INTEGRATION OF COMPREHENSION AND METACOMPREHENSION USING NARRATIVE TEXTS

by

Matt C. Keener

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Educational Psychology

The University of Utah

August 2011

ABSTRACT

The purpose of the present research was to investigate text comprehension of narrative texts at varying levels of comprehension and examine how metacomprehension varies as a function of the level of comprehension when making retrospective (posttest) confidence judgments of performance. Using Kintsch's construction-integration theory of text comprehension, three types of question were developed to probe textbase and situation model levels of text representation at three levels of difficulty: (a) textbase, literal (easiest), (b) situation model, temporal ordering (low difficulty inferences), and (c) situation model, propositional logic (high difficulty inferences). Differences in percent correct, response time in milliseconds per character, and max amplitude of pupil size confirmed the predicted difficulty of the three question types, except that there was no significant difference in pupil size between the literal and temporal ordering questions. The three types of questions were then used to examine the effect of question difficulty on metacomprehension judgments of confidence, absolute accuracy (calibration accuracy and bias), and relative accuracy (*Goodman-Kruskal gamma coefficient* or *G*).

Results showed that readers were sensitive to different levels of comprehension and showed different levels of metacomprehension confidence and accuracy depending on the type of question. As predicted, absolute accuracy showed the effects of anchoring-and-adjustment when making these judgments across question type. That is, subjects appeared to be anchoring on a moderate estimate of success that corresponded most

closely in this study to performance on literal questions and adjusted their confidence for temporal ordering and propositional logic questions. The results related to bias provided support for the hard-easy effect, with propositional logic questions (i.e., hard questions) showing overconfidence and literal questions (i.e., easy questions) showing no significant bias, although bias scores did not discriminate between temporal ordering and propositional logic questions. As predicted, relative accuracy (*G*) appeared to be stable across question types with no significant differences by question type. As with previous studies, the differences in the results concerning absolute versus relative accuracy suggest that the two types of accuracy are measuring different components of metacomprehension.

Dedicated to Carrie A. Beach,
for her *continued* love, support, understanding,
and gracious research assistance

TABLE OF CONTENTS

ACKNOWLEDGMENTS

I would like to thank at least some of those people who helped me to complete this dissertation. First, my parents brought me into this world. Therefore, one might argue that they merit at least some of the credit or the blame in this matter.

Doug Hacker, my committee chair and mentor in the learning sciences. He endured countless discussions and should be commended for his patience and passion for research, although he made a questionable choice of mentee. He tried to teach me a great deal about processes involved in writing and metacognition.

Kirsten Butcher was a committee member who provided helpful feedback about the dissertation, especially about reading processes and the construction-integration model—and sometimes at her own considerable inconvenience. Toward completion of the dissertation, she was traveling out of the country and she texted me messages from the airplane before takeoff and after landing in order to help me make a deadline.

Anne Cook was a committee member who tried to teach me a great deal about processes involved in reading and different methods of examining them. She was a second mentor to me who provided important guidance as I learned to teach, during a period of time when Doug abandoned me and disappeared (i.e., “sabbatical”).

Denise Francis-Montano has been a good friend for years. She provided feedback and support as a peer researcher while I worked on my masters thesis. Then, years later, she helped to schedule my dissertation defense at the last minute while Carrie and I were

in Boston getting married—and she took great care of the cats.

John Kircher should have been on my committee and I do not remember him having a good excuse. He provided more feedback than anyone except for Doug and Adrienne. He helped in the design of FORTRAN programs that were necessary to conduct the research. He also tried to teach me a great deal about statistics and psychophysiology.

Adrienne Splinter was a cohort doctoral student in the learning sciences program who provided feedback and support throughout the entire process of developing and conducting the research. She has been a good friend. I was fortunate that she shared the journey of the doctoral program with me and I hope that we will work together further in developing other assessments of reading comprehension (and metacomprehension).

Dan Woltz was a committee member who tried to teach me a great deal about processes involved in memory, as well as statistics and psychological measurement. He provided a lot of important feedback and raised measurement issues that were not explored in the dissertation study. I hope to explore those issues eventually.

Finally, I would like to thank Sean Casey, Takashi Furuhata, Natalie Harris, Michael Johnson, Brian Kuhlman, Dan Olympia, Pooja Patnaik, Mark St. André, Dave Strayer, Kristin Swenson, Andrea Webb, and Yanli Yuan for offering feedback and support during the process of developing the dissertation and/or conducting the research.

