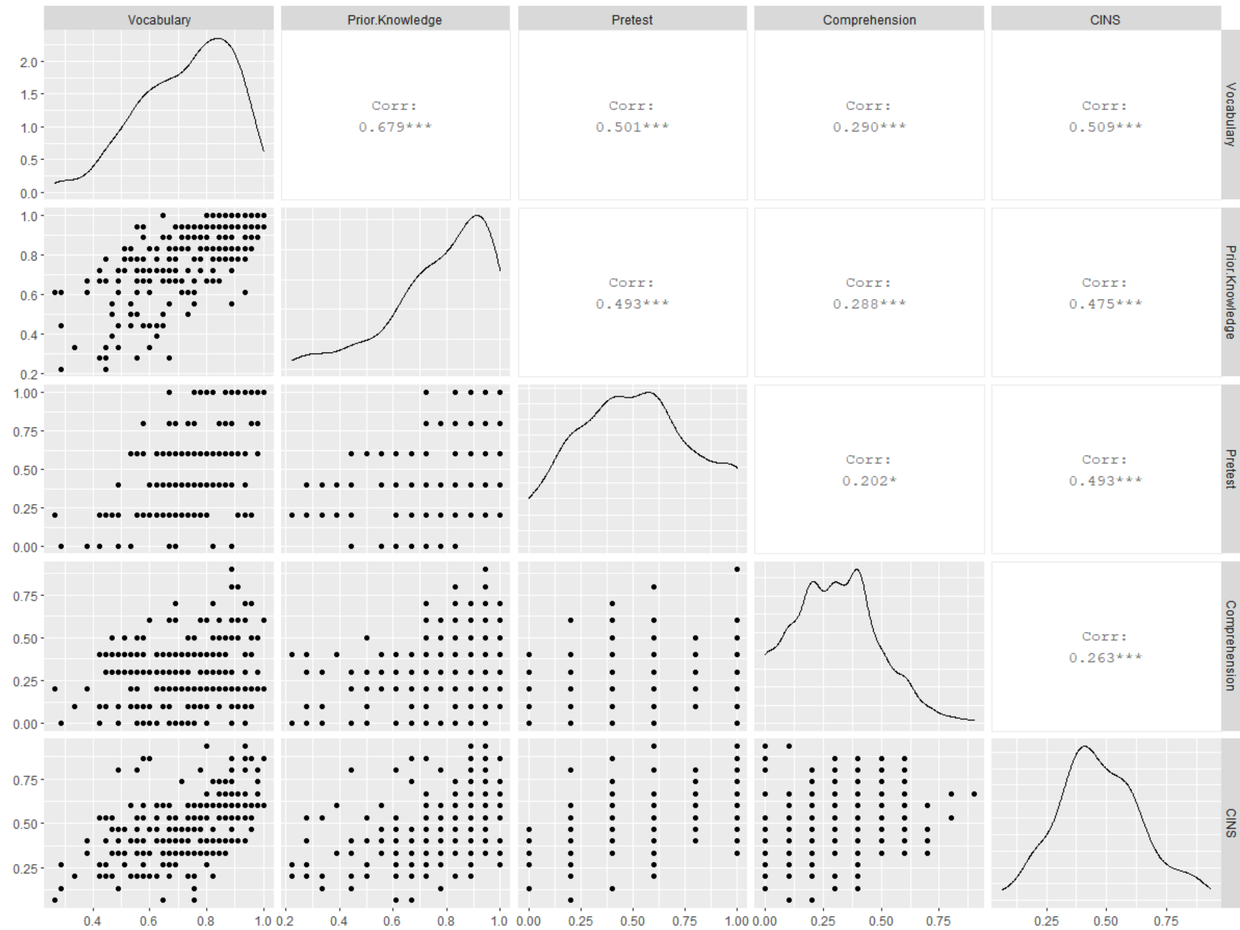


Figure 1. Scatterplots, distributions, and correlations of vocabulary, prior knowledge, and linguistic indices (*pretest n = 141).

* = $p < 0.05$; *** = $p < 0.001$

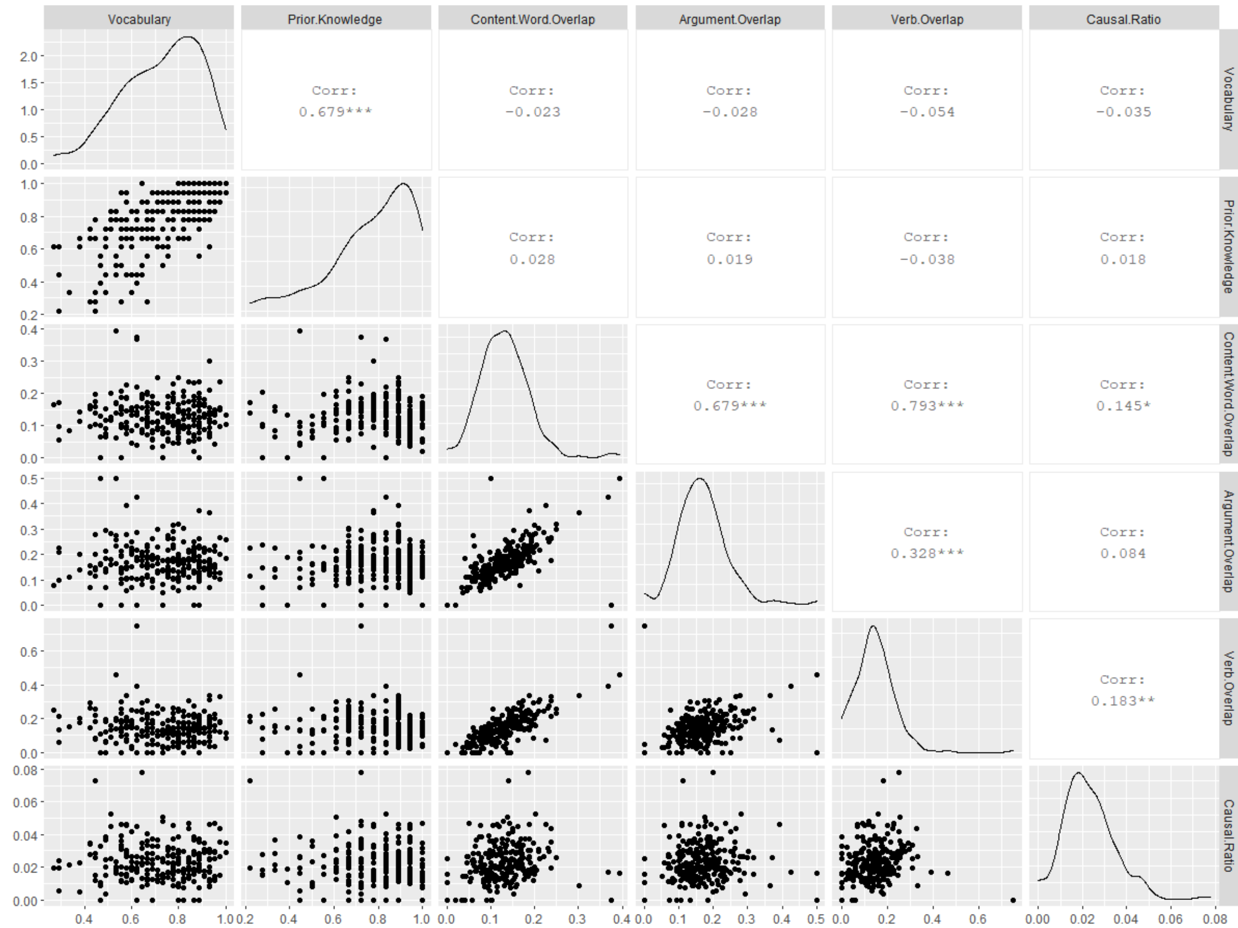


Figure 2. Scatterplots, distributions, and correlations of vocabulary, prior knowledge, and linguistic indices.

* = $p < 0.05$; *** = $p < 0.001$

Only a subset of participants (n = 141) were given the pretest, so the pretest was used only to examine a priori misconceptions at the condition level. A 2(prompt: self-explanation, think aloud) x 2(text: refutational, non-refutational) ANOVA was conducted to test differences in prior misconceptions on natural selection. There were no significant differences between groups, $F(3,137) < 1$ (see Appendix D for means and standard deviations as a function of condition). This finding indicates that there were no differences in a priori misconceptions as a function of condition.

All participants were assessed on prior science knowledge and vocabulary. Two 2(prompt: self-explanation, think aloud) x 2(text: refutational, non-refutational) ANOVAs were conducted to test for differences in prior knowledge and vocabulary as a function of condition. There were no significant differences in prior knowledge, $F(3,236) < 2$, or vocabulary, $F(3,236) < 2$, as a function of condition. This ANOVA indicates that there were no differences in prior knowledge or vocabulary as a function of condition.

The high correlation between prior knowledge and vocabulary indicated the possibility of multicollinearity if both were included in analyses. A linear regression predicting CINS scores with prior knowledge and vocabulary was conducted and a variance inflation factor (VIF) was calculated. The VIF of prior knowledge and vocabulary was 1.8. A similar VIF was obtained in a regression predicting comprehension scores with prior knowledge and vocabulary. Based on these results, it was determined that prior knowledge and vocabulary would not violate multicollinearity, thus both were included in the analyses.

Misconceptions

Table 7 presents the means and standard deviation of CINS score as a function of condition. Effect sizes were calculated for the differences between conditions, and for all comparisons, the effect size was small (Cohen's $d < 0.2$).

Table 7.

Means and standard deviations: Proportion of correct answers on the conceptual inventory of natural selection as a function of condition (n = 240)

	Think-Aloud	Self-Explain
Non-Refutational	0.47 (0.20)	0.47 (0.17)
Refutational	0.49 (0.17)	0.45 (0.17)

Note: Standard Deviations are in parentheses

Table 8 shows the results of a linear mixed effects (LME) model predicting CINS score from vocabulary score and prior knowledge score as fixed effects, and participant and question entered as random effects. Both vocabulary and prior knowledge accounted for significant variance in students' performance on the CINS.

Table 8.

LME results predicting score on the conceptual inventory of natural selection with vocabulary and prior knowledge

Fixed Effects	Estimate	SE	*t*	*p*
Intercept	0.03	0.04	0.58	0.55
Vocabulary Score	0.38	0.08	4.79	<0.01
Prior Knowledge Score	0.21	0.07	3.06	<0.01

Note: p values were obtained using Satterthwaite degrees of freedom.

Table 9 shows the LME model that was built to test the effects of condition on CINS score. Vocabulary, prior knowledge, text type, constructed response prompt and interactions entered as fixed effects, and participant and question as random effects. Vocabulary and prior knowledge accounted for significant variance in CINS score.

