Effects of Bold Key Terms on Judgments of Learning and Reading Comprehension

by

Andrew Scott Roberts

A dissertation submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama
May 7th, 2016

Keywords: metacomprehension, judgments of learning, perceptual fluency, reading ability

Approved by
Aimee Callender, Assistant Professor of Psychology
Alejandro Lazarte, Associate Professor of Psychology
Ana Franco-Watkins, Associate Professor of Psychology
Chris Correia, Associate Professor of Psychology

Abstract

People often use different types of available information as cues to make inferences about memory (Koriat, 1997). When these cues are unreliable predictors of memory (like font size), metacognitive illusions can ensue (Mueller, Dunlosky, Tauber, & Rhodes, 2014). Extending upon Rhodes and Castel's (2008) findings that perceptual fluency cues within word list can produce metacognitive illusions, this dissertation presents a pilot study and two experiments that attempted to determine whether the presence of bold key terms within reading passages produces similar metacognitive illusions.

The pilot study revealed that regardless of reading ability, bold font increased immediate confidence judgments (i.e., judgments of learning (JOLs)) without increasing comprehension. Building off these findings, Experiment 1 and Experiment 2 incorporated a delayed-JOL paradigm (see Maki 1998a) to attempt to improve judgment accuracy by reducing the likelihood that font would affect judgements. However, both experiments demonstrated low reading ability participants' judgments were still influenced by bold font even after a delay resulting in larger JOLs (there was a significant effect in Experiment 2). Although test performance was significantly better for bold terms in Experiment 1 when key terms were always correct answers during testing, bold did not improve test performance in Experiment 2 when each question's stem contain a specific key term because recall was necessary to answer questions. Across these studies, participants were not particularly accurate at evaluating their memory which suggests bold key terms can produce metacognitive illusions by leading to higher expectations of comprehension than those achieved.

Acknowledgements

This research would not have been possible without the contributions of several people. First and foremost, I would like express my gratitude to Dr. Aimee Callender who took a chance on me and has pushed me to grow over the last several years and achieve more than I thought possible. It may have not always been the easiest journey along the way, but I am glad you allowed me to find my own path and kept me on track to finish what I started.

Secondly, Dr. Ana Franco-Watkins for her ever-present optimism, practical advice, willingness to keep my company in her lab, and for enduring the countless times I referenced JOLs in discussion.

Third, Drs. Alejandro Lazarte and Christopher Correia for your willingness to be a part of my qualifying exam and dissertation committees. Your feedback, questions, advice, and generosity have been greatly appreciated.

Fourth, to my lifelong mentor, friend, teacher, and coach Major Tommie Lee Williams who set the precedent for the type of man I inspire to be and has always been a steadfast reminder of the power of perseverance and integrity. Your words and guidance have often seen me through adversity through the years.

Finally, to my wife and friend, Daniela: thank you for always believing in me and challenging me to dream bigger and better things. I know none of this would have been possible if you had not change me for the better. It is only through your love, support, and motivation that I made it to this point and I look forward to what the future may hold with you by my side.

Table of Contents

List of Tables

Pilot Study

Study 1

Study 2

List of Figures

Chapter 1: Introduction

Within the field of developmental psychology during the 1970s, insights about children's awareness of their own cognitive processes (or lack thereof) provided the origins of metacognition (Nelson, Narens, & Dunlosky, 2004). Described as "knowledge and cognition about cognitive phenomena" (Flavell, 1979, p. 906), *metacognition* is the mechanism individuals use to monitor their cognitive processes across various intellectual domains. Through the years, interest in metacognition has produced an extensive body of research and associated terminology (Veenman, Van Hout-Wolters, & Afflerbach, 2006).

One area where metacognitive research has gained a foothold is in reading comprehension. Within educational settings, reading texts for understanding is an essential form of instruction (Griffin, Jee, & Wiley, 2009; Wiley, Griffin, & Thiede, 2005). *Metacomprehension*, or the ability to accurately judge comprehension, is necessary for students to engage in self-regulated learning (Rawson & Dunlosky, 2002; Rawson, Dunlosky, & McDonald, 2002). As a special form of metacognition (Nelson & Narens, 1990; Weaver & Bryant, 1995), over two decades of research have demonstrated that people are very poor at accurately monitoring their metacomprehension (Dunlosky & Lipko, 2007; Glenberg, Sanocki, Epstein, & Morris, 1987)

Early work during the 1980s demonstrated students' judgments of their comprehension often exceed their actual performance on objective measures (Glenberg, Wilkinson, & Epstein, 1982). To reduce this bias, researchers have attempted to discover learning conditions and study strategies that will facilitate higher levels of metacomprehensive accuracy (Rawson, Dunlosky,

& Thiede, 2000). In spite of these learning interventions, attempts to improve metacomprehension have achieved limited success. One reason why improvements in accuracy have been elusive is individuals may rely on different types of metacognitive cues to predict performance (van Loon, de Bruin, van Gog, & van Merriënboer, 2013). Regardless of what types of cues individuals use, if these cues lack *cue diagnosticity*, they can affect metacognitive predictions without improving test performance (Pyc, Rawson, & Aschenbrenner, 2014) resulting in *illusions of knowing* (Epstein, Glenberg, & Bradley, 1984; Glenberg et al., 1982). In other words, when the types of information individuals use to predict their memory in the future do not accurately reflect what they will remember, these cues can mislead their self-evaluations by inflating or decreasing their confidence. As a result, illusions of knowing can emerge from relying on inappropriate cues to predict memory which can in turn interfere with students' ability to regulate their study choices.

Among the various types of cues that can be used to make metacognitive judgments, bold font (a typographic cue) was investigated in this dissertation to determine how it affects students' judgments and reading comprehension. Despite being widely used by authors to emphasize important content and draw attention to important concepts in textbooks (Lorch, Lorch, & Klusewitz, 1995), there has been limited research on whether these cues are beneficial to readers, especially when individual differences in reading ability are considered. However, recent work by Rhodes and Castel (2008) and Mueller, Dunlosky, Tauber, and Rhodes (2014) suggests that variations in font can produce illusions of knowing. In both studies, when words were easier to perceive because they were presented in larger fonts, participants believed they would be more likely to remember them. However, contrary to their expectations, participants were not better at remembering these larger words compared to words presented in smaller fonts.

Because comprehension depends on whether readers can direct their attention to the most important aspects of text (Gaddy, van den Broek, & Sung, 2001), and given the findings associated with the "font-size effect" (i.e., Rhodes & Castel, 2008), it is entirely plausible students could suffer from similar illusions when words are presented in bold font within text. Bearing this in mind, the goal of this dissertation was to investigate whether differences in fonts used to emphasize key terms within texts also produce metacognitive illusions. In this pursuit, we collected reading ability data from all participants and used this data to classify them into high and low reading ability groups. This allowed us to determine whether variations in reading comprehension skills would contribute to differences in susceptibility to illusions of knowing associated with bold font.

This dissertation includes a review of metacognitive theory and pertinent findings, a pilot study with analyses and discussion, and two additional experiments investigating the effects of key term font on metacognitive judgments and test performance. First, an overview of general and specific metacognitive theories and various measurement of metacognition are presented. Second, theory regarding cue utilization and the effects of different metacognitive cues is discussed. Third, reading comprehension theory and findings regarding individual differences in reading ability are provided. Fourth, a rationale behind this dissertation is provided followed by a pilot study that explored whether bold font is a salient enough cue to produce metacognitive illusions. Based on the pilot study's results and discussion, two additional experiments that incorporated reliable methods from the literature were conducted. Full analyses of these experiments are provided and interpretations of findings and their implications are discussed in length.

General metacognitive components

Within Flavell's (1979) original metacognitive framework, he made a distinction between metacognitive knowledge and metacognitive skills. This division between knowledge and skills has persisted through the years and continues to be the most agreed upon method for classifying metacognitive components today (Veenman et al., 2006).

Metacognitive knowledge. *Metacognitive knowledge* refers to what individuals know about their own cognitive processes and cognition in general (Schraw, 1998) and it is comprised of people's beliefs about how different types of factors interact with each other and influence the outcomes of various cognitive pursuits (Flavell, 1979). These factors include self-perceptions people have about their cognitive capabilities, beliefs about the processing demands imposed by different cognitive tasks, and theories about useful learning strategies (Flavell, 1999).

In addition to these factors, Flavell (1979) proposed that the amount of information available to people during cognitive activities can vary from task to task. As a result, one of the most important aspects of metacognitive knowledge is the ability to monitor what information is currently available while also knowing what variations in the amount of available cues imply about specific task demands. Individuals who can monitor and understand what this information (or lack thereof) implies, can choose more optimal strategies to complete a given task.

Although metacognitive knowledge may prove invaluable for strategizing and optimizing learning efficiency, whether this knowledge is useful will depend on whether individuals make correct assumptions about learning. Because this knowledge can be based on either correct or incorrect beliefs about learning processes (Veenman et al., 2006), individuals may fail to understand content they are trying to learn even when they pay attention to materials (Flavell, 1979). As a result, they can misunderstand and/or fail to develop coherent mental representations (which is an important aspect of reading comprehension and will be discussed below).

Metacognitive skills. Given that beliefs may be inherently incorrect, metacognitive knowledge by itself is not sufficient for achieving one's learning goals. As a result, *metacognitive skills* are equally important because they are comprised of specific activities individuals can engage in to better regulate their learning (Schraw, 1998) including specific procedures and problem-solving techniques (Veenman et al., 2006).

During the learning process, individuals can evaluate their progress and decide whether they are achieving their learning goals (Flavell, 1979). Based on these self-evaluations, metacognitive strategies can be implemented to increase learning by monitoring current progress and making adjustments where needed. Whereas metacognitive knowledge is resistant to change and does not provide feedback about learning progress, metacognitive skills can change and adapt over time because they have a built-in feedback mechanism (Veenman et al., 2006). Therefore, these skills can be acquire and refined through experience, which allows individuals to have more options available to them for pursuing their cognitive (i.e., acquiring new knowledge) and metacognitive goals (i.e., testing one's self) (Flavell, 1979).

Specific metacognitive components

Metamemory. Given the limited amount of time students can dedicate to their studies, it is paramount that they accurately evaluate their learning (Dunlosky & Rawson, 2012; Rawson & Dunlosky, 2007). If students can be taught to discriminate between materials they have learned from those they have not learned, they can allocate more time to mastering the content they understand less (Dunlosky & Lipko, 2007). However, in order to be knowledgeable about memory and use this knowledge to engage in controlled behaviors (like allocating study time), students must be able to accurately monitor the content currently available in memory (Leonesio

& Nelson, 1990).

To illustrate this point, consider the choices a typical student must make when preparing for an upcoming exam. First, he or she has to identify which content should be studied. This self-monitoring process requires *metamemory*, or the knowledge about one's own memory (Leonesio & Nelson, 1990; Nelson & Narens, 1990). Having identified what to study, the student must exert control over their study choices by determining how to allocate their study time appropriately and/or which strategies to use. Within this context, self-monitoring provides information to the student about what content has been learned, and based on the feedback acquired from gauging their own memory, they can decide which course(s) of action need to be pursued (Perfect & Schwartz, 2002).

Framework of metamemory. Extending this classic metacognitive example a step further, making effective study choices requires that there is an exchange of information between the learner and the material being studied. According to Nelson and Naren's (1990) *metamemory framework*, all cognitive processes are comprised of two interrelated levels— the meta-level and the object-level. The meta-level monitors and controls overall cognitive processing, whereas the object-level involves specific mental activities (i.e., allocating attention, strategy use, etc.) that occur when attempting to complete tasks (Zhao & Linderholm, 2008). Within this model, there is a dynamic exchange of information that occurs between levels, in which, metacognitive control and monitoring are represented based on the direction of the flow of information (Nelson & Narens, 1990)

Control. From a top-down perspective, the meta-level modifies the object-level resulting in changes in overall behaviors (Nelson & Narens, 1990). These changes that occur at the object-level could result in initiating new actions, continuing ongoing processes, or choosing to terminate a behavior. Referring to the student preparing for an exam, metacognitive control would relate to the specific study choices they make— begin studying, keep studying, or stop studying. Although these choices are important in completing their learning goals, what is absent at the meta-level is specific feedback about whether these choices are effective (Nelson & Narens, 1990).

Monitoring. Because metacognitive control does not provide feedback about the outcomes of the choices being made, Nelson and Narens (1990) proposed that the object-level provides information to the meta-level about cognitive progress in a bottom-up manner. As such, monitoring occurring at the object-level can result in changes in the processes occurring at the meta-level. When preparing for an exam, depending upon the feedback acquired from monitoring learning at the object-level (perhaps through self-testing or some similar mechanism), the student can decide whether his or her current study strategies are producing sufficient learning. In order to make this determination, several different types of metacognitive judgments may be used to monitor the extent of learning.

Judgments of learning. Although there are several different types of metacognitive judgments that can be used to monitor progress (i.e., feeling of knowing (FOKs), ease of learning (EOLs), etc.), researchers often require participants to make prospective metamnemonic judgments to predict their memory in the future (Schwartz, 1994). When these predictions are made during the process of studying (or shortly thereafter), they are referred to as *judgments of learning* (*JOLs)* (Leonesio & Nelson, 1990; Nelson & Dunlosky, 1991; Nelson et al., 2004; Schwartz, 1994). By measuring how well specific information will be available and/or accessible in memory later (Schwartz, 1994), JOLs provide an index of the ability to encode and retain information (Schraw, 2009b).

Even though metacognitive judgments (JOLs) can be made using dichotomous predictions, the most common approach is to measure confidence using a continuous scale ranging from no confidence to complete confidence (Schraw, 2009a). In evaluating reading comprehension (which will be discussed below), Likert scales with varying levels of understanding or confidence have been commonly utilitized by researchers (see Glenberg et al., 1982; Maki & Berry, 1984; Rawson & Dunlosky, 2007), but to provide more discriminative measures of confidence, continuous scales are often relied on to evaluate students' confidence in their exam performance within classroom environments (see Hacker, Bol, Horgan, & Rakow, 2000; Miller & Geraci, 2011).

Effects on study choices. JOLs are one of the most frequently used metacognitive judgments for investigating self-monitoring (Nelson et al., 2004; Schwartz, 1994). Regardless of differences in scale, these judgments provide a measurement of confidence based on what people *believe* they know while learning (Metcalfe & Finn, 2008). Through research, it has been established that JOLs provide valuable insight into metacognition's regulatory functions because these judgments predict overt changes in study behaviors as well as changes in subjective beliefs related to study materials. For instance, these judgments are negatively correlated with study times and perceived difficulty of to-be-learned materials (Cull & Zechmeister, 1994; Metcalfe, 2009; Thiede & Dunlosky, 1999). More time is often allocated to study items judged to be more difficult because people intuitively know that mastering difficult content requires more effort and time (Son & Metcalfe, 2000); however, when there is insufficient time to study, students will prioritize learning easier items first (Son & Metcalfe, 2000, Experiments 1 & 3).

One of the unique findings about JOLs is these judgments actually have direct effects on study choices students make independent of their actual test performance (Metcalfe & Finn, 2008). When given the opportunity to restudy to-be-learned word pairs, students based their study choices on previous JOLs and disregarded their prior test performance. Therefore, when students *believed* they knew these word pairs (i.e., made large JOLs initially), they were less likely to choose to restudy them even when their previous test performance was poor!

Consistent with these findings, Hacker et al. (2000) observed an almost identical pattern of behavior in a classroom setting. Throughout the semester, students judged their performance on each exam. Surprisingly, compare to prior judgments, actual performance had less of an effect on judgments made on subsequent exams. As a result, many students did not adjust their study strategies enough to improve their performance and continued to make inaccurate self-assessments, which is analogous to illusions of knowing.

These findings by Metcalfe and Finn (2008) and Hacker et al. (2000) lend credibility to Nelson and Dunlosky's (1991) assertion that "the accuracy of JOLs is critical because if JOLs are inaccurate, the allocation of subsequent study time will correspondingly be less than optimal" (p. 267). Both examples illustrate students' failure to accurately monitor their memory which contributed to inefficient self-regulated learning. Following testing, feedback about metacognitive inaccuracies should have facilitated corrective behaviors on subsequent tests, but this feedback was disregarded. Consequently, students did not sufficiently adjust their expectations (lowered JOLs) or adopt more effective study strategies to match (or exceed) their perceived levels of competence, which represents a failure in metacognition.

Metacomprehension. In addition to metamemory, another metacognitive component that researchers are interested in is *metacomprehension* (Rawson & Dunlosky, 2002). Self-regulated learning from texts requires the ability to accurately judge comprehension (Rawson & Dunlosky, 2002; Rawson et al., 2002), which makes metacomprehension an important research area.

Historically, “metamemory for text” has been described as a special form of metacognition (Nelson & Narens, 1990; Weaver & Bryant, 1995) because metacomprehension utilizes both control and monitoring processes. In fact, Maki and Serra (1992) demonstrated that Nelson and Narens’ (1990) *theoretical framework of metamemory* can be successfully applied to evaluate metacomprehension. That is, as students obtained more knowledge while reading, their predictions of future test performance became more accurate because they were evaluating the amount of information they acquired from reading (Maki & Serra, 1992; Weaver & Bryant, 1995).

