Effects of Keyword Generation and Peer Collaboration on Metacomprehension Accuracy in Middle School Students

Lisa S. Pao

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

Columbia University

2014

ABSTRACT

Effects of Keyword Generation and Peer Collaboration on Metacomprehension Accuracy in Middle School Students

Lisa S. Pao

Metacomprehension refers to the ability to judge one's own comprehension. Studies in the literature have shown that generating keywords after reading helps adults and children make comprehension judgments that are better correlated with their actual comprehension. Researchers have also found that when metacomprehension is framed in terms of confidence, there is an effect of ability, where individuals with low ability tend to be overconfident in their judgments, while those with high ability tend to be underconfident. This paper describes two experiments investigating metacomprehension in seventh graders.

Experiment 1 sought to replicate and extend the finding that generating keywords after reading improves the accuracy of comprehension judgments and the effectiveness of study choices. To account for potential effects of time on task, participants in the control condition were asked to read passages twice in lieu of generating keywords. Two measures of metacomprehension accuracy (signed differences and gamma correlations) were based on comprehension judgments taken at two time points (pre-test and post-test). The moderating effects of reading ability were also examined.

The results of Experiment 1 showed that participants were overconfident in their judgments of their own comprehension. Overconfidence was greater for pre-test predictions than for post-test reflections, and it was also greater for participants with lower reading ability.

Generating keywords caused participants to become significantly less overconfident– or more accurate– from pre-test to post-test in their comprehension judgments, but it did not actually boost comprehension scores. In other words, generating keywords helped participants know that they did not know; it did not, however, help them know more.

In Experiment 2, the investigation of generating keywords and rereading text was situated within a new context incorporating practice test questions. Studies have shown that practice testing is an effective study strategy. Additionally, since researchers have found that learners can use information about peer performance as a basis for making judgments about themselves, Experiment 2 also asked whether peer collaboration might increase metacomprehension accuracy. Participants were randomly assigned to four conditions: individual/keyword, individual/reread, collaborate/keyword, and collaborate/reread. All participants answered practice test questions; participants in the individual conditions worked on the questions alone, while participants in the collaborative conditions discussed the questions with a partner.

As in Experiment 1, participants in Experiment 2 were also overconfident in judging their own comprehension. Again, there was an effect for time of judgment, such that predictions were more overconfident than were reflections. Surprisingly, peer collaboration was found to lead to greater overconfidence in comprehension judgments. Participants who collaborated with a peer were more overconfident than participants who worked alone. Experiment 2 showed that in the presence of practice testing and peer collaboration, the interactive effect of keyword generation and time of judgment was minimized. Within the keyword group, participants who collaborated and participants who worked alone did not differ in overconfidence. Within the reread group, however, participants who collaborated were significantly more overconfident than those who worked alone.

Taken together, these two studies suggest that middle school students are generally overconfident in their judgments of comprehension. However, the results indicate that study strategies designed to enhance comprehension and learning can be effective in reducing students' overconfidence about themselves.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor and sponsor, Joanna Williams, for the support and guidance you have given me over the course of my time here. From the moment I first arrived on campus five years ago, you have exercised a firm and steady hand in shaping my development as a scholar and a researcher, and I am grateful for having had the opportunity to learn from you. It has been an honor to be your student.

Thank you to the rest of my committee: Janet Metcalfe, Elizabeth Tipton, Stephen Peverly, and Linnea Ehri. Your questions, comments, and insights have made all the difference. I am grateful for having had the chance to learn from each one of you in classes, lab meetings, office hours, and, of course, hearings and defenses. Thank you.

Last but not least, I am so grateful for the support I have had from my friends and family throughout my time in graduate school. To Laurie Hemberger, Lisa Simmons, and the rest of the girls: thank you for all of the very happy hours that have helped keep it real and keep me sane. To my family and the Portage la Prairie community: thank you for being my cheerleaders from afar. And finally, to Seth, who has been and remains my biggest supporter- thank you.

-L.P.

To my family– Pao and Tardiff.
Thank you for making this possible.

L.P.

CHAPTER 1

INTRODUCTION

As Charles Darwin (1871) cautioned in *The Descent of Man*, "ignorance more frequently begets confidence than does knowledge" (p. 3). Darwin's comment was directed at those who continued to insist that the origin of humankind was a problem that could never be solved by science. Nearly a century and a half later, Nate Silver (2012) issued the following warning in *The Signal and the Noise*: "the amount of confidence someone expresses in a prediction is not a good indicator of its accuracy- to the contrary, these qualities are often inversely correlated" (p. 203). Silver was discussing failures of predictions among stock traders, political pundits, weather forecasters, and sports bettors.

Both Darwin and Silver were referring to a specific case of confidence– *overconfidence*– in which individuals rate their own competence more highly than is warranted by their actual accomplishment. Researchers studying cognitive bias and decision making have suggested that overconfidence is rooted in two different sources: a lack of knowledge or skill, and also a misperception about the self and about others. With regards to the first source, the *double-curse hypothesis* holds that individuals who lack the skills to produce a correct answer in the first place also necessarily lack the skills to determine whether that answer is right or wrong (Kruger & Dunning, 1999). With regards to the second source, studies from the social psychology literature have found an *above-average effect*, where the majority of people tend to (incorrectly *and* impossibly) believe that they are better than the average person (e.g., Alicke, 1985; Brown, 1986; Dunning, Meyerowitz, & Holzberg, 1989).

One way to account for overconfidence as well as underconfidence is to attribute both to a failing of metacognition. Simply put, *metacognition* can be defined as thinking about thinking, or knowing about knowing (Flavell, 1979). Most definitions of metacognition consist of two components: knowledge of cognitive processes, and control of cognitive processes (Hacker, 1998; Jacobs & Paris, 1987, Kuhn & Dean, 2004). Others have invoked the notion of levels of processing, suggesting that metacognition involves the interplay between a cognitive level, where memory, attention, comprehension, and learning occur, and a metacognitive level, which monitors and controls the activity occurring at the cognitive level (Nelson & Narens, 1990; 1994). Ultimately, the importance of metacognition is that it allows a learner to control his or her own learning and understanding (Nelson & Narens, 1990).

Cognitive psychologists often separate metacognition into two types of knowing: knowing when you know and knowing when you do not know (Metcalfe & Finn, 2013). In a now (in)famous news briefing given in 2002, former Secretary of Defense Donald Rumsfeld added a third when he made the following statement: "as we know, there are known knowns; there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns -- the ones we don't know we don't know." Social psychologists studying optimism and pessimism in relation to performance have contributed a fourth: unknown knowns- things we do not realize we know, and that we may therefore be pleasantly surprised to discover that we actually *do* know (Krueger & Mueller, 2002; Shepperd, Ouellette, & Fernandez, 1996; Wedell & Parducci, 2000).

The current research focuses on metacognition in relation to reading comprehension– commonly referred to as *metacomprehension*. Metacomprehension involves people's ability to judge their comprehension of text (Dunlosky & Lipko, 2007). Like metacognition,

metacomprehension involves two components: evaluation of the quality of comprehension, and regulation of strategy use to address comprehension gaps (Huff & Nietfeld, 2009). Both components are essential to effective learning from text: metacognitive readers can detect instances in which their comprehension breaks down due to inadequate skills, and then select and apply appropriate strategies to repair their comprehension (Williams & Atkins, 2009). Additionally, if learners can distinguish between well-learned and less well-learned material, they can focus their efforts on unlearned material and avoid wasting time (Dunlosky & Lipko, 2007).

While most studies in the metacomprehension literature have involved adults, de Bruin, Thiede, Camp, and Redford (2011) investigated metacomprehension accuracy in elementary and middle school students. Students were randomly assigned to either a keyword condition, in which they were asked to generate a list of keywords about a text after reading but before testing, or to a no-keyword control condition. In the experiment, participants read several text passages, either generated or did not generate keywords, made comprehension judgments, and took comprehension tests. The researchers found that for the middle school students, metacomprehension accuracy was significantly greater in the keyword group than in the no-keyword group; however, elementary students in both conditions made many inaccurate judgments, and the two groups did not differ. When asked to select texts for restudy, the middle school students based their study choices on their comprehension judgments, choosing texts they perceived as less well-learned. The elementary students also based their study choices on their comprehension judgments, but because those judgments were inaccurate to begin with, they actually ended up choosing the wrong texts to restudy (de Bruin et al., 2011).

In addition to being the first study in the literature to examine metacomprehension in children, the work of de Bruin and her colleagues (2011) demonstrates the importance of metacomprehension accuracy: learners who are accurate– that is, neither overconfident nor underconfident– in their judgments of their own comprehension can study much more effectively than learners who are inaccurate. For students, the inability to make accurate judgments can have serious consequences, such as studying the wrong material or terminating study prematurely. According to Metcalfe and Dunlosky (2008), one of the most compelling reasons for investigating metacognitive and metamemory judgments is that those judgments presumably guide learners' subsequent study behavior and learning gains. Given the reading scores from the most recent administration of the National Assessment of Educational Progress, where only 35 percent of fourth graders and 36 percent of eighth graders performed at or above the *proficient* level (National Center for Education Statistics, 2013), we might ask whether this low performance is due to poor metacomprehension, and if so, whether and how metacomprehension accuracy can be improved.

Evidence from social psychology and cognitive psychology suggests that two factors that could help address these questions are ability and collaboration. Kruger and Dunning (1999) linked inaccurate judgments about oneself with both academic achievement and peer perception, proposing that while low-achieving students tend to incorrectly assess their own performance, high-achieving students know how well they themselves performed on a test or task but overestimate how well others are doing on the same task or test. In other words, overconfidence among low achievers stems from an error about the *self*, while underconfidence among high achievers stems from an error about *others*. Tversky and Kahneman (1974) theorized that when individuals make estimates, they employ an *anchoring and adjustment heuristic*, in which they

use existing information as a starting point and then modify the value of their estimate until a plausible result is obtained. In accordance with *cognitive anchoring theory* (Tversky & Kahneman, 1974), collaborating with others may provide learners with more accurate anchors on which to base their comprehension judgments.

Overview of the Current Research

The current research comprises two experiments that asked whether or not keyword generation affects the accuracy of seventh graders' comprehension judgments and also whether or not those judgments can then be applied in order to make effective study choices.

The purpose of Experiment 1 was to extend the finding that generating keywords after reading improves the accuracy of comprehension judgments and the effectiveness of study choices. A goal of Experiment 1 was to compare keyword generation to a stronger control condition than existing studies have used. Since the literature has shown that rereading– that is, restudying text again after an initial reading– is one of the most frequently used study techniques (e.g., Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013), it was adopted for use as a control condition. Two measures of metacomprehension accuracy (signed differences and gamma correlations) were based on comprehension judgments taken at two time points (pre-test and post-test). The moderating effects of reading ability were also examined.

In Experiment 2, the comparison of keyword generation to rereading text was situated within a new study context incorporating practice testing. A new variable, peer collaboration, was added. Of primary interest was whether the effect of keyword generation on metacomprehension accuracy and study regulation would continue to hold in the presence of practice testing and peer collaboration.

CHAPTER 2

REVIEW OF THE LITERATURE

The current research seeks to investigate the effect of keyword generation on metacomprehension accuracy and study regulation in middle school students. Two experiments were designed to explore whether or not keyword generation affects the accuracy of seventh graders' comprehension judgments and also whether or not those judgments can then be applied in order to make effective study choices. This chapter reviews the literature on metacomprehension. It will begin with an overview of the pioneering work on calibration of comprehension that to this day remains the paradigm for metacomprehension research. Next, it will discuss two different measures of metacomprehension accuracy commonly used within the literature. Then, it will review a series of studies that have sought to experimentally manipulate metacomprehension accuracy. Finally, it will discuss the relationship between metacomprehension and collaboration.

Pioneering Metacomprehension Research

Glenberg and Epstein (1985) are widely credited as being the very first researchers to study the construct now referred to as *metacomprehension*. In a series of studies (e.g., Epstein, Glenberg, & Bradley, 1984; Glenberg, Wilkinson, & Epstein, 1982), the researchers examined participants' ability to detect contradictions or inconsistencies embedded in text passages in relation to their confidence ratings about their comprehension and also to their actual performance on tests assessing comprehension of the passages. In those studies, the researchers found an *illusion of knowing* effect, in which, despite having provided high confidence ratings

about their understanding of a text, participants failed to detect the logical flaw that had been implanted in the text. In other words, participants' positive appraisals of their own comprehension were incongruous with their low performance on the actual measures of comprehension.

Recognizing that the illusion of knowing was an incomplete construct, since in addition to the possibility of overconfidence in the absence of knowledge, there was also the option of underconfidence when knowledge was in fact present, Glenberg and Epstein (1985) adopted the terms "well calibrated" and "poorly calibrated" to refer to learners, such that well calibrated individuals know when they know and know when they don't know, while poorly calibrated individuals don't know when they know and also don't know when they don't know.

Glenberg and Epstein (1985) conducted a series of three experiments in order to situate the illusion of knowing effect within a calibration framework. In the experiments, participants read short expository texts, made confidence judgments predicting their ability to draw correct inferences about each text, and answered true-or-false test questions requiring them to verify inferences about each text. Based on the confidence ratings and the test scores, point-biserial correlations were computed in order to obtain quantitative measures of calibration of comprehension.

The first of these experiments investigated the effect of varying the time interval between reading a text and making a confidence rating about that text. The researchers hypothesized that delaying the confidence judgments would increase calibration accuracy because participants would be forced to base their judgments on comprehension of the text rather than on memory for the text. Participants were assigned to either an immediate-rating condition or a delayed-rating condition. The results showed that the difference between the mean point-biserial correlations for

the two conditions was not significant, and neither of the mean correlations was significantly different from zero (r = .07 for the immediate condition; r = .04 for the delayed condition). The researchers concluded that overall calibration of comprehension was very poor and that the elapsed time between reading texts and making confidence judgments did not appear to have an effect on calibration accuracy (Glenberg & Epstein, 1985).

The second experiment in the series examined how participants' expectations about a forthcoming test would affect their calibration accuracy. Participants were assigned to either a familiarization condition, in which participants completed a practice session with the text paragraphs and the test questions, or a control condition, in which participants simply read the practice texts without exposure to the test questions. As in the first experiment, the difference between the mean point-biserial correlations for the two conditions was not significant, and neither of the mean correlations was significantly different from zero (r = .06 for the control condition; r = .12 for the familiarization condition). Again, calibration was fairly poor, and the experimental manipulation did not appear to have an effect on calibration accuracy.

In the third experiment, Glenberg and Epstein (1985) tested the hypothesis that the act of taking the inference test affected the accuracy of participants' confidence judgments (as opposed to the confidence judgments affecting test performance). A new procedure was developed, in which after reading the texts, participants completed the following five steps: (1) made pre-test confidence judgments predicting their ability to draw correct inferences about each text, (2) answered true or false inference verification test questions about each text, (3) made post-test confidence judgments about their performance on the just-completed inference tests, (4) made new confidence judgments predicting their ability to correctly verify a new set of inferences about each of the previously-read and previously-judged texts, and (5) answered a new set of true

or false inference verification test questions. Essentially, Experiment 3 added a set each of retrospective judgments, new prospective judgments, and new inference questions to the procedure used in Experiments 1 and 2.

For each of the participants in the study, three calibration scores were computed: *initial calibration of comprehension*, which involved initial confidence judgments about the texts and performance on the first inference test (1 and 2 in the above list); *calibration of performance*, which involved performance on the first inference test and retrospective confidence judgments about test performance (2 and 3); and *recalibration of comprehension*, which involved the second set of confidence judgments about the texts and performance on the second inference test (4 and 5). As expected from the results of the first two experiments, mean initial calibration was very low (r = .04) and was not significantly different from zero. Mean calibration of performance (r = .23) was significantly different from zero and from mean initial calibration. Mean recalibration (r = .19) was significantly different from zero and from mean initial calibration.

Glenberg and Epstein (1985) concluded that the results of Experiment 3 were consistent with the hypothesis that taking the inference test affected the calibration of participants' confidence judgments. One explanation the researchers offered for this finding is that verifying the first set of inferences afforded participants the opportunity to detect flaws or inconsistencies in their cognitive representations of the texts, which in turn allowed them to be better informed when making the second set of confidence judgments (Glenberg & Epstein, 1985). In other words, in those cases where calibration was poor, the assessment of comprehension was likely to be based on incorrect or irrelevant information. However, once learners were encouraged to activate appropriate mental representations and to base their judgments accordingly, calibration was improved.

The lasting contribution of Glenberg and Epstein (1985) is that most metacomprehension research to date has been modeled after their paradigm. Experiment 3 revealed the existence of unanticipated causal pathways between metacognitive judgments and test performance. Rather than it being the case that metacognitive judgments affected or predicted the magnitude of test scores, the evidence suggested that the metacognitive judgments were susceptible to influence from previous testing. Several important lessons can be drawn from the work of Glenberg and Epstein (1985): first, calibration of comprehension is often very poor but can be improved; second, the ordering of steps in the procedure is critically important in metacomprehension studies; third, calibration accuracy is subject to the effects of practice and self-assessment; and finally, the difference between prospective and retrospective judgments is a promising area for future study.

Measuring Metacomprehension Accuracy

Measures of metacomprehension accuracy are based on the relationship between comprehension judgments and test scores. One method commonly used to operationalize metacomprehension accuracy is to compute a correlation between judgments of perceived comprehension and actual performance on a comprehension test. Because comprehension judgments often take the form of ratings or rankings, which are measured on ordinal rather than interval scales, non-parametric measures of correlation are thought to be more appropriate than the Pearson product-moment correlation (Goodman & Kruskal, 1954; Nelson, 1984). Although Glenberg and Epstein (1985) used point-biserial correlations, as the tests they used contained only dichotomous true-or-false items, most studies in the literature have since used multiple-

choice items with more than two answer choices and, accordingly, have instead used gamma correlations[1].

Within the literature, metacomprehension accuracy as operationalized via gamma correlations is often referred to as *resolution* or *relative accuracy*. Resolution measures the degree to which an individual's judgments correlate with his or her performance across texts (Dunlosky & Lipko, 2007), and assesses whether judgments increase monotonically with performance (Maki, 1998a; Schraw, 2009). Gamma ranges from –1 to +1, with stronger positive correlations indicating greater accuracy. To illustrate, suppose an individual read five texts, and responded to the prompt, "Given six test questions about each text, how many questions do you think you will answer correctly for each text?" with the following predictions: 4, 3, 5, 3, and 2. Suppose that he or she then took the tests and obtained the following scores: 5, 3, 4, 4, and 3. The gamma for this individual is 0.71, a moderately strong positive correlation. The individual here provided higher comprehension judgments for passages on which he or she obtained higher scores.

Alternatively, metacomprehension accuracy can be operationalized as the signed differences between judgments and performance. The literature uses the terms *confidence* and *absolute accuracy* to refer to this measure. Signed differences provide an index of the degree to which an individual is underconfident or overconfident (Maki, 1998a; Schraw, 2009). Positive values indicate overconfidence, negative values indicate underconfidence, and zeroes indicate

[1] Gamma is a non-parametric statistic that can be used when data are on an ordinal scale and when there are many ties in the data (Goodman & Kruskal, 1954; Nelson, 1984). Gamma is computed with the formula, $G = (C – D)/(C + D)$, where a concordance (C) between two variables (e.g., metacomprehension judgment and test performance) occurs when one variable is increasing from one text to another and the other variable is also increasing across this same pair of texts, and where a discordance (D) occurs when one variable is increasing from one text to another and the other variable is decreasing across this same pair of texts (Nelson, 1984).

accuracy– neither overconfidence nor underconfidence. For the individual described in the preceding paragraph, subtracting each of the test scores from the corresponding prediction results in the following signed differences: -1, 0, +1, -1, and -1. Most of the values are negative, and the mean of the values is -0.4, indicating that the individual is underconfident in his or her predictions.

There are different arguments for and against using each measure. Many studies in the literature have used gammas, so they can be compared across studies (Metcalfe & Finn, 2013). Another advantage is that gamma does not require that comprehension judgments and test scores be in the same units (Maki, 1998a). However, a problem with using gamma is that values are often very low: in two separate reviews of metacomprehension studies, researchers found that mean gammas did not exceed 0.30 (Dunlosky & Lipko, 2007; Maki, 1998b). Likewise, Kelemen, Frost, and Weaver (2000) computed test-retest reliability and found that while test scores, comprehension judgments, and signed differences were stable over time and across tasks, gamma correlations were not. A further disadvantage of using gamma correlations is that, unlike signed differences, they do not indicate overconfidence or underconfidence. On the other hand, a problem with signed differences is that overconfident and underconfident judgments can cancel each other. Squaring the values is an inadequate solution, since the sign of the deviation is as important as its magnitude.

It seems that, alone, neither measure is sufficient. As further illustration, consider an individual who makes the predictions, 6, 5, 6, 6, and 5, and obtains the scores 3, 2, 5, 3, and 2. Because the judgments increased monotonically with the test scores, this individual has perfect resolution or relative accuracy with a gamma of 1. Subtracting each test score from the corresponding prediction results in the following values for confidence or absolute accuracy: +3,

+3, +1, +3, and +3. In this case, the mean of the values is +2.6, indicating that this individual is overconfident. Ignoring either measure would lead to very different interpretations. How could an individual be considered "accurate" despite having provided clearly incorrect and overconfident judgments? Maki (1998a) recommended that, when possible, researchers should include both measures in order to adequately understand the phenomenon.

Maki, Shields, Wheeler, and Zacchilli (2005) investigated the effects of text difficulty and verbal ability on signed differences and gamma correlations. Participating college students read texts, made pre-test comprehension judgments, took comprehension tests, and made post-test judgments about their test performance. The researchers randomly assigned participants to a hard-text condition, in which they read texts from a GRE preparation manual; a revised-text condition, in which they read texts that had been revised to increase readability; and a mixed-text condition, in which they read both hard and revised texts. The participants were also classified into low, medium, and high ability groups based on their scores on the verbal portion of either the SAT or the ACT.