CHAPTER I

INTRODUCTION

Overview

In recent years, researchers have made attempts to integrate theories (or models) of text comprehension with theories of metacomprehension of text, or the ability to judge one's own comprehension of a text (Dunlosky, Rawson, & Hacker, 2002; Wiley, Griffin, & Thiede, 2005). Much of this research has been exploratory, but current findings have shown that accuracy of metacomprehension judgments is tied not only to general text comprehension but also to specific kinds of text comprehension (Dunlosky, Rawson, & Middleton, 2005; Salmen, 2004). The concept of comprehension implies more than one possible level of comprehension, ranging from literal memory of text to inferential processing (Kintsch, 1998), and therefore, the concept of metacomprehension implies more than one possible level of metacomprehension. The purpose of the present research was to investigate text comprehension at literal and inferential levels and examine how metacomprehension varies as a function of level of comprehension.

According to the construction-integration model (Kintsch, 1998), a reader's comprehension of text material may vary at different levels of comprehension. For example, reading at a textbase level of comprehension may require the simple retrieval of information from memory about the text, whereas reading at a situation model level of comprehension may require the generation of inferences based on information from the

text and the reader's background knowledge. Although some types of inference appear to be relatively automatic for readers, others require greater cognitive effort (Graesser, Louwerse, McNamara, Olney, Cai, & Mitchell, 2007; Zwaan & Radvansky, 1998).

Examining varied levels of comprehension has important implications for metacomprehension of text (Dunlosky et al., 2002; Wiley et al., 2005). Although the construction-integration model has been applied only in limited ways in the research on metacomprehension (Dunlosky et al., 2002; Salmen, 2004), bringing theories of comprehension together with theories of metacomprehension provides a promising new area of research that can potentially inform both areas. Varied levels of comprehension should arguably impact how readers monitor their comprehension and how accurately they can monitor. Wiley et al. (2005) suggested that comprehension at the situation model level should provide a better measure of comprehension than the textbase level because it is the integration of textual information with background knowledge that is the desired goal of reading. With this in mind, knowing how readers monitor their comprehension at the situation model level and whether they can accurately monitor across all types of inferences are important questions to be researched. In addition, because most research of metacomprehension has focused on expository texts, examining these questions with narrative texts can potentially expand our knowledge of metacomprehension.

This study followed an integrated approach to assess comprehension and metacomprehension using narrative texts and three types of questions that measure varied levels of comprehension: literal questions (textbase), temporal ordering inference questions (situation model, low difficulty), and propositional logic inference questions

(situation model, high difficulty). To examine metacomprehension, posttest confidence judgments of performance and metacomprehension accuracy were measured for each of these question types.

Review of the Literature

Comprehension of Text

According to the construction-integration model (Kintsch, 1998; Kintsch & van Dijk, 1978), a reader's comprehension of text involves the construction of an internal representation of the information presented in the text and the subsequent integration of this information with existing knowledge. Kintsch (1998) proposed that "a context-insensitive construction process is followed by a constraint-satisfaction, or integration, process that yields if all goes well, an orderly mental structure out of initial chaos" (p. 5). The reader's text representation may be separated into three levels of comprehension: (a) the lexical, or surface features of text (words, syntax); (b) the textbase, or basic propositions found in text; and (c) the situation model, or propositions that elaborate into a whole and extend to general world knowledge and personal experience.

Assessments of comprehension have included the use of different types of questions to measure different levels of comprehension. Lexical comprehension can be measured with word-level types of questions, although decoding of words becomes automatic for most readers relatively quickly. Therefore, most research has focused on textbase and situation model levels of comprehension. Textbase comprehension can be measured with literal questions because the required information appears explicitly in the text or can be derived by making basic inferences that are required for maintaining the local coherence of a text. Finally, a situation model level of comprehension can be

measured with inferential questions because comprehension at this level requires the integration of textual information with the reader's background knowledge; therefore, some information required to answer the question does *not* appear directly in the text. Correctly answering these questions would indicate that readers were successful at retaining some information from the text and constructing the remaining information through inferential processing within the context of their background knowledge. Whether the inference is constructed during reading (i.e., online) or only after reading (i.e., offline) when prompted by the question is not captured by this type of comprehension assessment.

Kintsch (1998) provided a simplified way to classify reading inferences (see Table 1). In this typology, all inferences can be classified in one of four categories defined by two axes: automatic versus controlled, and retrieval versus generation. *Automatic inferences* require less cognitive effort in comparison to *controlled inferences*; and *retrieval* implies that the information needed for the inference may be easily accessed

Table 1

A Classification System for Inferences in Text Comprehension

(Adapted from Kintsch, 1998, p. 189)

	Retrieval	Generation
Automatic processes	A. bridging inferences, associative elaborations	C. transitive inferences in a familiar domain
Controlled processes	B. search for bridging knowledge	D. *logical inferences* * [emphasis added]

* Including inductive and deductive types of logic

from long-term memory whereas *generation* requires the generation of new information. For Kintsch, the inferences that best demonstrate a situation model level of comprehension require *controlled generation*—for example, inferences that involve "deductive reasoning" (p. 192). In theory, the nature of the inference corresponds to the nature of the situation model being applied; therefore, Kintsch has inferred the importance of establishing the specific nature of situation models in addition to simply distinguishing between textbase and situation model levels of comprehension.