However, unlike predicting memory using word lists or paired associates (see Arbuckle & Cuddy, 1969; Leonesio & Nelson, 1990; Nelson & Dunlosky, 1991), predicting memory for texts is a more complex process because it entails judging comprehension (metacomprehension) as well as predicting future memory (metamemory) (Maki & Berry, 1984; Rawson et al., 2002). Although metamemory and metacomprehension share similar conceptual roots, Wiley et al. (2005) advocated that if memory of text and comprehension are distinct psychological phenomena (as they are proposed to be), then researchers interested in metacomprehension need to consider additional methodological issues not addressed within the scope of metamemory research.

With that being said, metacomprehension researchers have often adopted methodologies similar to those used to investigate metamemory. By examining the relationship between JOLs and test performance, these researchers have emphasized the need for accurate monitoring processes in order to promote more efficient, self-regulated learning (Chiang, Therriault, & Franks, 2010; Dunlosky, Rawson, & Middleton, 2005). Although successful monitoring can improve learning, inaccurate self-monitoring attempts can produce poor study choices by compromising efficiency (Maki, Willmon, & Pietan, 2009). Overconfident students may study insufficiently while underconfident students may waste time studying information they already know (Maki et al., 2009; Son & Metcalfe, 2000; Thiede & Dunlosky, 1994; Thiede & Dunlosky, 1999; Wiley et al., 2005).

Gamma. Heeding Nelson's (1984) recommendation, it is common practice to evaluate the accuracy of metamemory and metacomprehension judgments by examining the correspondence between individual judgments and performance on corresponding items by computing nonparametric gamma correlations (Maki, Shields, Wheeler, & Zacchilli, 2005). Gamma can range from -1 to 1 with -1 indicating a perfect negative relationship between JOLs and performance and +1 indicating perfect predictive accuracy (Maki, 1998a; Nelson & Dunlosky, 1991). When high JOLs are provided for remembered items and low JOLs are given for items forgotten, gamma will be positive (Serra & Metcalfe, 2009). But if high JOLs correspond with forgetting and low JOLs predict remembering, the resulting gamma will be negative indicating an imperfect relationship between judgment and performance. As a result, gamma measures *relative accuracy* by examining whether judgments correlate with test performance (Dunlosky & Lipko, 2007).

To approximate how students naturally prepare for exams (Linderholm, Zhao, Therriault, & Cordell-McNulty, 2008), most metacomprehension studies have investigated the relationship between control and monitoring by requiring participants to read a series of passages, rate their comprehension for each one, and then complete a comprehension test (Rawson et al., 2000). It has become clear from these research studies that people are generally poor at monitoring what they know (Dunlosky et al., 2005). Metacomprehension accuracy is typically poor (Dunlosky & Lipko, 20007; Glenberg et al., 1987) with gamma means hovering around .25 out of a max of 1.0

(Griffin, Wiley, & Thiede, 2008; Maki, 1998b; Thiede & Anderson, 2003; Thiede, Griffin, Wiley, & Anderson, 2010). Students are particularly bad at accurately judging their comprehension when reading expository texts (Thiede, Wiley, & Griffin, 2011; Weaver & Bryant, 1995) with correlations between their judgments and performance rarely exceeding +.40 (Dunlosky et al., 2005).

Bias**.** Despite their endorsement by Nelson (1984) and prevalence within metacognitive literature, gamma correlations have some limitations. Specifically, gamma provides a measurement of the ability to discriminate between materials that produce higher performance from those that produce poorer performance (Maki et al., 2005). However, gamma does not indicate whether individuals are overconfident or underconfident when making metacognitive judgments.

Consequently, some researchers have argued that measurements like gamma obtained through laboratory studies may not accurately reflect how students evaluate or “calibrate” their study behaviors in classroom settings (Hacker, Bol, & Kenner, 2008). Given that overconfidence is a robust phenomenon (see Fischhoff, Slovic, & Lichtenstein, 1977; Keren, 1991; Koriat, Lichtenstein, & Fischhoff, 1980), calculating *bias* is also common practice within educational research because it provides a measurement of *absolute accuracy*, or the degree to which people are over- or underconfident in their cognitive abilities (Dunlosky & Lipko, 2007; Maki et al., 2005).

By subtracting actual performance from JOLs (Maki et al., 2005; Schraw, 2009a), bias produces a range of scores in both positive and negative directions (Schraw, 2009a). Positive values indicate overconfidence because JOLs exceeded actual performance, while negative values demonstrate underconfidence because JOLs are lower than performance (Maki et al., 2005). Compared to gamma, bias demonstrates both the direction and the magnitude of discrepancies that exist between individuals' JOLs and performance (Schraw, 2009a).

To be clear, measurements of relative and absolute accuracy assess different aspects of metacognition (Dunlosky & Lipko, 2007). Applying these within Nelson and Naren's (1990) model, gamma evaluates metacognition at the object-level (i.e., item-by-item basis) while bias examines metacognition at the meta-level (i.e., whether overall perceptions match overall performance). In agreement with Schraw's (2009a) assertion that researchers should use multiple measures whenever possible, the experiments conducted in this dissertation have investigated metacomprehension accuracy using both gamma and bias to determine whether bold typographic cues affect JOLs at the concept-level and/or overall comprehension.

Illusions of knowing. Glenberg et al. (1982) suggest that self-assessments of comprehension from texts often exceed performance on objective measures of comprehension. In their experience, students are often surprised by their lower-than-expected performance despite claiming to be prepared for their exams. When such disparities exist between subjective beliefs about comprehension and actual test performance, students demonstrate illusions of knowing (Epstein et al., 1984; Glenberg et al., 1982).

The occurrence of illusions of knowing illustrate students' propensity to be overconfident when evaluating their comprehension (Glenberg & Epstein, 1985). This overconfidence is particularly prevalent among the poorest performing students and persists even when they are provided revised texts designed to improve comprehension (Maki et al., 2005). Even though these texts improve test performance, they also increased the magnitude of overconfidence among all readers. That is, revised texts increase students' confidence more than it improved their test performance.

In each of these examples, overconfidence is a consequence of inadequate reading comprehension skills. Many readers fail to notice when contradictory information is present within passages (Glenberg et al., 1982) and do not make accurate predictive inferences (Glenberg & Epstein, 1985). Making predictive inferences while reading is important for comprehension (Long, Oppy, & Seely, 1994) and similarly, self-correcting is an essential metacomprehension process (Maki, Jonas, & Kallod, 1994). Therefore, illusions of knowing for texts demonstrate some readers may lack fundamental comprehension skills that prevent them from making accurate judgments of their comprehension. In this context, whatever cues they rely on to judge comprehension are insufficient indicators of their overall level of understanding. As a result, these readers fail to make the adjustments necessary to improve their comprehension.

Cue-utilization: judgments of learning

According to the *cue-utilization framework of metacognitive monitoring* (Koriat, 1997), poor metacognitive accuracy results from individuals attempting to predict future performance by relying on cues that lack predictive validity (van Loon et al., 2013). Although this theory was originally conceptualized to account for findings within metamemory paradigms, it has provided a useful account for the potential mechanisms that may underlie metacomprehension accuracy (Jaeger & Wiley, 2014). Bearing this in mind, illusions of knowing can emerge when JOLs are influenced by available cues that are not reliable predictors of future memory (Mueller et al., 2014; Rhodes & Castel, 2008).

Given the importance of evaluating comprehension, Rawson and Dunlosky (2002) proposed that identifying types of cues that affect metacomprehension judgments is imperative. If metacognitive judgments are inferential assessments derived from available cues, the more these cues overlap with subsequent criterion performance, the more accurate these assessments will be (Koriat, 1993; Rawson & Dunlosky, 2002).

JOLs can be influenced by three different types of cues: intrinsic, extrinsic, and mnemonic (Koriat, 1997). Intrinsic cues emerge from properties and characteristics of the content being studied that discloses information about the ease or difficulty of learning (Castel, 2008; Koriat, 1997; Pyc et al., 2014). Based upon "inherent attributes" that exist independent of experimental conditions (Koriat, 1997), these cues affect judgments of learning for paired-associates (i.e., Arbuckle & Cuddy, 1969; Nelson & Leonesio, 1988; Underwood, 1966) and self-evaluations of reading comprehension (i.e., Rawson & Dunlosky, 2002; McCabe, Kraemer, Miller, Parmar, & Rusica, 2006; Serra & Dunlosky, 2010; Weaver & Bryant, 1995).

With regard to reading comprehension, the visual format and configuration of texts is an intrinsic cue that affects metacognitive judgments. More visually appealing texts produce larger JOLs compared to less visually appealing texts (McCabe et al., 2006; Serra & Dunlosky, 2010). Additionally, the inherent difficulty of texts (typically measured by readability formulas) can also systematically influence judgments (Rawson & Dunlosky, 2002; Weaver & Bryant, 1995). Because text difficulty is associated with ease of processing and coherence—how well readers make connections between propositions within their mental representations of text (Benjamin, 2012)—easier to read texts are often given larger JOLs. Considering less skilled readers may require more cohesive texts in order to understand content (Vitale & Romance, 2007), improving readability by revising texts can increase comprehension and improve monitoring accuracy (Weaver & Burns, 1990; Weaver, Bryant, & Burns, 1995).

In contrast to intrinsic cues, extrinsic cues are based on factors related to either the learning conditions present or the types of encoding strategies used (Koriat, 1997; Zaromb, Karpicke, & Roediger, 2010). In many cases, these cues are directly related to the experimental manipulations being investigated by researchers including rereading texts (Rawson et al., 2000), summarizing texts (Thiede & Anderson, 2003), and generating lists of key words (Thiede, Anderson, & Therriault, 2003; Thiede, Dunlosky, Griffin, & Wiley, 2005). Although extrinsic cues should provide more diagnostic cues of memory because they are based on encoding processes, these cues are often discounted more relative to intrinsic cues (Koriat, 1997; Pyc et al., 2014; Zaromb et al., 2010). As a result, people pay more attention to their beliefs about the ease/difficulty of materials while studying and less to how they actually study when making JOLs.

From a theoretical perspective, the underlying logic behind manipulating experimental conditions is to encourage participants to attend to more mnemonic cues in order to improve metacognitive accuracy (Koriat, 2012). Based on subjective experiences related to learning (Pyc et al., 2014), mnemonic cues include processing fluency, accessibility of information, and ease of retrieval (Koriat, 2011). Although these cues should indicate whether materials have been learned and can be recalled in the future, mnemonic cues may be overshadowed by intrinsic and/or extrinsic cues that have indirect effects on subjective perceptions about learning (Koriat, 1997). As a result, the accuracy of metacognitive judgments may vary from situation to situation depending on which cues these judgments are based on.

To illustrate the relationship between these different types of cues, consider the following situation. The inherent difficulty of a text (intrinsic cue) could force a student to engage in specific learning strategies (extrinsic cues) and depending on the extent to which information is retained (mnemonic cue), the magnitude and subsequent accuracy of their JOLs for this content could vary immensely. The problem with relying on currently accessible cues to predict future memory is these cues may not be accessible during retrieval attempts (Koriat, 1997; Koriat &

Bjork, 2005). If JOLs exceed memory, people are described as being overconfident in their ability to remember (Koriat & Bjork, 2005; Nelson & Dunlosky, 1991) which could be attributed to them relying on available cues that are inaccessible during future retrieval attempts.

Processing fluency. Among the various types of intrinsic cues that can influence JOLs, processing fluency, or the ease with which information is processed and/or retrieved (Hilbig, Erdfelder, & Pohl, 2011; Lanska, Olds, & Westerman, 2014), is a particularly influential metacognitive cue (Alter & Oppenheimer, 2009b). Processing fluency varies across all cognitive tasks (Alter & Oppenheimer, 2009a), and although this cue can be informative, it can also misguide JOLs because individuals may make incorrect inferences about learning (Koriat & Ma'ayan, 2005). These inferences may stem from naïve theories they have which are used to interpret the effects of fluency (Schwarz, 2004). In turn, these interpretations can impact subjective learning experiences and alter underlying metacognitive processes (Koriat, Nussinson, & Ackerman, 2014).

When information is processed more fluently, it may evoke feelings of familiarity (Schwarz, 2004). These feelings of familiarity may depend upon the types of inferences people make about the ease of processing (Jacoby & Whitehouse, 1989). Research has demonstrated that familiarity influences metacognitive judgments (Koriat 1997; Metcalfe, Schwartz, & Joaquim, 1993) because when information is presented in a more fluent manner— it is judged to be more familiar (Miele & Molden, 2010). However, familiarity may not be indicative of whether actual memory for the content being judged exists (Jacoby, Kelley, & Dywan, 1989; Whittlesea, 1993).

Perceptual fluency. One method that has been used to investigate the effects of processing fluency is through manipulating *perceptual fluency*, or the ease with which information can be perceived during study (Lanska et al., 2014). For example, Rhodes and Castel (2008) demonstrated that differences in the font-size of words within word lists affect metacognitive judgments independent of memory. Specifically, significantly larger JOLs were given to words presented in 48 pt. font compared to words presented in 18 pt. font. In spite of these differences in JOLs, there were no differences in recall based on font size. Thus, font size provided a fluent cue that was not diagnostic of recall and resulted in metacognitive illusions.

Mueller et al. (2014) determined that the "font-size effect" observed by Rhodes and Castel (2008) was not caused by differences in processing fluency per se, but rather it is due to participants' beliefs about the effects of processing fluency. They reported that study times did not vary as a function of font and there were no differences in recognition rates due to font during a lexical decision making task. These findings indicate font size provided no benefits to either encoding or retrieval fluency. However, self-reports from participants demonstrated they believed they would be better at recalling larger words because they are easier to remember. In other words, participants had pre-existing beliefs about the effects of font size on their memory (i.e. encoding fluency) and these beliefs persisted even when JOLs were made before each word was presented.

Encoding vs. retrieval fluency. In light of the effects of fluency on metacognition, Koriat and Ma'ayan (2005) proposed that JOLs are influenced by two specific types of processing fluency— encoding fluency and retrieval fluency. Encoding fluency affects the ease of learning and is associated with the amount of time and effort dedicated to studying materials (Koriat & Ma'ayan, 2005; Koriat, Ma'ayan, & Nussinson, 2006). In contrast, retrieval fluency depends on how easily information can be accessed from memory (Benjamin & Bjork, 1996).

Depending on when JOLs are made, individuals may rely more heavily on encoding fluency or retrieval fluency (Koriat & Ma'ayan, 2005). Immediate JOLs made after studying are susceptible to the effects of encoding fluency and how learners interpret their study effort (Koriat et al., 2014) because this is the most accessible cue (Koriat & Ma'ayan, 2005). But following delays, memory related to study effort decays and individuals shift toward retrieval fluency—whether or not information is accessible when making JOLs.

Direct-access vs. inferential view. The divergent effects of fluency on metacognition are manifested in the types of information (cues) used to evaluate memory. According to the *direct access view of metacognition*, monitoring judgments are sensitive to the strength of the memory traces about the items/content being judged (Schwartz, Benjamin, & Bjork, 1997). This theory assumes that when monitoring memory, people attempt to retrieve information directly related to a specific target (i.e., target-based information) (Schwartz, 1994). In terms of JOLs, direct access implies that variations in encoding create differences in memory strength which then contribute to variability in judgment magnitude (Cohen, Sandler, & Keglevich, 1991; Koriat, 1997). Bearing this in mind, direct-access depends on retrieval fluency and the types of mnemonic cues used to evaluate memory (Koriat & Ma'ayan, 2005). Assuming these cues remain accessible over time, study methods designed to improve memory should also improve metacognitive accuracy (Schwartz et al., 1997).

In contrast, the *inferential view* does not assume that metacognitive judgments depend on whether people have direct access to information related to the content being judged (Schwartz et al., 1997). Instead other sources of available information (i.e., ease of processing, familiarity, etc.) are attended to and used to make inferences about the extent to which content will be remembered (Koriat, 1997; Schwartz et al., 1997). Because these inferences can be based on encoding fluency, JOLs depend on people's beliefs about how different available cues relate to future memory (Koriat, 1997). Therefore, the more these cues are predictive of memory, the more accurate JOLs will be.

Typographical cues. An important question of interest in this dissertation is whether manipulating encoding fluency by using a common typographical cue (i.e., bold font) affects metacognitive judgments and reading comprehension. Typographical cues (i.e., underlining, italicizing, and/or printing words in bold) are designed to make certain words more distinctive in order to make it easier for readers to identify main points (Gaddy et al., 2001).

Readers learn to use different cues within texts to aid comprehension (Kintsch & Yarbrough, 1982) and even though typographic cues seem to improve recall for cued information (Gaddy et al., 2001; Lorch et al., 1995), what is not currently clear is how these cues affect metacomprehension or the propensity to which students depend on these cues while reading. Although bold font may make reading individual words easier (Krulee & Novy, 1986) and improve recall in some situations (Foster & Coles, 1977; Margolin, 2013; Marks, 1966), some individuals may be more proficient at monitoring surface memory for text itself (metamemory) and less effective at monitoring their comprehension (Wiley et al., 2005). Bearing this in mind, typographical cues could contribute to metacognitive illusions if these cues are not predictive of overall comprehension.