Using signed differences, Maki and her colleagues (2005) found that, overall, low-ability students were overconfident while high-ability students were underconfident. In particular, low-ability participants were overconfident in their pre-test predictions but accurate in their post-test judgments, while medium- and high-ability participants were accurate in their predictions but underconfident in their post-test judgments. Using gamma correlations, the researchers obtained values ranging from -0.02 to 0.64. However, the only significant main effect was for judgment type: post-test judgment gammas were significantly higher than pre-test judgment gammas, but neither verbal ability nor text difficulty had an effect on gamma (Maki et al., 2005).

Maki and her colleagues (2005) found that not only were signed differences and gamma correlations affected differently by text difficulty and verbal ability, the two measures were also uncorrelated. The researchers interpret this finding as an indication that the two measures may tap different processes altogether, or at least, may tap different aspects of metacomprehension (Maki et al., 2005). However, the researchers note that the gammas obtained in their study followed consistent patterns: almost all of the gammas were significantly different from zero, all of the post-test gammas were significantly greater than the pre-test gammas, and mean gammas for hard texts were higher (but not significantly so) than for revised texts. Maki and her colleagues (2005) conclude that, taken together, these results do not support the assertion that gamma is an unreliable and unstable measure (e.g., Kelemen et al., 2000).

Following the example of Maki and her colleagues (2005), the current research includes signed differences as well as gamma correlations in an effort to obtain a more complete understanding of metacomprehension. Even if the two measures capture different aspects of the construct, assuming that both do in fact measure metacomprehension, the resulting interpretations should be consistent with one another. The current research also includes post-test reflections as well as pre-test predictions in order to replicate and extend the finding that post-test judgments were more accurate than pre-test judgments for both measures of metacomprehension accuracy.

Manipulating Metacomprehension Accuracy

Another group of metacomprehension researchers has focused on the effects of keyword generation. Thiede, Anderson, and Therriault (2003) investigated the role of monitoring accuracy in regulation of learning. Participants in the study were asked to read several texts, make pre-test

comprehension judgments about each text, and take a comprehension test for each text. Monitoring accuracy was manipulated by having participants generate keywords immediately after reading each text, generate keywords at a delay after reading all of the texts, or not generate any keywords at all. After completing the tests, participants were told the number of items they had answered correctly. They were then given the opportunity to select texts for optional rereading and were tested once again. As in previous studies, metacomprehension accuracy was operationalized as the gamma correlations between the judgments and the test scores.

The researchers found that the gamma correlations between pre-test comprehension judgments and test scores were significantly higher for the delayed-keyword group (M = .70) than for the immediate-keyword group (M = .23) or for the no-keyword group (M = .36). Additionally, the delayed-keyword group was significantly more likely than either of the other groups to select less-learned texts for restudy instead of better-learned texts. It was also the only group with significant score increases from the first test to the second test.

Thiede and his colleagues (2003) suggest that the superior metacomprehension accuracy and study regulation of the delayed-keyword group can be explained by spreading activation theories of text comprehension (e.g., Britton & Gülgöz, 1991). They argue that when writing keywords immediately after reading, an individual may have information active in his or her working memory even for a text that was not well understood. However, when writing keywords at a delay, the individual is more dependent on long-term memory, and therefore, is in a better position to realize that he or she may have little to draw on when attempting to retrieve information about a poorly understood text.

In a follow-up study, Thiede, Dunlosky, Griffin, and Wiley (2005) investigated several hypotheses for the delayed-keyword effect on metacomprehension accuracy. One potential

explanation involves the reading-keyword lag, or the elapsed time between reading and keyword generation; according to this account, delaying keyword generation forces a participant to access a situation model for a given text. A *situation model* is a mental representation that connects ideas within the text to other ideas contained in the text and to the reader's prior knowledge about the text (Kintsch, 1988; 1998). A second possibility involves the keyword-keyword lag, or the time elapsed between generating keywords for one text and then another text; according to this explanation, the successive generation of keywords causes participants to judge the texts relative to each other. A third option involves the keyword-judgment lag; this explanation holds that minimizing the lag between generating keywords and judging comprehension inhibits forgetting of keyword-relevant information. The researchers conducted a series of experiments in order to systematically evaluate the three hypotheses, and found evidence in favor of the reading-keyword lag hypothesis (Thiede et al., 2005). They found that spacing keyword generation and delaying comprehension judgments did not have an effect on metacomprehension accuracy, thus disconfirming the keyword-keyword lag and the keyword-judgment lag explanations for the delayed-keyword effect.

Anderson and Thiede (2008) sought to determine whether the delayed-keyword effect would extend to summary writing in place of keyword generation. The researchers used a counterbalanced within-subjects design with three conditions: delayed summary, immediate summary, and no summary. Participants read several texts, made pre-test comprehension judgments, and took a test. As in previous studies, metacomprehension accuracy was operationalized as the gamma correlations between the judgments and the test scores. The researchers found that the participants had significantly higher metacomprehension accuracy when they wrote summaries at a delay than when they wrote summaries immediately or when

they did not write summaries. Anderson and Thiede (2008) concluded that the delayed-keyword effect does in fact extend to summary writing, and that the results of their study provide further support for the situation model hypothesis.

The evidence from the work of Thiede and his colleagues suggests that the delayed-keyword (and delayed-summary) effect is robust, and that the delayed-keyword manipulation should be included as a baseline condition in future metacomprehension studies. Indeed, the gamma correlation of .70 observed in the delayed-keyword condition of the Thiede et al. (2003) study is higher than any other value of gamma observed in any of the other studies included in this literature review. Additionally, the support for the situation model hypothesis indicates the importance of choosing text passages and test questions that require individuals to create and access situation models while reading.

Metacomprehension and Collaboration

One important issue that metacomprehension researchers have not considered is the role of interpersonal interactions in learning and studying. Literacy learning is oftentimes social in nature, as students work with peers and teachers, asking and answering questions, arguing and defending opinions, and discussing and critiquing ideas. Even when working independently, readers must attempt to make sense of the motives, beliefs, and values of authors and characters, while writers must try to produce texts that will be coherent and convincing to others. Proponents of collaboration in the classroom argue that it encourages active participation, teaches children to work cooperatively, and prepares them for the transition into the workplace and society (De Lisi & Goldbeck, 1999; Fawcett & Garton, 2005). In fact, a key subset of standards within the newly adopted Common Core State Standards for English language arts is grouped together under the

heading, “Comprehension and Collaboration” (National Governors Association Center for Best Practices, Council of Chief State School Officers, 2010).

One type of a metacognitive skill that is social in nature is theory of mind. *Theory of mind* refers to an individual’s ability to construe people, including the self and others, as agents acting in accordance with their own desires, beliefs, and emotions (Wellman, 2011). In other words, theory of mind involves understanding that one has a mind and that others have their own minds as well. Flavell (2000; 2004) noted that the most salient difference between metacognition and theory of mind is that metacognition is idiosyncratic and introspective, while theory of mind involves an individual’s ability to understand others. Other researchers have argued that theory of mind is an essential social skill in that it allows individuals to regulate their relationships with others (Watson, Linkie-Nixon, Wilson, & Capage, 1999).

Harris, de Rosnay, and Pons (2005) use children’s comprehension of the story of Little Red Riding Hood to illustrate the development of theory of mind. In their example, children hear the story and are then asked two questions about it: (1) Who does Little Red Riding Hood think will answer the door- Grandma, or the Wolf, and, (2) How does she feel while knocking on the door- happy, or afraid? According to the researchers, children of different ages provide different answers that reflect their understanding of theory of mind: 3-year-olds incorrectly answer that Little Red Riding Hood expects the wolf to answer the door, 4- and 5-year-olds correctly answer that she expects Grandma to answer the door but incorrectly report that Little Red Riding Hood feels scared while knocking, and 6-year-olds can correctly answer that she expects Grandma to answer the door and that (therefore) Little Red Riding Hood is not scared as she knocks (Harris et al., 2005).

In a study involving four- to six-year old children and their parents, Mar and his colleagues investigated whether exposure to children's storybooks, movies, or television would predict children's theory-of-mind development (Mar, Tackett, & Moore, 2010). Using measures based on the Author Recognition Test (ART) developed by Stanovich and West (1989), Mar and his colleagues presented parents with lists of storybook, movie, and television program titles containing both real items and foils, and asked them to identify the real items. The researchers found that children whose parents were better at recognizing storybook titles performed better on a battery of theory-of-mind tasks; in a hierarchical regression analysis, parents' recognition of storybook titles predicted 26% of the variance in children's theory-of-mind scores above and beyond age, sex, vocabulary, and parental income (Mar et al., 2010). Interestingly, recognition of movie titles also increased prediction of theory-of-mind scores by 33%; however, recognition of television program titles did not predict any additional variance.

Another area of research that illustrates the relationship between metacognition and collaboration is *cognitive anchoring theory*. Tversky and Kahneman (1974) suggested that when individuals make estimates, they often employ an *anchoring and adjustment heuristic*, in which they use existing numerical information as an anchor or starting point and then adjust the value of their estimate accordingly until a plausible result is obtained. Collaboration can provide such an anchor in that an individual can use his or her appraisal of another as a point of comparison for his or her own performance. Piaget (1959) emphasized the comparative nature of collaboration, arguing that by exposing the differences between one's own knowledge and that of others, peer interactions promote cognitive conflict and change. When working with others, students encounter information about how others are performing, which can affect their self-judgments (Zhao & Linderholm, 2011). According to Zhao and Linderholm (2011), with regards

to making metacomprehension judgments, collaborating with others may provide individuals with more accurate anchors on which to base their judgments about themselves.

In two experiments, Zhao and Linderholm (2011) investigated whether providing participants with information about typical peer performance would produce anchoring effects on their metacomprehension judgments. In the first experiment, college-aged participants were randomly assigned to either a high-anchor condition, where they were told that mean peer performance was 85%, a low-anchor condition, where they were told that mean peer performance was 55%, or a no-anchor condition, where no information about peer performance was provided. The researchers found that participants in the high-anchor and the no-anchor conditions made judgments that were significantly higher in magnitude than those made by the low-anchor group; in terms of accuracy, the high-anchor group was significantly less accurate (overconfident) than the low-anchor group. Zhao and Linderholm (2011) also found that the mean judgment given by participants in the no-anchor condition was 77%, or about a C/C+ on a typical college grading scale.

In the second experiment, the researchers tested whether adjusting the values of the anchors would produce more symmetrical anchoring effects. Using the mean judgment of the no-anchor group as a midpoint, the researchers set the high anchor as 95% and the low anchor as 55%. The results supported the symmetrical anchoring hypothesis, with the high-anchor group providing significantly higher judgments and the low-anchor group providing significantly lower judgments relative to the no-anchor group. As in the first Experiment , the low-anchor group had the highest absolute accuracy of the three groups. The researchers suggest that this might be because mean performance across the three groups on the comprehension test was generally low; 50% in Experiment 1 and 45% in Experiment 2.

Summary

This chapter presented a review of the literature on metacomprehension. It reviewed the seminal studies that have informed and inspired much of the research on metacomprehension. It compared and contrasted two widely used measures of metacomprehension accuracy: signed differences and gamma correlations. It described the development and investigation of the keyword generation manipulation. It discussed the relationship between metacomprehension and collaboration.

CHAPTER 3

EXAMINING THE EFFECTS OF KEYWORD GENERATION (EXPERIMENT 1)

By middle school, much of the learning that students are expected to do will occur through independent reading of textual material. As students complete assigned readings and review class notes, they must assess, control, and repair their ongoing comprehension of the to-be-learned information. The ability to monitor and regulate comprehension of text is often referred to as *metacomprehension*. Students who are better at knowing what they do and do not comprehend– that is, students who have greater metacomprehension accuracy– will be better able to learn from text. They will be able to selectively reread, memorize, and restudy as needed, and will consequently be more effective at regulating their study than students who are unable to be metacognitive about their comprehension.

De Bruin, Thiede, Camp, and Redford (2011) investigated metacomprehension accuracy and study regulation in elementary and middle school students. Students read several text passages, either generated or did not generate keywords, made judgments about how well they comprehended the text, and took comprehension tests. The researchers computed gamma correlations between comprehension judgments and test scores. For the middle school students in the study, gamma correlations were significantly higher in the keyword group than in the no-keyword group, indicating that the keyword group had greater metacomprehension accuracy. However, elementary students in both conditions made many inaccurate judgments, and there was no difference in gamma correlations between the two conditions. When asked to select texts for restudy, the middle school students based their study choices on their comprehension judgments, choosing texts they perceived as less well comprehended. The elementary students

also based their study choices on their comprehension judgments, but because those judgments were inaccurate to begin with, they actually ended up choosing the wrong texts to restudy. This is the first study in the metacomprehension literature to involve children, and it shows the importance of metacomprehension accuracy for children as well as for adults: learners who are accurate in their judgments of their own comprehension can study more effectively than learners who are inaccurate.

In a study involving college students, Maki, Shields, Wheeler, and Zacchilli (2005) investigated the effects of verbal ability on metacomprehension accuracy. Participants read several passages, made pre-test judgments about their comprehension, took comprehension tests, and made post-test judgments about their comprehension. The researchers computed signed differences between comprehension judgments and test scores. They found that although students in all ability groups were generally inaccurate in their comprehension judgments, there was an effect of ability: students with lower verbal ability tended to be more overconfident while those with higher verbal ability were more underconfident. Time of judgment was also found to have an effect, such that overconfidence was higher for judgments made at pre-test than for judgments made at post-test.

Experiment 1 sought to extend the work of de Bruin and her colleagues. Their design, procedure, and materials were adopted for use here. Several changes were made. First, in order to control for time on task, Experiment 1 compared keyword generation to rereading text rather than to not generating keywords. Additionally, based on the work of Maki and her colleagues, Experiment 1 incorporated the following modifications into the paradigm developed by de Bruin and her colleagues: (1) participants made comprehension judgments after testing as well as before testing, (2) signed differences between comprehension judgments and test scores were

computed in addition to gamma correlations, and (3) the effect of reading ability on metacomprehension accuracy and study regulation was also examined.

Questions and Hypotheses

Experiment 1 compared the effects of two strategies, keyword generation and rereading text, on metacomprehension accuracy and study regulation. It also examined the effects of time of judgment and reading ability. The following questions were addressed:

1. Do middle school students tend to be accurate when making judgments about their own comprehension? If they are inaccurate, do they tend to be underconfident or overconfident?
2. Do strategy condition and time of judgment affect whether middle school students are underconfident, accurate, or overconfident in their comprehension judgments?
    a. Do strategy condition and time of judgment affect metacomprehension accuracy as operationalized via signed differences?
    b. Do strategy condition and time of judgment affect metacomprehension accuracy as operationalized via gamma correlations?
3. Do strategy condition and time of judgment affect study regulation among middle school students?
4. Does reading ability have an effect on whether middle school students tend to be underconfident, accurate, or overconfident in their comprehension judgments?
5. Do the effects of strategy condition and time of judgment on metacomprehension accuracy differ by ability group?

    a. Does ability have a moderating effect when metacomprehension accuracy is operationalized via signed differences?
    b. Does ability have a moderating effect when metacomprehension accuracy is operationalized via gamma correlations?
6. Do the effects of strategy condition and time of judgment on study regulation differ by ability group?

A review of the literature on metacomprehension accuracy led to the following hypotheses for Experiment 1:

1. **Middle school students will be overconfident in their judgments of comprehension.** De Bruin and her colleagues (2011) found that fourth, sixth, and seventh graders provided comprehension judgments that were higher than their actual test scores. Other studies of metacognitive monitoring have shown that children tend to overestimate their performance when making judgments of learning (e.g., Koriat & Shitzer-Reichert, 2002). Thus, it is expected that participants will be overconfident in their judgments.
2. **Strategy condition and time of judgment will interact in their effects on metacomprehension accuracy.** De Bruin and her colleagues (2011) found that generating keywords led to greater metacomprehension accuracy than not generating keywords. Maki and her colleagues (2005) found that post-test reflections were more accurate than pre-test predictions. It is expected that generating keywords will lead to better metacomprehension accuracy than rereading text, and that post-test reflections will be more accurate than pre-test predictions.
3. **Strategy condition and time of judgment will interact in their effects on study regulation.** Improving metacomprehension accuracy via keyword generation will also

lead to an improvement in study regulation, as de Bruin and her colleagues (2011) found. Increasing the accuracy of a learner's comprehension judgments will enable that learner to make more effective decisions about whether or not further study is required.

4. **Reading ability will have an effect on the accuracy of comprehension judgments.** According to the double-curse hypothesis, individuals with low ability are doubly cursed in that not only do they lack the skills to produce a correct answer, they also lack the ability to determine whether an answer is correct or incorrect (Dunning, Johnson, Ehrlinger, & Kruger, 2003). Thus, it is expected that participants with low scores on the comprehension test will also have poor metacomprehension accuracy.
5. **The effects of strategy condition and time of judgment on metacomprehension accuracy will differ by ability group.** De Bruin and her colleagues (2011) found that generating keywords improved metacomprehension accuracy for sixth and seventh graders but not for fourth graders. It is expected that, like the fourth graders, older struggling readers in this study may not benefit from keyword generation.
6. **The effects of strategy condition and time of judgment on study regulation will differ by ability group.** Because participants in the low ability group are expected to make inaccurate comprehension judgments, it is also expected that their subsequent study behavior will also be negatively impacted.

Method

Participants

A total of 109 seventh grade students (61 male and 48 female) from three demographically similar public schools in New York City participated in Experiment 1.

Design

A posttest-only control-group design was used. Participants were randomly assigned to one of two strategy conditions: (1) keyword or (2) reread. There were 55 participants in the keyword condition and 54 in the reread condition.

Procedure

Participants completed the following activities in order in a single 70-minute-long experimental session: (1) they read five passages, (2) they either generated keywords about the passages or reread the passages (the experimental manipulation), (3) they made pre-test predictions, (4) they answered test questions about the passages, (5) they made post-test reflections, and (6) they chose passages for restudy. The activities were administered by the author in a typical classroom setting with the classroom teacher present.

Participants in the keyword condition were asked to generate five keywords that captured the meaning of each passage. Participants were given the following instructions: "A keyword is a word that captures the meaning or the essence of a text. For example, on a text entitled *Titanic*, possible keywords might be ship, iceberg, lifeboat, sink, and tragedy." They did not receive any other training in keyword generation. Participants in the reread condition were presented with a list of the passage titles and were asked to check off the titles as they finished rereading.

Materials

The materials were the same as those used by de Bruin and her colleagues (2011) in their study. They were obtained by writing to Anique de Bruin and Keith Thiede. All materials were presented to participants in a single printed test booklet. In order to account for potential effects of passage order, five different randomized orderings of the reading passages were generated

using a Latin square design. For each ordering of the passages, two versions of the test booklets were created, one for the keyword condition and one for the reread condition, resulting in a total of ten different versions of the test booklets. The ordering was maintained throughout the entire experimental procedure, such that participants who read Passage 1 first also generated keywords, answered test questions, made comprehension judgments, and made restudy choices about Passage 1 first.

Reading passages. Five cause-effect expository text passages covering topics related to life science were used. Word counts for the passages ranged from 294 to 418, and Flesch-Kincaid reading grade levels for the passages ranged from 6.4 to 8.2. The passages had been developed so that important causal relationships among the propositions in the text were not explicitly stated; this was done to elicit deep comprehension on the level of the situation model (de Bruin et al., 2011; Wiley, Griffin, & Thiede, 2005). Appendix A contains the five passages, along with a table of word counts and readability statistics.

Test questions. A comprehension test with six multiple-choice questions accompanied each reading passage. There were thirty test questions in total (6 questions x 5 passages). One point was awarded for a correctly answered question; possible scores on each test ranged from 0 to 6. Appendix B contains the test questions.

Comprehension judgment prompts. Participants were asked to make two different comprehension judgments about each reading passage: (1) pre-test predictions, which were made after reading but before testing, and (2) post-test reflections, which were made after reading and after testing. Participants made five predictions and five reflections, for a total of ten comprehension judgments. Predictions were prompted with the query, "Please circle how many of the six test questions *you think you will answer correctly* about the text entitled, [*title*]."

Reflections were prompted with, "Please circle how many of the six test questions *you think you answered correctly* about the text entitled, [*title*]." In response to the prompts, participants were asked to circle a number between 0 and 6. The comprehension judgment prompts are presented in Appendix C.

Study regulation prompts. Participants were asked to indicate whether or not they would spend additional time restudying a passage if they were to be given additional test questions. Study choices were prompted with the query, "Suppose we were to give you another test about the passages you have read and answered questions about. Would you want to reread or restudy this passage?" Participants were not required to actually restudy the passages; they were simply asked to indicate whether or not they would restudy each passage by circling "yes" or "no." The study regulation prompts can be found in Appendix D.

Dependent Measures

There were two dependent measures: (1) metacomprehension accuracy and (2) study regulation. *Metacomprehension accuracy* was operationalized in two different ways: (1a) as the signed differences between comprehension judgments and test scores, and (1b) as the gamma correlations between comprehension judgments and test scores. *Study regulation* was operationalized as the gamma correlations between restudy choices and comprehension judgments.

Data Analysis

The data from Experiment 1 were analyzed using a mixed 2 x 2 ANOVA, with *strategy condition* treated as a between-subjects variable (2 levels: *keyword* and *reread*) and *time of judgment* treated as a within-subjects variable (2 levels: *prediction* and *reflection*). A subsequent

analysis included *ability* (3 levels: *high*, *medium*, and *low*) as a between-subjects moderating variable within a 2 x 2 x 3 mixed ANOVA design.