Table 9.

LME results predicting score on the conceptual inventory of natural selection with vocabulary, prior knowledge, text type, and constructed response prompt

Fixed Effects	Estimate	SE	*t*	*p*
Intercept	0.05	0.04	1.132	0.43
Vocabulary Score	0.39	0.02	17.88	<0.01
Prior Knowledge Score	0.21	0.02	11.38	<0.01
Text (Refutational vs non-refutational)	0.02	0.03	0.66	0.51
Prompt (Self-Explanation vs Think-Aloud)	0.01	0.03	0.45	0.65
Text * Prompt	0.001	0.04	0.03	0.97

Note: *p* values were obtained using Satterthwaite degrees of freedom.

A likelihood ratio test showed the first model was a better fit for the data, $\chi^2(3) = 1.43$, $p = 0.69$. A final set of two LME models were built to examine the interaction between conditions and individual differences. One model retained vocabulary and its interaction terms, and one model retained prior knowledge and its interaction terms. Neither the model with vocabulary and interactions, $\chi^2(2) = 0$, $p = 1$, nor prior knowledge and interactions, $\chi^2(2) = 0$, $p = 1$, better fit the data. An LME model with all predictors and all interactions was tested but did not converge. Because five of the questions were used as a pretest, these models were also tested using question type (pretest, non-pretest) as a within-subjects factor, and the pattern of results was the same.

Comprehension Test

Table 10 shows the means of the average comprehension test scores as a function of condition.

Effect sizes were calculated for the differences between conditions, and for all comparisons, the effect size was small (Cohen's *d* < 0.2).

Table 10.

Means and standard deviations: Average comprehension score as a function of condition (n = 240)

	Think-Aloud	Self-Explain
Non-Refutational	0.27 (0.18)	0.32 (0.17)
Refutational	0.33 (0.21)	0.29 (0.17)

Note: Standard Deviations are in parentheses

Table 11 shows a linear mixed effects (LME) model predicting comprehension score with vocabulary score and prior knowledge score as fixed effects, and participant and question entered as random effects. Consistent with our hypotheses, both vocabulary and prior knowledge accounted for significant variance in students' performance on the comprehension questions. This finding indicates that students with high-knowledge and high vocabulary knowledge had better text comprehension compared to low-knowledge students.

Table 11.

LME results predicting comprehension score with vocabulary and prior knowledge

Fixed Effects	Estimate	SE	*t*	*p*
Intercept	0.02	0.06	0.32	0.75

Table 11.

LME results predicting comprehension score with vocabulary and prior knowledge

Fixed Effects	Estimate	SE	*t*	*p*
Vocabulary Score	0.19	0.09	2.07	0.03
Prior Knowledge Score	0.17	0.08	2.15	0.03

Note: *p* values were obtained using Satterthwaite degrees of freedom.

Table 12 shows the second LME that was used to test the effects of constructed response prompt and text type. Vocabulary, prior knowledge, text type, constructed response prompt and interactions were entered as fixed effects, and participant and question as random effects. Vocabulary and prior knowledge accounted for significant variance in comprehension score.

Table 12.

LME results predicting comprehension score with vocabulary, prior knowledge, text type, and constructed response prompt

Fixed Effects	Estimate	SE	*t*	*p*
Intercept	0.12	0.06	0.54	0.15
Vocabulary	0.19	0.09	2.06	0.03
Prior Knowledge	0.18	0.08	2.12	0.03
Text (Refutational vs Non-refutational)	-0.01	0.03	0.38	0.70
Prompt (Self-Explanation vs Think-aloud)	0.05	0.03	1.48	0.14
Text (Ref) * Prompt (SE)	0.07	0.05	1.52	0.13

Note: *p* values were obtained using Satterthwaite degrees of freedom.

A likelihood ratio test showed the first model was a better fit for the data than the first model , $\chi^2(3) = 3.71$, $p = 0.29$. A final set of two LME models were built to examine the interaction between conditions and individual differences. One model retained vocabulary and its interaction terms, and one model retained prior knowledge and its interaction terms. Neither the model with vocabulary and interactions, $\chi^2(2) = 0$, $p = 1$, nor prior knowledge and interactions, $\chi^2(2) = 0$, $p = 1$, better fit the data. An LME model with all predictors and all interactions was tested but did not converge.

Cohesion

A series of 12 linear regressions were conducted to determine the effects of condition (text type, constructed response prompt) and individual differences on students' referential and causal cohesion. Three regressions models were used for each cohesion index. Separate regressions for the interactions of individual differences with conditions were required to ensure adequate power.

Content word overlap. The first set of regressions predicted content word overlap. Table 13 presents the linear regression predicting content word overlap from main effects and interactions of text type and constructed response prompt, while holding number of words, vocabulary, and prior knowledge constant. The model accounted for a significant amount of variance in content word overlap, $R^2 = 0.04$, $F(6,233) = 2.46$, $p = 0.02$. There was a main effect of prompt such that students had greater content word overlap in their constructed responses when prompted to self-explain compared to think-aloud.

Table 13.

Linear regression predicting content word overlap as a function of condition with covariates

Source	β	SE	*t*	*p*
Intercept	0.19	0.03	7.01	<0.01
Number of words	-0.001	0.001	-0.34	0.74
Vocabulary	-0.01	0.04	-0.31	0.75
Prior Knowledge	0.01	0.03	0.26	0.80
Prompt (Self-explain vs Think-aloud)	0.04	0.01	3.01	<0.01
Text (Refutational vs Non-refutational)	-0.01	0.01	-0.42	0.67
Text (Ref) * Prompt (SE)	-0.03	0.02	-1.55	0.12

Table 14 presents the linear regression predicting content word overlap from main effects and interactions of vocabulary, text type, and constructed response prompt, while holding prior knowledge constant. The model accounted for a significant portion of the variance in content word overlap, $R^2 = 0.11$, $F(8, 231) = 4.71$, $p < 0.01$. There was a main effect of prompt such that students had greater content word overlap in their constructed responses when prompted to self-explain compared to think-aloud. There was a main effect of text such that students had greater content word overlap in their constructed responses when reading a refutational text compared to reading a non-refutational text. Figures 3-5 present a 3-way interaction such that students with low vocabulary prompted to self-explain a non-refutational text had more content word overlap than students prompted to think-aloud while reading a non-refutational text.

Table 14.

Linear regression predicting content word overlap as a function of condition and interactions with vocabulary

Source	β	SE	*t*	*p*
Intercept	0.10	0.05	2.11	0.03
Prior Knowledge	0.02	0.04	0.51	0.61
Vocabulary	0.09	0.07	1.35	0.17
Prompt (Self-explain vs Think-aloud)	0.20	0.07	2.90	<0.01
Text (Refutational vs Non-refutational)	0.21	0.07	3.27	<0.01
Vocabulary * Prompt (SE)	-0.21	0.09	-2.32	0.02
Vocabulary * Text (Ref)	-0.30	0.09	-3.44	<0.01
Text (Ref) * Prompt (SE)	-0.42	0.09	-4.63	<0.01
Vocabulary * Text (Ref) * Prompt (SE)	0.54	0.12	4.42	<0.01

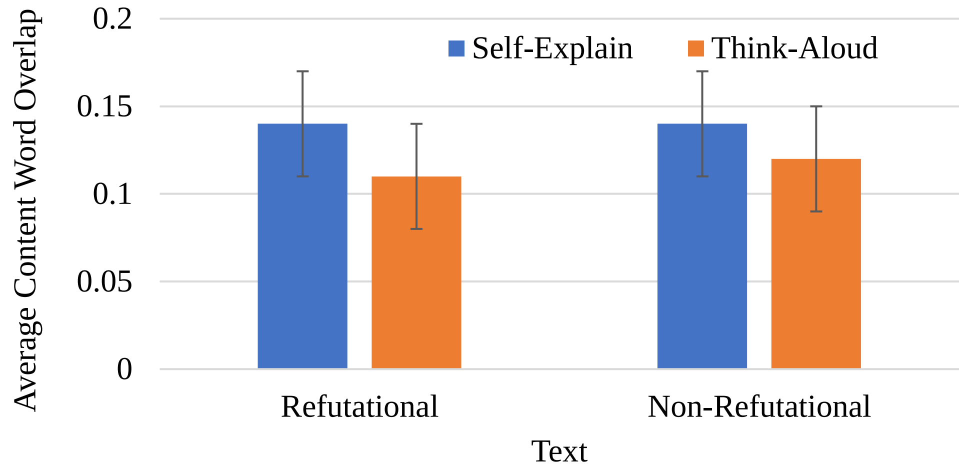


Figure 3. Content word overlap of responses as a function of condition at 1 standard deviation above the mean of vocabulary, holding prior knowledge constant.

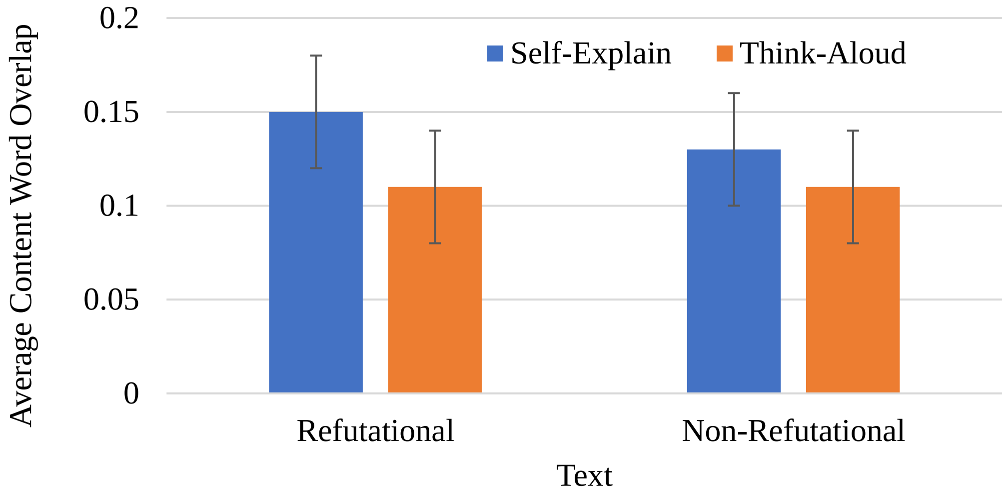


Figure 4. Content word overlap of responses as a function of condition at the mean of vocabulary, holding prior knowledge constant.