Construction-integration model (CI)

Despite the fact metamemory and metacomprehension share similar basic memory processes, understanding texts requires additional complex comprehension and encoding processes (Kintsch & Welsch, 1991). Regarded as the most prominent and comprehensive model of reading comprehension (Graesser, 2007; Wiley et al., 2005), Kintsch's (1988) *construction integration model (CI)* proposes that texts are processed in series of cycles during which individuals construct mental representations by integrating different “chunks” of text together to form larger propositional networks (Singer & Kintsch, 2001).

During the initial processing of a text, an exact replication of the phrasing of sentences is retained (Graesser, 2007). From this surface code, specific propositions that preserve the meaning of a text are extracted to create a textbase. Although the surface code and textbase are sufficient for reproducing a text, remembering a text and learning from it are not equivalent learning outcomes because comprehension requires a deeper level of understanding (Kintsch, 1994).

To achieve comprehension, a situation model must be built by making connections between expressed relationships from within a text and integrating them with prior knowledge (Wiley et al., 2005). Because comprehension relies on the interaction of bottom-up and top-down processes that occur concurrently, prior knowledge is integral to interpreting the overall meaning of a text (Kintsch, 2005). Once a situation model is obtained for a text, this newly acquired knowledge can then be applied to novel situations (Wiley et al., 2005).

Reading ability. Differences in reading ability can affect comprehension. Whereas skillful readers are adept at remembering information stated explicitly within text (Hannon, 2012; Masson & Miller, 1983) and applying prior knowledge to make inferences (Hannon & Daneman, 1998; Singer & Ritchot, 1996), less skillful readers struggle to integrate new ideas presumably because they do not construct accurate propositional networks (Long et al., 1994). As a result, less skillful readers’ mental representation are more fragmented because they shift their focus too often when reading (Gernsbacher, Varner, & Faust, 1990) while quickly losing access to recently read information (Gernsbacher & Faust, 1991; Sachs, 1967). As a consequence, these readers do not have access to well-developed and organized situational models when reading (Murray & Burke, 2003) which can prevent them from achieving comprehension.

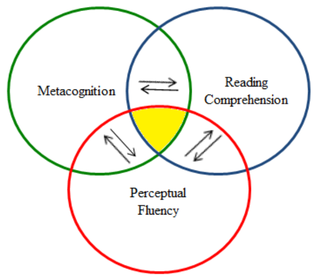
In support of these claims, Griffin et al. (2008) demonstrated that metacomprehension accuracy among typical college readers varied as a function of reading ability, which resulted in an interaction between comprehension and rereading on monitoring accuracy. During a single reading attempt, low ability readers' monitoring was substantially poorer compared to high ability readers. When low ability readers were allowed to reread texts, their metacomprehension accuracy improved to levels comparable to high ability readers. Although rereading did not improve test performance, it did enable less proficient readers, who struggle with basic comprehension processes, to acquire and access more information from the object-level in order to evaluate their comprehension.

Given these differences in reading comprehension processes between readers, if individuals rely initially on inappropriate cues to predict their performance, their metacomprehension judgments will be inaccurate (Thiede et al., 2010). This is indeed true of at-risk adult readers (students enrolled in remedial reading courses), who compared to proficient college readers, frequently reported using surface cues (i.e., text readability, vocabulary difficulty, and passage length) when making metacognitive judgments (Thiede et al., 2010).

Although Thiede et al. (2010) demonstrated at-risk readers' propensity to rely on surface cues to predict comprehension, higher ability readers may also use similar surface-level cues to make metacomprehension judgments (Griffin et al., 2008; Thiede & Anderson, 2003). For example, Roberts and Callender (2014) determined that both high and low reading ability participants' JOLs were significantly larger for key terms printed within passages in a big-bold font (28 pt. bold Arial) compared to standard font (12 pt. Arial). Even though high reading ability participants recalled more content than low reading ability participants, there were no differences in recall based on font type for either group. In other words, both groups' JOLs were influenced by surface cues (typographical font) which resulted in metacognitive illusions.

Rationale for current studies

Figure 1. Representation of the relationship between metacognition, reading comprehension and perceptual fluency



The purpose of the following studies is to extend upon previous findings to investigate how perceptual fluency cues affect metacognitive judgments during reading comprehension. As illustrated in Figure 1, metacognition, reading comprehension, and perceptual fluency each share a bidirectional relationship with adjoining concepts; thus, one area can influence the other and vice versa. Although metacognition can guide reading comprehension processes, the types of metacognitive strategies and cues a reader uses to evaluate their comprehension may depend on reading ability. Conversely, perceptual fluency can help facilitate comprehension by signaling important concepts, but if readers attend to these cues at the expense of processing surrounding content, they may develop incomplete situation models. Lastly, when monitoring one's knowledge, perceptual fluency cues can provide a basis for judging understanding, but if these cues are unreliable predictors of future memory, then readers may develop illusions of knowing.

Ultimately, the area highlighted in the figure above represents the intersection between metacognition, reading comprehension, and perceptual fluency cues. By manipulating key term font within reading passages, the experiments contained in this dissertation will investigate how perceptual fluency cues may (or may not) influence JOLs and whether such cues affect reading comprehension. Depending on the results obtained, these studies could provide significant implications ranging from designing optimal reading materials to providing students with more effective reading comprehension strategies.

Pilot Study

Although Roberts and Callender (2014) demonstrated typographical cues influence judgments of metacomprehension, it is unclear whether the metacognitive illusions observed would persist with a less extreme typographic cue. Therefore, a pilot study was conducted to investigate whether presenting key terms in bold without changing font size would result in similar increases in JOLs without improving memory. By incorporating a more traditional typographical cue, it is possible to determine whether the previously observed metacognitive illusions from Rhodes and Castel (2008) and Roberts and Callender (2014) were attributable to the extreme disparity in font size between the to-be-judged content.

Method

Participants

Fifty undergraduate psychology students received extra credit for participating in the pilot study. A 2 x 2 mixed design with two levels of reading ability (high vs. low) as a between-subject factor and two levels of key term font (bold vs. standard) as a within-subject factor was used. Dependent variables of interest were global JOLs, key-term specific JOLs, final multiple-choice test performance, bias (difference between JOLs and performance), and gamma.

Materials

MMCB. Reading ability was evaluated using a computer-based version of the reading comprehension section from the Multi-media Comprehension Battery (MMCB; Gernsbacher &

Varner, 1998). The MMCB has good predictive validity and provides a measure of general comprehension skills (Maki & Maki, 2002) and scores from the reading comprehension section correlate highly with verbal SAT (Gernsbacher et al., 1990). The version administered in the current study was comprised of four short stories with a series of twelve multiple choice questions immediately following each story. Scores on this measure could range from 0 to 48 and using a median score of 33, participants were classified into either a low reading ability group (scores of 33 and below) or high reading ability group (scores of 34 and up).

Reading passages. Twelve nonfiction passages (≈ 250 words per passage) on a variety of topics (famous people, places, diseases, etc.) and corresponding test items obtained from the Graduate Record Examination (GRE) and Test of English as a Foreign Language (TOEFL) previously used by Roediger and Marsh (2005) were modified for use in the current study. Within each passage, four key terms were presented in either 12 point bold Arial (bold) or 12 point Arial (standard). These terms could occur anywhere within a given passage. Key term font was counterbalanced across all passages (six passages per font) and divided into two equal blocks. Each block contained three passages per font and passage order was randomized.

Final test. A 48-item multiple choice comprehension test (one question per term; four questions per passage) was administered to all participants. Key terms were always the correct answer to corresponding test items.

Procedure

During an initial session, all participants completed the MMCB to evaluate their reading ability. Based on a median split, scores on this measure were used to classify participants into high and low reading ability groups used during subsequent analyses.

After consent was obtained, participants were randomly assigned to one of two counterbalanced conditions with equal number of passages with and without bold key terms. Participants received verbal instructions that they would read several short passages and make confidence judgments based on what they read following each passage. Reading was self-paced although there was a one hour time limit to complete the experiment.

Each passage was presented individually. Immediately after reading each passage, the following prompt appeared on-screen: "*You will now be asked to make several predictions about your understanding of specific concepts from the passages you just read. For each concept, indicate how confident you are in your ability to answer a question about this concept using any number on a scale from 0 to 100 with a 100 indicating 100% confident in your ability to answer a question about this concept."*

After receiving these instructions, participants were asked to make an assessment of their comprehension for the entire passage (i.e., global-level JOLs) as well as for each key term (i.e. term-specific JOLs) using a method adapted from Rawson and Dunlosky (2007). For global JOLs, participants were provided a Likert scale (0%, 25%, 50%, 75%, and 100%) to choose from to indicate their confidence in the ability to successfully answer questions about a specific reading topic. The values provided on this scale were intended to match the actual level of performance that could be achieved for a given topic based on the four corresponding test items.

For the term-specific JOLs, each key term was presented one-at-a-time on-screen in standard 12 pt. font (no bold fonts were used during these prompts to prevent biasing these judgments). Participants then submitted a confidence judgment for each term before the next term appeared. Once all key terms from a particular passage were judged, a new reading passage would appear and this procedure repeated until all 12 passages were completed.

After providing judgments for the last passage, participants were dismissed from lab and reminded to return two days later. During the return visit, a multiple choice comprehension test was given to participants. Each question appeared on-screen with four answer options and participants had to respond to each question before the next question would appear. When participants had completed this assessment, they were dismissed from the lab.

Results

Prior to analyzing the data, a median split based on MMCB scores (median = 33) was conducted to classify participants into high reading ability (HRA) (n = 26) and low reading ability (LRA) (n = 24) groups used during subsequent analyses.

JOLs

Global JOLs. To determine whether the presence of bold font within texts affected overall assessments of comprehension, average global JOLs per font were calculated for each participant (see Table 1 for means). A 2 (Font: bold vs. standard) x 2 (Reading Ability: high vs low) mixed Analysis of Variance (ANOVA) revealed no significant effects of font, $F(1, 48) = .75$, $p = .39$, $\eta_p^2 = .02$, or reading ability, $F(1, 48) = 1.24$, $p = .27$, $\eta_p^2 = .03$, and no significant interactions between font and reading ability, $F(1, 48) = .53$, $p = .47$, $\eta_p^2 = .01$. The presence of bold font within passages had no effect on HRA's global JOLs and produced only a slight increase in LRA's global JOLs.

Table 1
Global Judgments of Learning (JOLs) by Reading Ability and Font

	Bold		Standard	
Reading Ability	*M*	*SE*	*M*	*SE*
High	71.64	2.52	71.47	3.12
Low	68.15	2.62	66.25	3.24

Term-specific JOLs. To investigate whether font affected metacomprehension judgments for key terms, average JOLs per font were calculated for each participant (see Table 2 for means). A 2 (Font: bold vs. standard) x 2 (Reading Ability: high vs. low) mixed ANOVA revealed a significant main effect of Font, $F(1, 48) = 15.96$, $p < .001$, $\eta_p^2 = .25$, but no significant effect of reading ability, $F(1, 48) = .98$, $p = .33$, $\eta_p^2 = .02$, or interaction, $F(1, 48) = .70$, $p = .41$, $\eta_p^2 = .01$. Overall, JOLs were significantly larger for bold key terms (*M* = 76.64, *SE* = 1.97) compared to standard key terms (*M* = 72.16, *SE* = 2.41) regardless of reading ability.

Table 2
Term-specific Judgments of Learning (JOLs) by Reading Ability and Font

	Bold		Standard	
Reading Ability	*M*	*SE*	*M*	*SE*
High	79.22	2.73	73.80	3.34
Low	74.06	2.84	70.51	3.48

Test Performance

Items on the final test were divided into two groups based on font (24 items per font). Due to a programming error, the correct answer to one test item was not a key term and was excluded from analysis. Depending on which counterbalancing sequence a participant was assigned to, they would have responded to either 23 bold and 24 standard key term questions or 24 bold and 23 standard key term questions. Test performance was averaged across participants by font (see Table 3). A 2 (Font: bold vs. standard) x 2 (Reading Ability: high vs. low) mixed ANOVA on test performance revealed a significant effect of reading ability, $F(1, 48) = 5.71$, $p = .02$, $\eta_p^2 = .11$. HRA participants (*M* = 73.88, *SE* = 3.17) outperformed LRA participants (*M* = 62.93, *SE* = 3.30). No other effects or interactions were significant, $p > .05$.

Table 3

Test Performance by Reading Ability and Font

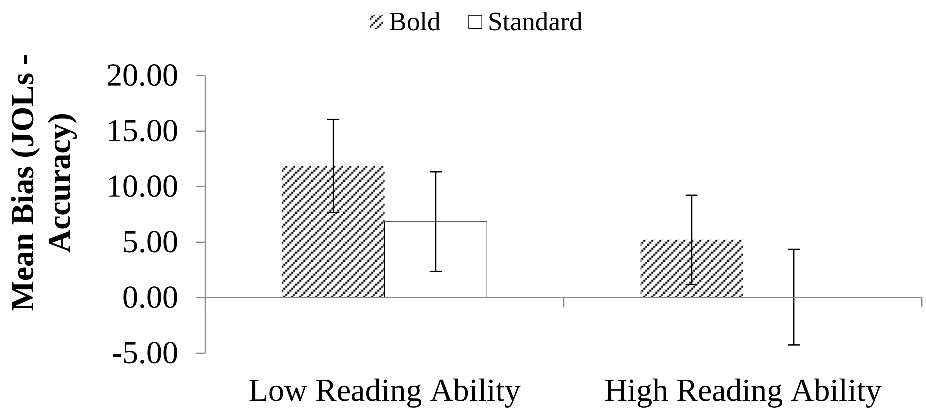
	Bold		Standard	
Reading Ability	*M*	*SE*	*M*	*SE*
High	74.00	3.29	73.75	3.41
Low	62.19	3.43	63.66	3.55

Bias

For each font, a bias score was calculated by subtracting mean test performance from mean term-specific JOLs for each participant. These scores were then averaged across all participants. Positive scores indicate overconfidence (JOLs > performance) while negative scores indicate underconfidence (JOLs < performance). A 2 (Font: bold vs. standard) x 2 (Reading Ability: high vs. low) mixed (ANOVA) was conducted. Results revealed a main effect of font, $F(1, 48) = 9.76$, $p < .01$, $\eta_p^2 = .17$. There were no other significant effects or interactions, $p > .05$.

Participants were significantly more overconfident for bold key terms ($M = 8.54$, $SE = 2.90$) compared to standard key terms ($M = 3.45$, $SE = 3.11$). Because bias could vary as a function of font and/or reading ability, planned comparisons with Bonferroni corrections were conducted. Significant differences in bias across fonts were observed for LRA, $F(1, 48) = 4.55$, $p = .04$, $\eta_p^2 = .09$, and HRA, $F(1, 48) = 5.24$, $p = .03$, $\eta_p^2 = .10$. As depicted in Figure 2, the presence of bold key terms within texts resulted in significant increases in bias regardless of reading ability.

Figure 2. Comparison of bias by reading ability and font



Gamma

Gamma correlations were calculated to analyze participants' ability to discriminate between levels of performance on an item-by-item basis for key terms by comparing each term-specific JOL with its' corresponding test item. One HRA participant was excluded from this analysis because gamma could not be calculated for bold font pairs due to them answering all bold term questions correctly. A 2 (Font: bold vs. standard) x 2 (Reading Ability: high vs. low) mixed ANOVA revealed no significant effects of font, $F(1, 47) = .74$, p = .39, $\eta_p^2 = .02$, or reading ability, $F(1, 47) = .08$, p = .78, $\eta_p^2 < .01$ interactions between font and reading ability, $F(1, 47) = .06$, p = .81, $\eta_p^2 < .01$. Mean gammas (reported in Table 4) indicate that participants were not very accurate at predicting performance on an item-by-item basis, but their accuracy was slightly better for standard key terms.

Table 4

Gamma Correlations by Reading Ability and Font

	Bold		Standard	
Reading Ability	*M*	*SE*	*M*	*SE*
High	.21	.09	.31	.09
Low	.21	.09	.26	.09

Pilot Study Discussion

The pilot study provided an important extension of the metacognitive illusions observed by Rhodes and Castel (2008) and Roberts and Callender (2014). By using a more traditional typographical cue, these results demonstrate metacognitive illusions are not limited to the "font-size effect" (Mueller et al., 2014). These findings provide an important contribution to the metacognitive illusions research that has often relied on extreme disparities in stimuli to produce these illusions. In this study, presenting key terms in bold font within texts without increasing font size produced significantly larger JOLs and did not improve memory. Furthermore, HRA and LRA participants were overconfident in their assessments (although bold font increased the magnitude of overconfidence among LRA) indicating both groups were susceptible to illusions of knowing resulting from different typographic cues.

Based on these findings, the use of bold font to emphasize concepts within textbooks may need to be reconsidered. As a typographical cue, bold is successful at drawing readers' attention to specific concepts (as evident by increased JOLs for bold terms), but capturing readers' attention by increasing perceptual fluency does not produce improvements in comprehension for these concepts (as evident by the lack of differences in test performed based on font). If students rely extensively on these cues to predict comprehension, their JOLs may become more susceptible to illusions of knowing.