Results

An exploratory analysis on comprehension judgments and test scores revealed three extreme outliers in the data. One was identified by inspection of a boxplot for values that were more than 3 box-lengths from the edge of the box, and two were identified as having Studentized residuals greater than +3 or less than -3. These three outliers were excluded from the analyses that follow, resulting in a total sample size of N = 106 with n = 53 participants in each condition.

Table 1 presents the characteristics of the participants by strategy condition.

Table 1
Characteristics of the Participants by Strategy Condition

		Keyword (n = 53)	Reread (n = 53)
School	A	34	31
	B	13	14
	C	6	8
Age	11 years	4	1
	12 years	48	46
	13 years	1	6
Gender	Male	17	29
	Female	36	24

Chi-square tests were conducted to detect differences in participant characteristics across the keyword and reread conditions. There was a significant difference in assignment to strategy condition for gender, $\chi^2(1, N = 106) = 5.53$, $p = .019$, such that the proportion of females was

higher in the keyword condition than in the reread condition[2]. There were no significant differences for age, $\chi^2(2, N = 106) = 5.41$, $p = .067$, or for school, $\chi^2(2, N = 106) = 0.46$, $p = .79$.

Table 2 presents the means and standard deviations for test scores and comprehension judgments.

Table 2
Test Scores and Comprehension Judgments by Strategy Condition

Condition		Prediction		Test Score		Reflection	
		Mean	*SD*	Mean	*SD*	Mean	*SD*
Keyword	(n = 53)	0.77	0.13	0.58	0.18	0.68	0.15
Reread	(n = 53)	0.72	0.16	0.56	0.18	0.71	0.15

Levene's tests showed that the assumption of homogeneity of variances held for mean prediction ($p = .10$), mean test score ($p = .97$), and mean reflection ($p = .94$). Shapiro-Wilk's tests were used to determine whether the assumption of normality held. Within the keyword condition, the assumption of normality was satisfied for mean test score ($p = .20$), but not for mean prediction ($p = .009$) or for mean reflection ($p = .046$). Within the reread condition, the assumption of normality was satisfied for mean prediction and mean reflection ($p = .17$ and $p = .22$, respectively), but not for mean test score ($p = .024$). Since ANOVA has been found to be robust to violations of normality (e.g., Feir-Walsh & Toothaker, 1974), we chose to proceed with the analyses with caution.

An ANOVA on mean predictions revealed an effect of strategy condition, $F(1, 105) = 4.06$, $p = .046$, $\eta_p^2 = .038$, such that mean prediction was significantly higher in the keyword condition than in the reread condition ($M = 0.77$ vs. $M = 0.72$). There was no difference between

[2] Because the exploratory analysis revealed an effect for gender, an additional set of analyses was carried out to determine whether the effects of strategy condition and time of judgment on metacomprehension accuracy and study regulation differed by gender. Results indicated that gender did not moderate the effects of strategy condition and time of judgment on either metacomprehension accuracy or study regulation. The complete analysis is presented in Appendix E.

the keyword and reread conditions on mean test scores ($M = 0.58$ vs. $M = 0.56$), $F(1, 105) = 0.38$, $p = .54$, or on mean reflections ($M = 0.68$ vs. $M = 0.71$), $F(1, 105) = 1.25$, $p = .27$.

Metacomprehension Accuracy

Signed differences. Group means of the intra-individual signed differences between comprehension judgments (scored as a proportion out of 6 points) and test scores (scored as a proportion out of 6 points) across the five passages were computed. The means and standard errors for signed differences are presented in Figure 1. Zero values indicate accuracy, positive values indicate overconfidence, and negative values indicate underconfidence.

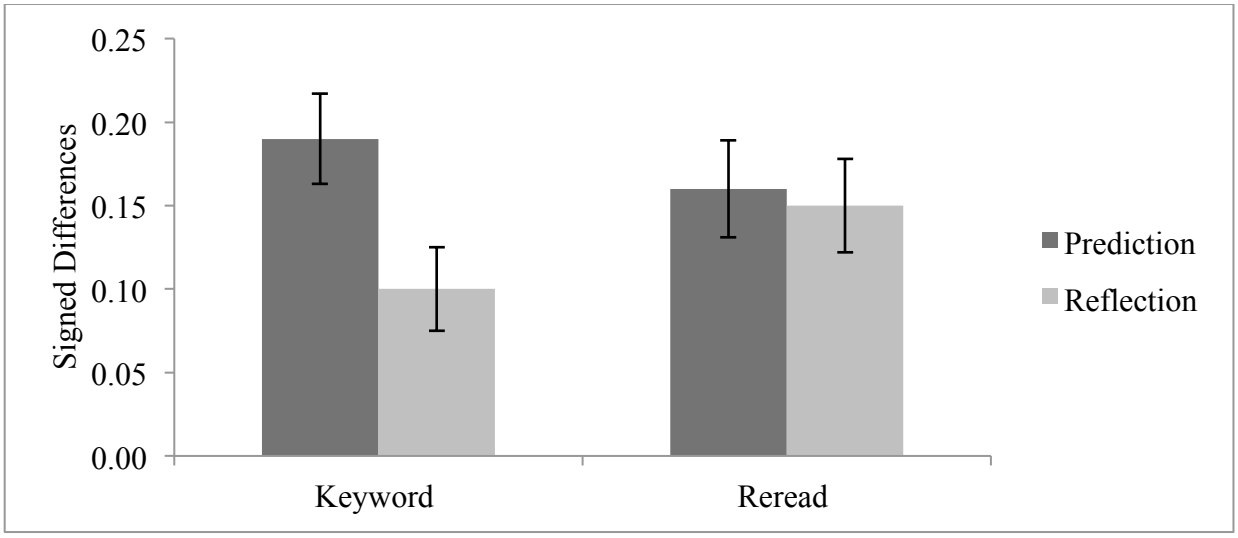


Figure 1. Signed differences between comprehension judgments and test scores.

Question 1. Do middle school students tend to be accurate when making judgments about their own comprehension? If they are inaccurate, do they tend to be underconfident or overconfident?

Hypothesis 1. Middle school students will be overconfident in their judgments of comprehension.

As shown in Figure 1, all of the signed differences between comprehension judgments and test scores were positive, indicating that participants were overconfident. One-sample t-tests were conducted to determine whether the signed differences were significantly different than zero. Within the keyword group, signed differences between test scores and predictions (M = 0.19, SEM = 0.028) were significantly different from zero, $t(52) = 6.89$, $p < .001$, as were those between test scores and reflections (M = 0.10, SEM = 0.025), $t(52) = 3.92$, $p < .001$. Similarly, within the reread group, signed differences between test scores and predictions (M = 0.16, SEM = 0.029) were significantly different from zero, $t(52) = 5.31$, $p < .001$, as were those between test scores and reflections (M = 0.15, SEM = 0.028), $t(52) = 5.49$, $p < .001$.

Hypothesis 1 was supported. The non-zero values for signed differences show that the comprehension judgments were inaccurate. That all of the signed differences were positive shows that students were overconfident in their judgments of comprehension.

Question 2a. Do strategy condition and time of judgment affect metacomprehension accuracy as operationalized via signed differences?

Hypothesis 2a. Strategy condition and time of judgment will interact in their effects on metacomprehension accuracy as operationalized via signed differences.

An ANOVA on signed differences showed a statistically significant interaction between time of judgment and strategy condition, $F(1, 104) = 14.32$, $p < .001$, $\eta_p^2 = .12$. Because the interaction term was significant, simple main effects analyses were conducted.

There was a statistically significant simple main effect of time of judgment for the keyword group, $F(1, 52) = 28.81$, $p < .001$, $\eta_p^2 = .36$. Within the keyword group, there was a bigger difference between predictions and test scores than between reflections at test scores (M = 0.19 vs. M = 0.10), $p < .001$. In other words, overconfidence was greater for predictions than for

reflections. There was no effect of time of judgment within the reread group (M = 0.16 vs. M = 0.15), $F(1, 52) = 0.024$, $p = .88$.

An analysis of simple main effects for strategy condition showed that the keyword and reread groups did not differ on overconfidence of predictions (M = 0.19 vs. M = 0.16), $F(1, 105) = 0.77$, $p = .38$, or of reflections (M = 0.10 vs. M = 0.15), $F(1, 105) = 2.08$, $p = .15$.

Hypothesis 2a was supported. Time of judgment and strategy condition interacted in their effects on metacomprehension accuracy as measured via signed differences. The accuracy of predictions differed from the accuracy of reflections within the keyword group but not within the reread group. Specifically, the keyword group became less overconfident from pre-test to post-test while the reread group remained consistently overconfident.

Gamma correlations. Group means of the intra-individual gamma correlations between comprehension judgments and test scores across the five passages were computed. Potential values for gamma range from -1 to +1. Stronger positive values for gamma indicate better correlation between judgments and test scores, zero values indicate a lack of a relationship, and negative values indicate an inverse relationship.

Gamma correlations could not be computed for 3 students in the keyword group and 7 students in the reread group because these students provided the same comprehension judgment for all five passages. As a result, the analysis was completed with n = 50 students in the keyword group and n = 46 in the reread group. Figure 2 presents the mean gamma correlations along with the standard errors for the keyword and reread groups.

As illustrated in Figure 2, all of the gamma correlations were small but positive, indicating that comprehension judgments were weakly related to test scores. Additionally, relative to the gammas, the standard errors were quite large.

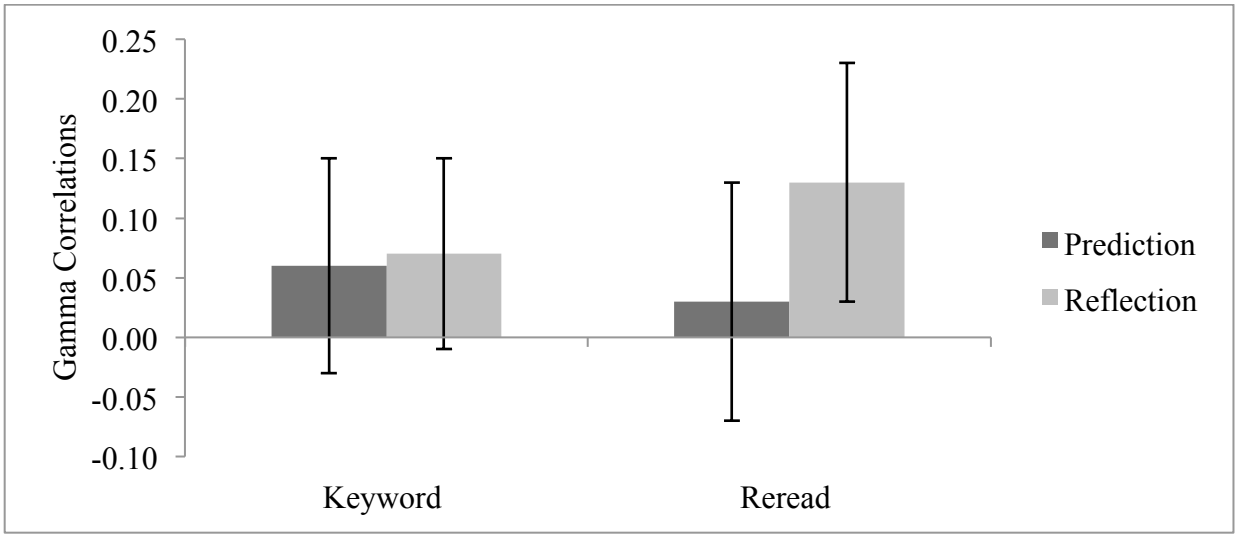


Figure 2. Gamma correlations between comprehension judgments and test scores.

One-sample t-tests showed that within the keyword group, gammas between test scores and predictions ($M = 0.061$, $SEM = 0.095$) were not significantly different from zero, $t(49) < 1$, $p = .53$, and neither were gammas between test scores and reflections ($M = 0.068$, $SEM = 0.083$), $t(51) < 1$, $p = .42$. Similarly, within the reread group, gammas between test scores and predictions ($M = 0.015$, $SEM = 0.10$) were not significantly different from zero, $t(47) < 1$, $p = .89$, and neither were gammas between test scores and reflections ($M = 0.14$, $SEM = 0.10$), $t(48) = 1.32$, $p = .19$. That none of the gamma correlations were significantly different than zero provides further support for the hypothesis that middle school students are inaccurate in their judgments about their comprehension.

Question 2b. Do strategy condition and time of judgment affect metacomprehension accuracy as operationalized via gamma correlations?

Hypothesis 2b. Strategy condition and time of judgment will interact in their effects on metacomprehension accuracy as operationalized via gamma correlations.

An ANOVA on gamma correlations indicated that the interaction between time of judgment and strategy condition was not statistically significant, $F(1, 94) = 0.30$, $p = .59$. The main effect of time of judgment was not statistically significant, $F(1, 94) = 0.47$, $p = .49$, nor was the main effect of strategy condition, $F(1, 94) = 0.017$, $p = .90$.

Hypothesis 2b was not supported. Time of judgment and strategy condition did not interact in their effects on metacomprehension accuracy as measured via gamma correlations. Neither time of judgment nor strategy condition had an effect on metacomprehension accuracy.

Study Regulation

Study regulation was operationalized as the mean of the intra-individual gamma correlations between comprehension judgments (which ranged from 0 to 6) and restudy choices (which were scored as either 0 or 1, with 1 indicating that a text was selected for restudy and 0 indicating that it was not). A stronger negative correlation indicates more effective study regulation in that participants chose to restudy texts perceived as less well learned (and did not choose to restudy texts perceived as well learned).

Gammas could not be computed for 9 students in the keyword group and 17 students in the reread group; these students made the same restudy choice for all five passages. Thus, the analysis was completed with n = 44 students in the keyword group and n = 36 in the reread group.

Figure 3 illustrates the study regulation gammas (means and standard errors) for the two groups.

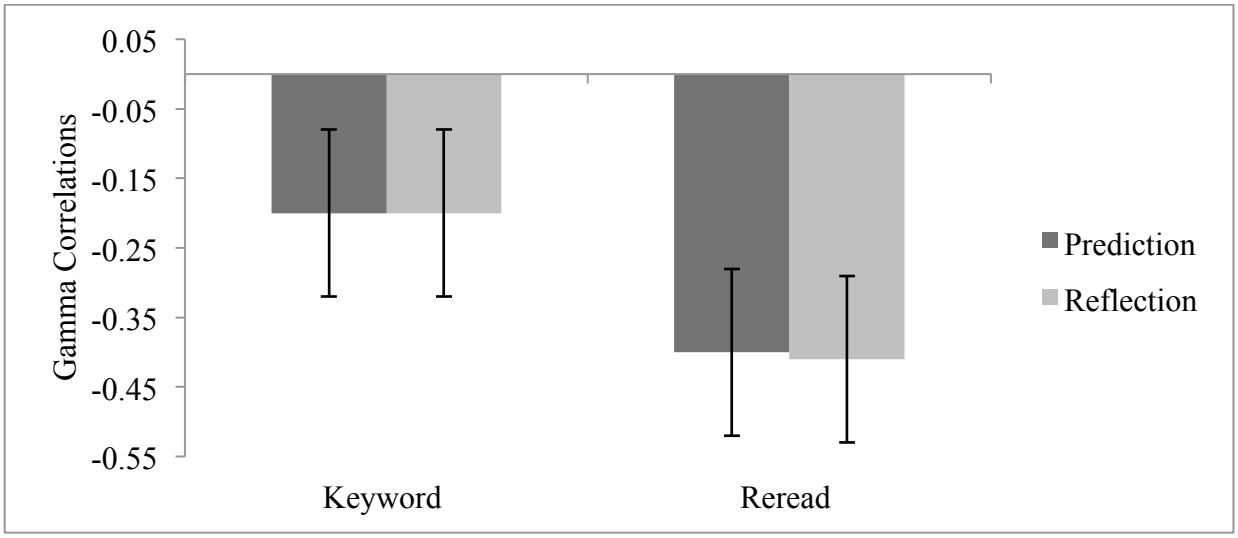


Figure 3. Study regulation gammas by strategy condition.

All of the obtained gammas were negative, indicating that participants based their study choices on their comprehension judgments. One-sample t-tests showed that within the keyword group, gammas between study choices and predictions (M = -0.21, SEM = 0.12) were not significantly different from zero, $t(43)$ = -1.68, p = .10, and neither were gammas between study choices and reflections (M = -0.20, SEM = 0.12), $t(45)$ = -1.60, p = .12. However, within the reread group, gammas between study choices and predictions (M = -0.40, SEM = 0.12) were significantly different from zero, $t(37)$ = -3.41, p = .002, as were the gammas between study choices and reflections (M = -0.41, SEM = 0.12), $t(38)$ = -3.51, p = .001.

Strongly negative gammas are typically interpreted as showing more effective study regulation in that participants chose to restudy texts perceived as less well learned and did not choose to restudy texts perceived as well learned. An exploration of the test scores and restudy choices suggests that the latter half of the previous statement is the more likely explanation. Study regulation gammas were larger for participants in the reread group, who were less likely to

choose to restudy and who obtained lower scores on the comprehension tests. Table 3 presents the percentage of students in each condition choosing to restudy, along with the mean test score.

Table 3
Restudy Choices and Test Scores by Strategy Condition

Condition		Percent Choosing to Restudy		Mean Test Score	
		Mean	*SD*	Mean	*SD*
Keyword	(n = 53)	46.4%	0.25	0.58	0.18
Reread	(n = 53)	43.0%	0.27	0.56	0.18

Question 3. Do strategy condition and time of judgment affect study regulation among middle school students?

Hypothesis 3. Strategy condition and time of judgment will interact in their effects on study regulation.

An ANOVA on study regulation gammas indicated that the interaction between time of judgment and strategy condition was not statistically significant, F(1, 78) = 0.010, $p = .92$. The main effect of time of judgment was not statistically significant, $F(1, 78) = 0.013$, $p = .91$. The main effect of strategy condition was not statistically significant, $F(1, 78) = 2.98$, $p = .088$.

Hypothesis 3 was not supported. Time of judgment and strategy condition did not interact in their effects on study regulation. Neither time of judgment nor strategy condition had an effect on study regulation.

Reading Ability

Classification. A tertile split on the sum of comprehension test scores across the five reading passages was used to classify participants into ability groups. There were three levels: *high* (n = 37), *medium* (n = 35), and *low* (n = 34). Table 4 presents the means and standard deviations for test scores and comprehension judgments by strategy condition and ability.

Table 4

Test Scores and Comprehension Judgments by Strategy Condition and Ability

Condition		Prediction		Test Score		Reflection	
		Mean	*SD*	Mean	*SD*	Mean	*SD*
Keyword/High	(n = 18)	0.79	0.11	0.78	0.07	0.74	0.10
Keyword/Medium	(n = 19)	0.81	0.10	0.57	0.06	0.71	0.14
Keyword/Low	(n = 16)	0.72	0.16	0.37	0.07	0.58	0.16
Reread/High	(n = 19)	0.72	0.17	0.75	0.06	0.74	0.17
Reread/Medium	(n = 16)	0.78	0.11	0.58	0.04	0.74	0.10
Reread/Low	(n = 18)	0.65	0.17	0.35	0.08	0.66	0.17

An ANOVA on mean predictions did not show a significant interaction between strategy condition and ability, $F(2, 105) = 0.22$, $p = .80$. There was a main effect of ability on mean predictions, $F(2, 105) = 5.22$, $p = .007$, $\eta_p^2 = .095$. Pairwise comparisons revealed that mean predictions differed between the low ability group and the medium ability group, $M = 0.68$ ($SEM = 0.02$) vs. $M = 0.79$ ($SEM = 0.02$), $p = .004$. There was no main effect for strategy condition on mean predictions, $F(1, 105) = 3.59$, $p = .061$.

An ANOVA on mean test scores indicated that the interaction between strategy condition and ability was not significant, $F(2, 105) = 1.12$, $p = .33$. There was a main effect of ability on mean test scores, $F(2, 105) = 350.68$, $p < .001$, $\eta_p^2 = .88$. Pairwise comparisons showed that mean test scores differed between the low ability group and the medium ability group, $M = 0.36$ ($SEM = 0.01$) vs. $M = 0.57$ ($SEM = 0.01$), $p < .001$; between the low ability and the high ability groups, $M = 0.36$ vs. $M = 0.76$ ($SEM = 0.01$), $p < .001$; and between the medium ability and high ability groups, $M = 0.57$ vs. $M = 0.76$, $p < .001$. There was no main effect for strategy condition on mean test scores, $F(1, 105) = 1.83$, $p = .18$.

An ANOVA on mean reflections did not show a significant interaction between strategy condition and ability, $F(2, 105) = 0.87$, $p = .42$. There was a main effect of ability on mean reflections, $F(2, 105) = 6.67$, $p = .002$, $\eta_p^2 = .12$. Pairwise comparisons revealed that mean

predictions differed between the low ability group and the medium ability group, $M = 0.62$ ($SEM = 0.03$) vs. $M = 0.72$ ($SEM = 0.02$), $p = .015$, and between the low ability group and the high ability group $M = 0.62$ vs. $M = 0.74$ ($SEM = 0.02$), $p = .004$. There was no main effect for strategy condition on mean predictions, $F(1, 105) = 1.80$, $p = .18$.

Metacomprehension accuracy (signed differences). Figure 4 presents the signed differences between comprehension judgments and test scores by ability group (means and standard errors). Zero values indicate accuracy, positive values indicate overconfidence, and negative values indicate underconfidence.