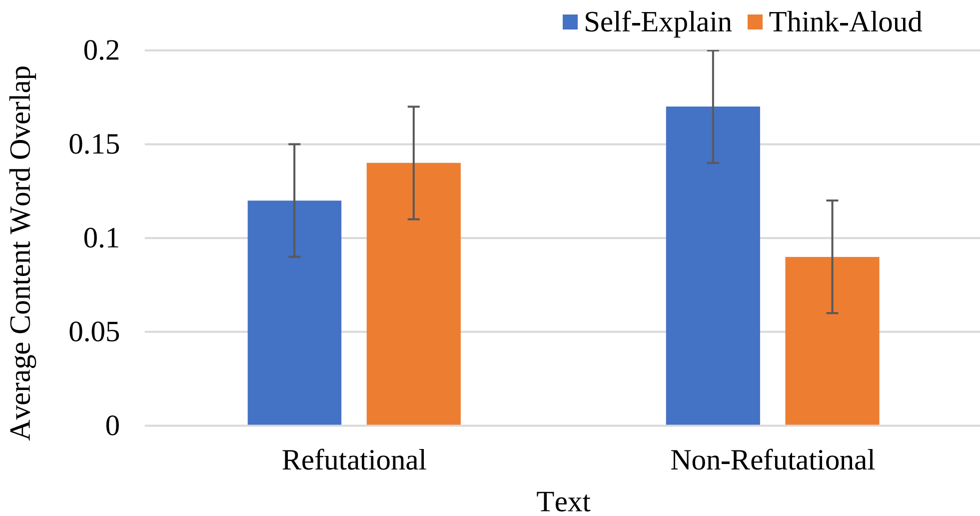


Figure 5. Content word overlap of responses as a function of condition at 1 standard deviation below the mean of vocabulary, holding prior knowledge constant.

Table 15 presents the linear regression predicting content word overlap from main effects and interactions of prior knowledge, text type, and constructed response prompt, while holding vocabulary constant. The third accounted for a significant portion of the variance in content word overlap, $R^2 = 0.09$, $F(8, 231) = 4.16$, $p < 0.01$. There was a main effect of prompt such that students had greater content word overlap in their constructed responses when prompted to self-explain compared to think-aloud. There was a main effect of text such that students had greater content word overlap in their constructed responses when reading a refutational text compared to reading a non-refutational text. Figures 6-8 show the 3-way interaction such that students with low prior knowledge prompted to self-explain a non-refutational text had more content word overlap than students prompted to think-aloud while reading a non-refutational text.

Table 15

Linear regression predicting content word overlap as a function of condition (text, prompt) and prior knowledge

Source	β	SE	*t*	*p*
Intercept	0.09	0.04	2.05	0.04
Vocabulary	-0.01	0.04	-0.16	0.87
Prior Knowledge	0.12	0.06	2.08	0.04
Prompt (Self-explain vs Think-aloud)	0.22	0.06	3.42	<0.01
Text (Refutational vs Non-refutational)	0.18	0.06	3.10	<0.01
Prior Knowledge * Prompt (SE)	-0.22	0.08	-2.87	<0.01
Prior Knowledge * Text (Ref)	-0.24	0.08	-3.28	<0.01
Text (Ref) * Prompt (SE)	-0.38	0.09	-4.41	<0.01
Prior Knowledge * Text (Ref) * Prompt (SE)	0.44	0.11	4.16	<0.01

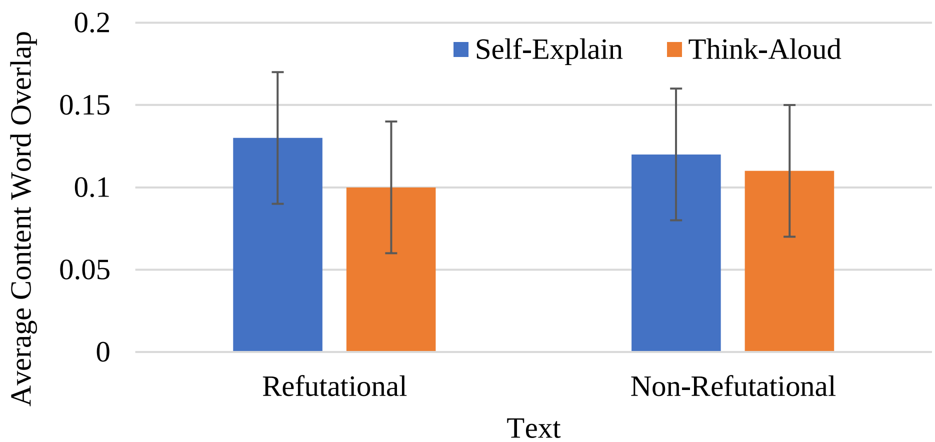


Figure 6. Content word overlap of responses as a function of condition at 1 standard deviation above the mean of prior knowledge, holding vocabulary constant.

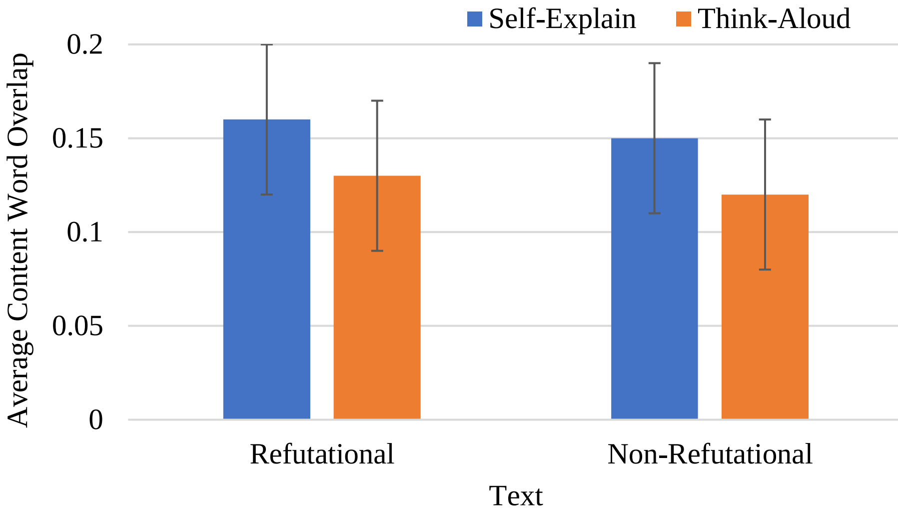


Figure 7. Content word overlap of responses as a function of condition at the mean of prior knowledge, holding vocabulary constant.

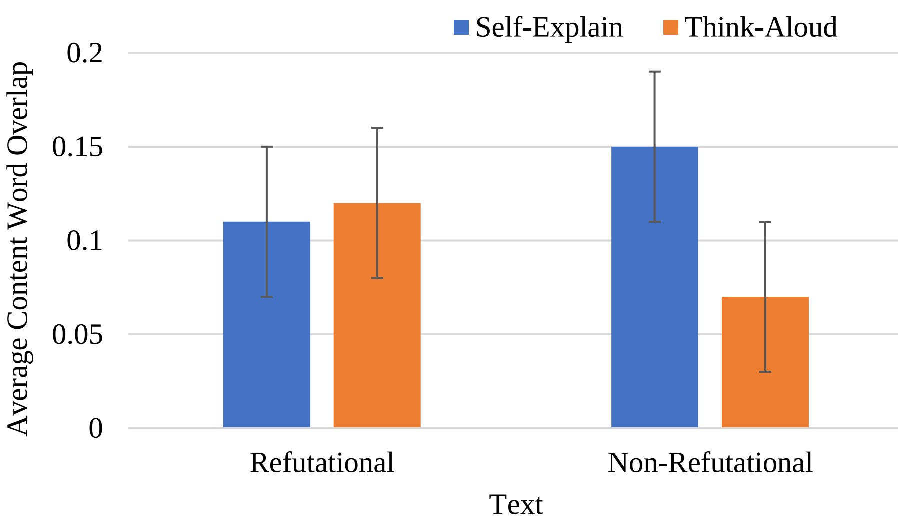


Figure 8. Content word overlap of responses as a function of condition at 1 standard deviation below the mean of prior knowledge, holding vocabulary constant.

Argument overlap. The second set of regressions predicted argument overlap. The first regression model predicted argument overlap from text type and constructed response prompt, holding number of words, vocabulary, and prior knowledge constant.

The model did not account for a significant amount of variance in argument overlap, R^2 = 0.001, F(6,233) = 1.07, p = 0.37.

The second model predicted argument overlap from main effects and interactions of vocabulary, text type and constructed response prompt, holding prior knowledge constant. The model did not account for a significant portion of variance in argument overlap, R^2 = 0.02, F(8, 231) = 1.88, p = 0.06.

The third model predicted argument overlap from main effects and interactions of prior knowledge, text type and constructed response prompt, holding vocabulary constant. The model did not account for a significant portion of variance in argument overlap, R^2 = 0.02, F(8, 231) = 1.69, p = 0.1.

Verb overlap. The third set of regressions predicted verb overlap. The first regression model predicted verb overlap from text type and constructed response prompt, holding number of words, vocabulary, and prior knowledge constant. The model did not account for a significant amount of variance in verb overlap, R^2 = 0.001, F(6,233) = 1.37, p = 0.22.

Table 16 presents the linear regression predicting verb overlap from main effects and interactions of vocabulary, text type, and constructed response prompt, while holding prior knowledge constant. The model accounted for a significant portion of the variance in verb overlap, R^2 = 0.05, F(8, 231) = 2.58, p = 0.01. There was a main effect of prompt such that students had greater verb overlap in their constructed responses when prompted to self-explain compared to think-aloud. There was a main effect of text such that students had greater verb overlap in their constructed responses when reading a refutational text compared to reading a non-refutational text. Figures 9-11 show the 3-way interaction such that students with low vocabulary prompted to self-explain a non-refutational text had more verb overlap than students prompted to think-aloud while reading a non-refutational text.

Table 16.