Other examinations of inferences have expanded on the kinds of inference that are associated with text comprehension. For example, Graesser, Louwerse, McNamara, Olney, Cai, and Mitchell (2007) classified 13 types as a "landscape of inferences" that might help researchers to address the whole landscape rather than ignoring large sections of it (p. 290). They also provided four general categories of inferences based mainly on increasing levels of cognitive effort: (a) *automatic*, generated very reliably and quickly (within about 500 ms) with minimal effort; (b) *routine*, more effort but still reliable and relatively quick (within about a second); (c) *strategic*, being sensitive to reader's goals and strategies but moderately quick; and (d) *off-line*, made only after reading of text with considerable time and effort.

In related research, Zwaan and colleagues (Zwaan, Langston, & Graesser, 1995; Zwaan & Radvansky, 1998) classified five dimensions that appear to be monitored automatically while reading narrative texts: (a) *time*, the temporal order of narrative events, (b) *space*, the most immediate area(s) in which events occur, (c) information related to the *protagonist*, (d) *causality* of relevant events, and (e) *intentionality* of relevant characters. Although this research was not concerned specifically with

inferences, these narrative dimensions may be assessed with either literal or inferential questioning. For example, if the temporal ordering of narrative events is monitored automatically by readers, then related inferential questions should be relatively easy.

According to the construction-integration model (Kintsch, 1998), inferences that require controlled generation (see Table 1, cell D) provide the best measure of a reader's situation model, and inferences that fall within this category include logical inductive and deductive inferences (Graesser, Millis, & Zwaan, 1997; Kintsch, 1993; Kintsch, 1994). Inductive logic involves using premises to form generalizations that appear probable but may not be true; for example, if every raven seen by a person were black, that person might conclude that all ravens were black even though it is logically possible that non-black ravens exist. Deductive logic involves using premises to form necessary conclusions; for example, if it is given that all ravens are black, then it follows necessarily that any bird identified as a raven must also be black. This study focused on deductive rather than inductive logic because necessary conclusions should produce clearer responses than probable conclusions. More specifically, the study involved propositional logic. Propositional logic is a specific type of deduction that has been applied to the "natural" or everyday propositions that exist as complex constructions in text and other discourse, in contrast to other basic forms of deductive logic (Braine, Reiser, Rumain, 1998; Braine et al., 1995).

Although these types of inference are generally thought of as belonging to special cases of logic and are not "naturally" occurring, readers of narrative text appear to be able to make propositional logic inferences (Franks, 1997) and under certain conditions they may do so automatically (Lea, 1995). Lea showed that when presented with five-to-six

contiguous propositions in narrative text, readers routinely made propositional logic inferences and often did not realize that they were making them. Franks (1997) found that readers of varying age and ability can make propositional logic inferences to some degree of success and that the relative increases in performance for different types of logic and content across age groups were generally consistent with expectations of cognitive development. Also, Rader and Sloutsky (2002) found that readers may be primed for propositional logic inferences regardless of whether those inferences were logically valid. Thus, under certain conditions, readers can make these kinds of inference, and this process may be automatic when reading short narrative texts or short contiguous propositions that have been constructed in a way that is conducive for an automatic direct-reasoning approach (Lea, 1995). Lea, Mulligan, and Walton (2005) also investigated whether readers make propositional logic inferences that were presented at some distance from one another, and found that these inferences appeared to be made more often when a later contextual cue was presented with the second premise to reactivate the first premise; however, these texts were presented a sentence at a time. It remains to be explored whether readers make propositional logic inferences when reading more naturalistic texts that are longer than the typically short texts used in prior studies, as the present study does, and when they are not being cued for the inferences in proximity to the premises.

It is important to note that some researchers (e.g., Gerrig & O'Brien, 2005) argue that automatic inferences are those that only require memory retrieval and that classifications of inferences may be misleading because of exceptions to the rule. However, my purpose in using these classifications was to serve as a foundation for the

generation of texts and questions that would present varying levels of difficulty. It is also important to note that many studies of inference generation in reading comprehension have been concerned with *online inferences*, meaning the inferences were assumed to be made while in the process of reading the text (e.g., Lea, 1995; Zwaan et al., 1995). In such studies, the presentation of text materials have been tightly controlled in order to look for evidence of inference generation during or immediately after the reading of necessary information. In contrast, this study used longer, more naturalistic texts and each of these was followed by an assessment of comprehension with questions that may have actively promoted the generation of *offline inferences,* or inferences that were made after reading the text. For this reason, the specific timing of inference generation was not established in this study. Nevertheless, given that temporal ordering inferences with narrative texts are typically easier to generate online than propositional logic inferences, it seems likely that this would be the case for similar inferences generated offline.