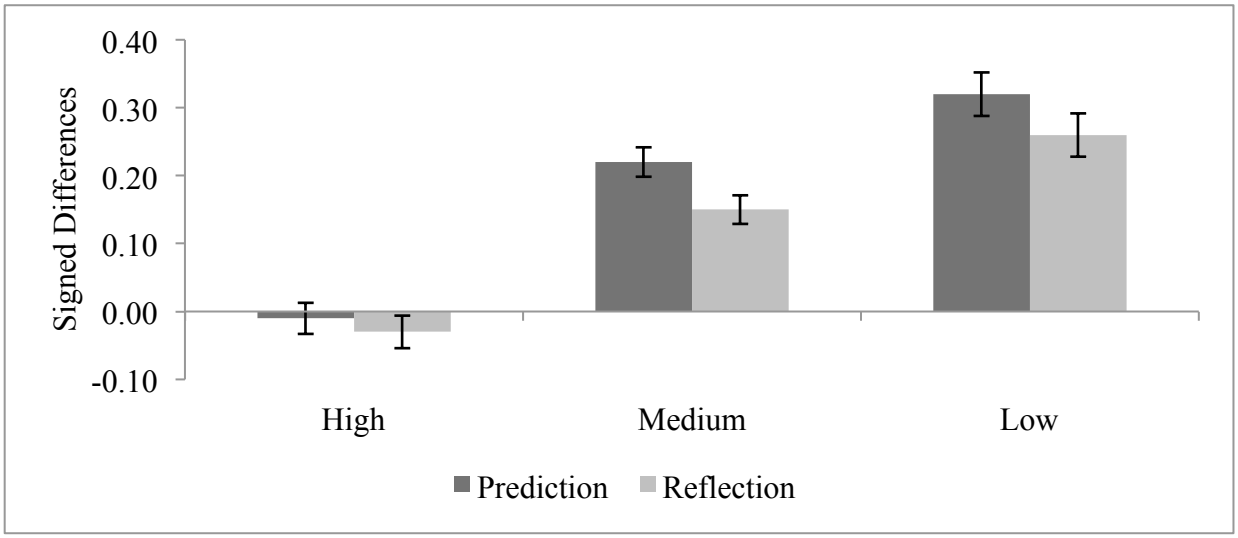


Figure 4. Signed differences by ability group.

Question 4. Does reading ability have an effect on whether middle school students tend to be underconfident, accurate, or overconfident in their comprehension judgments?

Hypothesis 4. Reading ability will have an effect on the accuracy of comprehension judgments.

An ANOVA on signed differences showed an effect of reading ability, $F(2, 103) = 46.15$, $p < .001$, $\eta_p^2 = .47$. Post hoc tests for reading ability showed that the low ability group was more

overconfident than the medium ability group, $M = 0.29$ ($SEM = 0.024$) vs. $M = 0.19$ ($SEM = 0.024$), $p = .007$, and the high ability group, $M = 0.29$ vs. $M = -0.018$ ($SEM = 0.023$), $p < .001$. The medium ability group was more overconfident than the high ability group, $M = 0.29$ vs. $M = -0.018$, $p < .001$.

Within the high ability group, signed differences between test scores and predictions ($M = -0.011$, $SEM = 0.023$) were not significantly different from zero, $t(36) < 1$, $p = .65$, and neither were the signed differences between test scores and reflections ($M = -0.026$, $SEM = 0.024$), $t(36) = -1.11$, $p = .27$. These results indicate that the high ability group was accurate in both predictions and reflections.

There was an effect of time of judgment, $F(1, 103) = 14.83$, $p < .001$, $\eta_p^2 = .13$. Predictions were more overconfident than reflections, $M = 0.18$ ($SEM = 0.015$) vs. $M = 0.13$ ($SEM = 0.015$).

The interaction between reading ability and time of judgment was not significant, $F(2, 103) = 1.85$, $p = .16$.

Hypothesis 4 was supported. Reading ability had an effect on the accuracy of comprehension judgments. The low and medium ability groups were overconfident in their comprehension judgments, while the high ability group was accurate. Additionally, overconfidence was higher for predictions than for reflections.

Figure 5 presents the signed differences between comprehension judgments and test scores by reading ability and strategy condition (means and standard errors).

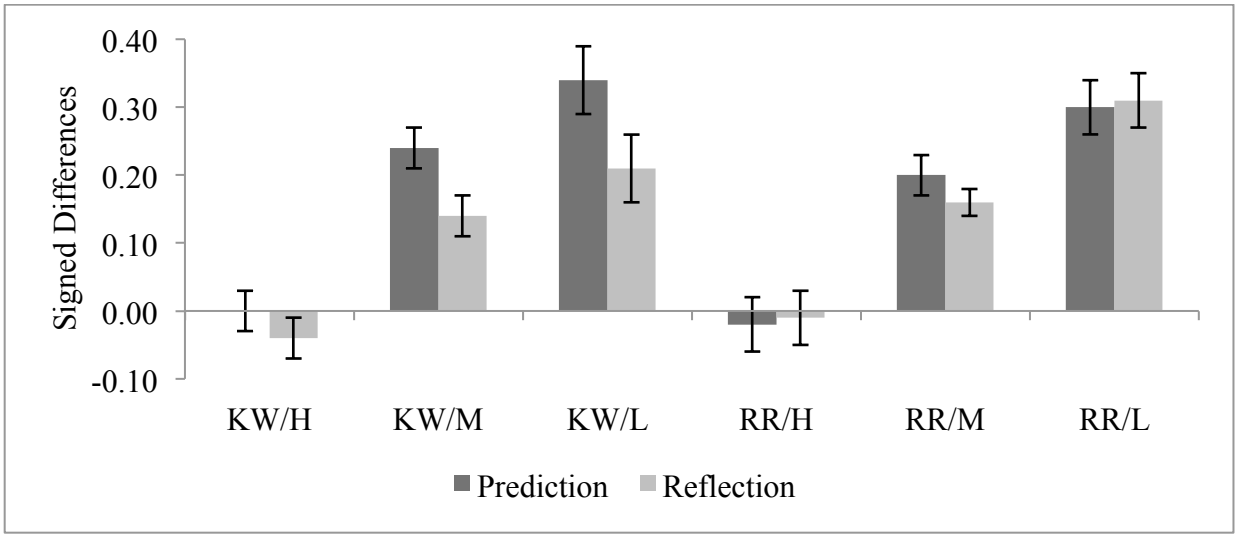


Figure 5. Signed differences by strategy condition and ability group.

Question 5a. Does reading ability have a moderating effect when metacomprehension accuracy is operationalized via signed differences?

Hypothesis 5a. The effects of strategy condition and time of judgment on metacomprehension accuracy as operationalized via signed differences will differ by ability group.

An ANOVA on signed differences showed that the three-way interaction was not significant, $F(2, 100) = 1.69$, $p = .19$. The two-way interaction between time of judgment and strategy condition was statistically significant, $F(1, 100) = 14.66$, $p < .001$, $\eta_p^2 = .13$. The two-way interaction between time of judgment and ability group was not significant, $F(2, 100) = 2.04$, $p = .14$. The two-way interaction between strategy condition and ability group was not significant, $F(2, 100) = 0.19$, $p = .83$. There was a significant effect of ability group, $F(2, 100) = 44.69$, $p < .001$, $\eta_p^2 = .47$.

Hypothesis 5a was not supported. Time of judgment, strategy condition, and ability group did not interact in their effects on signed differences. Reading ability did not moderate the

interaction between time of judgment and strategy condition. However, the two-way interaction between time of judgment and strategy condition remained significant, and there was a significant effect of ability group.

Metacomprehension accuracy (gamma correlations). Figure 6 presents the gamma correlations between comprehension judgments and test scores by strategy condition and ability group (means and standard errors). Stronger positive values for gamma indicate better correlation between judgments and test scores, zero values indicate a lack of a relationship, and negative values indicate an inverse relationship.

Question 5b. Does reading ability have a moderating effect when metacomprehension accuracy is operationalized via gamma correlations?

Hypothesis 5b. The effects of strategy condition and time of judgment on metacomprehension accuracy as operationalized via gamma correlations will differ by ability group.

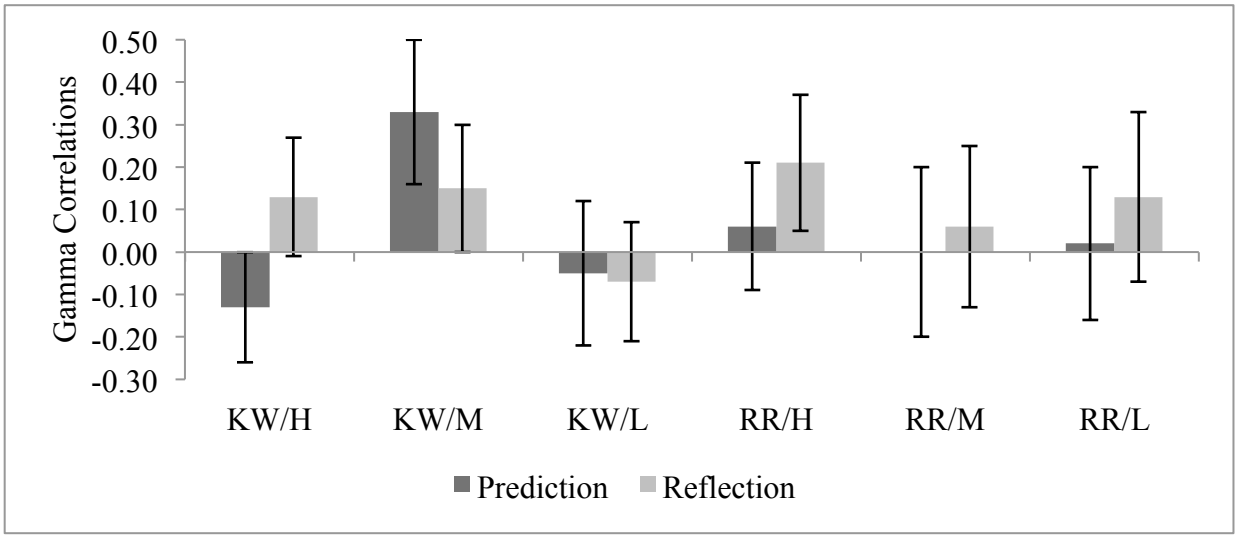


Figure 6. Gamma correlations by strategy condition and ability group.

An ANOVA on gamma correlations showed that the three-way interaction was not significant, $F(2, 90) = 0.38$, $p = .68$. None of the two-way interactions were significant ($p > .05$ in all cases). None of the main effects were significant ($p > .05$ in all cases).

Hypothesis 5b was not supported. Time of judgment, strategy condition, and ability group did not interact in their effects on gamma correlations.

Study regulation. Figure 7 presents the gamma correlations between restudy choices and comprehension judgments by strategy condition and ability group. A stronger negative gamma correlation indicates more effective study regulation in that participants chose to restudy texts perceived as less well learned (and did not choose to restudy texts perceived as well learned).

Question 6. Do the effects of strategy condition and time of judgment on study regulation differ by ability group?

Hypothesis 6. The effects of strategy condition and time of judgment on study regulation will differ by ability group.

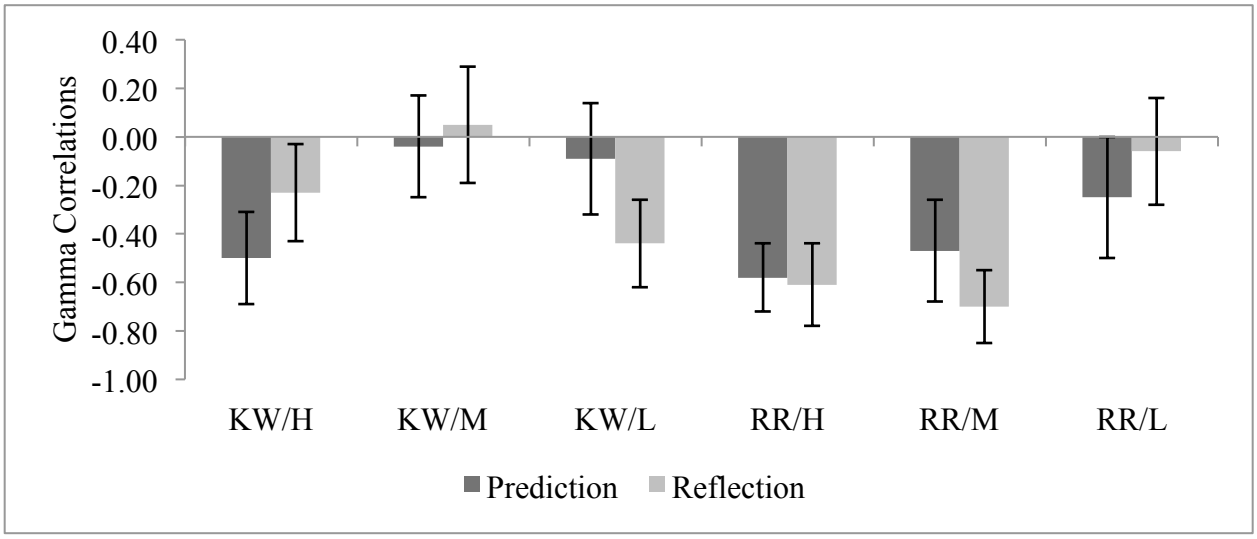


Figure 7. Study regulation gammas by strategy condition and ability group.

An ANOVA on study regulation gammas showed that the three-way interaction was not significant, $F(2, 74) = 2.70$, $p = .074$. None of the two-way interactions were significant ($p > .05$ in all cases). None of the main effects were significant ($p > .05$ in all cases).

Hypothesis 6 was not supported. Time of judgment, strategy condition, and ability group did not interact in their effects on study regulation.

Discussion

Experiment 1 showed that middle school students tend to be inaccurate in their judgments of comprehension. In general, students in the study were overconfident about their comprehension. Time of judgment and reading ability were found to have effects on the degree of overconfidence. Predictions were more overconfident than reflections, and lower reading ability was linked to higher overconfidence. Participants who were classified as having high reading ability were actually quite accurate in judging their own comprehension.

Strategy condition and time of judgment interacted in their effects on metacomprehension accuracy as operationalized via signed differences. For the keyword group, there was a bigger difference between predictions and test scores than between reflections and test scores, indicating that predictions were more accurate than were reflections. For the reread group, there was no difference due to time of judgment; participants were equally overconfident in their reflections as in their predictions.

However, when metacomprehension accuracy was operationalized using gamma correlations, time of judgment and strategy condition did not interact, and neither time of judgment nor strategy condition had an effect on its own. There was no effect of reading ability.

A similar pattern of results was obtained for study regulation, which was also operationalized using gamma correlations.

The different patterns of results obtained using the two measures of metacomprehension accuracy are surprising. Additional research is needed to determine why this may be. One possibility is that signed differences and gamma correlations measure different aspects of metacomprehension. A second possibility is that gamma correlations are an inappropriate measure of metacomprehension accuracy in the context of the current experimental paradigm. The latter explanation is supported by the number of cases in which gamma could not be computed (i.e., the instances in which participants made the same rating or the same restudy choice for all passages) and also by the large standard errors for the gammas obtained in this study.

Surprisingly, although keyword generation led to an increase in metacomprehension accuracy from pretest to posttest, it did not appear to have an effect on actual comprehension. Scores on the comprehension test did not differ across the keyword and reread groups. These results suggest that while generating keywords may have helped participants recognize that their comprehension was poor, it did not appear to be effective in actually improving their comprehension.

Summary

The current research compared the effects of generating keywords and rereading text on metacomprehension accuracy and study regulation. The following hypotheses were tested (supported hypotheses appear in bolded font):

1. **Middle school students will be overconfident in their judgments of comprehension.**
2. **Strategy condition and time of judgment will interact in their effects on metacomprehension accuracy.**[3]
3. Strategy condition and time of judgment will interact in their effects on study regulation.
4. **Reading ability will have an effect on the accuracy of comprehension judgments.**
5. The effects of strategy condition and time of judgment on metacomprehension accuracy will differ by ability group.
6. The effects of strategy condition and time of judgment on study regulation will differ by ability group.

[3] Hypothesis 2 was supported when metacomprehension accuracy was operationalized via signed differences, but it was not supported when gamma correlations were used.

CHAPTER 4

IMPROVING THE UTILITY OF THE STUDY STRATEGIES (EXPERIMENT 2)

Experiment 1 showed that middle school students tend to be overconfident in their judgments of their own comprehension, and that the degree of overconfidence depends on what type of study strategy students used and also on when they made the comprehension judgment. Students who generated keywords made post-test reflections that were significantly less overconfident than were their pre-test predictions, while students who reread text remained as overconfident in their reflections as in their predictions. Surprisingly, although the two groups differed on metacomprehension, there was no difference in actual comprehension: mean test score was 58 percent for the keyword group and 56 percent for the reread group. That test scores were so low in both groups suggests that generating keywords and rereading text may be equally ineffective as comprehension strategies. Therefore, a goal of Experiment 2 was to embed the comparison of generating keywords and rereading text within a context involving an effective comprehension strategy in order to determine the effects on metacomprehension.

Dunlosky, Rawson, Marsh, Nathan, and Willingham (2013) recently reviewed and evaluated ten study strategies that have been the focus of extensive research by cognitive and educational psychologists. Each strategy was rated in terms of the empirical evidence supporting its effectiveness and of the practical considerations surrounding its implementation. A strategy was classified as having high utility if it was found to have robust and generalizable effects on learning and if it was determined to be easy to use and also likely to be used by students. Keywords were included among the reviewed strategies, but in the sense of associating keywords with mental imagery in order to learn vocabulary words or verbal materials and not as a

comprehension strategy. Rereading– restudying text again after an initial reading was found to be one of the strategies most frequently used by students (Dunlosky et al., 2013; Karpicke, Butler, & Roediger, 2009). However, based on their review of the empirical literature, Dunlosky and his colleagues (2013) classified both keywords and rereading as being of low utility.

In contrast, one of the study strategies rated as having high utility is practice testing, which Dunlosky and his colleagues (2013) define as any form of low- or no-stakes testing that students can complete on their own over to-be-learned material. Examples of practice testing include using flashcards to practice recall of target material, completing practice problems or questions after reading, and completing practice quizzes or practice tests accompanying textbooks (Dunlosky et al., 2013). In one of the reviewed studies, Butler (2010) had students read expository texts and then either reread the texts or take practice tests; performance on new test questions given one week later was higher following practice testing than following rereading. In Experiment 2, practice test questions were included in the experimental procedure in an attempt to improve comprehension as well as metacomprehension. All participants answered two practice test questions about each passage before either generating keywords or rereading passages.

In addition to examining the effects of study strategy, Experiment 2 also investigated the effects of a second factor: peer collaboration. Language arts classrooms are oftentimes highly social in nature, as teachers integrate collaboration and discussion with instruction in an attempt to stimulate and support deep comprehension and critical thinking (Cazden & Beck, 2003). Additionally, research on anchoring and estimation shows that college students can use information about peer performance as a basis for their judgments about themselves (Zhao &

Linderholm, 2011). Therefore, a second goal of Experiment 2 was to determine whether peer collaboration has an effect on either comprehension or metacomprehension.

Questions and Hypotheses

The purpose of Experiment 2 was to examine the effects of keyword generation on metacomprehension accuracy and study regulation in a context that was more similar to an authentic school situation. The comparison of keyword generation to rereading text was situated within a new study context incorporating practice testing. In addition, a new variable, peer collaboration, was introduced. The following questions were addressed:

7. Do middle school students tend to be accurate when making judgments about their own comprehension? If they are inaccurate, do they tend to be underconfident or overconfident?
8. Do strategy condition, collaboration condition, and time of judgment affect whether middle school students are underconfident, accurate, or overconfident in their comprehension judgments?
    a. Do strategy condition, collaboration condition, and time of judgment affect metacomprehension accuracy as operationalized via signed differences?
    b. Do strategy condition, collaboration condition, and time of judgment affect metacomprehension accuracy as operationalized via gamma correlations?
9. Do strategy condition, collaboration condition, and time of judgment affect study regulation among middle school students?

Based on a review of the literature on metacomprehension accuracy and on the results of Experiment 1, the following hypotheses were expected for Experiment 2:

7. **Middle school students will be overconfident in their judgments of comprehension.** As was found in Experiment 1, it is expected that participants will be overconfident in their judgments.
8. **Strategy condition, collaboration condition, and time of judgment will interact in their effects on metacomprehension accuracy.** Zhao and Linderholm (2011) found that college students can use peer performance as a basis for their judgments about themselves. Experiment 1 showed that generating keywords led to an increase in metacomprehension accuracy from pre-test to post-test. Therefore, it is expected that the combination of peer collaboration and keyword generation will lead to more accurate post-test reflections.
9. **Strategy condition, collaboration condition, and time of judgment will interact in their effects on study regulation.** Increasing the accuracy of a learner's comprehension judgments will enable that learner to make more effective decisions about whether or not further study is required.

Method

Participants

A total of 190 seventh grade students (90 male and 100 female) from four demographically similar public schools in New York City participated in Experiment 2. (One of the four schools had previously participated in Experiment 1, but no student who participated in Experiment 1 also participated in Experiment 2.)

Design

A 2 x 2 factorial design was used. Participants were randomly assigned to the following four conditions: individual/keyword (n = 43), collaborate/keyword (n = 54), individual/reread (n = 43), and collaborate/reread (n = 50).

Procedure

Participants completed the following steps in order in a single 80-minute-long experimental session: (1) they read five passages, (2) they answered practice test questions about the passages, (3) they either generated keywords about the passages or reread the passages, (4) they made pre-test predictions, (5) they answered test questions about the passages, (6) they made post-test reflections, and (7) they chose passages for restudy.

Participants in the two individual conditions completed all of the above steps independently. Participants in the collaborative conditions worked with a partner on the second and third steps. Each pair worked together to discuss the practice test questions, and pairs in the keyword condition also worked together to generate a list of five keywords about the gist of each passage. They completed the other five steps on their own. The steps were administered by the author in a typical classroom setting with the classroom teacher present.

Materials

The reading passages, test questions, comprehension judgment prompts, and study regulation prompts used in Experiment 1 were used again in Experiment 2.