Linear regression predicting verb overlap as a function of condition and interactions with vocabulary

Source	β	SE	*t*	*p*
Intercept	0.15	0.07	2.27	0.02
Prior Knowledge	-0.01	0.05	-0.25	0.80
Vocabulary	0.07	0.10	0.77	0.44
Prompt (Self-explain vs Think-aloud)	0.21	0.10	2.12	0.03
Text (Refutational vs Non-refutational)	0.25	0.09	2.69	0.01
Vocabulary * Prompt (SE)	-0.21	0.13	-1.66	0.09
Vocabulary * Text (Ref)	-0.33	0.12	-2.61	<0.01
Text (Ref) * Prompt (SE)	-0.45	0.13	-3.49	<0.01
Vocabulary * Text (Ref) * Prompt (SE)	0.57	0.17	3.24	<0.01

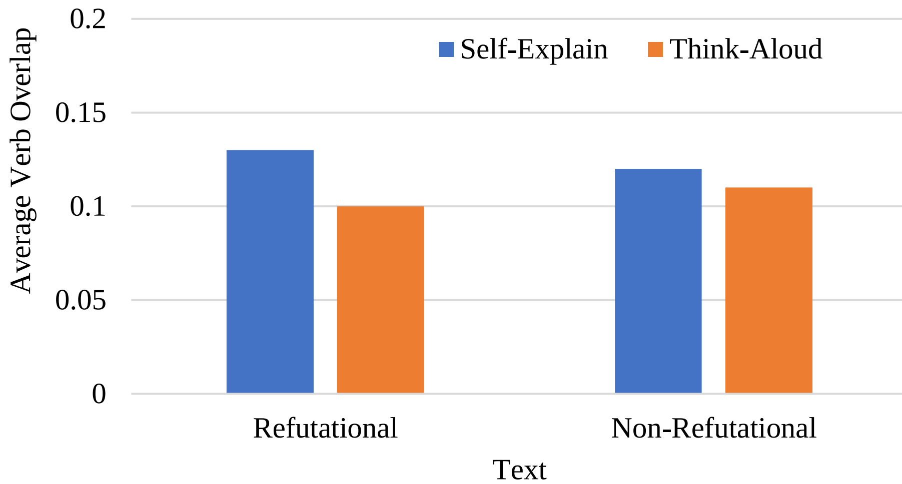


Figure 9. Verb overlap of responses as a function of condition at 1 standard deviation above the mean of vocabulary, holding prior knowledge constant.

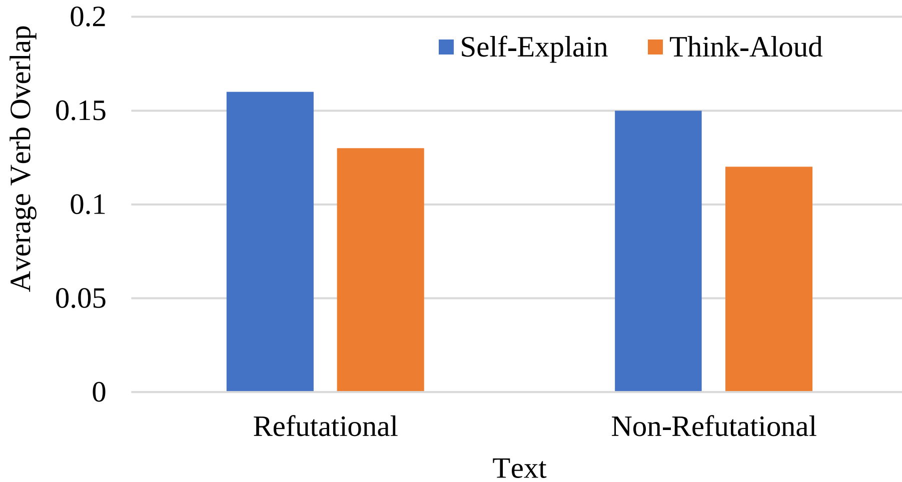


Figure 10. Verb overlap of responses as a function of condition at the mean of vocabulary, holding prior knowledge constant.

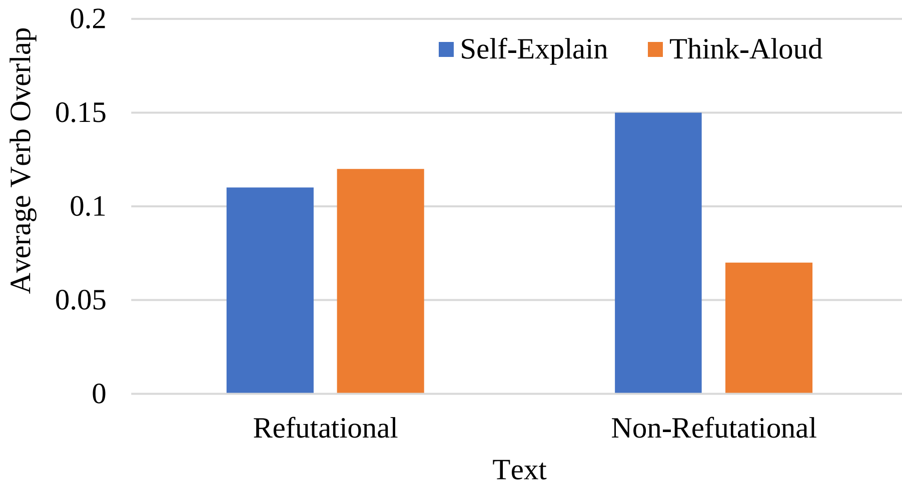


Figure 11. Verb overlap of responses as a function of condition at 1 standard deviation below the mean of vocabulary, holding prior knowledge constant.

Table 17 presents the linear regression predicting verb overlap from main effects and interactions of prior knowledge, text type, and constructed response prompt, while holding vocabulary constant. The model accounted for a significant portion of the variance in verb overlap, $R^2 = 0.06$, $F(8, 231) = 2.78$, $p < 0.01$. There was a main effect of prompt such that students had greater verb overlap in their constructed responses when prompted to self-explain compared to think-aloud. There was a main effect of text such that students had greater verb overlap in their constructed responses when reading a refutational text compared to reading a non-refutational text. Figures 12-14 show the 3-way interaction such that students with low vocabulary prompted to self-explain a non-refutational text had more verb overlap than students prompted to think-aloud while reading a non-refutational text.

Table 17.

Linear regression predicting verb overlap as a function of condition and interactions with prior knowledge

Source	β	SE	*t*	*p*
Intercept	0.18	0.06	2.73	<0.01
Vocabulary	-0.02	0.06	-0.46	0.64
Prior Knowledge	0.06	0.09	0.74	0.46
Prompt (Self-explain vs Think-aloud)	0.22	0.09	2.39	0.02
Text (Refutational vs Non-refutational)	0.21	0.09	2.46	0.01
Prior Knowledge * Prompt (SE)	-0.21	0.11	-1.88	0.06
Prior Knowledge * Text (Ref)	-0.25	0.11	-2.37	0.01
Text (Ref) * Prompt (SE)	-0.46	0.12	-3.77	<0.01
Prior Knowledge * Text (Ref) * Prompt (SE)	0.53	0.15	3.52	<0.01

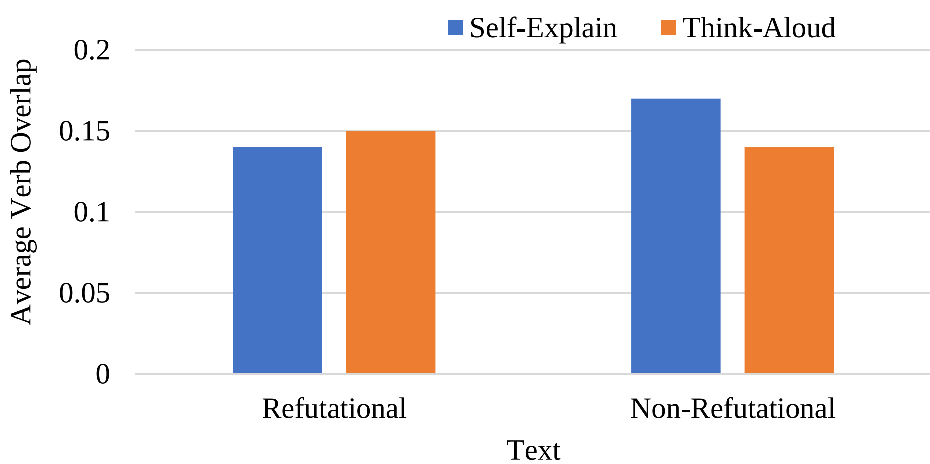


Figure 12. Verb overlap of responses as a function of condition at 1 standard deviation above the mean of prior knowledge, holding vocabulary constant.

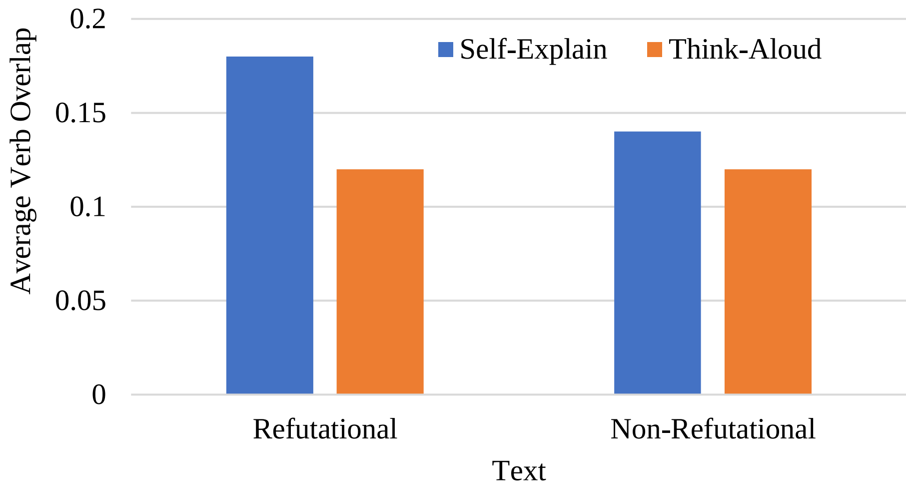


Figure 13. Verb overlap of responses as a function of condition at the mean of prior knowledge, holding vocabulary constant.

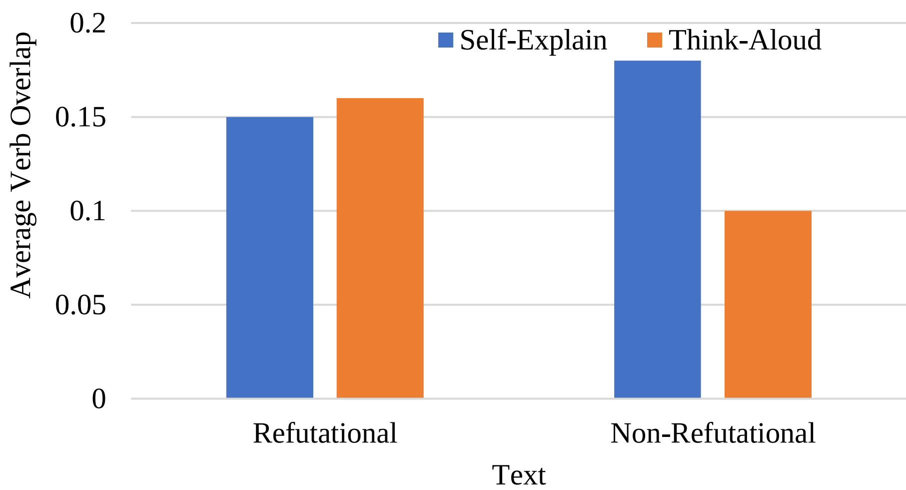


Figure 14. Verb overlap of responses as a function of condition at 1 standard deviation below the mean of prior knowledge, holding vocabulary constant.