Koriat (1997) hypothesized that immediate JOLs (like those used in this pilot) are particularly sensitive to intrinsic cues (ease/difficulty of materials), but are insensitive to extrinsic cues (learning context and encoding strategies). Reliance on intrinsic cues associated with processing fluency could potentially explain why overconfidence was observed for both reading ability groups. Repeated presentations of stimuli can induce subjective feelings of perceptual fluency (Reber, Winkielman, & Schwarz, 1998) and item-by-item JOLs are especially susceptible to these subjective experiences attributed to processing fluency (Besken & Mulligan, 2013; Koriat, Bjork, Sheffer, & Bar, 2004).

Although fluency is a metacognitive cue that is sometimes discounted because it is believed to not be a relevant source of information (Alter & Oppenheimer, 2009b; Oppenheimer 2004), given students' beliefs about the format of texts and fonts (see Serra & Metcalfe, 1992; Mueller et al., 2014), participants' JOLs may have been based upon their pre-existing beliefs about the benefits of bold font. Even though students received no instructions to pay more (or less) attention to bold key terms while reading and bold font was not present during JOL prompts, making immediate JOLs could have supported any naïve theories participants had about the importance of bold font within texts, thereby inflating their JOLs for bold terms.

Because intrinsic cues appear to interfere with metacognitive accuracy, Experiments 1 and 2 investigated whether delaying JOLs mitigated the occurrence of metacognitive illusions for bold key terms. Delaying JOLs for comprehension should produce more accurate judgments because surface memory quickly decays, which reduces the likelihood that JOLs will be influenced by surface level features (Thiede et al., 2010). Despite this theoretical basis, Maki's (1998a) results comparing the accuracy of immediate to delayed comprehension predictions suggests otherwise. Contrary to the typical improvements in accuracy following delayed JOLs for paired-associate learning (i.e., *delayed-JOL effect*; Nelson & Dunlosky, 1991), Maki reported delayed JOLs for comprehension were not more accurate than immediate JOLs.

To account for why delayed JOLs did not improve accuracy, Maki (1998a) proposed that the available cues used to predict performance differed from those available during testing. In support of this rationale, Thiede et al. (2010) suggested that in order for metacomprehension accuracy to improve, readers must use appropriate cues to judge comprehension. If JOLs are based on inappropriate cues such as surface-level memory like recognition or feelings of familiarity, poor accuracy will ensue especially when test questions emphasize situation-level information.

It is possible to improve metacomprehension accuracy following delays if readers are encouraged to engage in additional comprehension processes like writing summaries (Anderson & Thiede, 2008; Thiede & Anderson, 2003), taking practice tests (Maki & Serra, 1992), rereading texts (Rawson et al., 2000) and generating lists of keywords (Thiede et al., 2003). Thiede et al. (2005) contend that generating keywords following a delay provides more diagnostic cues of comprehension because information must be retrieved from long term memory to complete this task. Extending this logic to Experiments 1 and 2, delayed term-specific JOLs should be more accurate because bold font should not influence these judgments. Delaying judgments should eliminate the accessibility of surface cues (like font) from memory which will force readers to use other cues (hopefully mnemonic cues) to evaluate their understanding of concepts (Griffin et al., 2008). However, if bold font affects delayed metacognitive judgments, this would indicate students deliberately attempt to encode and retain these typographic cues in long term memory as a means to predict future comprehension.

Chapter 2: Experiments

Experiment 1

Experiment 1 expanded upon the pilot study by investigating whether font affected delayed JOLs by adopting methodologies to test delayed JOLs for comprehension developed by Maki (1998a) and Rawson and Dunlosky (2007). Because differences in reading ability could interact with the effects of font, it was hypothesized that there would be a significant interaction between font and reading ability on term-specific delayed JOLs. Low reading ability participants' JOLs would be significantly larger for bold terms because they attend to surface cues more while reading and are more likely to base their judgments on processing fluency. However, contrary to the pilot study's results, it was predicted that bold font would not affect high reading ability participants' delayed JOLs because instead of using encoding fluency (which affects immediate JOLs), these readers are more likely to rely on a combination of retrieval fluency and direct-access to target-specific information to make delayed JOLs.

As a byproduct of introducing delayed JOLs in Experiment 1, there should be a significant difference in absolute and relative metacomprehension accuracy based on reading ability. LRA participants are expected to be significantly overconfident (i.e., bias) and less accurate in the JOLs on an item-by-item basis (i.e., lower gamma) compared to HRA because these groups use different types of cues to make JOLs. Considering test performance is a component in calculating both bias and gamma, any differences in either measurement should be attributed to changes in participants' JOLs and not due to changes in test performance because the same final multiple choice test from the pilot study was used in this experiment.

Method

Participants

Sixty-nine undergraduate psychology students enrolled in psychology courses were recruited though SONA system website (https://auburn.sona-systems.com/) and received SONA credit for participation which could be used for extra credit in their psychology courses.

Materials

All materials were administered on five desktop computers using E-Prime 2.0. Participants' reading ability was evaluated using the reading comprehension section from the Multi-media Comprehension Battery (MMCB; see Gernsbacher & Varner, 1988) obtained during a prior session. The MMCB is comprised of four passages with 12 multiple questions per passage with scores ranging from 0 to 48.

Twelve nonfiction passages (≈ 250 words per passage) on a variety of topics (famous people, places, diseases, etc.) and corresponding test items obtained from the Graduate Record Examination (GRE) and Test of English as a Foreign Language (TOEFL) previously used by Roediger and Marsh (2005) were modified for use in the current study. Within each passage, four key terms were identified and presented in either 12 point bold Arial (bold) or 12 point Arial (standard). Across all passages, key term font was counterbalanced and divided into two equal blocks. Each block included three passages per font and their order was randomized. A 48-item multiple choice comprehension test was administered to all participants. Each correct response corresponded to a specific key term.

Procedure

All participants completed the MMCB in a prior 30 min session to evaluate their reading ability. Scores obtained on this measure were used to classify participants into reading ability groups during subsequent analyses (see below).

Consent was obtained from participants upon arrival to lab. They were randomly assigned to one of two reading conditions (alternative versions based on the counterbalancing procedure) and provided verbal instructions that they would be asked to read several short passages and make several confidence judgments about concepts they read about. Reading was self-paced with a one hour time limit for completion.

To evaluate the effects of font on delayed JOLs, the following procedures were used based on modifications to the methodologies developed by Maki (1998a) and Rawson and Dunlosky (2007). Participants read a block of six consecutive passages and then were asked to make an assessment of their comprehension for the entire passage (i.e., global-level JOL) as well as for each key term (i.e. term-specific JOL). For global JOLs, participants were provided a specific passage title (presented at the top of the screen), a Likert scale (0%, 25%, 50%, 75%, and 100%), and were asked the following question: "*In two days (48 hours) from now, how well will you be able to correctly answer multiple choice questions about this material?"*.

Immediately following each global JOL, each of the four key terms from a specific passage were presented one-at-a-time in the center of the screen in sequential order using standard 12 pt. Arial font. During the duration of these presentations, each passage's title was presented concurrently at the top of the screen to remind participants where these terms originated from. Using a scale from 0 to 100, participants were required to submit a term-specific JOL for each key term in response to the following prompt: *"How confident are you that you can correctly answer a multiple choice question about this concept two days (48 hours) from now?"*

Once participants had provided four term-specific JOLs associated with a specific passage, they proceeded to provide global JOLs and term-specific JOLs for all remaining passages from the first block. Participants then read another series of six consecutive passages (block 2) and repeated all judgment procedures again until all passages and terms had been judged. Afterwards, they were dismissed from lab and reminded to return two days later.

When participants returned to lab two days later, they were administered a computer-based multiple choice comprehension test. Each question appeared one-at-a-time and participants were required to select an answer before progressing to the next question. The first half of the test corresponded to passages from the first reading block and the second half corresponded to the second block's passages. After completing this assessment, participants were thanked for their participation and dismissed from lab.

Data Analyses

Performance on the MMCB was used to classify participants into high and low reading ability groups using a median split (median = 33) to determine if there were any group level differences in the effects of font. These groups were treated as a categorical variable for all analyses.

Adhering to Schraw's (2009a) recommendation that researchers use multiple types of measurement to investigate different aspects of metacognition, the effects of font on relative and absolute accuracy were investigated. The relative accuracy of JOLs was analyzed by calculating Goodman-Kruskal gamma correlations (G) and absolute accuracy was evaluated by calculating bias by subtracting mean JOLs from mean test performance per font for all participants.

Global JOLs, term-specific JOLs, test performance, bias (absolute accuracy), gamma (relative accuracy), and reaction time data were analyzed using a series of one-way repeated measures Analysis of Variance (ANOVAs). For each ANOVA, the within-subjects independent variable was key term font comprised of two levels (bold vs. standard) and the between-subjects categorical variable was comprised of two levels of reading ability (high reading ability vs. low reading ability); thus, each repeated measures ANOVA followed a 2 x 2 mixed design. Because each dependent variable was analyzed separately, Bonferroni corrections were used to correct for family-wise error.

Results

The median split based on MMCB scores (median = 33) was used to classify participants into high reading ability (HRA) (n = 31) and low reading ability (LRA) (n = 38) groups during subsequent analyses.

JOLs

Global JOLs. To investigate whether font affected participants' overall assessment of their comprehension for passages, average global JOLs per font were calculated for each participant (see Table 5 for means). A 2 (Font: bold vs. standard) x 2 (Reading Ability: high vs low) mixed Analysis of Variance (ANOVA) revealed no significant effects of font on global JOLs, $F(1, 67) = .02$, $p = .89$, $\eta_p^2 < .001$, and no significant interactions between font and reading ability, $F(1, 67) = .61$, $p = .44$, $\eta_p^2 = .01$, although there was a marginally significant effect of reading ability on global JOLs, $F(1, 67) = 3.71$, $p = .06$, $\eta_p^2 = .05$.

Table 5
Global Judgments of Learning (JOLs) by Reading Ability and Font

	Bold		Standard	
Reading Ability	*M*	*SE*	*M*	*SE*
High	60.38	3.18	62.50	2.69
Low	55.95	2.88	54.48	2.43

In general, HRA were more confident overall compared to LRA regardless of font, but font did not produce any significant differences in confidence within or between groups at the global-level.

Term-specific JOLs. To investigate whether font influenced key term-specific judgments, average JOLs per font were calculated for each participant (see Table 6 for means). A 2 (Font: bold vs. standard) x 2 (Reading Ability: high vs. low) mixed ANOVA revealed no significant main effects, $F(1, 67) = 1.03$, $p = .31$, effects of reading ability, $F(1, 67) = 2.89, p = .09, \eta_p^2 = .04\eta_p^2 = .02$, or any significant interactions, $F(1, 67) = 1.91$, $p = .17$, $\eta_p^2 = .03$.

Table 6
Term-specific Judgments of Learning (JOLs) by Reading Ability and Font

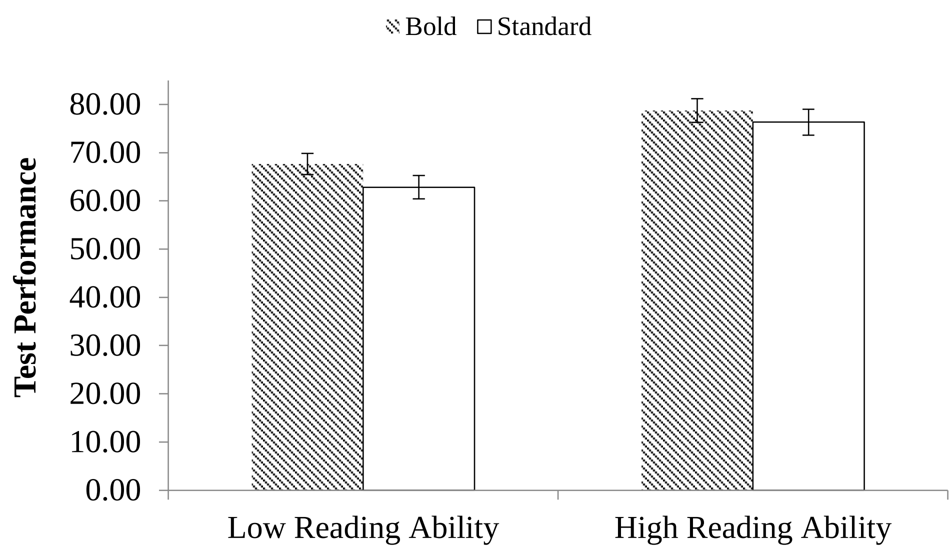
	Bold		Standard	
Reading Ability	*M*	*SE*	*M*	*SE*
High	56.44	2.98	57.18	2.78
Low	53.57	2.69	48.75	2.51

Because it was hypothesized that LRA's JOLs would be significantly larger for bold key terms while HRA's JOLs would not be affected by font, planned comparisons with Bonferroni corrections were conducted. As expected, there was no significant difference in JOLs across fonts for HRA, $F(1, 67) = .06$, $p = .81$, $\eta_p^2 < .01$. However, contrary to our hypothesis, there was only a marginal effect of font on LRA's JOLs, $F(1, 67) = 3.21$, $p = .08$, $\eta_p^2 = .05$. Although LRA's JOLs for bold terms were larger compared to standard terms, differences in key term font did not significantly affect LRA's delayed JOLs for these terms.

Test Performance

Participants' responses on the final test were divided into two groups based on key term font (24 items per font). Test performance was averaged across participants by font (see Figure 3). A 2 (Font: bold vs. standard) x 2 (Reading Ability: high vs low) mixed ANOVA revealed a main effect of font on performance, $F(1, 67) = 4.09$, $p = .05$, $\eta_p^2 = .06$. In addition, there was a significant effect of reading ability, $F(1, 67) = 17.24$, $p < .001$, $\eta_p^2 = .21$, but there was not a significant interaction between font and reading ability, $F(1, 67) = .45$, $p = .50$, $\eta_p^2 = .01$. Test performance was better overall for questions associated with bold key terms (*M* = 73.21, *SE* = 1.65) compared to standard key terms (*M* = 69.59, *SE* = 1.81) and HRA performed better than LRA (*M* = 77.55, *SE* = 2.21 vs. *M* = 65.24, *SE* = 1.99).

Figure 3. Comparison of test performance by reading ability and font

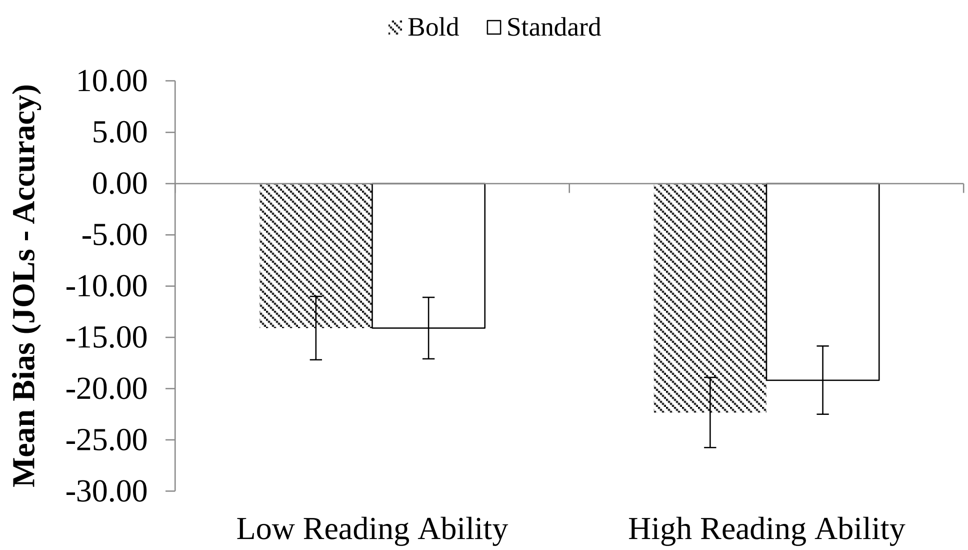


Although HRA outperforming LRA is to be expected given previous results, it is surprising that key term font had an effect on test performance two days later because this was not observed during the pilot study. To better understand these effects of font, post hoc contrasts with Bonferroni corrections were conducted to examine whether test performance varied significantly across fonts within reading ability groups. The results of this analysis revealed that LRA's test performance was significantly better for bold key terms (*M* = 67.65, *SE* = 2.21) compared to standard key terms (*M* = 62.83, *SE* = 2.43), $F(1, 67) = 4.04$, $p = .05$, $\eta_p^2 = .06$, but HRA's test performance for bold terms (*M* = 78.76, *SE* = 2.45) was only slightly better than standard key terms (*M* = 76.34, *SE* = 2.69), $F(1, 67) = .83$, $p = .37$, $\eta_p^2 = .01$. Based on these results, test performance for LRA benefitted from the combination of bold terms within texts followed by delayed JOLs. This finding will be discussed more below.

Bias.

A bias score was calculated for each font by subtracting mean test performance from mean JOLs for each participant. These scores were averaged across all participants. Positive scores indicate overconfidence (JOLs > performance) while negative scores indicate underconfidence (JOLs < performance). A 2 (Font: bold vs. standard) x 2 (Reading Ability: high vs. low) mixed ANOVA was conducted. The results of this analysis revealed no main effects, $F(1, 67) = .69$, $p = .41$, $\eta_p^2 = .01$, or interactions, $F(1, 67) = .69$, $p = .41$, $\eta_p^2 = .01$. Contrary to what was hypothesized, there was not a significant effect of reading ability on bias, $F(1, 67) = 2.61$, $p = .11$, $\eta_p^2 = .04$.