Practice test questions. For each passage, two free-response practice test questions were created that assessed comprehension of the causal relationships in the text. The practice test questions were designed to be congruent with the cause-effect structure of the passages, and were modeled after the ones used by Williams and her colleagues in their studies on cause-effect

text structure (e.g., Williams et al., 2009; Williams et al., 2014). For the passage entitled, "The Fight or Flight Response," the practice test questions were: (1) *Why is the endocrine system called the "fight or flight response"?* and (2) *What happens when adrenaline is released into the blood?* Appendix F presents the practice test questions.

Dependent Measures

There were two dependent measures: (1) metacomprehension accuracy and (2) study regulation. *Metacomprehension accuracy* was operationalized in two different ways: (1a) as the signed differences between comprehension judgments and test scores, and (1b) as the gamma correlations between comprehension judgments and test scores. *Study regulation* was operationalized as the gamma correlations between restudy choices and comprehension judgments.

Data Analysis

The data from Experiment 2 were analyzed using a mixed 2 x 2 x 2 ANOVA. *Strategy condition* was treated as a between-subjects variable (2 levels: *keyword* and *reread*). *Collaboration condition* was treated as a between-subjects variable (2 levels: *individual* and *collaborate*). *Time of judgment* was treated as a within-subjects variable (2 levels: *prediction* and *reflection*).

Results

An exploratory analysis on comprehension judgments and test scores revealed one extreme outlier in the data. This was assessed by inspection of a boxplot for values that were more than 3 box-lengths from the edge of the box. The outlier was excluded from the analyses

that follow, resulting in a total sample size of N = 189. (The outlier was in the keyword/collaborate condition.) Table 5 presents the characteristics of the participants by strategy condition and collaboration condition.

Chi-square tests were conducted to detect differences in participant characteristics across the experimental conditions. Following random assignment, there were no significant differences in assignment to the four conditions for age, $\chi^2(6, N = 189) = 2.09$, $p = .91$, gender, $\chi^2(3, N = 189) = 1.15$, $p = .77$, or for school, $\chi^2(9, N = 189) = 13.57$, $p = .14$.

Table 5
Characteristics of the Participants by Strategy Condition and Collaboration Condition

		Keyword/ Individual (n = 43)	Keyword/ Collaborate (n = 53)	Reread/ Individual (n = 43)	Reread/ Collaborate (n = 50)
School	A[4]	16	8	23	14
	D	5	7	8	10
	E	19	22	22	22
	F	3	6	0	4
Age	11 years	9	11	10	7
	12 years	30	38	31	39
	13 years	4	4	2	4
Gender	Male	21	22	22	25
	Female	22	31	21	25

Levene's tests revealed that the assumption of homogeneity of variances held for mean prediction ($p = .58$), mean test score ($p = .28$), and mean reflection ($p = .85$). Shapiro-Wilk's tests were used to determine whether the assumption of normality held. For mean predictions, the assumption of normality held for the keyword/individual group ($p = .084$), the reread/individual group ($p = .44$), and the reread/collaborate group ($p = .34$), but was violated for the keyword/collaborate group ($p = .004$). For mean test scores, the assumption of normality held for

[4] School A previously participated in Experiment 1 during the previous school year. No student who participated in Experiment 1 also participated in Experiment 2.

the keyword/individual group (p = .12), the reread/individual group (p = .59), and the keyword/collaborate group (p = .38), but was violated for the reread/collaborate group (p = .043). For mean reflections, the assumption of normality held for the keyword/collaborate group (p = .13), but was violated for the keyword/individual group (p = .028), the reread/individual group (p = .041), and the reread/collaborate group (p = .014). Since ANOVA has been found to be robust to violations of normality (e.g., Feir-Walsh & Toothaker, 1974), we chose to proceed with the analyses with caution.

Table 6 presents the means and standard deviations for test scores and comprehension judgments.

Table 6
Test Scores and Comprehension Judgments by Strategy and Collaboration Condition

Condition		Prediction		Test Score		Reflection	
		Mean	*SD*	Mean	*SD*	Mean	*SD*
KW/IND	(n = 43)	0.69	0.16	0.52	0.19	0.63	0.17
KW/COL	(n = 53)	0.72	0.14	0.55	0.16	0.68	0.15
RR/IND	(n = 43)	0.71	0.15	0.59	0.17	0.67	0.16
RR/COL	(n = 50)	0.75	0.14	0.55	0.17	0.71	0.16

An ANOVA on mean predictions showed no interaction between strategy condition and collaboration condition, $F(1, 188) < 1$, p = .95, and no main effects of strategy, $F(1, 188) = 1.28$, p = .26, or collaboration, $F(1, 188) = 2.48$, p = .12. An ANOVA on mean test scores showed no interaction between strategy condition and collaboration condition, $F(1, 188) = 2.22$, p = .14, and no main effects of strategy, $F(1, 188) = 1.97$, p = .16, or collaboration, $F(1, 188) < 1$, p = .88. An ANOVA on mean reflections showed no interaction between strategy condition and collaboration condition, $F(1, 188) < 1$, p = .82, and no main effects of strategy, $F(1, 188) = 2.44$, p = .12, or collaboration, $F(1, 188) = 3.39$, p = .07.

Metacomprehension Accuracy

Signed differences. Group means of the intra-individual signed differences between comprehension judgments (scored as a proportion out of 6 possible points) and test scores (scored as a proportion out of 6 possible points) across the five passages were computed. The signed differences are presented in Figure 8. Zero values indicate accuracy, positive values indicate overconfidence, and negative values indicate underconfidence.

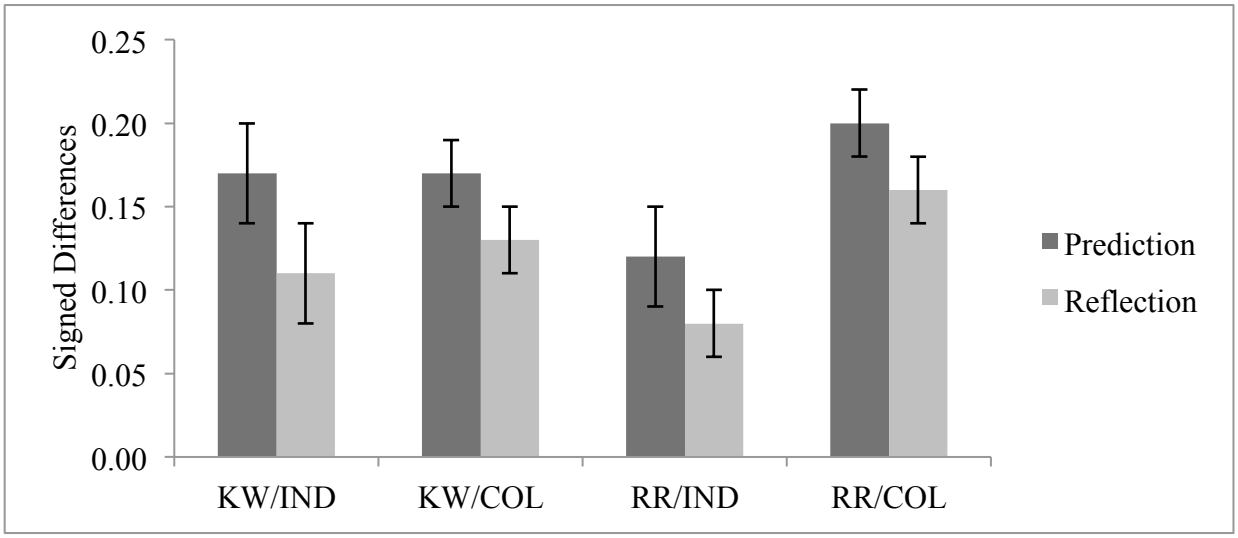


Figure 8. Signed differences between comprehension judgments and test scores.

Question 7. Do middle school students tend to be accurate when making judgments about their own comprehension? If they are inaccurate, do they tend to be underconfident or overconfident?

Hypothesis 7. Middle school students will be overconfident in their judgments of comprehension.

As shown in Figure 8, all of the signed differences were positive, indicating that participants were overconfident in their comprehension judgments.

Hypothesis 7 was supported. All of the signed differences were positive, showing that students were overconfident in their judgments of comprehension.

Question 8a. Do strategy condition, collaboration condition, and time of judgment affect metacomprehension accuracy as operationalized via signed differences?

Hypothesis 8a. Strategy condition, collaboration condition, and time of judgment will interact in their effects on metacomprehension accuracy as operationalized via signed differences.

An ANOVA on signed differences showed that the three-way interaction was not significant, $F(1, 185) < 1$, $p = .71$. The two-way interaction between time of judgment and strategy condition was not significant, $F(1, 185) < 1$, $p = .47$. The two-way interaction between time of judgment and collaboration condition was not significant, $F(1, 185) < 1$, $p = .58$. The two-way interaction between collaboration condition and strategy condition was not significant, $F(1, 185) = 2.41$, $p = .12$. The main effect of strategy condition was not significant, $F(1, 185) < 1$, $p = .85$.

There was a main effect for time of judgment, $F(1, 185) = 26.89$, $p < .001$, $\eta_p^2 = .127$. Post hoc tests showed that there was a bigger difference between predictions and test scores than between reflections at test scores, $M = 0.17$ ($SEM = 0.013$) vs. $M = 0.12$ ($SEM = 0.012$); in other words, overconfidence was greater for predictions than for reflections.

The main effect of collaboration condition approached significance, $F(1, 185) = 3.51$, $p = .062$, $\eta_p^2 = .019$. Therefore, post hoc tests were conducted. Results showed that overconfidence was higher in the collaborative group than in the individual group, $M = 0.16$ ($SEM = 0.015$) vs. $M = 0.12$ ($SEM = 0.017$). Within the reread group, overconfidence was higher in the collaborative group than in the individual group, $M = 0.18$ ($SEM = 0.021$) vs. $M = 0.10$ ($SEM =$

0.023), $F(1, 91) = 6.29$, $p = .014$, $\eta_p^2 = .065$. Within the keyword group, the collaborative and individual groups did not differ in overconfidence, $F(1, 94) < 1$, $p = .83$.

Hypothesis 8a was not supported. Strategy condition, collaboration condition, and time of judgment did not interact in their effects on metacomprehension accuracy. None of the two-way interactions were statistically significant. There was a main effect of time of judgment, such that predictions were more overconfident– or less accurate– than reflections. Overconfidence was higher among participants who collaborated than among those who worked individually, and the effect of collaboration on overconfidence was especially pronounced within the reread group.

Gamma correlations. Group means of the intra-individual gamma correlations between comprehension judgments and test scores across the five passages were computed. Figure 9 presents the gamma correlations.

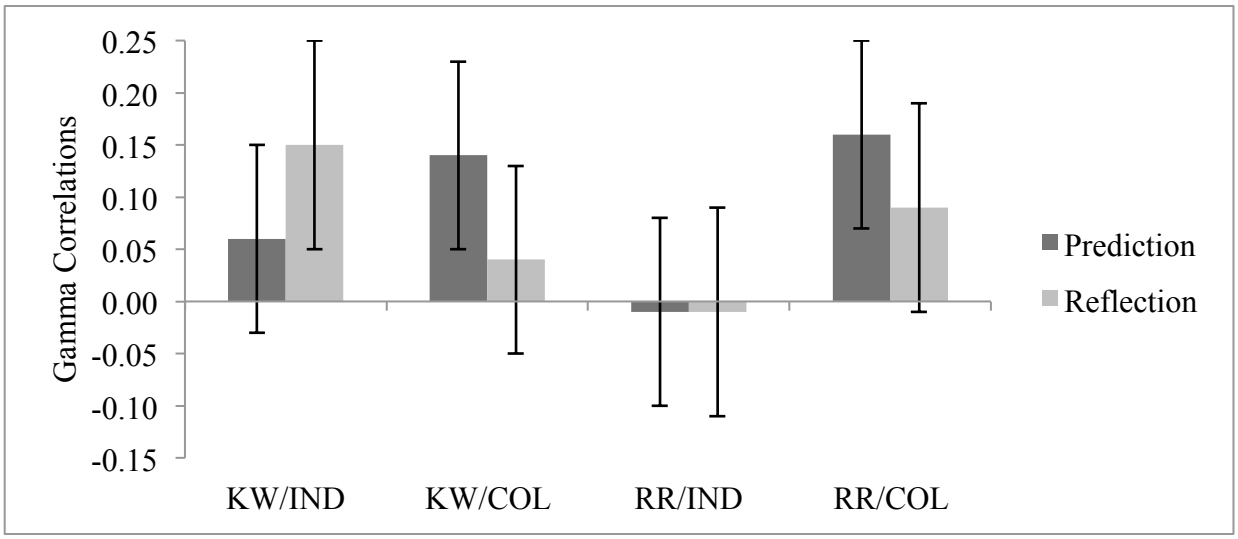


Figure 9. Gamma correlations between comprehension judgments and test scores.

Gammas could not be computed in instances where participants made the same judgment for all five passages. The analysis was completed with n = 40 students (out of 43) in the individual/keyword group, n = 48 (out of 53) in the collaborate/keyword group, n = 40 (out of

43) in the individual/reread group, and n = 43 (out of 43) in the collaborate/reread group. The gamma correlations were small, indicating that comprehension judgments were weakly related to test scores. Additionally, relative to the gammas, the standard errors were quite large.

Question 8b: Do strategy condition, collaboration condition, and time of judgment interact when metacomprehension accuracy is operationalized via gamma correlations?

Hypothesis 8b: Strategy condition, collaboration condition, and time of judgment will interact in their effects on metacomprehension accuracy as operationalized via gamma correlations.

An ANOVA on gamma correlations showed that the three-way interaction was not significant, $F(1, 161) < 1$, $p = .45$. None of the two-way interactions were significant ($p > .05$ in all cases). None of the main effects were significant ($p > .05$ in all cases).

Hypothesis 8b was not supported. Strategy condition, collaboration condition, and time of judgment did not interact in their effects on metacomprehension accuracy. None of the two-way interactions were significant, and none of the main effects were significant.

Study Regulation

Study regulation was operationalized as the mean of the intra-individual gamma correlations between comprehension judgments (which ranged from 0 to 6) and restudy choices (which were scored as either 0 or 1, with 1 indicating that a text was selected for restudy and 0 indicating that it was not). A stronger negative correlation indicates more effective study regulation in that participants chose to restudy texts perceived as less well learned (and did not choose to restudy texts perceived as well learned).

Gammas could not be computed in instances where participants made the same restudy choice for all five passages. The analysis was completed with n = 33 students (out of 43) in the individual/keyword group, n = 40 (out of 53) in the collaborate/keyword group, n = 36 (out of 43) in the individual/reread group, and n = 36 (out of 50) in the collaborate/reread group.. Figure 10 illustrates the study regulation gammas. In general, the gammas were small and negative, indicating that participants based their study choices on their comprehension judgments, although the relationship was weak. Standard errors were large.

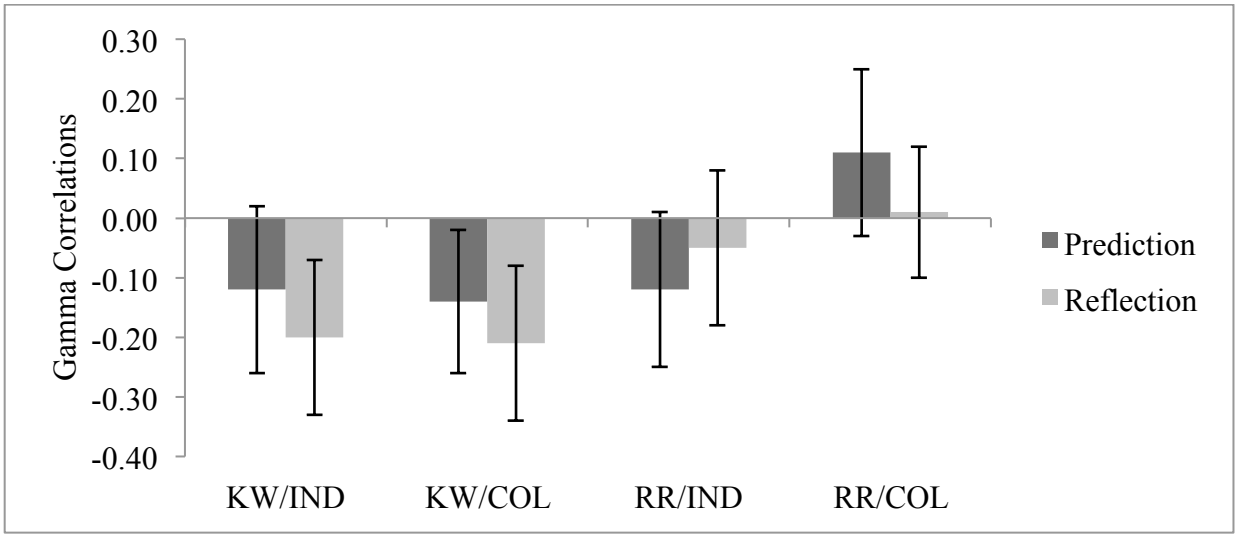


Figure 10. Study regulation gammas by strategy and collaboration condition.

Question 9: Do strategy condition, collaboration condition, and time of judgment interact in their effects on study regulation?

Hypothesis 9: Strategy condition, collaboration condition, and time of judgment will interact in their effects on study regulation.

An ANOVA on study regulation showed that the three-way interaction was not significant, $F(1, 136) < 1$, $p = .68$. None of the two-way interactions were significant ($p > .05$ in all cases). None of the main effects were significant ($p > .05$ in all cases).

Hypothesis 9 was not supported. Time of judgment, strategy condition, and collaboration condition did not interact in their effects on study regulation. None of the two-way interactions were significant, and none of the main effects were significant.

Discussion

As in Experiment 1, Experiment 2 showed that middle school students tend to be overconfident when making judgments about their own comprehension. Time of judgment was again found to have an effect on the degree of overconfidence, such that predictions were more overconfident than predictions.

Peer collaboration also had an effect on metacomprehension accuracy; the effect, however, was not quite statistically significant. Participants who collaborated with a peer were more overconfident than participants who worked alone. This finding is surprising, as it was hypothesized that collaborating with a peer would provide an anchor for the judgments an individual made about him or herself and would therefore lead to increased metacomprehension accuracy. Quite the contrary, peer collaboration appeared to inflate participants' confidence about how well they had comprehended what they read.

Interestingly, the effect of collaboration on overconfidence was especially pronounced within the reread group. Within the keyword group, participants who collaborated and participants who worked alone did not differ in overconfidence. Within the reread group, however, participants who collaborated were significantly more overconfident than those who worked alone. One possible explanation for this finding is that peer collaboration may be a source of distraction or confusion that prevents students from achieving the level of focused attention that is necessary to make an accurate metacognitive judgment.

The addition of practice test questions appeared to lead to less overconfidence than was observed in Experiment 1. For three out of the four groups in the current study, signed differences were lower than those from Experiment 2; the exception was the collaborate/reread group. Interestingly, as in Experiment 1, scores on the comprehension test were low, ranging from 0.52 to 0.59 (scores were 0.58 and 0.56 in Experiment 1). It appears that adding practice test questions was insufficient to boost comprehension.

Again, as was found in Experiment 1, the obtained gamma correlations for both metacomprehension accuracy and study regulation were small, especially relative to the standard errors. Together, the results of the two experiments suggest that gamma correlations may not be the most appropriate or powerful statistic given the design of the study and its materials. It bears noting that, in the current study, 11 students were excluded from the analysis on metacomprehension accuracy gammas, and 44 were excluded from the analysis on study regulation gammas.

Summary

The current research examined the effects of keyword generation and peer collaboration on metacomprehension accuracy and study regulation. The following hypotheses were tested (supported hypotheses appear in bolded font):

7. **Middle school students will be overconfident in their judgments of comprehension.**
8. Strategy condition, collaboration condition, and time of judgment will interact in their effects on metacomprehension accuracy.
9. Strategy condition, collaboration condition, and time of judgment will interact in their effects on study regulation.

CHAPTER 5

GENERAL DISCUSSION

The current research explored whether middle school students are able to make accurate comprehension judgments and whether they can apply those judgments to make effective study choices. Experiment 1 involved 109 seventh graders from New York City public schools and compared the effects of generating keywords and rereading text on metacomprehension accuracy and study regulation. Experiment 2 involved 190 seventh graders and asked whether the effect of keyword generation on metacomprehension accuracy and study regulation would continue to hold in the presence of practice testing and peer collaboration.

Summary of the Findings

The results of Experiment 1 showed that middle school students tend to be overconfident in their judgments of their own comprehension. Overconfidence was greater for pre-test predictions than for post-test reflections, and it was also greater for students with lower reading ability. Generating keywords caused participants to become significantly less overconfident– or more accurate– from pre-test to post-test in their comprehension judgments but it did not actually boost comprehension scores; in other words, generating keywords helped participants know that they did not know, but it did not, however, help them know more.

Similarly, participants in Experiment 2 were also overconfident in judging their own comprehension. Again, there was an effect for time of judgment, such that predictions were more overconfident than were reflections. Surprisingly, peer collaboration was found to lead to greater overconfidence in comprehension judgments. Participants who collaborated with a peer were

more overconfident than participants who worked alone. Experiment 2 showed that in the presence of practice testing and peer collaboration, the interactive effect of keyword generation and time of judgment was minimized. Within the keyword group, participants who collaborated and participants who worked alone did not differ in overconfidence. Within the reread group, however, participants who collaborated were significantly more overconfident than those who worked alone.

Implications

The finding that generating keywords led participants to become significantly less overconfident– or more accurate– from pre-test to post-test has important practical implications. It bears noting that the actual scores on the comprehension tests did not differ across groups. In Experiment 1, the mean percentage score across the five tests was 58% for the keyword group and 56% for the reread group. The discrepancy between comprehension judgments and test scores did, however, differ across groups and over time. For the keyword group, the gap between perceived and actual comprehension narrowed from pre-test to post-test; for the reread group, there was no change.