While not hypothesized, previous research (McNamara, 2004) suggests that the quality of students' constructed responses is related to their score on comprehension measures. A mediation model was constructed to examine the potential relationships and the result was non-significant. See Appendix F for the full model.

Causal ratio. The fourth set of regressions predicted causal ratio. The first regression predicted causal ratio from text type and constructed response prompt, holding number of words, vocabulary, and prior knowledge constant. The model did not account for a significant amount of variance in causal ratio, $R^2 = 0.001$, $F(6,233) = 0.94$, $p = 0.4$.

The second regression predicted causal ratio from main effects and interactions of vocabulary, text type and constructed response prompt, holding prior knowledge constant. The model did not account for a significant portion of variance in causal ratio, $R^2 = 0.001$, $F(8, 231) = 0.99$, $p = 0.4$.

The third regression predicted causal ratio from main effects and interactions of prior knowledge, text type and constructed response prompt, holding number of words, vocabulary constant. The model did not account for a significant portion of variance in causal ratio, $R^2 = 0.002$, $F(8, 231) = 1.09$, $p = 0.37$.

Discussion

The current study examined the extent to which prompting students to self-explain a non-refutational or refutational text affected their understanding of natural selection in comparison to thinking-aloud while reading a non-refutational text. Students were prompted to self-explain or think-aloud while reading a text (refutational, non-refutational). The students then completed misconception, comprehension and individual difference assessments, and their responses to the text were analyzed for referential and causal cohesion.

Students with high vocabulary knowledge and prior knowledge had fewer misconceptions and better comprehension after reading, regardless of text or prompt, compared to less knowledgeable students. However, there was no effects of reading prompt or text type on students' misconceptions or comprehension. For both the content word and verb overlap of the responses, students prompted to self-explain wrote responses with more overlap compared to students prompted to think-aloud. Furthermore, the advantage of self-explanation prompt was enhanced for low-knowledge students reading the non-refutational text.

This study attempted to answer the theoretical question of how students correct their misconceptions in science. Readers with high prior knowledge and reading skill tended to perform better on the comprehension assessment and have fewer misconceptions compared to less skilled students. The finding of the relationships between reading skill, prior knowledge, and students' comprehension test score is consistent with past research on comprehension (McNamara, 2004). Both comprehension test scores and prevalence of misconceptions are a function of the stability of students' mental representation of the text (Kendeou & van den Broek, 2007). According to the knowledge activation hypothesis, inferencing while reading is essential to building a stable mental representation of the text, and skilled students tend to generate more inferences while reading than less skilled students (McNamara, de Vega, & O'Reilly, 2007). Thus, the skilled students' advantage in generating inferences led to more stable mental representations in general compared to the less skilled students. Thus the relationship of reading skill and prior knowledge to students' comprehension of the text and misconception prevalence depended on the stability of their mental representation.

While the importance of prior knowledge and reading skill were confirmed, neither reading the refutational text nor being prompted to self-explain reduced misconceptions. This finding is inconsistent with the current research on refutational texts and the co-activation hypothesis. The underlying assumption of refutational texts is that providing explicit examples of the misconception and the correct concept affords the reader the opportunity to correct the error (Kendeou & van den Broek, 2007). The affordances of the refutational text were expected to be enhanced when readers were prompted to self-explain while reading. This hypothesis was drawn from the research on self-explanation showing that prompting self-explanation enhances readers' comprehension of science texts (McNamara, 2004). However, in this study, prompting self-explanation neither enhanced readers' comprehension nor reduced their misconceptions after reading. Overall, this study indicates that simply using refutational texts or simple reading interventions are not sufficient interventions for all misconceptions in science.

There are two explanations for the results. First, the current study may have been underpowered, as the effect sizes for the between groups comparisons were all small. Second, there was potentially a backfire effect. The direction of the effects suggests that the students prompted to self-explain while reading the refutational text had the worst comprehension and most misconceptions of the four groups. Kessler et al. (2019) described a backfire effect where students higher in flexible thinking acquired more misconceptions after reading about vaccination. Kessler et al interpreted this backfire effect as those students more deeply considering the alternative explanations. The equivalent interpretation for the current study is that students prompted to self-explain more deeply processed the incorrect information in the refutational text compared to students prompted to think-aloud. The deeper processing led to lower comprehension and more misconceptions because any learned, incorrect information was interfering with activation of the correct information. A backfire effect was not predicted, nor the findings significant in the current study. Therefore, a follow-up study is needed to address this question.

This study attempted to use linguistic measures of students' constructed responses to assess students' on-line processing while reading. The results were inconclusive, with only one of the two measures of causal cohesion showing that students prompted to self-explain had greater causal reasoning compared to students prompted to think-aloud. There was the same effect of self-explanation on one of the two measures of referential cohesion. The linguistic findings are inconsistent with the hypotheses that only students' causal reasoning would be affected by self-explanation.

This study does fit with new research showing the effects of self-explanation prompt on overall cohesion of constructed responses. For example, Creer et al. (2020) showed students prompted to self-explain had greater verb *and* noun overlap in constructed responses compared to students prompted to think-aloud. The findings of Creer et al and this study show that multiple dimensions of cohesion should be considered when using constructed responses to examine the on-line processing of readers.

This study confirms the importance of reading skill and prior knowledge to conceptual change and comprehension. This study also highlights the necessity to examine multiple dimensions of cohesion in constructed responses. One future direction is a study specifically focused on the potential backfire effect of combining self-explanation and refutational texts. The processes behind conceptual change rely on a complex framework including, but not limited to, prior knowledge activation, causal reasoning, and epistemology (Allen, McCrudden, & McNamara, 2015; Bråten & Strømsø, 2004; Kendeou et al., 2014). This study demonstrated that reading prompt and refutational texts are not foolproof methods to promote misconception revision. Therefore, it is necessary to use these tools in combination with other interventions to effectively enhance conceptual change.

REFERENCES

Allen, L. K., McCrudden, M. T., & McNamara, D.S., (2015). Change your mind: Investigating the effects of self-explanation in the resolution of misconceptions. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. Maglio, (Eds.), *Proceedings of the 37th Annual Meeting of the Cognitive Science Society (Cog Sci 2015)*, (pp. 78-83). Pasadena, CA: Cognitive Science Society.

Allen, L. K., Snow, E. L., & McNamara, D. S. (2015). Are you reading my mind? Modeling students' reading comprehension skills with natural language processing techniques. In P. Nlikdyrin, A. Merceron, & G. Siemens (Eds.), *Proceedings of the fifth annual international learning analytics and knowledge conference* (pp. 246-254). Poughkeepsie, NY.

Anderson, D. L., Fisher, K. M., & Norman, G. J. (2002). Development and evaluation of the conceptual inventory of natural selection. *Journal of research in science teaching*, 39(10), 952-978.

Ariasi, N., & Mason, L. (2011). Uncovering the effect of text structure in learning from a science text: An eye-tracking study. *Instructional science*, *39*(5), 581-601.

Baker, R. S., Corbett, A. T., & Koedinger, K. R. (2004, August). Detecting student misuse of intelligent tutoring systems. In *International conference on intelligent tutoring systems* (pp. 531-540). Springer, Berlin, Heidelberg.

Baker, R. S., D'Mello, S. K., Rodrigo, M. M. T., & Graesser, A. C. (2010). Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies*, *68*(4), 223-241.

Braasch, J. L., Goldman, S. R., & Wiley, J. (2013). The influences of text and reader characteristics on learning from refutations in science texts. *Journal of Educational Psychology*, *105*(3), 561.

Bråten, I., & Strømsø, H. I. (2004). Epistemological beliefs and implicit theories of intelligence as predictors of achievement goals. Contemporary Educational Psychology, 29(4), 371-388.

Bråten, I., & Strømsø, H. I. (2009). Effects of task instruction and personal epistemology on the understanding of multiple texts about climate change. Discourse Processes, 47(1), 1-31.

Brooke, J. (1996). SUS-A quick and dirty usability scale. *Usability evaluation in industry*, *189*(194), 4-7.

Broughton, S. H., Sinatra, G. M., & Reynolds, R. E. (2010). The nature of the refutation text effect: An investigation of attention allocation. *The Journal of Educational Research*, *103*(6), 407-423.

Conners, F., Olson, R., Balota, D., Flores d'Arcais, G., & Rayner, K. (1990). Comprehension processes in reading. *Reading comprehension in dyslexic and normal readers: A component skills analysis*, 557-579.

Coley, J. D., & Tanner, K. (2015). Relations between intuitive biological thinking and biological misconceptions in biology majors and nonmajors. *CBE—Life Sciences Education*, 14(1), ar8.

Coté, N., & Goldman, S. R. (1999). Building representations of informational text: Evidence from children's think-aloud protocols. *The construction of mental representations during reading*, 169-193.

Creer, S. D., McCarthy, K. S., Magliano, J. P., McNamara, D. S., & Allen, L. K. (2020). Self-Explanation vs. Think Aloud: What Natural Language Processing Can Tell Us (No. 3624). EasyChair.

Ericsson, K. A. (2006). The influence of experience and deliberate practice on the development of superior expert performance. *The Cambridge handbook of expertise and expert performance*, 38, 685-705.

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. Psychological review, 100(3), 363.

Ericsson, K. A., & Simon, H. A. (1998). How to study thinking in everyday life: Contrasting think-aloud protocols with descriptions and explanations of thinking. *Mind, Culture, and Activity*, 5(3), 178-186.

Feltovich, P. J., Coulson, R. L., & Spiro, R. J. (2001). Learners'(mis) understanding of important and difficult concepts: A challenge to smart machines in education. *Smart machines in education*, 349-375.

Gallup. (2017). "In U.S., Belief in Creationist Views of Humans at All-Time Low" May 22, 2017.

Graesser, A. C., Millis, K. K., & Zwaan, R. A. (1997). Discourse comprehension. *Annual review of psychology*, *48*(1), 163-189.

Graesser, A., Brennan, S. E., Goldman, S. R., & Schober, M. F. (Eds.). (2009). *Effects of Personal Involvement in Narrative Discourse*. Routledge.

Gregory, T. R. (2009). Understanding natural selection: essential concepts and common misconceptions. *Evolution: Education and Outreach*, *2*(2), 156.

Guzzetti, B. J., Snyder, T. E., Glass, G. V., & Gamas, W. S. (1993). Promoting conceptual change in science: A comparative meta-analysis of instructional interventions from reading education and science education. *Reading Research Quarterly*, 117-159.