Figure 4. Comparison of bias by reading ability and font



As depicted in Figure 4, delayed JOLs eliminated overconfidence (previously observed) and resulted in underconfidence regardless of font and reading ability. These results were unexpected but indicate delayed JOLs were effective at lowering participants' confidence regardless of their reading ability or key term font.

Gamma

Gamma correlations were calculated to analyze participants' ability to discriminate between levels of performance on an item-by-item basis by pairing each term-specific JOL with performance on its corresponding multiple choice question. Two HRA participants were excluded from this analysis because gamma could not be calculated for them due to their perfect performance on all test questions associated with a specific font. For the remaining 67 participants, a 2 (Font: bold vs. standard) x 2 (Reading Ability: high vs. low) mixed ANOVA revealed no main effects, $F(1, 65) = .00$, $p = 1.00$, $\eta_p^2 < .001$, or interactions, $F(1, 65) = .62$, $p = .44$, $\eta_p^2 = .01$. Contrary to what was hypothesized, there was not a significant effect of reading ability, $F(1, 65) = 1.11$, $p = .30$, $\eta_p^2 = .02$. As presented in Table 7, HRA participants were not more accurate than LRA. Even more surprising, LRA were actually more accurate than HRA at discriminating between levels of performance regardless of key term font.

Table 7

Gamma Correlations by Reading Ability and Font

	Bold		Standard	
Reading Ability	*M*	*SE*	*M*	*SE*
High	.26	.06	.22	.06
Low	.28	.06	.33	.05

Reading and Response Times

In light of potential effects of key term font on encoding fluency and/or retrieval fluency, reaction time data was collected for all participants for several different variables and analyzed.

Overall reading time. Total reading times were calculated for each participant per font and then analyzed using a 2 (Font: bold vs. standard) x 2 (Reading Ability: high vs low) mixed Analysis of Variance (ANOVA) (see Table 8). Results revealed no significant main effect of font, $F(1, 67) = 2.54$, $p = .12$, $\eta_p^2 = .04$, no significant effects of reading ability, $F(1, 67) = 1.84$, $p = .18$, $\eta_p^2 = .03$, and no significant interaction between font and reading ability, $F(1, 67) = .97$, $p = .33$, $\eta_p^2 = .01$.

Table 8

Total Reading Times by Reading Ability and Font

	Bold		Standard	
Reading Ability	*M*	*SE*	*M*	*SE*
High	536.55s	27.88s	480.38s	25.45s
Low	476.25s	25.18s	463.03s	22.99s

Global JOLs. Average reaction times for global JOLs for each font were calculated per participant and analyzed using a 2 (Font: bold vs. standard) x 2 (Reading Ability: high vs low) mixed Analysis of Variance (ANOVA) (see Table 9). Results indicated there was not a significant main effect of font, $F(1, 67) = 0.01$, $p = .94$, $\eta_p^2 < .001$, or a significant interaction between font and reading ability, $F(1, 67) = 0.51$, $p = .48$, $\eta_p^2 = .01$, and there was only a marginally significant effect of reading ability, $F(1, 67) = 3.21$, $p = .08$, $\eta_p^2 = .05$.

Table 9

Global JOLs Response Times by Reading Ability and Font

	Bold		Standard	
Reading Ability	*M*	*SE*	*M*	*SE*
High	7.68s	.42s	8.19s	.45s
Low	6.22s	.35s	6.76s	.38s

Term-specific JOLs. Average reaction times for term-specific JOLs for each font were calculated per participant and analyzed using a 2 (Font: bold vs. standard) x 2 (Reading Ability: high vs low) mixed Analysis of Variance (ANOVA) (see Table 10). Results revealed a significant effect of reading ability on JOLs, $F(1, 67) = 6.60$, $p = .01$, $\eta_p^2 = .09$, but no main effect of font, $F(1, 67) = .01$, $p = .92$, $\eta_p^2 < .001$, or significant interaction, $F(1, 67) = .30$, $p = .59$, $\eta_p^2 < .01$. In general, HRA took considerably longer when making term-specific JOLs.

Table 10
Term-specific JOLs Response Times by Reading Ability and Font

	Bold		Standard	
Reading Ability	*M*	*SE*	*M*	*SE*
High	5.59s	.25s	5.49s	.28s
Low	4.68s	.22s	4.74s	.25s

Test questions. Average reaction times to respond to multiple choice questions were calculated for each font per participant and analyzed using 2 (Font: bold vs. standard) x 2 (Reading Ability: high vs low) mixed Analysis of Variance (ANOVA) (see Table 11). Results of this analysis revealed a significant main effect of font, $F(1, 67) = 10.18$, $p < .01$, $\eta_p^2 = .13$, but no significant effects of reading ability, $F(1, 67) = 1.93$, $p = .17$, $\eta_p^2 = .03$, or a significant interaction between font and reading ability, $F(1, 67) = .97$, $p = .33$, $\eta_p^2 = .01$.

Table 11
Multiple Choice Response Times by Reading Ability and Font

	Bold		Standard	
Reading Ability	*M*	*SE*	*M*	*SE*
High	9.39s	.42s	9.70s	.47s
Low	8.45s	.38s	9.04s	.42s

As depicted in Table 11, response times were significantly faster on questions related to bold key terms. Because LRA may have paid more attention to bold terms while reading, post hoc contrasts with Bonferroni corrections were conducted to determine whether font had differential effects on reaction times within reading ability groups. Results of this analysis revealed HRA's reaction times were not significantly different across fonts, $F(1, 67) = 2.21$, $p = .14$, $\eta_p^2 = .03$, however, LRA's reactions times were significantly faster for bold questions compared to standard questions, $F(1, 67) = 9.70$, $p <.01$, $\eta_p^2 = .13$.

Experiment 1 Discussion

Delayed JOLs are more accurate than immediate JOLs (i.e., Nelson & Dunlosky, 1991; Thiede et al., 2005), but it is unclear whether illusions of knowing related to the "font-size effect" (i.e., Rhodes & Castel, 2009) are resistant to similar improvements. By testing if delayed JOLs could eliminate the metacognitive illusions observed in the pilot study and in previous work by Roberts and Callender (2014), the current experiment provided some evidence that even after a delay, bold font within reading passages can continue to influence JOLs.

Effects of Font of JOLs

Although there was not a main effect of font on term-specific JOLs, it was hypothesized that bold font would significantly affect LRA's JOLs, but it would not influence HRA's JOLs. Subsequent analyses revealed a marginally significant effect of font on LRA's term-specific JOLs. Despite failing to support our hypothesis, the underlying pattern associated with this marginal effect warrants further discussion.

Considering that JOLs were delayed in this study, LRA should have had limited access to perceptions related to ease of processing (i.e., encoding fluency) which may have biased their immediate JOLs in prior studies. Whereas the effects of bold font on JOLs in the pilot study could be attributed to encoding fluency due to the inherent temporal contiguity between reading and judgments, this explanation does not explain why LRA's delayed JOLs continued to be larger for bold terms.

Because poorer readers attend to surface cues more while reading (Thiede et al., 2010), there is a greater likelihood they attempt to retain these cues in memory. Given the current findings, LRA appear to retain access to these surface cues long after initial encoding because their term-specific JOLs for bold terms (*M* = 53.57, *SD* = 2.69) were larger than their term-specific JOLs for standard terms (*M* = 48.75, *SD* = 2.51). Even though there was a marginal effect of font, the fact of the matter is font should only affect immediate JOLs because these judgments are influenced by encoding fluency (Koriat et al., 2014) and people make inferences (i.e., inferential view) about how different available cues relate to memory (Koriat, 1997).

Assuming this is true, one possible explanation for why bold font affected LRA's JOLs is due to retrieval fluency. If JOLs are delayed, people should use retrieval fluency to make these judgments (Koriat & Ma'ayan, 2005) because retrieval fluency is affected by how easily information comes to mind from long-term memory (Benjamin & Bjork, 1996). Given that LRA's delayed JOLs for bold terms were trending toward significance, font may be the most accessible cue they have to use to judge comprehension.

This is not to imply that HRA do not have access to these same surface cues, but they discount them more following delays because their bold term-specific JOLs (*M* = 56.44, *SD* = 2.98) were nearly identical to standard term-specific JOLs (*M* = 57.18, *SD* = 2.78). In the pilot, bold affected HRA's immediate JOLs, perhaps because HRA relied more on encoding fluency (i.e., ease of processing) to make these JOLs. But following a delay, HRA appear to switch to a different set of cues to evaluate their comprehension.

Metacomprehension Accuracy

Compared to previous studies, delayed JOLs were much lower in Experiment 1 which is to be expected (i.e., Dunlosky et al., 2005; Glenberg & Epstein, 1985; Anderson & Thiede, 2008); however, the decrease in JOL magnitude did not produce more accurate judgments which is consistent with Maki's (1998a) findings. Whereas participants in the pilot study were significantly overconfident especially for bold key terms, delayed JOLs in the current study produced only slight improvements in gamma and resulted in substantial levels of underconfidence regardless of reading ability and key term font.

This complete reversal in bias was unexpected because previous studies that used similar judgment paradigms found that delayed JOLs exceeded performance (Dunlosky et al., 2005; Lefèvre & Lories, 2004). For example, Dunlosky et al. (2005) had participants read seven passages before they made delayed term-specific JOLs about their knowledge of key term definition. Immediately after making judgments for all key terms within a specific passage, participants completed a cued-recall test.

Although Dunlosky et al. (2005) used a different final test procedure, it is unlikely that differences in final tests and/or when they were administered can sufficiently explain why underconfidence was observed in Experiment 1. Specifically, the only change in this experiment from the pilot study (where overconfidence was observed which is consistent with prior findings) was when JOLs were made. Therefore, the underconfidence observed in Experiment 1 can only be attributed to the delayed-JOL procedure adopted in this study.

Test Performance

Both reading ability groups performed significantly better on the final test when correct answers corresponded to bold terms. This finding is not entirely unexpected because recall often improves when the cued information is necessary to answer subsequent test questions (Crouse & Idstein, 1972). In fact, test performance improved in all conditions in the current experiment compared to the pilot study; however, a main effect of font was not observed in the pilot despite using the same final multiple choice test. It is possible that re-exposing participants to key concepts after a delay could have produced conditions analogous to the testing effect (Karpicke & Roediger, 2008). Repeated testing often improves performance compared to re-exposure to materials via rereading (Carpenter & DeLosh, 2006; Roediger & Karpicke, 2006). Even though participants were not tested per se during delayed-JOL prompts, they could have attempted to retrieve information related to these key concepts.

With that being said, the retrieval processes they engaged in to make term-specific judgments could have been more complex or difficult than those needed to perform well on the final test (i.e., familiarity). Retention for materials improves when retrieval is more difficult initially (Pyc & Rawson, 2009), and delayed JOL prompts could have facilitated better memory for these concepts because participants could not rely on information in short-term memory to help them make predictions.

Gamma

Although test performance improved, the accuracy of gamma in Experiment 1 increased only slightly compared to gamma obtained during the pilot study. Contrary to what was hypothesized, delayed JOLs did not result in significantly more accurate gamma correlations among HRA compared to LRA. In fact, gamma actually increased slightly for LRA and decreased for HRA in this study. Although this lack of overall improvement in gamma is not surprising (see Maki 1998a), it is surprising that LRA were more accurate than HRA at predicting comprehension on an item-by-item basis. Theoretically, HRA's JOLs should have been more accurate following delays because they have greater access to more diagnostic cues (i.e., direct access to mnemonic cues) for judging comprehension. But despite having more comprehension (as evident by their test performance two days later), HRA struggled to evaluate what they *actually* knew during these judgments.

Limitations

A possible explanation for why limited improvements in gamma paired with overall levels of underconfidence were observed in this study may be due in part to the design of the multiple choice test. The correct answer to each test item was a specific key term. Although it is common practice to use multiple choice questions to measure reading comprehension (Maki & Berry, 1984; Ozuru, Briner, Kurby, & McNamara, 2013; Rawson et al., 2000; Weaver & Bryant, 1995), a limitation with using multiple choice items is that correct answers can be identified through familiarity (Migo et al., 2014; Ozuru et al., 2013).

In lieu of using recollection to actively search memory for relevant text cues to answer test questions (Ozuru et al., 2013), recognition tasks like multiple choice tests may encourage individuals to rely more on familiarity— especially, when these feelings are attributed to differences in processing fluency (Yonelinas, 2002). As a result, perceptually fluent information can induce feelings of familiarity which may make recognizing older materials easier (Johnston, Dark, & Jacoby, 1985). This could potentially explain why test performance was better for questions related to bold terms in Experiment 1. In fact, both groups were faster at responding to questions associated with bold terms (LRA were significantly faster responding to these questions), which indicates participants may have been relying on familiarity during testing.

Although familiarity can improve test performance, a consequence of this strategy is it can also simultaneously decrease the accuracy of metacognitive judgments (Thiede & Dunlosky, 1994). By comparing immediate and delayed JOLs for recognition and recall for paired associates, Thiede and Dunlosky (1994) reported that even though performance for recognition was superior to recall (which is common) and delaying judgments improved the accuracy of both recognition and recall JOLs, delayed JOLs for recognition were less accurate compared to delayed JOLs for recall (lower gamma values).

Citing the *test-received hypothesis*, Thiede and Dunlosky (1994) proposed that the inferior accuracy observed when predicting recognition is due to correct guesses on recognition tests. When they factored out correct responses that students self-reported they were guessing on, gamma values for recognition increased. Based on this finding, they concluded recognition tests can reduce metacognitive accuracy (i.e., gamma) because there is a greater likelihood that items that received lower judgments can still be correctly answered during testing.

Bearing this in mind, the gamma values observed during the pilot and Study 1 may not be diagnostic of participants' *actual* ability to discriminate between different levels of comprehension. Because key terms were always correct answers, it would be easier for participants to guess correctly during the final test if they relied exclusively on familiarity. As such, familiarity could interfere with our attempts to determine whether font has differential effects on metacognition following delays. This also could explain why underconfidence was observed which is contrary to what is typically observed (see Dunlosky et al., 2005; Lefèvre & Lories, 2004).

Experiment 2

Although we cannot completely eliminate correct guessing when using multiple choice tests to assess comprehension, test items can be constructed so readers cannot rely exclusively on familiarity to answer questions (Ozuru et al., 2013). To mitigate the potential effects of familiarity on test performance, key terms were no longer used as correct answers for the multiple-choice test used in Experiment 2. Instead, each question was reconstructed to contain a specific key term within the question's stem and the correct answer consisted of content associated with a particular term. Correctly answering these revised questions should depend more on retrieving information from memory rather than simply guessing. Because readers vary in their ability to construct accurate situation models due to differences in their underlying processing and encoding skills (Gernsbacher et al., 1990), reading comprehension questions that require retrieval of situation-level information will depend more on reading ability and less on familiarity.

Based on these proposed changes to the final test, a moderate to strong effect of reading ability was expected because the disparity in performance between HRA and LRA would be magnified due to HRA being more proficient at making predictive inferences while reading and creating more organized mental representations of the text (Gernsbacher & Faust, 1991). Although HRA's performance may decrease slightly due to the test's increased difficulty, LRA's performance should be severely affected. This anticipated decrement in LRA's performance can be attributed to a lack of conceptual knowledge and more fragmented mental representations. In other words, LRA may recognize key terms within each question but may not be able to successfully answer questions about these concepts. During the pilot study and Experiment 1, LRA could have benefitted from relying on feelings of familiarity to answer test questions and thus, improved their test performance. The revised test questions in Experiment 2 will require information that they may not have access to due to their less efficient reading comprehension processes which should produce much lower performance within this group.

Assuming test performance depends upon recollection, improvements in relative accuracy (compared to previous studies) are anticipated because the imposed congruency between HRA metacognitive processes during JOL prompts (i.e., retrieving information related to a key term) and the retrieval processes present while attempting to answer test questions. Even though different fonts may still be accessible after delays, HRA are more likely to rely on comprehension cues related to key terms when make predictions rather than these surface cues. By retrieving target-specific information about each key term, there should be greater overlap between the types of information accessible at the time of judgment and retrieval cues accessed during testing.

With the increased difficulty of the final test, there should also be a significant effect of reading ability on bias with LRA being more overconfident due to their test performance decreasing. Because we anticipate LRA's JOLs will be significantly larger for bold terms, the decrease in their test performance should also result in a significant interaction between reading ability and font on relative accuracy. Paying attention to key term font while making JOLs and performing worse during testing will produce more discordant pairs (i.e., high JOLs matched with inaccurate responses on test items) thereby reducing LRA's gamma correlations. In contrast, HRA's performance may drop slightly but because there is more similarity between the types of information they use to make JOLs and information they need to answer test questions, their gamma should increase.

All things being equal, we anticipate there will be a significant interaction between font and reading ability on test performance. Specifically, LRA will perform worse on questions related to bold terms because they pay too much attention to surface cues at the expense of developing situation models necessary to understand these concepts. In contrast, HRA should perform similarly regardless of font because they process texts more efficiently and develop more complete representations.