These data indicate that keyword generation has an effect on metacomprehension but not on comprehension. Generating keywords helped participants know that they did not know; it did not, however, help them know more. These results suggest that keyword generation may be of low utility as a comprehension or study strategy. Instead, the benefit may be diagnostic in nature, in that the ability to generate keywords as well as the inability to do so are taken into account when making comprehension judgments. The ability to make accurate judgments is especially

important during self-regulated study, when a learner must make decisions about when to start, stop, and continue studying.

Based on the results of Experiments 1 and 2, there is reason to believe that the effect of keyword generation may be analogous to that of practice testing. Used alone, keyword generation led to better metacomprehension accuracy than did rereading text (Experiment 1); when participants in both groups were given practice test questions to answer before generating keywords or rereading text, there was no difference in metacomprehension accuracy (Experiment 2). Generating keywords and practice testing are strategies that require a learner to retrieve and recall previously read information. Both strategies are generally used after reading but before testing, and both have the potential to inform and direct further study. One answer to the question of how and when keyword generation should be used is that it can be applied in lieu of or in addition to practice testing.

When metacomprehension accuracy was operationalized via gamma correlations, the keyword and reread groups did not differ. The gamma correlations obtained in Experiments 1 and 2 were generally low, and were actually higher (though not significantly so) in the reread group than in the keyword. These results are surprising, given the findings of other studies in the literature and the results obtained here using measures of confidence. One explanation for this involves the nature of gamma. Some researchers have argued that gamma is an unreliable and unstable measure of metacomprehension accuracy (e.g., Kelemen, Frost, & Weaver, 2000; Thompson & Mason, 1996). Alternatively, it may be that gamma correlations and signed differences measure different aspects of metacomprehension. Maki, Shields, Wheeler, and Zacchilli (2005) found that gamma correlations and signed differences were differently affected by text difficulty and verbal ability and were also uncorrelated.

The results of the current research provide support for the latter explanation. Gamma correlations measure two variables are monotonically associated, and are useful for assessing the consistency of judgments across a set of items (Maki, 1998a; Schraw, 2009). Signed differences indicate the degree to which individuals are under- or over-confident, and are useful for assessing the precision of judgments. If the measures assess different aspects of metacomprehension, then the finding that they are differently affected by experimental manipulations need not indicate that either measure is unreliable or unstable.

In general, study regulation gammas were higher for the reread group than the keyword group, indicating that participants in this group made restudy choices that were more strongly associated with their comprehension judgments. However, it is important to note that the comprehension judgments upon which the study choices were based were themselves inaccurate– or overconfident– for this group. In other words, participants who reread texts were more likely to rate their comprehension highly and less likely to choose to restudy than participants who generated keywords. To wit, an examination of the data from Experiment 1 shows that 47% of participants in the keyword group chose to restudy, compared to only 42% of participants in the reread group (although this difference is not significant).

The addition of practice testing in Experiment 2 led to a different pattern of results than was observed in Experiment 1. In the presence of practice test questions, keyword generation did not lead to better metacomprehension accuracy than did rereading text. Interestingly, the data indicate that within the reread condition, overconfidence was actually higher (though not significantly so) among participants who worked with a partner than among participants who worked individually; there was no difference within the keyword condition. In other words, answering practice questions individually before rereading led to better accuracy than did

discussing practice questions with a partner before rereading. Peer collaboration did not appear to improve metacomprehension accuracy. On the contrary, collaborating may have actually had an adverse impact by inflating comprehension judgments.

Future Research

The design of the current research did not include an experimental manipulation of practice testing. Consequently, it cannot be known for certain that the different results obtained in Experiments 1 and 2 are due to the inclusion of practice testing in Experiment 2. Additional research is needed to determine if the effects of keyword generation and peer collaboration together are the same in the presence of practice testing and also in the absence of practice testing.

Another area for further investigation involves keyword quality. The current research did not examine the quality of the keywords that were generated, and it did not analyze the effect of keyword quality on metacomprehension accuracy. It is possible that higher keyword quality would be associated with higher metacomprehension accuracy and perhaps higher comprehension scores as well. It is also possible that given adequate training, students can learn to generate keywords in a way that does lead to increased comprehension.

A third area for additional study involves pair type. In Experiment 2, participants in the collaborative condition were randomly paired with a partner without consideration of gender or ability. It is possible that whether pairs are homogenous or heterogeneous in terms of these variables might have an effect on metacomprehension accuracy. Additionally, future work could examine the accuracy of participants' judgments about their partners.

REFERENCES

Alicke, M. D. (1985). Global self-evaluation as determined by the desirability and controllability of trait adjectives. *Journal of Personality and Social Psychology, 49,* 1621-1630.

Anderson, M. C. M. & Thiede, K. W. (2008). Why do delayed summaries improve metacomprehension accuracy? *Acta Psychologica, 128*, 110–118.

Britton, B. K. & Gülgöz, S. (1991). Using Kintsch's computational model to improve instructional text: Effects of repairing inference calls on recall and cognitive structures. *Journal of Educational Psychology, 83(3),* 329-345.

Brown, J. D. (1986). Evaluations of self and others: Self-enhancement biases in social judgments. *Social Cognition, 4,* 353-376.

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

Cazden, C. & Beck, S. W. (2003). Classroom discourse. In A. Graesser, M. Gernsbacher, & S. Goldman (Eds.), *Handbook of Discourse Processes* (pp. 165-198). Mahwah, NJ: Erlbaum.

Darwin, C. (1871). *The descent of man.* London: John Murray.

De Bruin, A. B. H., Thiede, K. W., Camp, G., & Redford, J. (2011). Generating keywords improves metacomprehension and self-regulation in elementary and middle school children. *Journal of Experimental Child Psychology, 109*, 294-310.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing By Latent Semantic Analysis. *Journal of the American Society For Information Science, 41,* 391-407.

De Lisi, R. & Goldbeck, S. L. (1999). Implications of Piagetian theory for peer learning. In A. M. O'Donnell & A. King (Eds.), *Cognitive perspectives on peer learning.* New Jersey: Erlbaum.

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

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

Dunning, D., Johnson, K., Ehrlinger, J. & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science, 12,* 83-87.

Dunning, D., Meyerowitz, J. A., & Holzberg, A. D. (1989). Ambiguity and self-evaluation: The role of idiosyncratic trait definitions in self-serving assessments of ability. *Journal of Personality and Social Psychology, 57,* 1082-1090.

Epstein, W., Glenberg, A. M., & Bradley, M. (1984). Coactivation and comprehension: Contribution of text variables to the illusion of knowing. *Memory & Cognition, 12,* 355-360.

Everitt, B. S. (1977). *The analysis of contingency tables*. London: Chapman and Hall.

Fawcett, L. M. & Garton, A. F. (2005). The effect of peer collaboration on children's problem-solving ability. *British Journal of Educational Psychology, 75,* 157-169.

Feir-Walsh, B. J. & Toothaker, L. E. (1974). An empirical comparison of the ANOVA F-test, normal score test and Kruskal-Willis test under violation of assumptions. *Educational and Psychological Measurement, 34,* 789-799.

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

Flavell, J. H. (2000). Development of children's knowledge about the mental world. *International Journal of Behavioural Development, 24,* 15-23.

Flavell, J. H. (2004). Theory of mind development: retrospect and prospect. *Journal of Developmental Psychology, 50*(3), 274-290.

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

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

Goodman, L. A. & Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association, 49,* 732-764.

Hacker, D. J. (1998). Self-regulated comprehension during normal reading. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 165-191). Mahwah, NJ: Erlbaum.

Harris, P., de Rosnay, M., & Pons, F. (2005). Language and children's understanding of mental states. *Current Directions in Psychological Science, 14,* 69-73.

Huff, J. D. & Nietfeld, J. L. (2009). Using strategy instruction and confidence judgments to improve metacognitive monitoring. *Metacognition and Learning, 4,* 161-176.

Jacobs, J. E. & Paris, S. G. (1987). Children's metacognition about reading: Issues in definition, measurement, and instruction. *Educational Psychologist, 22(3 & 4),* 255-278.

Karpicke, J. D., Butler, A. C., & Roediger, H. L., III. (2009). Meta-cognitive strategies in student learning: Do students practice retrieval when they study on their own? *Memory, 17,* 471–479.

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

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review, 95*, 163-182.

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

Koriat, A., & Shitzer-Reichert, R. (2002). Metacognitive judgments and their accuracy: Insights from the processes underlying judgments of learning in children. In P. Chambres, M. Izaute, & P. J. Marescaux (Eds.), *Metacognition: Process, function, and use* (pp. 1–17). Dordrecht, Netherlands: Kluwer.

Krueger, J. & Mueller, R. A. (2002). Unskilled, unaware, both? The better-than-average heuristic and statistical regression predict errors of estimates of own performance. *Journal of Personality and Social Psychology, 82,* 180-188.

Kruger, J. & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77,* 1121-1134.

Kuhn, D. & Dean, D. Jr. (2004). Connecting scientific reasoning and causal inference. *Journal of Cognition and Development, 5(2),* 262-288.

Maki, R. H. (1998a). Predicting performance on text: Delayed versus immediate predictions and tests. *Memory and Cognition, 26,* 959–964.

Maki, R. H. (1998b). Test predictions over text material. In D.J. Hacker, J. Dunlosky, & A.C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 117–144). Mahwah, NJ: Erlbaum.

Maki, R. H., Shields, M., Wheeler, A. E., & Zacchilli, T. L. M. (2009). Individual differences in absolute and relative metacomprehension accuracy. *Journal of Educational Psychology, 97(4)*, 723-731.

Mar, R.A., Tackett, J. L., Moore, C. (2010). Exposure to media and theory-of-mind development in preschoolers. *Cognitive Development,* 25, 69–78.

Metcalfe, J. & Dunlosky, J. (2008). Metamemory. In J. H. Byrne (Ed.), *Learning and memory: A comprehensive reference* (pp. 349-362). Waltham, MA: Elsevier Academic Press.

Metcalfe, J. & Finn, B. (2013). Metacognition and control of study choice in children. *Metacognition & Learning, 8(1),* 19-46.

National Center for Education Statistics (2013). *The nation's report card: A first look: 2013 mathematics and reading (NCES 2014-451).* Washington, DC: Institute of Education Sciences, U.S. Department of Education.

National Governors Association Center for Best Practices, Council of Chief State School Officers (2010). *Common Core State Standards.* Washington, DC: National Governors Association Center for Best Practices, Council of Chief State School Officers.

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin, 95,* 109-133.

Nelson, T. O. & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. Bower (Ed.), *The psychology of learning and motivation* (Vol. 26, pp. 125–140). New York: Academic Press.

Nelson, T. O. & Narens, L. (1994). Why investigate metacognition? In J. A. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: knowing about knowing* (pp. 1–25). Cambridge, MA: MIT Press.

Piaget, J. (1959). *The language and thought of the child* (3rd edition). London: Routledge & Kegan Paul. (M. Gabain, trans.)

Rumsfeld, D. H. (2002). DoD news briefing- Secretary Rumsfeld and Gen. Myers. Retrieved from http://www.defense.gov/transcripts/transcript.aspx?transcriptid=2636.

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

Shepperd, J. A., Ouellette, J. A., & Fernandez, J. K. (1996). Abandoning unrealistic optimism: Performance estimates and the temporal proximity of self-relevant feedback. *Journal of Personality and Social Psychology, 70,* 844-855.

Silver, N. (2012). *The signal and then noise: Why so many predictions fail- but some don't.* New York: Penguin.

Stanovich, K. E. & West, R. F. (1989). Exposure to print and orthographic processing. *Reading Research Quarterly, 24,* 402–433.

Thiede, K. W., Anderson, M. C. M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology, 95,* 66–73.

Thiede, K. W., Dunlosky, J., Griffin, T. D., & Wiley, J. (2005). Understanding the delayed keyword effect on metacomprehension accuracy. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 1267–1280.

Thompson, W. B. & Mason, S. E. (1996). Instability of individual differences in the association between confidence judgments and memory performance. *Memory & Cognition, 24,* 226-234.

Tversky, A. & Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science, 185*, 1124-1130.

Watson, A. C., Linkie-Nixon, C., Wilson, A., Capage, L. (1999). Social interaction skills and theory of mind in young children. *Developmental Psychology, 35,* 386–391.

Wedell, D. H. & Parducci, A. (2000). Social comparison: Lessons from basic research on judgment. In J. Suls & L. Wheeler (Eds.), *Handbook of social comparison: Theory and research* (pp. 2230252). New York: Kluwer Academic/Plenum.

Wellman, H. (2011). Developing a theory of mind. In U. Goswami (Ed.), Handbook of childhood cognitive development (Blackwell). (2nd edition)

Williams, J. P. & Atkins, J. G. (2009). The role of metacognition in teaching reading comprehension to primary students. In D. J. Hacker, Dunlosky, J. & A. C. Graesser (Eds.), *Handbook of metacognition in education.* Routledge.

Williams, J. P., Pollini, S., Nubla-Kung, A. M., Snyder, A. E., Garcia, A., Ordynans, J. G., & Atkins, J. G. (2014). An intervention to improve comprehension of cause/effect through expository text structure instruction. *Journal of Educational Psychology, 106(1),* 1-17.

Zhao, Q. & Linderholm, T. (2011). Anchoring effects on prospective and retrospective metacomprehension judgments as a function of peer performance information. *Metacognition and Learning, 6,* 25-53.

APPENDIX A

Table A1

Word Counts, Reading Levels, and Text Characteristics for the Passages

	Causes of Extinction	*Cellular Respiration*	*Circulatory System*	*Digestion*	*Fight or Flight*
Word count	327	322	418	294	359
Reading level[a]	7.5	8.2	6.4	7.2	6.7
Narrativity	0.14	0.15	0.28	0.14	0.25
Syntactic simplicity	0.97	0.87	0.44	0.81	0.93
Word concreteness	0.89	0.75	0.96	0.88	0.95
Referential cohesion	0.35	0.53	0.99	0.77	0.61
Deep cohesion	0.99	0.41	0.18	0.38	0.91

[a] Values used for reading levels are Flesch-Kincaid Grade Levels.

Causes of Extinction

Extinction is the disappearance of a species. Natural events, such as a volcanic eruption, can cause extinction. Human activities can also cause extinction through habitat destruction, hunting, pollution, and introduction of new species.

Most extinction is from habitat destruction— the loss of a natural habitat. This can occur when forests are chopped down. Forests get chopped down to build towns or create grazing land. Plowing grasslands or filling in wetlands also changes the area. Some species may be unable to survive these changes.

Habitat fragmentation is when larger habitats are broken into smaller, isolated pieces. For example, building a road through a forest disrupts habitats. Wind is likely to damage the exposed trees. Plants may be unable to spread their seeds. Habitat fragmentation is also harmful to large mammals. These animals often need large areas of land to find enough food to live. A small area may not have enough food. They may also be hurt trying to cross to another area.

Poaching is the illegal killing or taking of wildlife from their habitats. Hunters hunt many animals for their skin, fur, teeth, horns, or claws. Hunters sell the animals they kill. The animal parts are used for making things such as medicines, jewelry, and clothing. People also illegally take animals from their habitats to sell them as exotic pets. Tropical fish, tortoises, and parrots are popular pets, making them valuable to poachers. Endangered plants are sometimes illegally dug up and sold as houseplants or medicines.

Pollutants that cause pollution also hurt some species. Animals may drink water or breathe air with pollutants. Pollutants may also be in the soil which will then go into plants. These pollutants increase through the food chain. For example, a small amount of pollutants may be in each blade of grass, but a grazing cow that eats a lot of grass will contain a lot of those pollutants. Pollutants may kill or weaken organisms or cause birth defects.

Cellular Respiration

You've been hiking all morning, and you are hungry. You get out a sandwich you packed and begin munching. Food supplies your body with glucose, an energy-rich sugar. Respiration is the process by which cells obtain energy from glucose. During cellular respiration, cells break down simple food molecules such as glucose and release the energy they contain. Most of the energy used by the cells in your body is provided by cellular respiration.

Energy stored in cells is something like money in a savings account. During photosynthesis, plants capture energy from sunlight and ''save'' it in the form of carbohydrates, including sugars and starches. When you eat, you add to your body's energy savings account. When cells need energy, they ''withdraw'' it by breaking down the carbohydrates in the process of respiration.

Cellular respiration is a two-stage process. The first stage takes place in the cytoplasm of the organism's cells. There, molecules of glucose are broken down into smaller molecules. Oxygen is not involved, and only a small amount of energy is released.

The second stage of cellular respiration takes place in the mitochondria. There, the small molecules are broken down into even smaller molecules. These chemical reactions require oxygen, and they release a great deal of energy. This is why the mitochondria are sometimes called the ''powerhouses'' of the cell.

Energy is released as a product in both stages of respiration. This is transferred to other molecules, which then carry the energy where it is needed for the activities of the cell. The rest of the energy is released as heat. Two other products of cellular respiration are carbon dioxide and water. These products diffuse out of the cell. In most animals, the carbon dioxide and some water leave the body during exhalation or breathing out. When you breathe in, you take in oxygen – a raw material for respiration. When you breathe out, you release carbon dioxide and water.

The Circulatory System

Your body's circulatory system - your heart, lungs, blood vessels, and blood - helps to keep you alive. It carries needed things to cells and carries waste away from cells. Your blood also has cells that fight disease.

Your heart has four chambers and works at the center of the system, pushing blood through two different loops through different parts of your body. Your heart has two sides, and each side has an upper chamber, the atrium, and a lower chamber, the ventricle. When blood from the body fills up the right top part of the heart, or right atrium, it has a little oxygen but a lot of carbon dioxide. This oxygen-poor blood is dark red. The blood then flows from the right atrium into the right ventricle. The right ventricle pumps the oxygen-poor blood into the arteries that go to the lungs.

As blood goes through the lungs, large blood vessels break into smaller ones. Blood goes through tiny vessels, or capillaries, that meet up with the air that comes into the lungs. The air in the lungs has more oxygen than the blood in the capillaries, so oxygen moves from the air into the blood. Because the blood has more carbon dioxide then the air in the lungs, carbon dioxide moves in the opposite direction – from the blood into the air. This oxygen-rich blood, which is bright red, goes into the left side of the heart to be pumped through the second loop.

The second loop begins with the left atrium that is filled with oxygen-rich blood from the lungs. The blood moves into the left ventricle. From the left ventricle, the oxygen-rich blood is pumped into the largest artery in the body—the aorta.

After passing through arteries, blood goes through tiny capillaries in different parts of your body, including your brain, liver, and legs. These vessels meet up with body cells. Oxygen moves out of the blood and into the body cells. At the same time, carbon dioxide moves from the blood cells and into the blood. This blood, which is low in oxygen, then moves back to the right atrium of the heart through veins, completing the second loop. Blood only moves in one direction, and the whole trip through your body takes less than one minute. When the muscles in your body really need oxygen, the whole process speeds up. Your heart beats faster, and the blood can go through the body in as little as 30 seconds.

Digestion

When you chew, you start the mechanical digestion of food. Your teeth break up the food into small pieces. At the same time, your body makes saliva that contains enzymes which break the food's complex starch molecules into simpler ones. This is called chemical digestion. When you swallow, your tongue pushes the food down your throat, where it travels down to your stomach.

Your stomach is a J-shaped, muscular pouch in your abdomen. Most mechanical digestion happens in the stomach. Three strong layers of muscle contract to make a stirring motion. This action mixes the food with digestive fluids made by cells in the stomach.

Chemical digestion also happens as the food mixes with digestive fluid. Digestive fluid has the enzyme pepsin, which digests the proteins in your food, breaking them down into short chains of amino acids. Digestive juice also has hydrochloric acid. Without this acid, your stomach would not work well. Pepsin works best in hydrochloric acid.

A few hours after you eat, the stomach completes mechanical digestion of the food. By that time, most of the proteins have turned into shorter chains of amino acids. The food, now a thick liquid, is sent to the small intestine where most chemical digestion takes place. The small intestine is where fats are broken down by the liver and pancreas, which also help to break down any remaining starches and proteins. Once broken down, these molecules are absorbed into the body through the wall of the small intestine.

The digestive fluids produced by the pancreas and other organs do not break down fiber. The fiber remains undigested and it, along with water, goes to the large intestine. In the large intestine, water is absorbed, and the remaining material is eliminated from the body.

The Fight or Flight Response

You're in the park on a hot afternoon. Without warning, dark clouds form. Suddenly, there's a flash of lightning. You hear loud thunder. Someone screams, you jump and everyone runs for cover. Your heart is pounding. Your body's reaction to the sudden storm is caused by your endocrine system.

The endocrine system produces chemicals that control many of the body's daily activities. The endocrine system is made up of glands. A gland is an organ that makes or releases a chemical. Some glands, like the ones that make saliva and sweat, release their chemicals into tiny tubes called ducts. The ducts deliver the chemicals to a specific place within the body or to the skin's surface. However, endocrine glands make and release their chemical products into the bloodstream. This chemical product is called a hormone. Blood carries hormones everywhere in the body. A hormone does not interact with all organs in the body. It interacts with only certain cells that recognize the hormone's chemical structure. These are called target cells. Hormones will travel through the bloodstream until they find target cells that they fit. Target cells then respond by turning on, turning off, speeding up, or slowing down the activities of different organs and tissues. Hormones can control activities in tissues and organs that are far away from the glands that made them.

The endocrine system controls the body's response to an exciting situation and hormones are released by nerve impulses from the brain. For example, when a person sees something that is scary, the brain sends a message to a specific endocrine gland, the adrenal gland. That gland releases the hormone adrenaline into the bloodstream. The target cells for adrenaline are in your heart and lungs. These target cells then respond to adrenaline by speeding up the heart and lungs. It also releases sugars in your muscles to give you energy in case you need to either "fight" or run away from a situation ("flight"). That is why this has been called the fight-or-flight response. The heart will continue to beat faster until the adrenaline in the blood drops to a normal level.