Jackson, T., Boonthum-Denecke, C., & McNamara, D. (2015). Natural language processing and game-based practice in iSTART. *Journal of Interactive Learning Research*, *26*(2), 189-208.

Jackson, G. T., & McNamara, D. S. (2013). Motivation and performance in a game-based intelligent tutoring system. *Journal of Educational Psychology*, *105*(4), 1036.

Johnson-Laird, P. N. (1983). A computational analysis of consciousness. *Cognition & Brain Theory*.

Kendeou, P., Braasch, J. L., & Bråten, I. (2016). Optimizing conditions for learning: Situating refutations in epistemic cognition. *The Journal of Experimental Education*, 84(2), 245-263.

Kendeou, P., Muis, K. R., & Fulton, S. (2011). Reader and text factors in reading comprehension processes. *Journal of Research in Reading*, *34*(4), 365-383.

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(7), 1567-1577.

Kendeou, P., Walsh, E. K., Smith, E. R., & O'Brien, E. J. (2014). Knowledge revision processes in refutation texts. *Discourse Processes*, *51*(5-6), 374-397.

Kessler, E. D., Braasch, J. L., & Kardash, C. M. (2019). Individual Differences in Revising (and Maintaining) Accurate and Inaccurate Beliefs About Childhood Vaccinations. *Discourse Processes*, 56(5-6), 415-428.

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

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge university press.

Legare, C. H., & Lombrozo, T. (2014). Selective effects of explanation on learning during early childhood. *Journal of Experimental Child Psychology*, 126, 198-212.

MacGinitie, W. H., & MacGinitie, R. K. (1989). *Gates–MacGinitie Reading Tests*. Chicago, IL: Riverside.

Magliano, J. P., & Graesser, A. C. (2012). Computer-based assessment of student-constructed responses. *Behavior Research Methods*, 44(3), 608-621.

Mason, L., & Gava, M. (2007). Effects of epistemological beliefs and learning text structure on conceptual change.

McCarthy, K. S., Likens, A. D., Kopp, K. J., Watanabe, M., Perret, C. A., & McNamara, D. S. (2018). The "LO"-down on grit: Non-cognitive trait assessments fail to predict learning gains in iSTART and W-Pal. In *Companion Proceedings of the 8th International Conference on Learning Analytics and Knowledge (LAK'18).* Sydney, Australia.

McCarthy, K.S., Kopp, K. J., Allen, L. K., Kopp, & McNamara, D. S. (in press). Methods of Studying Text: Memory, Comprehension, and Learning. In O. Hajime & B. Schwarz (Eds.), *Research Methods in Human Memory*. New York, NY: Routledge

McCloskey, M. (1983). Naive theories of motion. In D. Gentner & A. L. Stevens (Eds.), *Mental models* (pp. 299–324). Hillsdale, NJ: Erlbaum.

McCrudden, M. T., & Kendeou, P. (2014). Exploring the link between cognitive processes and learning from refutational text. *Journal of Research in Reading*, 37(S1).

McNamara, D. S. (2004). SERT: Self-explanation reading training. *Discourse processes*, 38(1), 1-30.

McNamara, D. S. (2011). Measuring deep, reflective comprehension and learning strategies: challenges and successes. *Metacognition and Learning*, 6(2), 195-203.

McNamara, D. S. (2015). Self-Explanation and Reading Strategy Training (SERT) Improves low-knowledge students' science course performance. *Discourse Processes*.

McNamara, D. S., Boonthum, C., Levinstein, I. B., & Millis, K. (2007). Evaluating self-explanations in iSTART: Comparing word-based and LSA algorithms. *Handbook of latent semantic analysis*, 227-241.

McNamara, D. S., Crossley, S. A., Roscoe, R. D., Allen, L. K., & Dai, J. (2015). A hierarchical classification approach to automated essay scoring. *Assessing Writing*, *23,* 35-59.

McNamara, D. S., & Kintsch, W. (1996). Learning from texts: Effects of prior knowledge and text coherence. *Discourse processes*, *22*(3), 247-288.

McNamara, D.S., Levinstein, I.B., & Boonthum, C. (2004). iSTART: Interactive strategy trainer for active reading and thinking. *Behavioral Research Methods, Instruments, & Computers*, 36, 222-233

McNamara, D.S. & Magliano, J.P. (2009). Towards a comprehensive model of comprehension. In B. Ross (Ed.), *The psychology of learning and motivation*. New York, NY: Elsevier Science.

Millis, K. K., Golding, J. M., & Barker, G. (1995). Causal connectives increase inference generation. *Discourse Processes*, 20(1), 29-49.

Nehm, R. H., & Schonfeld, I. S. (2008). Measuring knowledge of natural selection: A comparison of the CINS, an open-response instrument, and an oral interview. *Journal of Research in Science Teaching*, *45*(10), 1131-1160.

O'Reilly, T., Best, R., & McNamara, D. S. (2004, January). Self-explanation reading training: Effects for low-knowledge readers. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 26, No. 26).

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

O'Reilly, T., Sinclair, G. P., & McNamara, D. S. (2004, January). Reading strategy training: Automated verses live. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 26, No. 26).

Ozuru, Y., Briner, S., Best, R., & McNamara, D. S. (2010). Contributions of self-explanation to comprehension of high-and low-cohesion texts. *Discourse Processes*, *47*(8), 641-667.

Ozuru, Y., Briner, S., Kurby, C. A., & McNamara, D. S. (2013). Comparing comprehension measured by multiple-choice and open-ended questions. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 67(3), 215.

Phillips, L. M., Norris, S. P., Osmond, W. C., & Maynard, A. M. (2002). Relative reading achievement: A longitudinal study of 187 children from first through sixth grades. *Journal of Educational Psychology*, *94*(1), 3.

Pinker, Steven. "How the mind works. 1997." *NY: Norton* (1997).

Sinatra, G. M., & Broughton, S. H. (2011). Bridging reading comprehension and conceptual change in science education: The promise of refutation text. *Reading Research Quarterly*, *46*(4), 374-393.

Snow, E. L., Jacovina, M. E., Jackson, G. T., & McNamara, D. S. (2016). iSTART-2: a reading comprehension and strategy instruction tutor. In *Adaptive educational technologies for literacy instruction*. Taylor and Francis.

Snow, E. L., Likens, A. D., Jackson, T., & McNamara, D. S. (2013, July). Students' walk through tutoring: Using a random walk analysis to profile students. In *Educational Data Mining 2013*.

Sterman, J. D. (2008). Risk communication on climate: mental models and mass balance. *Science*, 322(5901), 532-533.

van den Broek, P., & Kendeou, P. (2008). Cognitive processes in comprehension of science texts: the role of co-activation in confronting misconceptions. *Applied Cognitive Psychology*, 22(3), 335-351.

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. *The psychology of science text comprehension*, 131-154.

Vosniadou, S. (2003). Exploring the relationships between conceptual change and intentional learning. *Intentional conceptual change*, 377-406.

APPENDIX A

REFUTATIONAL TEXT & NON-REFUTATIONAL TEXT SAMPLES

Sample Non-refutational Text	Sample Refutational Text
This complex design of different organs enables our bodies to adjust to a constantly changing environment and to overcome such changes with seemingly little or no effort. It is mind-boggling that our bodies can deal with the vast array of specific challenges that we face. This ability points us to the idea of replication. A replicator is something that can make a copy of itself, with most of its traits duplicated in the copy, including the ability to replicate. Let's look at a concrete example. Suppose there are three animals, two with cloudy lenses, and one with a clear lens. Having a clear lens (A), causes an eye to see well (B); seeing well helps the animal avoid predators and find mates, which enables the animal to reproduce. The offspring (AA) have clear lenses and can see well too. The offspring	This complex design of different organs <u>has led some people to believe that organs have been designed in advance with a specific function in mind. However, the view that organs must have been designed in advance for a specific purpose is incorrect. This view is incorrect because it fails to take into account the idea of replication.</u> A replicator is something that can make a copy of itself, with most of its traits duplicated in the copy, including the ability to replicate. Let's look at a concrete example. Suppose there are three animals, two with cloudy lenses, and one with a clear lens. Having a clear lens (A), causes an eye to see well (B); seeing well helps the animal avoid predators and find mates, which enables the animal to reproduce. The offspring (AA) have clear

have eyes because their parents' eyes saw well, which enabled the parents to survive and reproduce. Their eyes look similar to their parents' eyes.	lenses and can see well too. It looks like the offspring have eyes so that they can see well, but this idea is incorrect. Rather, the offspring have eyes because their parents' eyes saw well, which enabled the parents to survive and reproduce. Their eyes look similar to their parents' eyes.

Above are a sample of the refutational text and the text from which it was adapted. The underlined sections contain the language, which was added to make the text refutational.

APPENDIX B

CONSTRUCTED RESPONSE PROMPTS

Think-aloud Prompt	Self-explain Prompt
You will now be asked to read these texts. After you read, you will be asked questions about the texts. One way to help us learn how you read is to think-aloud. To help you with this task, we would like you to report your thoughts about the text while you read. Please report your thoughts that immediately come to mind regarding how you understand the meaning of the text. After some sentences, which are bolded, there is a blank box. In the space provided below the bolded segment, please read the text, and then write any thoughts that immediately come to mind. Please note that there are no "right" or "wrong" thoughts. Here is an example from a student: Text: Because long-distance food shipments promote fuel use and the exploitation of cheap labor, shifting back to a more locally sourced food economy is often touted as a fairly straightforward way to cut externalities, restore some measure of equity between producers and consumers,	You will now be asked to read these texts. After you read, you will be asked questions about the texts. One way to improve your comprehension is to self-explain. To help you with this task, we would like you to provide your own self-explanations of the text while you read. Please explain the meaning of the text, elaborating beyond your initial understanding of the text. After some sentences, which are bolded, there is a blank box. In the space provided below the bolded segment, please read the text, and then write your explanation for the meaning of the text. Please note that there are no "right" or "wrong" self-explanations. Here is an example from a student: Text: Because long-distance food shipments promote fuel use and the exploitation of cheap labor, shifting back to a more locally sourced food economy is often touted as a fairly straightforward way to cut externalities, restore some measure of equity between producers and consumers,

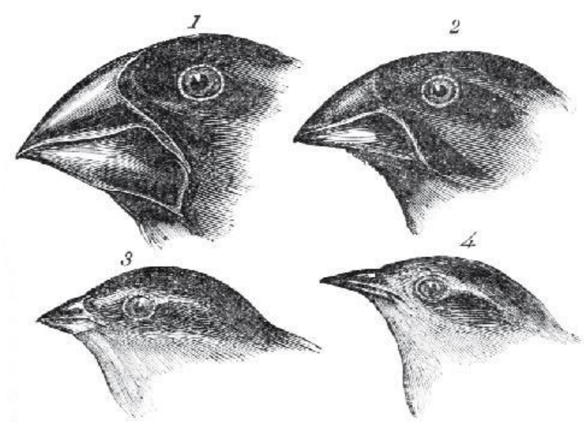
and put the food economy on a more sustainable footing. <u>Think-Aloud:</u> "That's really interesting! I didn't know that eating locally would make things cheaper and better for the economy." You can see that a think-aloud doesn't just restate the passage. It includes everything that comes to your mind. You can type anything that you think about the sentence as you read. You will now be asked to write your own think-aloud statements for each of the target segments within the text. Please respond to the text segments in the order in which they are presented.	and put the food economy on a more sustainable footing. <u>Self-Explanation:</u> "This sentence is saying eating food that is not produced locally is more expensive and that eating locally might actually reduce some costs and make the economy more sustainable. This makes sense since we spend so much money to get food shipped halfway around the world, when we could try to rely on closer food sources." You can see that an explanation doesn't just restate the passage. It explains what the passage means. You can use anything you know about the information in the sentence to explain it. You will now be asked to write your own self-explanations for each of the target segments within the text. Please respond to the text segments in the order in which they are presented.