Method

Participants

Sixty-one undergraduate psychology students enrolled in psychology courses were recruited though SONA system website (https://auburn.sona-systems.com/) and received SONA credit for participation which could be used for extra credit in their psychology courses.

Materials

The same materials used in Experiment 1 were used in Experiment 2 with the exception being the final multiple choice test was modified. Instead of correct answers to each multiple choice question corresponding to a specific key term (see Experiment 1), the multiple choice questions used in this study were written with a specific key term included in each question's stem and the correct answer being comprised of content associated with this term. A sample excerpt and corresponding question are presented in Figure 5.

Figure 5. Reading sample and modified test question

> …Once the Sun has used up its thermonuclear energy as a red giant, it will begin to shrink. After it shrinks to the size of the Earth, it will become a white dwarf star. The Sun may throw off huge amounts of gases in violent eruptions called **nova explosions** as it changes from a red giant to a white dwarf.
>
> Nova explosions occur when a red giant changes into a **_________**?
>
> a) yellow dwarf
> **b) white dwarf**
> c) black dwarf
> d) blue dwarf

Procedure

All procedures from Experiment 1 were used in Experiment 2. During a prior 30-minute session, all participants completed the MMCB to evaluate their reading ability. Scores obtained on this measure were used to classify participants into reading ability groups during subsequent analyses.

Consent was obtained from participants upon arrival to lab. They were then randomly assigned to one of two reading conditions (alternative versions based on the counterbalancing procedure). Participants were provided verbal instructions that they would be asked to read several short passages and make several confidence judgments about concepts they read about.

Reading was self-paced with a 1 hour time limit for completion.

In order to evaluate the effects of font on delayed JOLs, the following procedure was used based on modifications of methodologies developed by Maki (1998a) and Rawson and Dunlosky (2007). Participants were asked to make an assessment of their comprehension for the entire passage (i.e., global-level) as well as for each key term (i.e. term-specific) using a method adapted from Rawson and Dunlosky (2007). For global JOLs, participants were provided a specific passage title (presented at the top of the screen), a Likert scale (0%, 25%, 50%, 75%, and 100%), and were asked the following question: "*In two days (48 hours) from now, how well will you be able to correctly answer multiple choice questions about this material?"* The values provided on this scale were intended to match the actual level of performance that could be achieved for a given topic based on the four corresponding test items.

Immediately following each global JOL, each of the four key terms from a specific reading passage was presented one-at-a-time in the center of the screen in sequential order using standard 12 point Arial font. During the duration of these presentations, the passage's title was presented concurrently at the top of the screen to remind participants which passage a specific term was from. Using a 0-100 scale, participants were required to submit a JOL for each term (term-specific JOLs) in response to the following prompt: *"How confident are you that you can correctly answer a multiple choice question about this concept two days (48 hours) from now?"*

Once participants had provided four term-specific JOLs associated with a specific passage, they proceeded to provide global JOLs and term-specific JOLs for all remaining passages from the first block. After completing these steps for block 1, participants read another series of passages (block 2) and repeated all judgment procedures until all passages and terms had been judged. They were then dismissed from lab and reminded to return two days later.

When participants returned to lab, they were administered a computer-based multiple choice comprehension test. Each question appeared one-at-a-time and participants were required to select an answer before progressing to the next question. The first half of the test corresponded to passages from the first reading block and the second half corresponded to the second block of passages. After completing this assessment, participants were thanked for their participation and dismissed from lab.

Results

A median split based on MMCB scores (median = 33) was conducted to classify participants into high reading ability (HRA) (n = 25) and low reading ability (LRA) (n = 36) groups used during subsequent analyses.

JOLs

Global JOLs. To investigate whether bold font affected participants' overall assessment of their comprehension, average global JOLs per font was calculated for each participant (see Table 12 for means). A 2 (Font: bold vs. standard) x 2 (Reading Ability: high vs low) mixed Analysis of Variance (ANOVA) revealed no significant effects of font on global JOLs, $F(1, 59) = .04$, $p = .85$, $\eta_p^2 < .01$, no significant interactions between font and reading ability, $F(1, 59) = 1.54$, $p = .22$, $\eta_p^2 = .03$, or significant effect of reading ability, $F(1, 59) = 2.67$, $p = .11$, $\eta_p^2 = .04$.

Table 12
Global Judgments of Learning (JOLs) by Reading Ability and Font

	Bold		Standard	
Reading Ability	*M*	*SE*	*M*	*SE*
High	62.90	2.86	64.90	3.26
Low	59.56	2.39	56.80	2.72

Term-specific JOLs. To investigate whether font influenced metacomprehension judgments for key terms, average JOLs per font were calculated for each participant (see Table

13 for means). A 2 (Font: bold vs. standard) x 2 (Reading Ability: high vs. low) mixed ANOVA revealed a main effect of font, $F(1, 59) = 3.98$, $p = .05$, $\eta_p^2 = .06$, but no effects of reading ability, $F(1, 59) = .89$, $p = .35$, $\eta_p^2 = .02$, or significant interactions, $F(1, 59) = 2.33$, $p = .13$, $\eta_p^2 = .04$.

Table 13
Term-specific Judgments of Learning (JOLs) by Reading Ability and Font

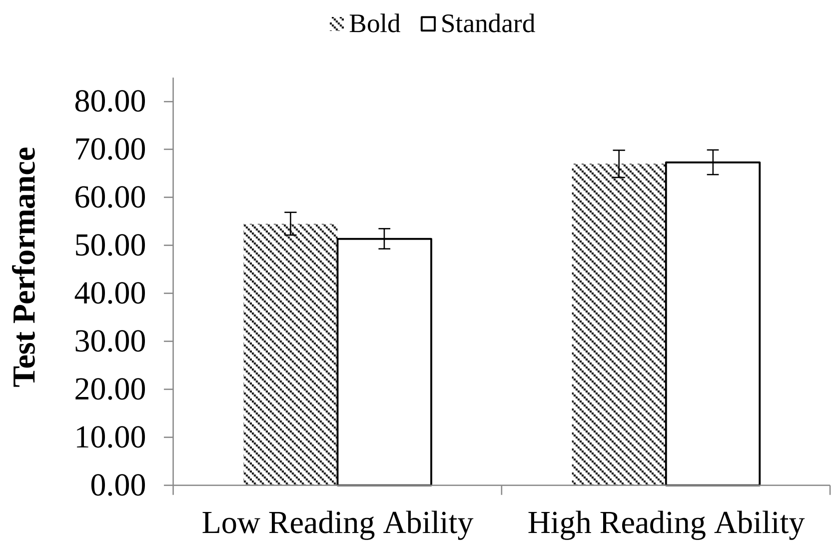
	Bold		Standard	
Reading Ability	*M*	*SE*	*M*	*SE*
High	58.11	2.97	57.15	3.03
Low	58.07	2.48	50.92	2.52

Because it was hypothesized that LRA's JOLs would be significantly larger for bold key terms while HRA's JOLs would not be affected by font, planned comparisons with Bonferroni corrections were conducted. There was no significant difference in JOLs across fonts for HRA, $F(1, 59) = .09$, $p = .76$, $\eta_p^2 < .01$, but there was a significant effect of font on LRA JOLs, $F(1, 59) = 7.56$, $p = .01$, $\eta_p^2 = .11$. As depicted in Table 13, LRA's JOLs for bold terms (*M* = 58.07, *SE* = 2.48) were significantly larger than their JOLs for standard terms (*M* = 50.92, *SE* = 2.52), but there was virtually no difference in HRA's JOLs (bold: *M* = 58.11, *SE* = 2.97; standard: *M* = 57.15, *SE* = 3.03). These results suggest that not only do LRA pay more attention to surface cues while reading, in doing so, these cues continue to have effects on their metacomprehension even after delays.

Test Performance

Each participant's responses on the final test were divided into two groups based on key term font (24 items per font). Test performance was averaged across participants by font (see Figure 6). A 2 (Font: bold vs. standard) x 2 (Reading Ability: high vs low) mixed ANOVA revealed neither a main effect of font on performance, $F(1, 59) = .60$, $p = .44$, $\eta_p^2 = .01$, and no significant interaction between font and reading ability, $F(1, 59) = .93$, $p = .34$, $\eta_p^2 = .02$. There was a significant effect of reading ability, $F(1, 60) = 22.18$, $p < .001$, $\eta_p^2 = .27$.

Figure 6. Comparison of test performance by reading ability and font

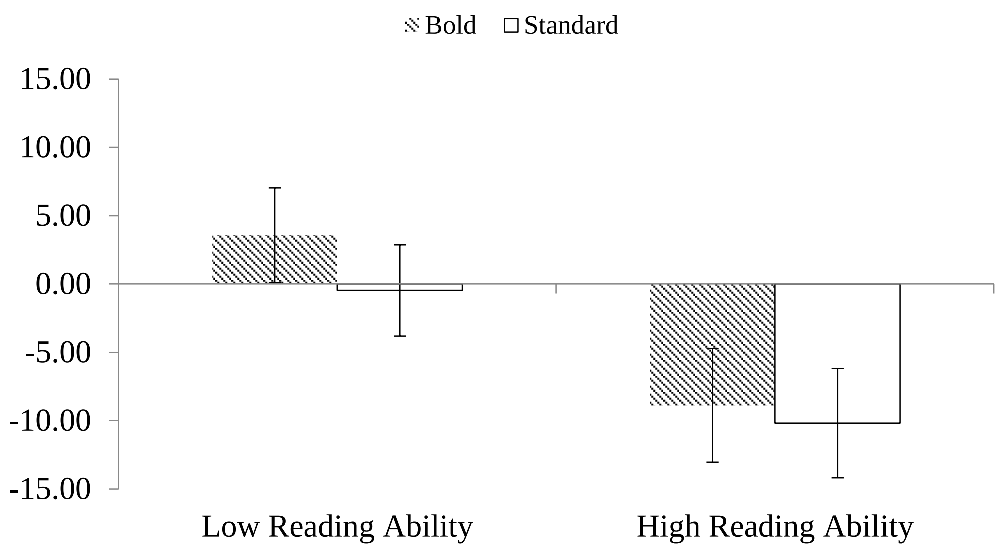


Although it was hypothesized that LRA would perform significantly worse on questions associated with bold terms, planned comparisons determined that key term font did not produce significant differences in test performance, $F(1, 59) = 1.84$, $p = .18$, $\eta_p^2 = .03$. In fact, LRA performed better on questions associated with bold terms ($M = 54.14$, $SE = 2.38$) compared to questions related to standard terms ($M = 51.39$, $SE = 2.12$). As expected, planned comparisons determined font had no effect on HRA's performance, $F(1, 59) = .02$, $p = .90$, $\eta_p^2 < .001$. Overall, HRA's bold test performance ($M = 67.00$, $SE = 2.85$) was nearly identical to their standard test performance ($M = 67.33$, $SE = 2.54$).

Bias.

A bias score was calculated for each font by subtracting mean test performance from mean JOLs for each participant. These scores were then averaged across all participants. Positive scores indicate overconfidence (JOLs > performance) while negative scores indicate underconfidence (JOLs < performance). A 2 (Font: bold vs. standard) x 2 (Reading Ability: high vs. low) mixed ANOVA was conducted. The results of this analysis revealed no significant main effect, $F(1, 59) = .82$, $p = .37$, $\eta_p^2 = .01$, or interactions, $F(1, 59) = .22$, $p = .64$, $\eta_p^2 < .01$. As hypothesized, there was a significant effect of reading ability on bias, $F(1, 59) = 6.28$, $p = .02$, $\eta_p^2 = .10$.

Figure 7. Comparison of bias by reading ability and font



In contrast to the underconfidence observed in Experiment 1 following delayed JOLs, the results from this study demonstrated delaying JOLs while increasing the difficulty of the final test had divergent effects on bias across groups. As depicted in Figure 7, HRA continued to be underconfident for both bold (*M* = -8.89, *SE* = 4.15) and standard fonts (*M* = -10.18, *SE* = 4.01). In contrast, LRA were well calibrated albeit slightly underconfident for standard font (*M* =-.47, *SE* = 3.34), but were definitely overconfident for bold terms (*M* = 3.56, *SE* = 3.46).

Gamma

Gamma correlations were calculated to analyze participants' ability to discriminate between levels of performance on an item-by-item basis (see Table 14) by pairing each term-specific JOL with performance on its corresponding multiple choice question. A 2 (Font: bold vs.

standard) x 2 (Reading Ability: high vs. low) mixed ANOVA revealed no main effects, *F*(1, 59) = .90, *p* = .35, η_p^2 = .02, or an interaction, *F*(1, 59) = .19, *p* = .67, η_p^2 < .01. Contrary to what was hypothesized, there was not a significant effect of reading ability, *F*(1, 59) = .07, *p* = .80, η_p^2 < .01. Overall, participants were not very accurate predicting performance on an item-by-item basis, which is to be expected given the increased difficulty of the final test.

Table 14

Gamma Correlations by Reading Ability and Font

	Bold		Standard	
Reading Ability	*M*	*SE*	*M*	*SE*
High	.08	.06	.16	.06
Low	.12	.05	.15	.05

Reading and Response Times

In light of potential effects of key term font on encoding fluency and/or retrieval fluency, reaction time data was collected for all participants for several different variables and were analyzed.

Overall reading time. Total reading times were calculated for each participants per font and then analyzed using a 2 (Font: bold vs. standard) x 2 (Reading Ability: high vs low) mixed Analysis of Variance (ANOVA) (see Table 15). Results revealed no significant main effect of font, *F*(1, 59) = .42, *p* = .52, η_p^2 = .01, no significant effects of reading ability, *F*(1, 59) = 1.00, *p* = .32, η_p^2 = .02, and no significant interaction between font and reading ability, *F*(1, 59) = .04, *p* = .85, η_p^2 < .01.

Table 15

Total Reading Times by Reading Ability and Font

	Bold		Standard	
Reading Ability	*M*	*SE*	*M*	*SE*
High	526.82s	39.95s	514.14s	28.83s
Low	495.89s	33.29s	472.91s	24.02s

Global JOLs. Average reaction times for global JOLs for each font were calculated per participant and analyzed using a 2 (Font: bold vs. standard) x 2 (Reading Ability: high vs low) mixed Analysis of Variance (ANOVA) (see Table 16). Results indicated there was not a significant main effect of font, $F(1, 59) = 2.45$, $p = .12$, $\eta_p^2 = .04$, or significant interaction between font and reading ability, $F(1, 59) = 0.00$, $p = .97$, $\eta_p^2 = .001$, but there was a significant effect of reading ability, $F(1, 59) = 9.85$, $p = .01$, $\eta_p^2 = .14$. In general, HRA took longer to make global JOLs (*M* = 7.93s, *SE* = .35s) compared to LRA (*M* = 6.49s, *SE* = .29s).

Table 16

Global JOLs Reaction Times by Reading Ability and Font

	Bold		Standard	
Reading Ability	*M*	*SE*	*M*	*SE*
High	7.68s	.44s	7.40s	.44s
Low	6.58s	.40s	6.80s	.40s

Term-specific JOLs. Average reaction times for term-specific JOLs for each font were calculated per participant and analyzed using a 2 (Font: bold vs. standard) x 2 (Reading Ability: high vs low) mixed Analysis of Variance (ANOVA) (see Table 17). Consistent with Experiment 1, results revealed a significant effect of reading ability on JOLs, $F(1, 59) = 6.87$, $p = .01$, $\eta_p^2 = .10$, but no main effect of font, $F(1, 59) = .99$, $p = .32$, $\eta_p^2 = .02$, or significant interaction, $F(1, 59) = .00$, $p = 1.00$, $\eta_p^2 < .001$. Overall, HRA took considerably longer when making term-specific JOLs.

Table 17
Term-specific JOLs Reaction Times by Reading Ability and Font

	Bold		Standard	
Reading Ability	*M*	*SE*	*M*	*SE*
High	5.74s	.37s	5.90s	.34s
Low	4.61s	.31s	4.77s	.28s

Test questions. Average reaction times to respond to multiple choice questions were calculated for each font per participant and analyzed using 2 (Font: bold vs. standard) x 2 (Reading Ability: high vs low) mixed ANOVA (see Table 18). Results of this analysis revealed no significant main effect of font, $F(1, 59) = .18$, $p = .67$, $\eta_p^2 < .01$, significant effects of reading ability, $F(1, 59) = 1.83$, $p = .18$, $\eta_p^2 = .03$, and no significant interaction between font and reading ability, $F(1, 59) = .36$, $p = .55$, $\eta_p^2 = .01$.

Table 18
Multiple Choice Response Times by Reading Ability and Font

	Bold		Standard	
Reading Ability	*M*	*SE*	*M*	*SE*
High	11.08s	.42s	11.36s	.51s
Low	10.51s	.35	10.46s	.42s

Discussion

The results from Experiment 2 complemented the Experiment 1's findings while also providing additional insight into the effects of key term font on test performance and JOLs.