APPENDIX B

Causes of Extinction

1. Which of the following would best deter poaching?
    a. Legalizing hunting of endangered or exotic animals
    b. Relocating endangered or exotic animals to new locations
    c. Introducing natural predators of endangered or exotic animals
    d. Decreasing demand for a particular animal or animal product

2. Most extinction results from:
    a. Habitat destruction
    b. Natural disasters
    c. Poaching
    d. Pollution

3. Which of the following would best minimize the harm done by habitat fragmentation?
    a. Legalizing the hunting of the constricted large mammals
    b. Expanding the territory of large mammals in adjacent areas
    c. Introducing natural predators of the large mammals
    d. Perform habitat fragmentation slowly so that animals can adjust

4. Poaching involves either the illegal killing or _____ of wildlife:
    a. Integration
    b. Introduction
    c. Production
    d. Removal

5. Which of the following causes of extinction are humans most responsible for?
    a. Poaching animals
    b. Causing natural disasters
    c. Introducing exotic plants
    d. Birth defects in animal parents

6. Chopping down a forest is an example of:
    a. Extinction through natural disasters
    b. Extinction through pollution
    c. Extinction through habitat fragmentation
    d. Extinction through poaching

Cellular Respiration

1. When you breathe into your hand and your palm becomes moist, this is caused by:
    a. The release of oxygen through your skin
    b. Your breath being warmer than the air around it
    c. Water from your cells leaving your body in your breath
    d. A normal breakdown of the cell's cytoplasm

2. Why would a runner take off her jacket on a cold day?
    a. She ate too much glucose before the run
    b. As her cells break down glucose, they release energy and heat
    c. As she runs, carbon dioxide combines with water to produce heat
    d. Starches combine with sugars to make her sweat

3. If glucose does not break down in your cells, you will feel:
    a. Tired
    b. Out of breath
    c. Hyperactive
    d. Overheated

4. Which of the following would improve an athlete's performance?
    a. Low-carbohydrate snack, low-oxygen environment
    b. Low-carbohydrate snack, high-oxygen environment
    c. High-carbohydrate snack, low-oxygen environment
    d. High-carbohydrate snack, high-oxygen environment

5. Which of the following describes the role of oxygen in cellular respiration:
    a. Oxygen is involved in the first stage only
    b. Oxygen is involved in the second stage only
    c. Oxygen is involved in both of the stages
    d. Oxygen is not involved in either stage

6. Cellular respiration in animals is analogous to _____ in plants:
    a. Water absorption
    b. Ripening
    c. Flowering
    d. Photosynthesis

The Circulatory System

1. Oxygen-poor blood is:
    a. Blue
    b. Bright red
    c. Dark red
    d. Colorless

2. Which of the following is true about the heart?
    a. It pushes blood into two different blood vessels
    b. It removes carbon dioxide from the blood
    c. It is a hollow bag like a balloon
    d. It collects oxygen to send to the lungs

3. The largest artery in the body is the:
    a. Aorta
    b. Atrium
    c. Vena cava
    d. Ventricle

4. Lack of oxygen in the air you breathe can cause:
    a. The heart to pump more slowly
    b. The direction of blood flow to reverse
    c. Higher levels of oxygen in the blood
    d. Higher levels of carbon dioxide in the blood

5. Which of the following causes the blood to flow faster through your body?
    a. Your stomach needs food
    b. Your brain needs blood
    c. Your muscles need oxygen
    d. Your body needs sleep

6. Why do you breath harder when you exercise?
    a. Because you need more carbon dioxide
    b. Because your heart pumps faster
    c. Because your blood becomes thinner
    d. Because you have more blood in your system

Digestion

1. Which of the following is a correct listing of the order of digestive functions?
    a. Absorption, elimination, chemical digestion
    b. Mechanical digestion, elimination, absorption
    c. Mechanical digestion, absorption, elimination
    d. Absorption, chemical digestion, elimination

2. Mechanical digestion of food is started by your:
    a. Esophageal muscles
    b. Tongue
    c. Saliva
    d. Teeth

3. Which of the following would prevent chemical digestion in your mouth?
    a. An absence of teeth
    b. An absence of pepsin
    c. An absence of hydrochloric acid
    d. An absence of saliva

4. Your stomach contains:
    a. Peroxide and formic acid
    b. Phosphorus and acetic acid
    c. Sulfur and nitric acid
    d. Pepsin and hydrochloric acid

5. What would happen if a person lost their liver and pancreas?
    a. They would be unable to digest water
    b. They would lose the ability to digest fiber
    c. They would be unable to digest fats
    d. They would be unable to produce mucus

6. Water is absorbed by the:
    a. Large intestine
    b. Small intestine
    c. Stomach
    d. Esophagus

The Fight or Flight Response

1. The endocrine system is made up of:
    a. Chemicals
    b. Ducts
    c. Glands
    d. Hormones

2. Which of the following best describes the endocrine system?
    a. It regulates organs using nerve impulses
    b. It is a system of muscles and fibers
    c. It is a system located in our arms and legs
    d. It works using chemical messages

3. If the product of the endocrine system does not fit a target cell, it will:
    a. Dissolve in the bloodstream
    b. Move on until it finds a cell that it fits
    c. Mutate itself in order to fit the cell
    d. Mutate the cell so that it fits

4. Which of the following is a result of releasing adrenaline into the bloodstream?
    a. Increase in breathing rate
    b. Decrease in sweat production
    c. Increase in drowsiness
    d. Decrease in heart rate

5. Hormones can regulate activities in organs:
    a. Close to the gland
    b. Far from the gland
    c. Either close to or far from the gland
    d. Only if they have ducts

6. What is the order of the process underlying the Fight or Flight response?
    a. Nerve impulse, release of hormone, increase in heart rate
    b. Increase in breathing rate, release of hormone, brain impulse
    c. Recognition of danger, activation of target cells, release of hormone
    d. Nerve impulse, decrease in heart rate, activation of target cells

APPENDIX C

Please circle how many of the six (6) test questions **you think you will answer correctly** about the text entitled ***Causes of Extinction***:

0 1 2 3 4 5 6

Please circle how many of the six (6) test questions **you think you will answer correctly** about the text entitled ***Cellular Respiration***:

0 1 2 3 4 5 6

Please circle how many of the six (6) test questions **you think you will answer correctly** about the text entitled ***The Circulatory System***:

0 1 2 3 4 5 6

Please circle how many of the six (6) test questions **you think you will answer correctly** about the text entitled ***Digestion***:

0 1 2 3 4 5 6

Please circle how many of the six (6) test questions **you think you will answer correctly** about the text entitled ***The Fight or Flight Response***:

0 1 2 3 4 5 6

Please circle how many of the six (6) test questions **you think you answered correctly** about the text entitled ***Causes of Extinction***:

0 1 2 3 4 5 6

Please circle how many of the six (6) test questions **you think you answered correctly** about the text entitled ***Cellular Respiration:***

0 1 2 3 4 5 6

Please circle how many of the six (6) test questions **you think you answered correctly** about the text entitled ***The Circulatory System***:

0 1 2 3 4 5 6

Please circle how many of the six (6) test questions **you think you answered correctly** about the text entitled ***Digestion***:

0 1 2 3 4 5 6

Please circle how many of the six (6) test questions **you think you answered correctly** about the text entitled ***The Fight or Flight Response***:

0 1 2 3 4 5 6

APPENDIX D

Directions:

Listed below are the passages you have read and answered questions about. Suppose we were to give you another test about the passages.

Which passages would you want to reread or restudy?

(You will NOT actually restudy the texts or answer more questions- we just want to know which texts you THINK YOU SHOULD read or study again in order to improve your test score.)

Please choose the texts that you would want to reread or restudy.

Passage	**Restudy?**	
Causes of Extinction	Yes	No
Cellular Respiration	Yes	No
The Circulatory System	Yes	No
Digestion	Yes	No
The Fight or Flight Response	Yes	No

APPENDIX E

EFFECTS OF GENDER, STRATEGY CONDITION, AND TIME OF JUDGMENT

Because the initial exploration of the data revealed a statistically significant difference in assignment of participants to strategy condition by gender, the analyses were repeated to determine whether the effects of strategy condition and time of judgment differed by gender. A mixed 2 x 2 x 2 ANOVA was used, with *strategy condition* and *gender* as between-subjects factors and *time of judgment* as a within-subjects factor. Table 4 presents the test scores and comprehension judgments by strategy condition and gender.

Table E1
Test Performance and Comprehension Judgments by Strategy Condition and Gender

Condition		Prediction		Test Score		Reflection	
		Mean	*SD*	Mean	*SD*	Mean	*SD*
Keyword/Male	(n = 17)	0.85	0.08	0.63	0.15	0.75	0.11
Keyword/Female	(n = 36)	0.74	0.13	0.56	0.19	0.65	0.15
Reread/Male	(n = 29)	0.70	0.13	0.57	0.19	0.69	0.15
Reread/Female	(n = 24)	0.73	0.19	0.55	0.16	0.74	0.15

An ANOVA on mean predictions showed a significant interaction between strategy condition and gender, $F(1, 105) = 6.71$, $p = .011$, $\eta_p^2 = .062$. An analysis of simple effects of strategy condition on mean predictions revealed a difference between males who generated keywords and males who reread text ($M = 0.85$ vs. $M = 0.70$), $F(1, 102) = 11.92$, $p = .001$, $\eta_p^2 = .11$. An analysis of simple effects of gender on mean predictions revealed a difference between males who generated keywords and females who generated keywords ($M = 0.85$ vs. $M = 0.74$), $F(1, 102) = 7.20$, $p = .009$, $\eta_p^2 = .066$.

An ANOVA on mean test scores showed no interaction effect, $F(1, 105) = 0.61$, $p = .44$, and no main effects for strategy condition, $F(1, 105) = 0.95$, $p = .33$, or for gender, $F(1, 105) = 1.45$, $p = .23$.

An ANOVA on mean reflections showed a significant interaction between strategy condition and gender, $F(1, 105) = 5.88$, $p = .017$, $\eta_p^2 = .054$. An analysis of simple effects of strategy condition on mean reflections revealed a difference between females who generated keywords and females who reread text ($M = 0.65$ vs. $M = 0.74$), $F(1, 102) = 5.27$, $p = .024$, $\eta_p^2 = .049$. An analysis of simple effects of gender on mean reflections revealed a difference between females who generated keywords and males who generated keywords ($M = 0.65$ vs. $M = 0.75$), $F(1, 102) = 5.18$, $p = .025$, $\eta_p^2 = .048$.

Metacomprehension Accuracy

Signed differences. Figure E1 presents the signed differences between comprehension judgments and test scores by gender and strategy condition.

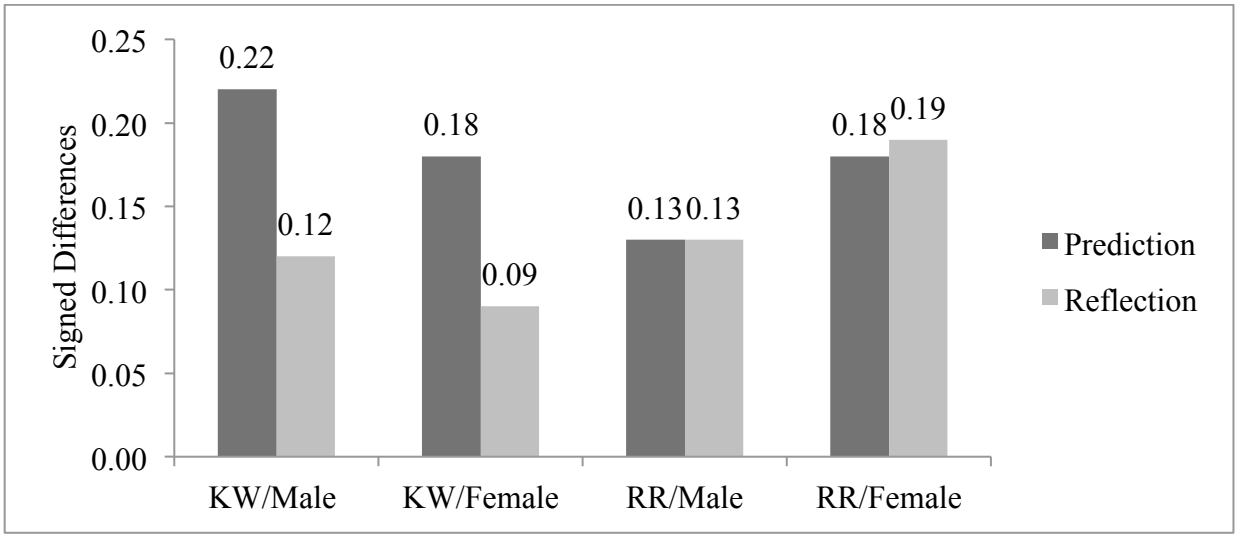


Figure E1. Signed differences by strategy condition and gender.

An ANOVA on signed differences showed that the three-way interaction was not significant, $F(1, 102) = 0.002$, $p = .96$. The two-way interaction between time of judgment and strategy condition was statistically significant, $F(1, 102) = 13.83$, $p < .001$, $\eta_p^2 = .12$. The two-way interaction between time of judgment and gender was not statistically significant, $F(1, 102) = 0.19$, $p = .66$. The two-way interaction between strategy condition and gender was not statistically significant, $F(1, 102) = 1.36$, $p = .25$. Time of judgment, strategy condition, and gender did not interact in their effects on signed differences. Gender did not moderate the interaction between time of judgment and strategy condition.

Gamma correlations. Figure E2 presents the gamma correlations between comprehension judgments and test scores by gender and strategy condition.

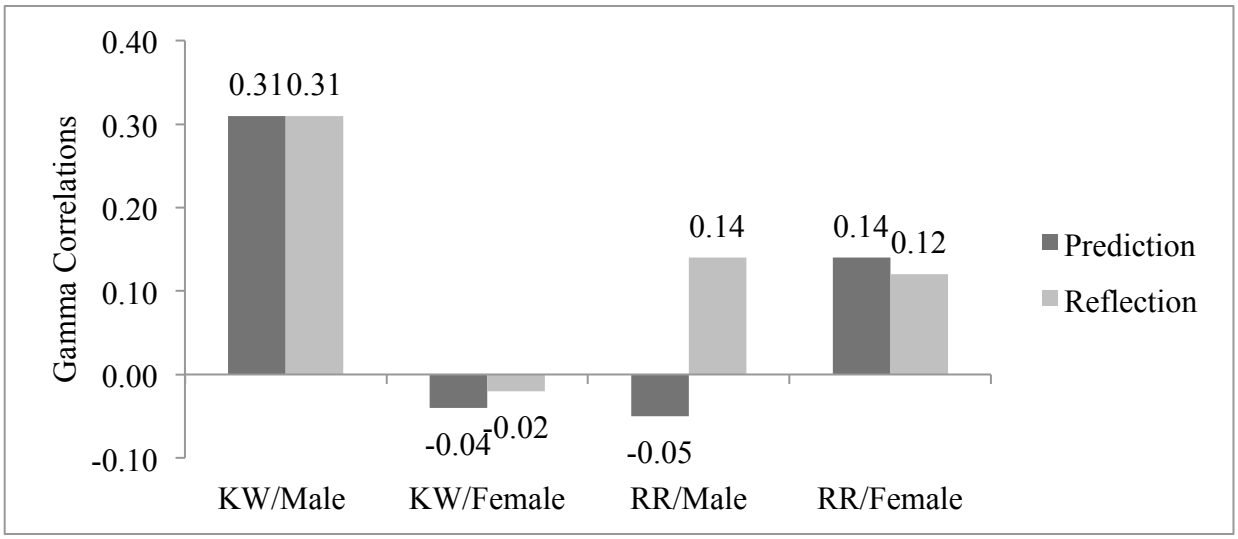


Figure E2. Gamma correlations by strategy condition and gender.

An ANOVA on gamma correlations showed that the three-way interaction was not significant, $F(1, 92) = 0.36$, $p = .55$. None of the two-way interactions were significant ($p > .05$

in all cases). None of the main effects were significant ($p > .05$ in all cases). Time of judgment, strategy condition, and gender did not interact in their effects on gamma correlations.

Study Regulation

Figure E3 presents the gamma correlations between restudy choices and comprehension judgments by gender and strategy condition.

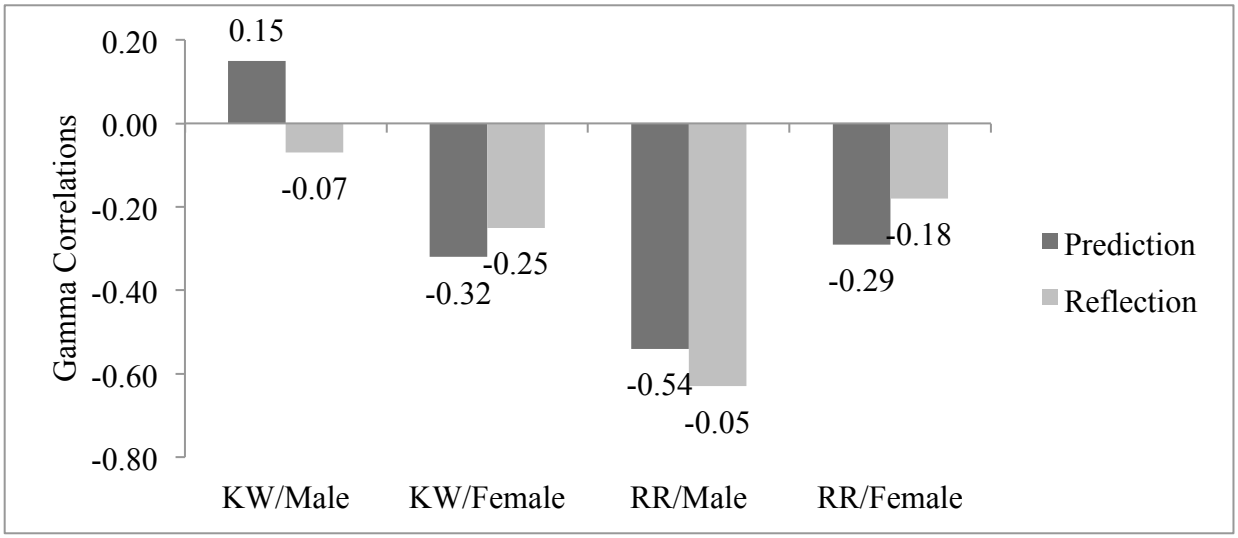


Figure E3. Study regulation by strategy condition and gender.

An ANOVA on study regulation showed that the three-way interaction was not significant, $F(1, 76) = 0.063$, $p = .80$. The two-way interaction between strategy condition and gender was significant, $F(1, 76) = 4.49$, $p = .037$, $\eta_p^2 = .056$. Simple main effects were analyzed. For prediction/restudy gammas, there was a statistically significant simple effect of strategy condition for males, $F(1, 78) = 5.09$, $p = .027$, $\eta_p^2 = .061$, but not for females, $F(1, 78) = 0.042$, $p = .84$. Among males in the study, restudy choices were better correlated with predictions in the reread condition than in the keyword condition, $M = -0.47$ ($SD = 0.16$) vs. $M = 0.15$ ($SD = 0.23$).

The two-way interaction between time of judgment and gender was not statistically significant, $F(1, 76) = 1.71$, $p = .20$. The two-way interaction between time of judgment and strategy condition was not statistically significant, $F(1, 78) = 0.20$, $p = .66$.

Time of judgment, strategy condition, and gender did not interact in their effects on study regulation. Strategy condition and gender interacted in their effects on study regulation. There was a simple main effect of strategy condition on prediction/restudy gammas for males.

Causes of Extinction

1. What are some ways in which humans cause extinction?
2. What can humans do to prevent or minimize extinction?

Cellular Respiration

1. What happens when humans eat food?
2. Why are mitochondria called “powerhouses”?

The Circulatory System

1. What determines the color of blood?
2. Why does our heart beat faster during exercise?

Digestion

1. Why is the stomach important in digestion?
2. What happens during chemical digestion?

The Fight or Flight Response

1. Why is the endocrine system called the fight-or-flight response?
2. What happens when adrenaline is released into the blood?