APPENDIX C

CONCEPTUAL INVENTORY OF NATURAL SELECTION

Your answers to these questions will assess your understanding of the Theory of Natural Selection. Please choose the answer that best reflects how a biologist would think about each question.

Scientists have long believed that the 14 species of finches on the Galapagos Islands evolved from a single species of finch that migrated to the islands one to five million years ago (Lack, 1940). Recent DNA analyses support the conclusion that all of the Galapagos finches evolved from the warbler finch (Grant, Grant & Petren, 2001; Petren, Grant & Grant, 1999). Different species live on different islands. For example, the medium ground finch and the cactus finch live on one island. The large cactus finch occupies another island. One of the major changes in the finches is in their beak sizes and shapes, as shown in the figure.



1. Geospiza magnirostris　　2. Geospiza fortis
3. Geospiza parvula　　4. Certhidea olivacea

Finches from Galápagos Archipelago

1. What would happen if a breeding pair of finches was placed on an island under ideal conditions with no predators and unlimited food so that all individuals survived? Given enough time…
    a. The finch population would stay small because birds only have enough babies to replace themselves.
    b. The finch population would double, and then stay relatively the same.
    c. The finch population would increase dramatically.
    d. The finch population would grow slowly and then level off.
2. Finches on the Galapagos Islands require food to eat and water to drink…

a. When food and water are scarce, some birds may be unable to obtain what they need to survive.
b. When food and water are limited, the finches will find other food sources, so there is always enough.
c. When food and water are scarce, the finches all eat and drink less so that all birds survive.
d. There is always plenty of food and water on the Galapagos Islands to meet the finches' needs.

3. Once a population of finches has lived on a particular island for many years…
    a. The population continues to grow rapidly.
    b. The population remains relatively stable, with some fluctuations.
    c. The population dramatically increases and decreases each year.
    d. The population will decrease steadily.
4. In the finch population, what are the primary changes that occur gradually over time?
    a. The traits of each finch within a population gradually change.
    b. The proportions of finches having different traits within a population change.
    c. Successful behaviors learned by finches are passed on to offspring.
    d. Mutations occur to meet the needs of finches as the environment changes.
5. Depending on their beak size and shape, some finches get nectar from glowers, some eat grubs from bard, some eat small seeds, and some eat large nuts. Which statement best describes the interactions among the finches and the food supply?
    a. The changes in the finches' beak size and shape occurred because of their need to be able to eat different kinds of food to survive.
    b. Changes in the finches' beaks occurred by chance, and when there was a good match between beak structure and available food, those birds had more offspring.
    c. The changes in the finches' beaks occurred because the environment induced the desired genetic changes.
    d. The finches' beaks changed a little bit in size and shape with each successive generation, some getting larger and some getting smaller.
6. How did the different beak types first arise in the Galapagos finches?
    a. The changes in the finches' beak size and shape occurred because of their need to be able to eat different kinds of food to survive.
    b. Changes in the finches' beaks occurred by chance, and when there was a good match.
    c. The changes in the finches' beaks occurred because the environment induced the desired genetic changes.
    d. The finches' beaks changed a little bit in size and shape with each successive generation, some getting larger and some getting smaller.

7. What type of variation in finches is passed to the offspring?
    a. Any behaviors that were learned during the finch's lifetime.
    b. Only characteristics that were beneficial during a finch's lifetime.
    c. <u>All characteristics that are genetically determined.</u>
    d. Any characteristics that were positively influenced by the environment during a finch's lifetime.
8. What caused populations of birds having different beak shapes and sizes to become distinct species distributed on the various islands?
    a. <u>The finches were quite variable, and those features were best suited to the available food supply on each island reproduced most successfully.</u>
    b. All finches are essentially alike and there are <u>not</u> really fourteen different species.
    c. Different foods are available on different islands and for that reason, individual finches on each island gradually developed the beaks they needed.
    d. Different lines of finches developed different beak types because they needed them in order to obtain the available food.

Guppies are small fish found in streams in Venezuela. Male guppies are brightly colored, with black, red, blue and iridescent (reflective) spots. Males cannot be too brightly colored or they will be seen and consumed by predators, but if they are too plain, females will choose other males. Natural selection and sexual selection push in opposite directions. When a guppy population lives in a stream in the absence of predators, the proportion of males that are bright and flashy increases in the population. If a few aggressive predators are added to the same stream, the proportion of bright colored males decreases within about five months (3-4 generations). The effects of predators on guppy coloration have been studied in artificial ponds with mild, aggressive, and no predators, and by similar manipulations of natural stream environments (Endler, 1980).

9. A typical natural population of guppies consists of hundreds of guppies. Which statement best describes the guppies of a single species in an isolated population?
    a. The guppies share all of the same characteristics and are identical to each other.
    b. The guppies share all of the essential characteristics of the species; the minor variations they display don't affect survival.
    c. The guppies are all identical on the inside, but have many different features.
    d. <u>The guppies share many essential characteristics, but also vary in many features.</u>

10. Fitness is a term often used by biologists to explain the evolutionary success of certain organisms. Which feature would a biologist consider to be the most important in determining which guppies were the "most fit"?
    a. Large body size and ability to swim quickly away from predators.
    b. Excellent ability to compete for food.
    c. <u>High number of offspring that survived to reproductive age.</u>
    d. High number of matings with many different females.
11. Assuming ideal conditions with abundant food and no predators, what would happen if a pair of guppies were placed in a large pool?
    a. The guppy population would grow slowly, as guppies would have only the number of babies that are needed to replenish the population.
    b. <u>The guppy population would grow slowly at first, then would grow rapidly and thousands of guppies would fill the pond.</u>
    c. The guppy population would never become very large, because only organisms such as insects and bacteria reproduce in that manner.
    d. The guppy population would continue to grow slowly over time.
12. Once a population of guppies has been established for a number of years in a real (not ideal) pond with other organisms including predators, what will likely happen to the population?
    a. <u>The guppy population will continue to stay about the same size.</u>
    b. The guppy population will continue to rapidly grow in size.
    c. The guppy population will gradually decrease until no more guppies are left.
    d. It is impossible to tell because populations do not follow patterns.
13. In guppy populations, what are the primary changes that occur gradually over time?
    a. The traits of each individual guppy within a population gradually change.
    b. <u>The proportions of guppies having different traits within a population gradually change.</u>
    c. Successful behaviors learned by certain guppies are passed on to offspring.
    d. Mutations occur to meet the needs of the guppies as the environment changes.

The Canary Islands are seven islands just west of the African continent. The islands gradually became colonized with life: plants, lizards, birds, etc. Three different species of lizards found on the islands are similar to one species found on the African continent (Thorpe & Brown, 1989). Because of this, scientists assume that the lizards traveled from Africa to the Canary Islands by floating on tree trunks washed out to sea.

14. Lizards eat a variety of insects and plants. Which statement describes the availability of food for lizards on the Canary Islands?
    a. Finding food is not a problem since food is always in abundant supply.
    b. Since lizards can eat a variety of foods, there is likely to be enough food for all of the lizards at all times.
    c. Lizards can get by on very little food, so the food supply does not matter.
    d. <u>It is likely that sometimes there is enough food, but at other times there is not enough food for all of the lizards.</u>
15. What do you think happens among the lizards of a certain species when the food supply is limited?
    a. The lizards cooperate to find food and share what they find.
    b. The lizards fight for the available food and the strongest lizards kill the weaker ones.
    c. Genetic changes that would allow lizards to eat new food sources are likely to be induced.
    d. <u>The lizards least successful in the competition for food are likely to die of starvation and malnutrition.</u>
16. Populations of lizards are made up of hundreds of individual lizards. Which statement describes how similar they are likely to be to each other?
    a. All lizards in the population are likely to be nearly identical.
    b. All lizards in the population are identical to each other on the outside, but there are differences in their internal organs such as how they digest food.
    c. <u>All lizards in the populations share many similarities, but there are differences in features like body size and claw length.</u>
    d. All lizards in the population are completely unique and share no features with other lizards.
17. Which statement could describe how traits in lizards pass from one generation of lizards to the next generation?
    a. Lizards that learn to catch a particular type of insect will pass the new ability to offspring.
    b. Lizards that are able to hear, but have no survival advantage because of hearing, will eventually stop passing on the "hearing" trait.
    c. Lizards with stronger claws that allow for catching certain insects have offspring whose claws gradually get stronger during their lifetime.
    d. <u>Lizards with a particular coloration and pattern are likely to pass the same trait on to offspring.</u>

	Lizard A	Lizard B	Lizard C	Lizard D
Body Length	20cm	12cm	10cm	15cm
Offspring surviving to adulthood	19	28	22	26

Age at death	4 years	5 years	4 years	6 years
Comments	Lizard A is very healthy, strong, and clever.	Lizard B has mated with many lizards.	Lizard C is dark, colored, and very quick.	Lizard D has the largest territory of all the lizards.