Effects of Font on JOLs

A main effect of font on term-specific JOLs was observed in this experiment, but it was not found in Experiment 1. Most likely the reason this effect occurred in one study but not the other is due to there being some variability even among LRA in the extent to which they relied on font as a cue when making judgments. Simply put, there was a marginal effect of bold font on

LRA's term-specific JOLs in Experiment 1 and as hypothesized, term-specific JOLs for bold terms were significantly larger compared to JOLs for standard terms in Experiment 2.

These results, when considered in junction with the fact font had no effect on HRA's JOLs in either studies, indicate LRA place more emphasis on font as a predictor of future memory and these cues remain accessible following delays. To the experimenter's knowledge, this is a unique finding because key term font should only influence immediate JOLs according to encoding fluency (see Koriat & Ma'ayan, 2005). However, because key term font persisted to influence LRA's delayed JOLs, this indicates typographic cues can also affect retrieval fluency. To remedy the metacognitive illusions associated with bold font, LRA readers may need to be provided with more explicit instructions about which diagnostic cues to pay attention to when assessing comprehension.

Test Performance and Response Times

As anticipated, introducing the modified multiple choice test in Experiment 2 did decrease performance. HRA performance continued to be superior to LRA and font did not affect performance or produce significant differences in reaction times during testing. All reaction times were slower and LRA were no longer significantly faster at responding to questions related to bold terms. Based on these results, we believe the revised multiple choice test was successful in reducing participants' ability to rely on familiarity as a response strategy.

Despite these changes from the previous study, there was not a significant interaction between font and reading ability on test performance as hypothesized. Instead of performing significantly worse on questions related to bold terms, LRA continued to perform slightly better on these questions (M = 54.14, SE = 2.38) compared to standard questions (M = 51.39, SE = 2.12). In spite of this slight increase in performance for bold terms, LRA comprehended less compared to HRA because HRA performed significantly better than them on the final test which is a common finding (Griffin et al., 2008; Maki et al., 2005).

Bias

The combination of reading texts with bold terms and making delayed JOLs (which could be similar enough to being re-exposed to material through rereading), may have helped LRA retain more access to information related to these concepts. However, the analysis related to bias suggests that using bold font within texts to emphasize concepts may adversely affect readers with less proficient comprehension skills.

In comparison to Experiment 1, using a more difficult final test in Experiment 2 resulted in a significant effect of reading ability on bias. Whereas all participants were underconfident during Experiment 1 (presumably because their test performance was inflated due to familiarity), the more difficult final test used in this experiment reduced the magnitude of bias across groups while also producing divergent effects of font on bias. HRA continued to be underconfident regardless of font (although to a lesser extent compared to Experiment 1), but LRA were overconfident in their performance associated with bold terms and were slightly overconfident for standard terms. Based on this finding, using standard font in texts (not bold), delaying JOLs, and increasing final test difficult is a potentially useful method for eliminating overconfidence among LRA without adversely impacting HRA's test performance. This difference in bias due to font among LRA provides further evidence that bold font is a typographic cue that can affect retrieval fluency without improving retrieval (i.e., test performance).

Gamma

Despite reducing the magnitude of bias among groups, there were no improvements in relative accuracy obtained in Experiment 2. Gamma correlations were considerably lower for both reading ability groups. It is surprising that HRA were slightly worse than LRA at predicting their performance given that HRA have more efficient reading processes and should be more adept at making delayed JOLs. Furthermore, LRA's gamma for standard terms was only slightly more reliable than their gamma for bold terms which failed to support our hypothesis that bold font would produce significantly lower gamma for LRA compared to standard font.

Overall, this lack of improvement in gamma may support Thiede et al. (2010) assertion that readers need to use more appropriate cues to judge comprehension in order for metacomprehension accuracy to improve. In this case, simply judging key terms following a delay may not be an adequate means to evaluate comprehension compared to other deliberate retrieval processes (i.e., generating key terms, writing summaries, taking tests, etc.) which may provide more diagnostic cues of comprehension.

General Discussion

The results from the pilot study and two experiments demonstrated that typographic cues within texts (i.e., bold font used to emphasis key terms) can influence metacomprehension judgments. Participants read passages with and without bold key terms and made JOLs either immediately after reading (pilot study) or following a delay (Experiments 1 & 2). In the pilot study, a main effect of font on term-specific JOLs was observed which resulted in significantly larger judgments for bold key terms. Despite this increased confidence, font did not produce any significant differences in test performance thereby resulting in a main effect of font on bias (participants were significantly more overconfident in their memory for bold terms).

Encoding Fluency

The pattern of results from this pilot can be attributed to the effects of encoding fluency. The reading and JOL procedure used created temporal contiguity which may have produced a greater sensation of familiarity for bold terms. Because immediate JOLs are susceptible to the effects of encoding fluency (Koriat et al., 20014) and participants may possess pre-existing beliefs or “naïve theories” about how specific cues relate to memory(see Schwarz, 2004), it is highly likely participants relied on beliefs about the ease of processing to make immediate JOLs.

Although participants were not asked explicitly to disclose their beliefs about font, they were significantly overconfident in their ability to answer questions related to bold terms which indicates they believed bold items were easier to remember/learn. Bearing in mind authors frequently use typographic cues to emphasize concepts (Lorch et al., 1995), it is safe to assume participants would have had plenty of experience reading textbooks with bold terms where they could have develop these “naïve theories.” Students’ beliefs about text format and fonts affect metacognitive judgments (Serra & Metcalfe, 1992; Mueller et al., 2014) and participants could have relied on their naïve theories based on experience to make inferences (i.e., inferential view; Koriat, 1997) regarding the perceived benefits of bold font.

Retrieval Fluency

To mitigate the effects of encoding fluency attributable to key term font, Experiment 1 and 2 measured metacomprehension using delayed JOLs by adopting and modifying procedures used by Maki (1998a) and Rawson and Dunlosky (2007). Because surface memory decays quickly (Thiede et al., 2010), delayed JOLs should have reduced the possibility that participants could rely on encoding fluency (i.e., key term font) to make inferences about memory. As a result, their JOLs should be more accurate as they rely on retrieval fluency and mnemonic cues to assess memory (Koriat & Ma’ayan, 2005).

Despite incorporating delayed JOLs, both Experiment 1 and 2 provided evidence that key term font remained an accessible cue that exerted influence over LRA’s JOLs. Experiment 1 produced a marginal effect of bold font on LRA's term-specific JOLs while Experiment 2 produced a main effect of font on these judgments which was driven by a significant effect of bold on LRA's term-specific JOLs.

These are particularly interesting findings because participants were only exposed to bold font briefly while reading each passage. The subsequent delay between reading and making JOLs was intended to reduce the likelihood font could influence these judgments, thereby eliminated the effects of encoding fluency. Although LRA took less time to read passages, their reading times were not significantly different from HRA, which indicates bold's effects on term-specific JOLs among LRA cannot be attributed to differences in temporal contiguity between groups.

In spite of the lack of temporal contiguity, the argument could be made that the delayed-JOL procedure was not long enough to eliminate encoding fluency. However, given that term-specific JOL magnitude decreased substantially from the pilot study, this explanation is unlikely. Especially, because information from short-term memory quickly decays (Nelson & Dunlsoky, 1994; Peterson & Peterson, 1959) which means access to surface cues associated with ease of learning should be eliminated following a delay.

The more likely explanation for why bold affect LRA's JOLs is these readers paid more attention to this bold font while reading. In doing so, they intentionally attempted to retain these typographic cues in memory. Because variations in encoding can produce differences in memory traces which affect the magnitude of JOLs, (Cohen et al., 1991; Koriat 1997), the effects of bold font on LRA's delayed JOLs were due to differences in retrieval fluency.

According to the *direct-access theory*, people try to retrieve information directly related to specific targets (Schwartz, 1994). Although retrieval fluency affects how easily information can be accessed from memory (Benjamin & Bjork, 1996), it is not always a reliable source of information for making metacognitive judgments (see Benjamin, Bjork, & Schwartz, 1998 for an explanation). In the context of Experiment 1 and 2, paying more attention to bold terms while reading may have made bold font the most fluent cue LRA could access during delayed JOLs. If LRA attempted to judge whether a term appeared in bold previously prior to making their JOLs, this process could have activated any “naïve theories” they possess which would explain why their confidence was inflated for bold terms.

Test Performance

With the exception of Experiment 1, font did not significantly affect test performance which indicates typographic cues have more effects on metacognition than comprehension. Even though LRA did perform significantly better on test questions associated with bold fonts in Experiment 1 and slightly better on Experiment 2, relying on processing fluency is not a particularly effective learning strategy. As previously discussed, LRA’s performance on Experiment 1 was likely inflated due to feelings of familiarity because a) they were significantly faster to respond to questions associated with bold terms and b) they took equally long to respond to the revised test questions during Experiment 2 regardless of font

In contrast to LRA, HRA’s performance was not affected by bold font regardless of final test format. They consistently outperformed LRA participants which we anticipated observing. Considering font had no effect on HRA, reading passages and making JOLs resulted in all key terms and associated content becoming equally salient to these readers. Even when key terms were no longer correct answers (Experiment 2), HRA continued to excel at comprehension which further demonstrates these readers have more proficient reading comprehension skills.

Bias

Despite not providing readers any specific metacognitive strategies, the data related to metacognitive accuracy resulted in some unexpected results. With respect to bias, which measures absolute metacognitive accuracy, Experiment 1 completely reversed the pilot study's findings of overconfidence and resulted in underconfidence for all groups (greater than 14 points difference between JOLs and test performance). Considering both studies used the same final comprehension test where key terms were correct answers, we did not anticipate this result which can be explained based on the overall decrease in participants' JOLs following delays. Instead of LRA being more overconfident overall (like what was observed in the pilot study), LRA were actually underconfident, but to a lesser extent than HRA.

There were also no significant differences in bias between these groups which failed to support our hypothesis. Although it is disappointing not to observe any group differences, we are most surprised that LRA were not overconfident (especially for bold terms) because we expected that their JOLs would be less resistant to change following a delay due to them attending to these typographic cues more while reading.

Previous work by Thiede et al. (2010) using a delayed judgment procedure adopted from Maki (1998a) produced overconfidence among at-risk college readers. We used this same procedure in Experiment 1, which is why we anticipated that LRA would be overconfident in judging bold terms. Even though the scope of judgments Thiede et al. elicited from readers was more akin to the global JOLs we used, given they reported at-risk readers relied more on surface cues to assess comprehension, we believed this pattern of overconfidence would extend to term-specific JOLs because these judgments are not significantly more accurate than global JOLs (Dunlosky et al., 2002; Rawson & Dunlosky, 2007).

In Experiment 2, we did support the hypotheses that there would be a significant effect of reading ability on bias. We believed LRA would be significantly more overconfident than HRA based on bold terms. Even though the revised multiple choice test reduced test performance, HRA continued to be underconfident (but to a lesser extent than Experiment 1) while LRA became overconfident for bold terms. This underconfidence among HRA paired with overconfidence among LRA was expected based on previous results (Thiede et al., 2010).

Gamma

Although the results from Experiment 2 were expected, we were surprised that LRA were better calibrated (less biased) than HRA and their relative accuracy (gamma) was also higher regardless of font. These results failed to support our hypothesis about relative accuracy in Experiment 1 and Experiment 2. Although improvements in metacomprehension accuracy following delays are difficult to obtain (Maki, 1998a), several studies demonstrated that delayed JOLs can be more accurate (Thiede & Anderson, 2003; Thiede et al., 2010). For example, proficient college readers are significantly more accurate (higher gamma values) at evaluating comprehension compared to at-risk readers following delays (Thiede et al., 2010) but the current experiments failed to corroborate these findings.

Limitations

Whereas we have evidence indicating LRA rely on typographic cues to predict comprehension, it is difficult to determine what types of information HRA used when making term-specific JOLs. HRA were not particularly well-calibrated (they were underconfident) and did not do a very good job of discriminating between levels of performance (their gammas were lower than LRA's) in either Experiment 1 or 2.

The underlying assumptions behind why term-specific JOLs should be more accurate is that when participants are presented a key-term, they use this prompt as a cue to retrieve target-specific information (Rawson & Dunlosky, 2007). Through these retrieval attempts, they should access information that will be diagnostic of future retrieval and use this content as a basis for making metacognitive judgments.

Although Rawson and Dunlosky (2007) acknowledged that subsequent work demonstrated these assumptions are not true, these studies did not consider individual differences in reading ability. Even if individuals do not automatically self-test prior to making delayed JOLs (Dunlosky et al., 2005), more proficient readers would be better at judging their comprehension for key terms (Thiede et al., 2010). According to the *situation model approach to metacomprehension* (Griffin et al., 2008), tests that require access to a text's situation model should improve metacomprehension (Thiede et al., 2010; Wiley et al., 2005). If this is true, then HRA should have achieved higher levels of accuracy during Experiment 2 when the more difficult test was used.

A possible explanation for why these discrepancies in gamma and bias emerged between reading ability groups is due to differences in HRA and LRA's beliefs about their own forgetting. According to Koriat et al. (2004), experienced-based JOLs that rely on cues like perceptual fluency and/or retrieval fluency do not take in consideration forgetting rates. In contrast, theory-based judgments depend exclusively on how people apply metacognitive knowledge to evaluate their own proficiencies including their ability to remember (Dunning, Johnson, Ehrlinger, & Kruger 2003; Koriat et al., 2004).

If we assume that LRA participants relied on experience-based JOLs and ignored forgetting rates (as evident by differences in term-specific JOLs based on font), then it may be possible to explain HRA's underconfidence as a byproduct of them relying on theory-based JOLs. During the delayed-JOL prompts, participants were ask to indicate how confident they were in the ability to answer a question about a key term two days later. If HRA paid more attention to how long information would need to be retained (not forgotten), this could produce lower JOLs. For example, Koriat et al. (2004) demonstrated that participants can access their knowledge about forgetting rates when making JOLs, but only when retention intervals are manipulated or JOLs are framed in terms of forgetting rather than memory. That is, people know that recall declines with the passage of time.

Ultimately, both groups of readers in the current experiments could have benefitted from being provided more explicit guidelines for which cues to use during delayed JOLs (instead of being permitted to judge key terms using their own criteria). All things being equal, judging key terms through delayed-JOL prompts is not equivalent to self-generated key-terms (see *delayed-keyword effect*, Thiede et al., 2005). To improve accuracy, readers may require additional generative comprehension processes designed to provide them with more diagnostic cues for comprehension like writing summaries (Thiede & Anderson, 2003), taking practice tests (Maki & Serra, 1992), and/or generating key terms (Thiede et al., 2003). Perhaps integrating similar procedures that have shown to improve accuracy or framing JOL prompts differently could have reduced LRA’s bias toward bold terms and/or helped improve metacomprehension for all groups. These are interesting question that warrants future investigation.

Conclusion

Building upon previous work regarding metacognitive illusions (i.e., font-size effect), these studies demonstrated that the presence of bold key terms within texts can affect metacomprehension. Depending on both when judgments occur (immediately or delayed) and reading ability, font can produce differential effects on performance and confidence. This is an important contribution to the metacognitive illusion literature because these effects were obtained using more ecologically valid materials (i.e., reading passages with bold terms).

Although more research is needed to determine the extent to which bold font affects metacomprehension, the current results indicate that less proficient readers rely more on surface cues to predict memory and these cues do not improve subsequent test performance.

References

Alter, A. L., & Oppenheimer, D. M. (2009a). Suppressing secrecy through metacognitive ease: Cognitive fluency encourages self-disclosure. *Psychological Science*, *20*(11), 1414-1420.

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

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

Arbuckle, T. Y., & Cuddy, L. L. (1969). Discrimination of item strength at time of presentation. *Journal of Experimental Psychology*, *81*, 126-131.

Benjamin, R. G. (2012). Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review*, *24*, 63-88.

Benjamin, A. S., & Bjork, R. A. (1996). Retrieval fluency as a metacognitive index. In L. Reder (Ed.), *Implicit memory and metacognition* (pp. 309–338). Hillsdale, NJ: Erlbaum.

Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General*, *1*, 55-68.

Besken, M., & Mulligan, N. W. (2013). Easily perceived, easily remembered? Perceptual interference produces a double dissociation between metamemory and memory performance. *Memory & Cognition*, *41*(6), 897-903.

Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition*, *34*, 268-276.

Castel, A. D. (2008). Metacognition and learning about primacy and recency effects in free recall: The utilization of intrinsic and extrinsic cues when making judgments of learning. *Memory & Cognition, 36*, 429-437.

Chiang, E. S., Therriault, D. J., & Franks, B. A. (2010). Individual differences in relative metacomprehension accuracy: Variation within and across task manipulations. *Metacognition Learning*, *5*, 121-135.

Cohen, R. L., Sandler, S. P., & Keglevich, L. (1991). The failure of memory monitoring in a free recall task. *Canadian Journal of Psychology, 45*, 523–538. doi:10.1037/h0084303

Crouse, J. H., & Idstein, P. (1972). Effects of encoding cues on prose learning. *Journal of Educational Psychology*, *63*(4), 309-313.

Cull, W. L., & Zechmeister, E. B. (1994). The learning ability paradox in adult metamemory research: Where are the metamemory differences between good and poor learners? *Memory & Cognition*, *22*, 249-257.

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

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

Dunlosky, J., Rawson, K. A., & Middleton, E. L. (2005). What constrains the accuracy of metacomprehension judgments? Testing the transfer-appropriate-monitoring and accessibility hypotheses. *Journal of Memory and Language*, *52*, 551-565.