APPENDIX G

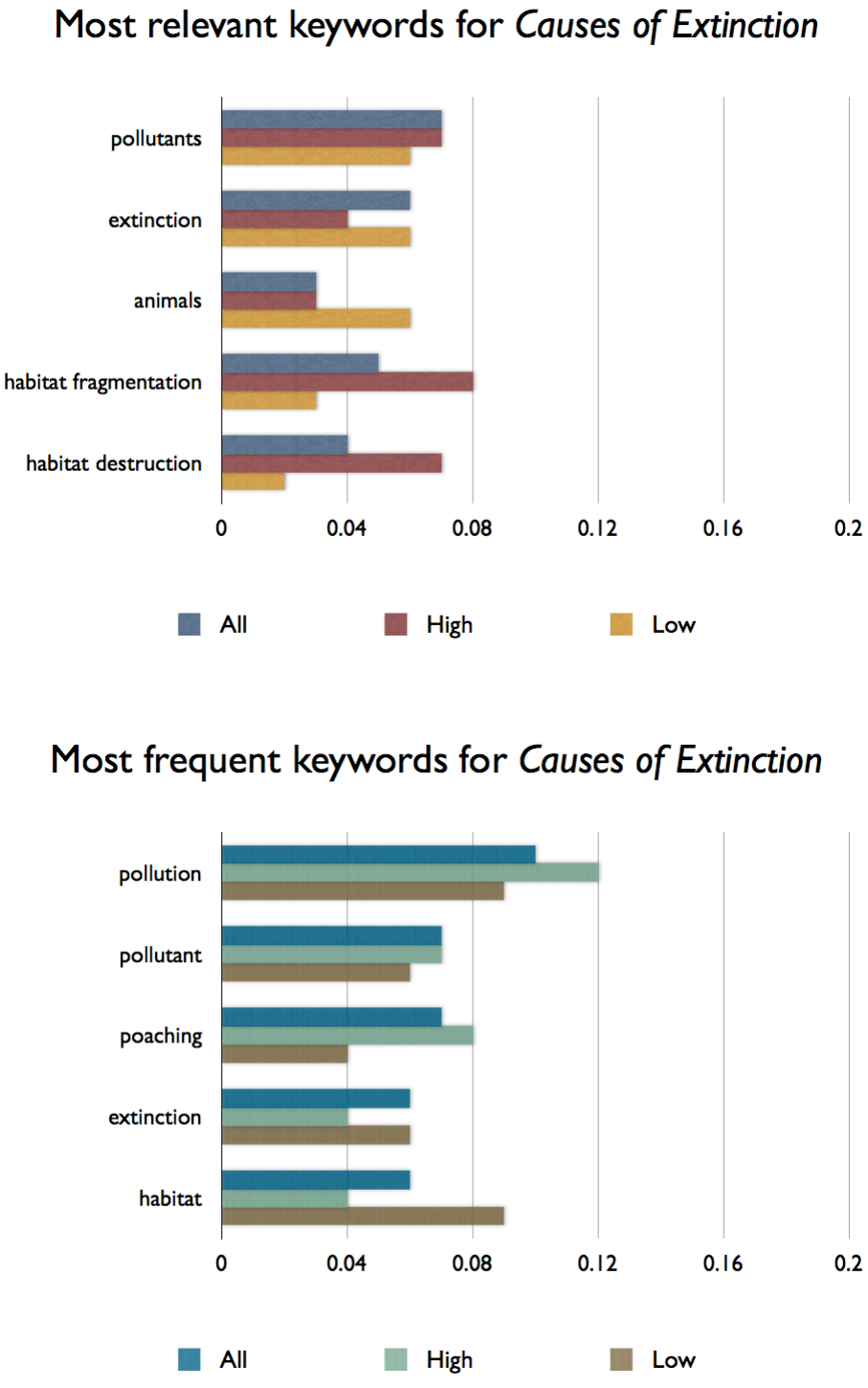
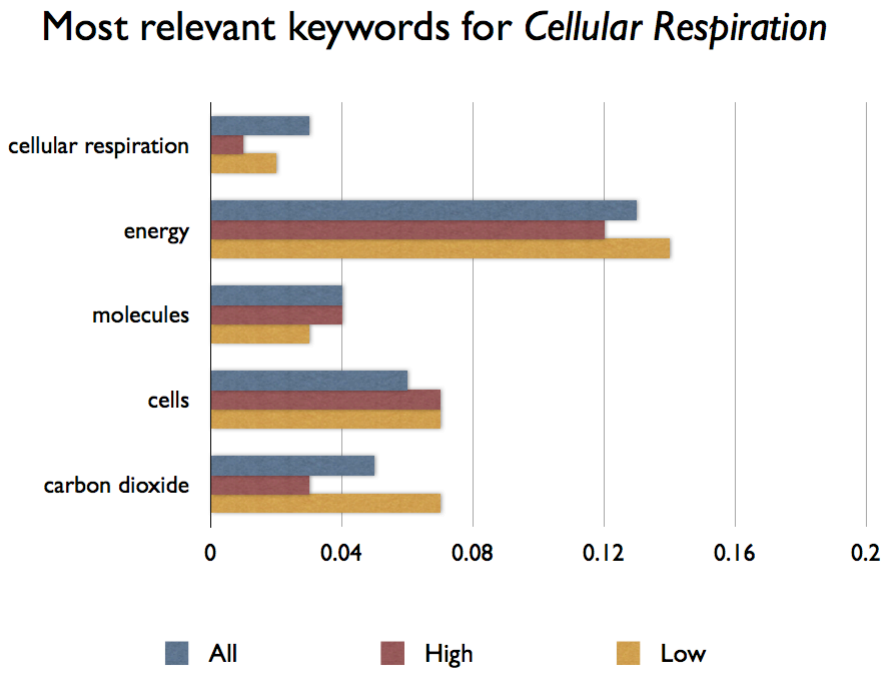


Figure G1. Keyword analysis for *Causes of Extinction.*



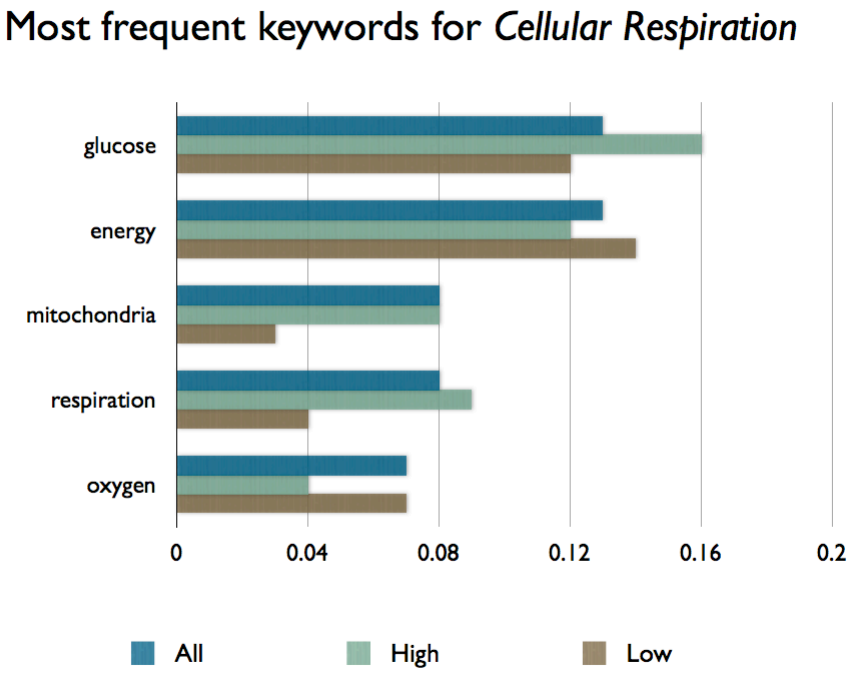
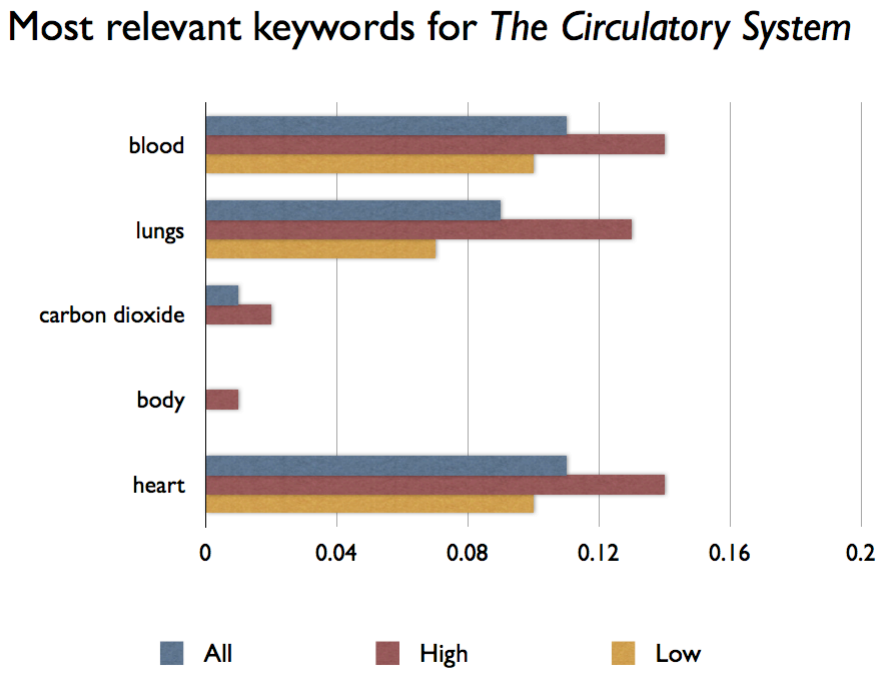


Figure G2. Keyword analysis for *Cellular Respiration.*



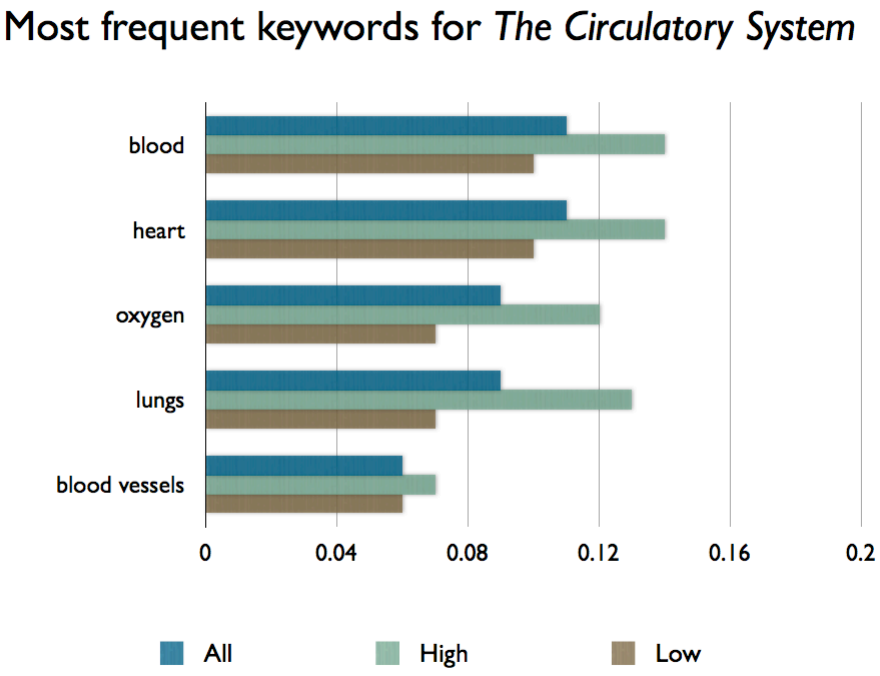


Figure G3. Keyword analysis for *The Circulatory System.*

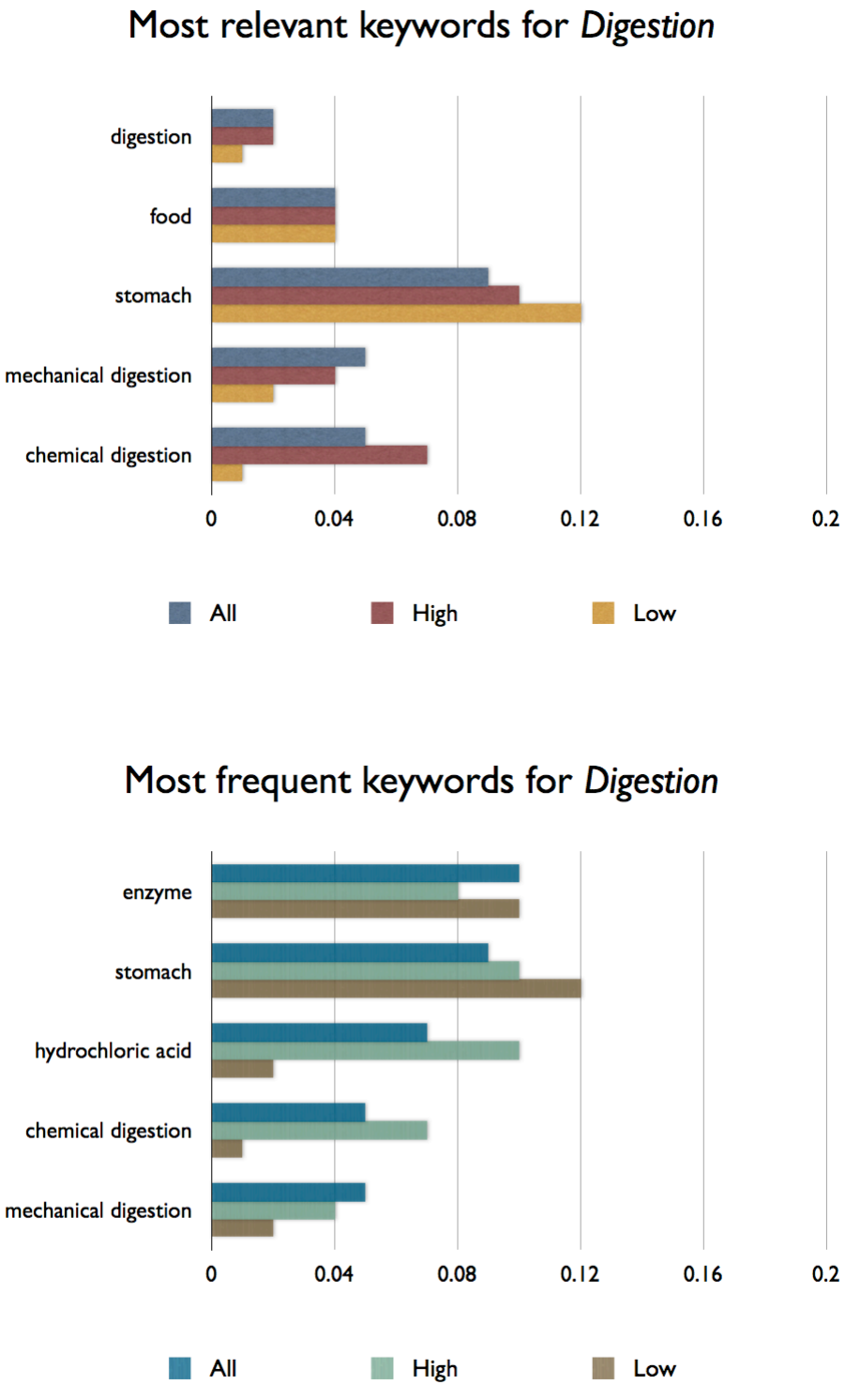


Figure G4. Keyword analysis for *Digestion*.

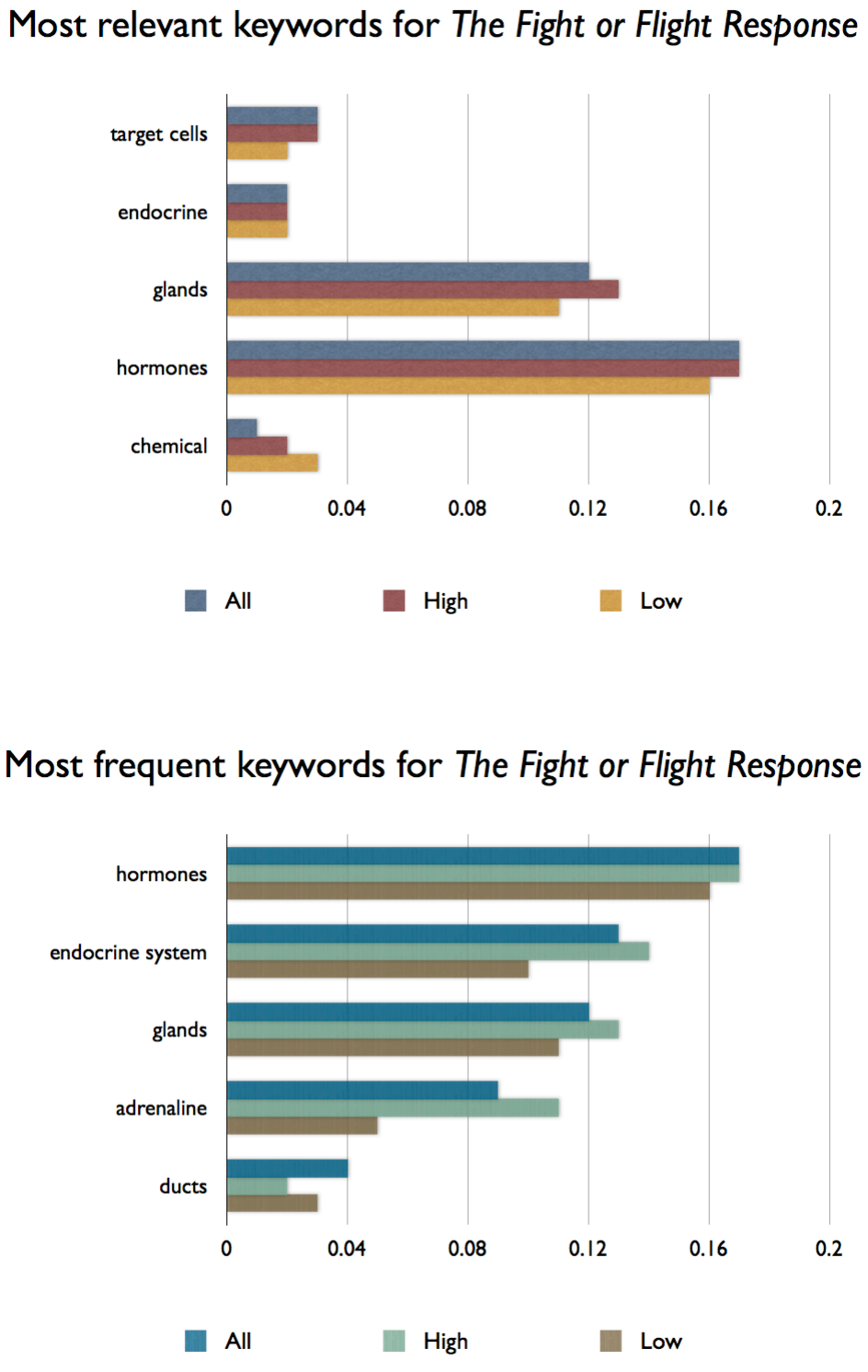


Figure G5. Keyword analysis for *The Fight or Flight Response.*

APPENDIX H

Latent Semantic Analysis (LSA) is a statistical technique for measuring the relationships between words, phrases, and passages (Deerwester, Dumais, Furnas, Landauer & Harshman, 1990). Common educational applications of LSA include selecting appropriate texts for a given learner based on level of reading ability or background knowledge, automatically scoring the content of essays and summaries, and helping students effectively summarize textual material.

Essentially, LSA works by first computing the match between a word and a larger body of text and then by using factor analysis to decompose the matrix of word-by-context data into a smaller set of dimensional factors (Deerwester et al., 1990). For each word that is analyzed using LSA, a value is computed that represents the similarity between that word and a user-specified passage or context.

Using LSA, the keywords generated by students in Experiment 1 were analyzed. Each of the keywords was entered into the one-to-many LSA analyzer (accessible at http://lsa.colorado.edu) along with the corresponding passage. For example, the passage, *Causes of Extinction*, was pasted into the analyzer as a comparison text along with the student-generated keywords for that passage (e.g., "destruction," "pollutants," "poaching," "habitat fragmentation," "hunting," etc.). Each keyword was compared to the passage, and a similarity score was computed for each keyword. Possible similarity scores ranged from -1 to +1.

Table H1 presents the mean similarity score, the range of similarity scores, the mean score on the comprehension test, and the Flesch-Kincaid reading grade level for each of the passages used in the current research.

Table HI
Similarity scores and reading levels for the passages

	Mean Similarity Score	*Range of Similarity Scores*	*Mean Test Score*	*Passage Reading Level*
Causes of Extinction	0.28	(0.07, 0.41)	0.59	7.5
Cellular Respiration	0.44	(0.03, 0.61)	0.51	8.2
The Circulatory System	0.60	(0.00, 0.84)	0.68	6.4
Digestion	0.58	(0.02, 0.81)	0.64	7.2
Fight or Flight Response	0.41	(0.02, 0.53)	0.48	6.7

APPENDIX I

Table I1
Correlations between Predictions and Scores by Study Condition for Experiment 1

Measure	Keyword			Reread		
	n	*M*	*SD*	*n*	*M*	*SD*
Gamma	52	0.09	0.68	49	0.01	0.69
Spearman	52	0.01	1.77	49	0.21	2.11
Kappa	55	-0.04	0.14	54	-0.03	0.16
Pearson	52	-0.42	2.07	49	-0.10	1.23
Fisher Z	52	0.10	0.80	49	-0.05	0.79

Table I2
Correlations between Reflections and Scores by Study Condition for Experiment 1

Measure	Keyword			Reread		
	n	*M*	*SD*	*n*	*M*	*SD*
Gamma	54	0.08	0.60	50	0.15	0.72
Spearman	54	-0.01	0.90	50	0.03	0.96
Kappa	55	0.06	0.23	54	0.00	0.18
Pearson	54	0.04	2.06	50	-0.08	1.35
Fisher Z	52	0.14	0.59	50	0.16	0.77

Gamma

Gamma is a non-parametric measure of correlation that is appropriate for measuring association among ordinal data (Everitt, 1977). Gamma is based on the difference between concordant pairs (C) and discordant pairs (D), and can be computed using the following formula:

$$G = \frac{\text{C - D}}{\text{C + D}}$$

Values for gamma range from -1 to +1, with -1 indicating a perfect negative correlation and +1 a perfect positive correlation. Gamma defines perfect association as weak monotonicity.

Spearman

Spearman's rank order correlation coefficient measures the strength of non-linear, monotonic relationships. It is computed on the ranks of paired measurements on two variables using the following formula, where $d_i = (x_i - y_i)$:

$$R_s = 1 - \frac{6 \Sigma d_i^2}{n(n^2 - 1)}$$

Kappa

Cohen's kappa coefficient is a measure of agreement for categorical items. It measures the agreement between two raters who classify N items into C mutually exclusive categories. Kappa can be computed with the following formula:

$$K = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

$\Pr(a)$ is the observed agreement among raters, and $\Pr(e)$ is the hypothetical probability of chance agreement. If there is complete agreement, then $K = 1$. If there is no agreement, then $K = 0$.

Fisher Z

The Fisher Z-transformation can be applied to Pearson's R when X and Y follow a bivariate normal distribution. The transformation is defined by the following:

$$Z = 0.5 \ln \frac{(1 + r)}{(1 - r)}$$