18. Fitness is a term often used by biologists to explain the evolutionary success of certain organisms. Above are descriptions of four fictional female lizards. Which lizard might a biologist consider to be "the most fit"?
    a. Lizard A
    b. <u>Lizard B</u>
    c. Lizard C
    d. Lizard D
19. According to the theory of natural selection, where did the variations in body size in the three species of lizards most likely come from?
    a. The lizards needed to change in order to survive, so beneficial new traits developed.
    b. The lizards wanted to become different in size, so beneficial new traits gradually appeared in the population.
    c. <u>Random genetic changes and sexual recombination both created new variations.</u>
    d. The island environment caused genetic changes in the lizards.
20. What could cause one species to change into three species over time?
    a. Groups of lizards encountered different island environments so the lizards needed to become new species with different traits in order to survive.
    b. Groups of lizards must have been geographically isolated from other groups and random genetic changes must have accumulated in these lizard populations over time.
    c. There may be minor variations, but all lizards are essentially alike and all are members of a single species.
    d. In order to survive, different groups of lizards needed to adapt to the different islands, and so all organisms in each group gradually evolved to become a new lizard species.

APPENDIX D

TEXT COMPREHENSION QUESTIONS

Mean, standard deviation, skew, and kurtosis of all piloted comprehension questions

	Mean (SD)	Skew	Kurtosis
<u>How does a round lens enable animals to reproduce?</u>	<u>0.25(0.39)</u>	<u>1.11</u>	<u>-0.46</u>
Why can't a single replicator fill the earth with its copies?	0.13(0.31)	2.21	3.22
<u>Why is good vision disproportionately passed down from parents to offspring?</u>	<u>0.31(0.34)</u>	<u>0.65</u>	<u>-0.74</u>
What are three things living things can do that separates them from non-living things?	0.11(0.25)	2.13	3.72
Why are copies of a replicator not identical?	0.14(0.35)	2.21	3.22
What is one example of the complex design of organs?	0.12(0.29)	2.26	3.72
<u>How do replicators adapt to the environment?</u>	<u>0.22(0.35)</u>	<u>1.27</u>	<u>0.13</u>
Why do replicators accumulate changes for the better?	0.08(0.18)	1.92	1.72
Why do the great, great grand copies of a replicators have the same good traits as early replicators?	0.07(0.22)	3.23	9.65
<u>How do random errors in the copying process increase the probability of replication?</u>	<u>0.31(0.33)</u>	<u>0.57</u>	<u>-0.73</u>
<u>What was the first replicator a product of?</u>	<u>0.26(0.36)</u>	<u>1.06</u>	<u>-0.22</u>
Once a replicator has a clear lens and round eyeballs, what else can change to improve their vision?	0.12(0.28)	2.14	2.39
What are two reasons an animal with cloudy lenses might be unable to reproduce?	0.11(0.26)	2.24	4.04
What process enables organisms to adjust to a constantly	0.05(0.17)		

changing environment?	3.51	12.65

Underlined questions were included in the study.

APPENDIX E

PRIOR SCIENCE KNOWLEDGE TEST

You will now answer questions that assess your prior knowledge of science. We anticipate that you will not have enough experience with the topics to answer all of the questions correctly. Therefore, simply answer the questions as accurately as you possibly can.

Q1 Changes in species over time is called

1. Fitness
2. <u>Evolution</u>
3. Diversity
4. Relative dating

Q2 Which of these has a positive charge and is found in the nucleus?

1. Neutrons
2. <u>Protons</u>
3. Electrons
4. Elements

Q3 Which of these causes ocean tides on Earth?

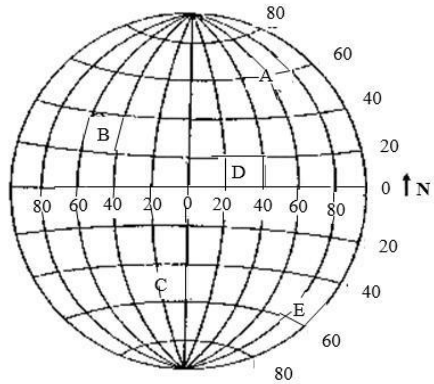
1. <u>The gravitational pull of the moon</u>
2. The revolution of the Earth around the sun
3. Differences in wind speed around the Earth
4. The tilt of the Earth's axis

Q4 A man shopping for a car wants to calculate the average price for a car at seven different dealers. How can he do this?

1. Add the number together
2. Identify the most frequent number
3. Sort the numbers and identify the fourth lowest number
4. <u>Add the numbers together and divide by seven</u>

Q5 Which three-week period produced the most dramatic rise in sales?

1. Weeks 5, 6, and 7
2. <u>Weeks 3, 4, and 5</u>
3. Weeks 1, 2, and 3
4. Weeks 2, 3, and 4



Q6 According to the adjacent illustration, which of the lettered points is found at 50 degrees south latitude and 20 degrees west longitude?

1. A
2. B
3. <u>C</u>
4. D

Q7 The basic unit of structure and function in living things is the

1. <u>Cell</u>
2. Tissue
3. Organ
4. Organ system

Q8 What are the recorded observations in an experiment called?

1. Apparatus
2. <u>Data</u>
3. Hypothesis

4. Variables

Q9 What fraction of a yard is two feet?

1. 1/10
2. <u>2/3</u>
3. 3/4
4. 1/3

Q10 Molten rock beneath the Earth's crust is called

1. <u>Magma</u>
2. Liquicite
3. Lava
4. Igneous

Q11 According to the protoplanet hypothesis, the solar system began as which of the following?

1. A star
2. A vacuum
3. <u>A huge cloud of dust and gas</u>
4. A group of comets

Q12 Michael paid $320 dollars for a bicycle at a 20% off sale. What was the original price of the bike?

1. $420 dollars
2. <u>$400 dollars</u>

3. $256 dollars

4. $300 dollars

Q13 What represents a chemical changes in matter?

1. Carbon dioxide undergoing sublimination

2. Water dissolving salt to form a solution

3. Water undergoing evaporation

4. <u>Metal post beginning to rust</u>

Q14 The New Wave Swim Team uses a 50-meter pool during the summer to prepare for long-course events. Josh competes in the 1,500 meter race. How many kilometers are in this event?

1. <u>1.5</u>

2. 15

3. 50

4. 150

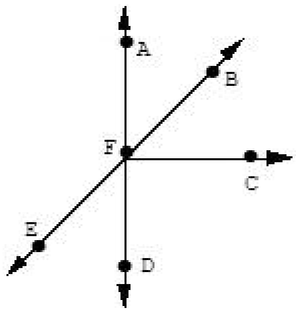
Q15 What is the name given to the science that studies the atmosphere?

1. Oceanography

2. Atmospherology

3. <u>Meteorology</u>

4. Weatherology



Q16 Which of the above pairs of line segments is perpendicular?

1. AD and EB
2. <u>AD and FC</u>
3. EB and FC
4. EF and FB

Q17 DNA is a

1. Carbohydrate
2. Lipid
3. <u>Nucleic Acid</u>
4. Sterol

Q18 Which of the following is not an element?

1. <u>Water</u>
2. Carbon
3. Oxygen
4. Hydrogen

APPENDIX F

PRELIMINARY MEDIATION ANALYSES

Previous research (see McNamara, 2004; Ozuru, Briner, Best, & McNamara, 2010) indicates that the quality of students' constructed responses is related to students' comprehension of a text. Thus, the relationship between constructed response prompt and CINS score may be mediated by the quality of the constructed response. We explored this potential relationship by conducting a mediation analysis using verb overlap as a proxy for response quality. Previous work has demonstrated verb overlap correlates with human ratings of writing quality (see Crossly, Kyle, & McNamara, 2018; Varner, Roscoe, & McNamara, 2013).

Figure 15 shows the results of the mediation analysis. The effect of prompt on CINS score was not mediated by verb overlap. There were no significant effects in the mediation analyses.

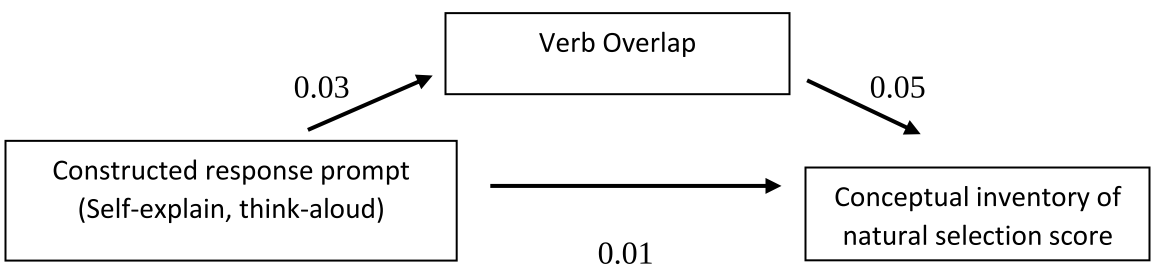


Figure 15. Regression coefficients for the relationship between constructed response prompt, verb overlap of the constructed responses, and score on the conceptual inventory of natural selection.

APPENDIX G

UNIVERSITY APPROVAL FOR HUMAN SUBJECTS RESEARCH

APPROVAL: EXPEDITED REVIEW

Danielle McNamara
Psychology
480/727-5690
dsmcnama@asu.edu

Dear Danielle McNamara:

On 9/19/2018 the ASU IRB reviewed the following protocol:

Type of Review:	Initial Study
Title:	Developing a Deeper Understanding of Cognitive Processes Driving Multiple Document Comprehension
Investigator:	Danielle McNamara
IRB ID:	STUDY00008716
Category of review:	(7)(b) Social science methods, (8)(a) Long-term follow-up, (7)(a) Behavioral research
Funding:	Name: DOEd: Institute of Education Sciences (IES), Grant Office ID: FP00012432
Grant Title:	FP00012432;
Grant ID:	FP00012432;
Documents Reviewed:	• Study1_Flier.pdf, Category: Recruitment Materials; • Study2 Consent.pdf, Category: Consent Form; • Study3 Consent.pdf, Category: Consent Form; • Study1 Consent.pdf, Category: Consent Form; • MD proposal 8-15-17 final4.docx, Category: Sponsor Attachment; • Study3 Recruitment Letter.pdf, Category: Recruitment Materials; • Study2_Flier.pdf, Category: Recruitment Materials; • Multi-Doc IRB Final.docx, Category: IRB Protocol; • ClarificationQuestions_9_6_18.pdf, Category: Other (to reflect anything not captured above); • Multi-Doc Appendix A Question Examples.pdf, Category: Measures (Survey questions/Interview questions /interview guides/focus group questions);

The IRB approved the protocol from 9/19/2018 to 9/18/2019 inclusive. Three weeks before 9/18/2019 you are to submit a completed Continuing Review application and required attachments to request continuing approval or closure.

If continuing review approval is not granted before the expiration date of 9/18/2019 approval of this protocol expires on that date. When consent is appropriate, you must use final, watermarked versions available under the “Documents” tab in ERA-IRB.

In conducting this protocol you are required to follow the requirements listed in the INVESTIGATOR MANUAL (HRP-103).

Sincerely,

IRB Administrator

cc: Micah Watanabe

Danielle McNamara