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.

Epstein, W., Glenberg, A. M., & Bradley, M. M. (1984). Coactivation and comprehension:

Contribution of text variables to the illusion of knowing. *Memory & Cognition*, *12*(4), 355-360.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1977). Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology*: *Human Perception and Performance*, *3*(4), 552-564.

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

Flavell, J. H. (1999). Cognitive development: Children's knowledge about the mind. *Annual Review of Psychology*, *50*, 21-45.

Foster, J., & Coles, P. (1977). An experimental study of typographic cueing in printed text. *Ergonomics*, *20*(1), 57-66.

Gaddy, M. L., van den Broek, P., &, Sung, Y. C. (2001). The influence of text cues on the allocation of attention during reading. In T. Sanders, J. Schilperoord, & W. Spooren (Eds.), *Text representation: Linguistic and psycholingusitic aspects* (pp. 89-110). Amsterdam: John Benjamins.

Gernsbacher, M. A., & Faust, M. E. (1991). The mechanism of suppression: A component of general comprehension skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(2), 245-262.

Gernsbacher, M. A., & Varner, K. R. (1988). *The multi-media comprehension battery* (Tech. Rep. No. 88-03). Eugene: University of Oregon, Institute of Cognitive and Decision Sciences.

Gernsbacher, M. A., Varner, K. R., & Faust, M. E. (1990). Investigating differences in general comprehension skill. *Journal of Experimental Psychology: Learning, Memory, and*

Cognition, *16*, 430-445.

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

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

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

Graesser, A. C. (2007). An introduction to strategic reading comprehension. In D. S. McNamara (Ed.), *Reading comprehension strategies: Theories, interventions, and technologies* (pp. 3-26). New York: Lawrence Erlbaum Associates.

Griffin, T. D., Jee, B. D., & Wiley, J. (2009). The effects of domain knowledge on metacomprehension accuracy. *Memory & Cognition*, *37*(7), 1001-1013.

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

Hacker, D. J., Bol, L., Horgan, D. D., & Rakow, E. A. (2000). Test prediction and performance in a classroom context. *Journal of Educational Psychology*, *92*(1), 160-170.

Hacker, D. J., Bol, L., & Keener, M. C. (2008). Metacognition in education: A focus on calibration. In J. Dunlosky & R. A. Bjork (Eds.), *Handbook of metamemory and memory* (pp. 429-455). New York: Taylor & Francis.

Hannon, B. (2012). Understanding the relative contributions of lower-level word processes, higher-level processes, and working memory to reading comprehension performance in proficient adult readers. *Reading Research Quarterly*, *47*(2), 125-152.

Hannon, B. & Daneman, M. (1998). Facilitating knowledge-based inference in less-skilled readers. *Contemporary Educational Psychology*, *23*, 149-172.

Hilbig, B. E., Erdfelder, E., & Pohl, R. F. (2011). Fluent, fast, and frugal? A formal model evaluation of the interplay between memory, fluency, and comparative judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(4), 827-839.

Jacoby, L. L., Kelley, C. M., & Dywan, J. (1989). Memory attributions. In H. L. Roediger & F. I. M. Craik (Eds.), *Varieties of memory and consciousness: Essays in honor of Endel Tulving* (pp. 391-422). Hillsdale, NJ: Erlbaum.

Jacoby, L. L., & Whitehouse, K. (1989). An illusion of memory: False recognition influenced by unconscious perception. *Journal of Experimental Psychology: General*, *118*(2), 126-135.

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

Johnston, W. A., Dark, V. J., & Jacoby, L. L. (1985). Perceptual fluency and recognition judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*, 3-11.

Karpicke, J. D., & Roediger, H. L., III. (2008). The critical importance of retrieval for learning. *Science, 319*, 966-968.

Keren, G. (1991). Calibration and probability judgments: Conceptual and methodological issues. *Acta Psychologica*, *77*, 217-273.

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

Kintsch, W. (1994). Text comprehension, memory, and learning. *American Psychologist*, *49*, 292-303.

Kintsch, W. (2005). An overview of top-down and bottom-up effects in comprehension: The CI perspective. *Discourse Processes*, *39*, 125-128.

Kintsch, W., & Welsch, D. M. (1991). The construction-integration model: A framework for studying memory for text. In W. E. Hockley & S. Lewandowsky (Ed.), *Relating theory and data: Essays on human memory in honor of Bennett B. Murdock* (pp. 367-385). Hillsdale, NJ.: Erlbaum

Kintsch, W., & Yarbrough, J. C. (1982). Role of rhetorical structure in text comprehension. *Journal of Educational Psychology*, *74*(6), 828-834.

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review, 100*, 609-639. doi:10.1037/0033-295X.100.4.609

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, *126*(4), 349-370.

Koriat, A. (2011). Subjective confidence in perceptual judgments: A test of the self-consistency model. *Journal of Experimental Psychology: General, 140,* 117-139. doi:10.1037/a0022171

Koriat, A. (2012). The self-consistency model of subjective confidence. *Psychological Review*, 119, 80-113.

Koriat, A., & Bjork, R. A. (2005). Illusions of competence in monitoring one's knowledge during study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(2), 187-194.

Koriat, A., Bjork, R. A., Sheffer, L., & Bar, S. K. (2004). Predicting one's own forgetting: The role of experience-based and theory-based processes. *Journal of Experimental Psychology: General*, *133*, 643-656. doi:10.1037/0096-3445.133.4.643

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory, 6,* 107-118.

Koriat, A., & Ma'ayan, H. (2005). The effects of encoding fluency and retrieval fluency on judgments of learning. *Journal of Memory and Language, 52*, 478-492.

Koriat, A., Ma'ayan, H., & Nussinson, R. (2006). The intricate relationship between monitoring and control in metacognition: Lessons for the cause-and-effect relation between subjective experience and behavior. *Journal of Experimental Psychology: General, 135*, 36-69.

Koriat, A., Nussinson, R., & Ackerman, R. (2014). Judgments of learning depend on how learners interpret study effort. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(6), 1624-1637.

Krulee, G. K., & Novy, F. (1986). Word processing and effects of variability in type fonts. *Perceptual and Motor Skills*, *62,* 999-1010.

Lanska, M., Olds, J. M., & Westerman, D. L. (2014). Fluency effects in recognition memory: Are perceptual fluency and conceptual fluency interchangeable? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(1), 1-11.

Lefèvre, N., & Lories, G. (2004). Text cohesion and metacomprehension: Immediate and delayed judgments. *Memory & Cognition*, *32*(8), 1238-1254.

Leonesio, R. J., & Nelson, T. O. (1990). Do different metamemory judgments tap the same underlying aspects of memory? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(3), 464-470.

Linderholm, T., Zhao, Q., Therriault, D. J., & Cordell-McNulty, K. (2008). Metacomprehension effects situation within an anchoring and adjustment framework. *Metacognition*

Learning, *3*, 175-188.

Long, D. L., Oppy, B. J., & Seely, M. R. (1994). Individual differences in readers' sentences- and text-level representations. *Journal of Memory and Language*, *36*, 129-145.

Lorch, R. F., Lorch, E. P., & Klusewitz, M. A. (1995). Effects of typographical cues on reading and recall of text. *Contemporary Educational Psychology*, *20*, 51-64.

Maki, R. H. (1998a). Predicting performance on text: Delayed versus immediate prediction and tests. *Memory & Cognition*, *26*(5), 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). Hillsdale, NJ: LEA.

Maki, R. H., & Berry, S. L. (1984). Metacomprehension of text material. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(4), 663-679.

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

Maki, W. S., & Maki, R. H. (2002). Multimedia comprehension skill predicts differential outcomes of web-based and lecture courses. *Journal of Experimental Psychology: Applied*, *8*(2), 85-98.

Maki, R. H., Shields, M., Wheeler, A. E., & Zacchilli, T. L. (2005). Individual differences in absolute and relative metacomprehension accuracy. *Journal of Educational Psychology, 97*(4), 723-731

Maki, R. H., & Serra, M. (1992). The basis of test predictions for text material. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 18*, 116-126.

Maki, R. H., Willmon, C., & Pietan, A. (2009). Basis of metamemory judgments for text with multiple-choice, essay, and recall tests. *Applied Cognitive Psychology*, *23*, 204-222.

Margolin, S. J. (2013). Can bold typeface improve readers' comprehension and metacomprehension of negation? *Reading Psychology*, *34*, 85-99.

Marks, M. B. (1966). Improving reading through better format. *Journal of Educational Research*, *60*(4), 147-151.

Masson, M. E. J., & Miller, J. A. (1983). Working memory and individual differences in comprehension and memory of text. *Journal of Educational Psychology*, *75*(2), 314-318.

McCabe, P. P., Kraemer, L. A., Miller, P. M., Parmer, R. S., & Ruscica, M. B. (2006). The effect of text format upon underachieving first year college students' self-efficacy for reading and subsequent reading comprehension, *Journal of College Reading and Learning*, *37*(1), 19-42.

Metcalfe, J. (2009). Metacognitive judgments and control of study. *Current Directions in Psychological Science*, *18*(3), 159-163.

Metcalfe, J., & Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin & Review*, *15*, 174-179.

Metcalfe, J., Schwartz, B. L., & Joaquim, S. G. (1993). The cue-familiarity heuristic in metacognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(4), 851-861.

Miele, D. B., & Molden, D. C. (2010). Naïve theories of intelligence and the role of processing fluency in perceived comprehension. *Journal of Experimental Psychology: General*, *139*(3), 535-557.

Migo, E. M., Quamme, J. R., Holmes, S., Bendell, A., Norman, K. A., Mayes, A. R., & Montaldi, D. (2014). Individual differences in forced-choice recognition memory:

Partitioning contributions of recollection and familiarity. *The Quarterly Journal of Experimental Psychology*, *67*(11), 2189-2206.

Miller, T. M., & Geraci, L. (2011). Unskilled but aware: Reinterpreting overconfidence in low-performing students. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(2), 502-506.

Mueller, M. L., Dunlosky, J., Tauber, S. K., & Rhodes, M. G. (2014). The font-size effect on judgments of learning: Does it exemplify fluency or reflect people's beliefs about memory? *Journal of Memory and Language*, *70*, 1-12.

Murray, J. D., & Burke, K. A. (2003). Activation and encoding of predictive inferences: The role of reading skill. *Discourse Processes*, *35*(2), 81-102.

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

Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "delayed JOL effect." *Psychological Science*, *2*(4), 267-270.

Nelson, T. O., & Leonesio, R. J. (1988). Allocation of self-paced study time and the "labor-in-vain effect." *Journal of Experimental Psychology: Learning, Memory and Cognition*, *14*, 676-686.

Nelson, T. O. & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.), *The psychology of learning and motivation*, *26*, pp. 125-141. New York: Academic Press, Inc.

Nelson, T. O., Narens, L., & Dunlosky, J. (2004). A revised methodology for research on metamemory: Pre-judgment recall and monitoring (PRAM). *Psychological Methods*, *9*, 53-69.

Oppenheimer, D.M. (2004). Spontaneous discounting of availability in frequency judgment tasks. *Psychological Science*, *15*, 100-105.

Ozuru, Y., Briner, S., Kurby, C. A., & McNamara, D. S. (2013). Comparing comprehension measured by multiple-choice and open-ended questions. *Canadian Journal of Experimental Psychology*, *67*(3), 215-227.

Perfect, T. J., & Schwartz, B. L. (Eds.). (2002). *Applied metacognition*. New York: Cambridge University Press.

Peterson, L. R., & Peterson, M. J. (1959). Short-term retention of individual verbal items. *Journal of Experimental Psychology*, *58*, 193-198.

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language*, *60*, 437-447.

Pyc, M. A., Rawson, K. A., & Aschenbrenner, A. J. (2014). Metacognitive monitoring during criterion learning: When and why are judgments accurate? *Memory & Cognition*, *42*, 866-897.

Rawson, K. A., & Dunlosky, J. (2002). Are performance predictions for text based on ease of processing? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(1), 69-80.

Rawson, K. A., & Dunlosky, J. (2007). Improving students' self-evaluation of learning for key concepts in textbook materials. *European Journal of Cognitive Psychology*, *19*, 559-579.

Rawson, K. A., Dunlosky, J., & McDonald, S. L. (2002). Influences of metamemory on performance predictions for text. *The Quarterly Journal of Experimental Psychology*, *55*(2), 505-524.

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

Reber, R., Winkielman, P., & Schwarz, N. (1998). Effects of perceptual fluency on affective judgments. *Psychological Science*, *9*(1), 45-48.

Rhodes, M. G., & Castel, A. D. (2008). Memory predictions are influenced by perceptual information: Evidence for metacognitive illusions. *Journal of Experimental Psychology: General*, *137*(4), 615-625.

Roberts, A. S. & A. A. Callender (2014). *Big and Bold Illusions: Effects of Font Size on Metacognitive Judgments*. Oral presentation given at 85th Annual Meeting of the Eastern Psychological Association, Boston, MA

Roediger, H. L., III, & Karpicke, J. D. (2006). The power of testing memory: Basic research and implication for educational practice. *Perspectives on Psychological Science*, *1*, 181-210.

Roediger, H.L., III, & Marsh, E.J. (2005). The positive and negative consequence of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 1155–1159

Sachs. J. S. (1967). Recognition memory for syntactic and semantic aspects of connected discourse. *Perception & Psychophysics, 2*, 437-442

Schraw, G. (1998). Promoting general metacognitive awareness. *Instructional Science*, *26*, 113-125.

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

Schraw, G. (2009b). Measuring metacognitive judgments. In D. J. Hacker, J. Dunlosky, & A. C.

Graesser (Eds.), *Metacognition in educational theory and practice.* (pp. 415-429). Hillsdale, NJ: LEA.

Schwartz, B. L. (1994). Sources of information in metamemory: Judgments of learning and feelings of knowing. *Psychonomic Bulletin & Review*, *1*(3), 357-375.

Schwartz, B. L., Benjamin, A. S., & Bjork, R. A. (1997). The inferential and experimental bases of metamemory. *Current Directions in Psychological Science*, *6*(5), 132-137.

Schwarz, N. (2004). Meta-cognitive experiences in consumer judgment and decision making. *Journal of Consumer Psychology*, *14*, 332-348.

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

Serra, M. J., & Metcalfe, J. (2009). Effective implementation of metacognition. In D. Hacker, J. Dunlosky, & A. Graesser (Eds.), *Handbook of metacognition in education* (pp. 278-298). New York: Psychology Press.

Singer, M., & Kintsch, W. (2001). Text retrieval: A theoretical exploration. *Discourse Processing*, *31*(1), 27-59.

Singer, M., & Ritchot, K. F. M. (1996). The role of working memory capacity and knowledge access in text inference processing. *Memory & Cognition*, *24*(6), 733-743.

Son, L. K. & Metcalfe, J. (2000). Metacognitive and control strategies in study-time allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 204-221.

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

Thiede, K. W., Anderson, M. C. M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology, 95*(1), 66-73.

Thiede, K. W., & Dunlosky, J. (1994). Delaying students' metacognitive monitoring improves their accuracy in predicting their recognition performance. *Journal of Educational Psychology*, *86*(2), 290-302.

Thiede, K. W., & Dunlosky, J. (1999). Toward a general model of self-regulated study: An analysis of selection of items for study and self-paced study time. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 1024-1037.

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*(6), 1267-1280.

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

Thiede, K. W., Wiley, J., & Griffin, T. D. (2011). Test expectancy affects metacomprehension accuracy. *British Journal of Educational Psychology*, *81*, 264-273.

Underwood, B. J. (1966). Individual and group differences of item difficulty for free learning. *Journal of Experimental Psychology*, *71*(5), 673-679.

van Loon, M. H., de Bruin, A. H., van Gog, T., & van Merriënboer, J. G. (2013). Activation of inaccurate prior knowledge affects primary-school students' metacognitive judgments and calibration. *Learning and Instruction*, *24*, 15-25.

Veenman, M. V. J., Van Hout-Wolters, B. H. A. M., & Afflerbach, P. (2006). Metacognition and learning: Conceptual and methodological considerations. *Metacognition and Learning*, *1*, 3-14.

Vitale, M. R., & Romance, N. R. (2007). A knowledge-based framework for unifying content-area reading comprehension and reading comprehension strategies. In D. McNamara

(Ed.), *Reading comprehension strategies: Theory, interventions, and technologies*. Mahwah, NJ: Erlbaum.

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

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

Weaver, C. A., III, & Burns, K. D. (1990). The effects of text revision on comprehension and comprehension monitoring. Paper presented at the meeting of the Psychonomic Society, New Orleans.

Whittlesea, B. W. A. (1993). Illusions of familiarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*, 1235-1253.

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

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language,* 46, 441-517.

Zaromb, F. M., Karpicke, J. D., & Roediger, H. L., III. (2010). Comprehension as a basis for metacognitive judgments: Effects of effort after meaning on recall and metacognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(2), 552-557. doi: 10.1037/a0018277

Zhao, Q., & Linderholm, T. (2008). Adult metacomprehension: Judgment processes and accuracy constraints. *Educational Psychology Review*, *20*, 191-206.