In sum, I conducted the first and second pilot studies to generate narrative texts and questions that would require retrieval of literal information from the texts and making temporal ordering and propositional logic inferences, which would differ on levels of difficulty as measured by percent correct and response time. Once these questions were generated, I then moved on to the main focus of this research, which was to examine metacomprehension on these differing levels of comprehension.

Metacomprehension of Text

Construction-integration (Kintsch, 1998) is a model of text comprehension; however, it has been justifiably applied to the metacomprehension of text (Dunlosky, Rawson, & Hacker, 2002; Wiley, Griffin, & Thiede, 2005), which is typically defined as

the ability to judge one's own learning or comprehension of text material (Dunlosky & Lipko, 2007; Maki & Berry, 1984). In the following section, I provide a brief general discussion of metacomprehension and follow with the implications for the construction-integration model.

Metacomprehension is a specific application of metacognition, which may be defined simply as thinking about one's own thoughts. Two basic characteristics of metacognition are relevant to metacomprehension. First, abundant evidence supports a distinction between automatic and controlled processes in metacognition, with automatic processes being implicit and fast compared with controlled processes being explicit and slower (Efklides, 2008). Second, controlled metacognitive processes may become activated by task difficulty (Alter, Oppenheimer, Epley, & Eyre, 2007), with more difficult tasks requiring greater control over cognitive processes.

Metacomprehension accuracy has been measured using several different measures. Each one provides insights into different aspects of metacomprehension, and in general, these measures may be assigned to either absolute accuracy or relative accuracy (Maki, Shields, Wheeler, & Zacchilli, 2005; Schraw, 2009). Three common measures were chosen: (a) calibration accuracy and (b) the bias score as measures of absolute accuracy, and (c) Gamma as a measure of relative accuracy.

Absolute accuracy, also known as calibration, is described as the degree of fit between a person's probability judgments of performance and his or her overall performance (Keren, 1991; Maki et al., 2005). Investigations of absolute accuracy often have included measures of accuracy before and after performance on a comprehension test. Readers make predictive comprehension judgments following the reading of text

and prior to answering test questions about it, with the purpose of predicting comprehension as measured by performance on the test, for example, "What percentage of items on this test of comprehension do you think you will get correct?" Posttest confidence judgments are made after answering questions on the text with the purpose of judging actual performance on the answered questions, for example, "What percentage of items on this test do you think you got correct?

The absolute value of the difference between judged and actual performance provides a measure of the magnitude of accuracy (i.e., calibration accuracy), with values closer to zero indicating perfect accuracy (Hacker, Bol, & Keener, 2008). Some researchers (e.g, Yates, 1990) have proposed the use of the Brier (or quadratic) score as a measure of calibration accuracy, which is calculated by squaring the difference between judged and actual performance, because this has been adjusted to better predict probabilities of outcomes (Brier, 1950). However, the Brier score contains components in addition to calibration (e.g., refinement) (Blattenberger & Lad, 1985) and may therefore be a hybrid measure of accuracy (Schraw, 2009); therefore, in the present study, calibration accuracy was measured more simply using the absolute value of the difference between judgments of performance and actual performance.

The bias score has been included in the present study as another measure of absolute accuracy, because it provides information about absolute accuracy that differs from the magnitude of accuracy. Bias is calculated as the signed difference between prediction and performance, and this provides a measure of overconfidence or underconfidence, with negative values indicating underconfidence and positive values indicating overconfidence. With posttest confidence judgments, there have been

interesting and somewhat counter-intuitive results found regarding systematic overconfidence or underconfidence in the judgments. An observed effect has become known as the "hard-easy effect" wherein difficult test items tend to produce overconfident judgments whereas easy items tend to produce underconfident judgments (Kleiter, Doherty, & Brake, 2003; Suantak, Bolger, & Ferrell, 1996), although the reasons for the effect are still a matter of dispute (cf. Juslin, Winman, & Olsson, 2000). And in another study, verbal ability was linked to levels of confidence: Readers with higher verbal ability were underconfident in judging performance on hard texts, whereas other readers showed little over- or underconfidence on hard texts and were generally quite accurate for easy texts (Maki et al., 2005).

Absolute accuracy, whether measured by a form of calibration accuracy or the bias score, tends to be more accurate on posttest confidence judgments than predictive judgments of comprehension. This is likely due in part to the increased specificity of posttest judgments (Hacker, Bol, Horgan, & Rakow, 2000; Maki, 1998) and this has been referred to as *the testing effect* (Pressley & Ghatala, 1990).

Relative accuracy has been the primary measure of metacomprehension of text (Dunlosky & Lipko, 2007). Relative accuracy measures the degree to which a person's judgments can predict the likelihood of correct performance of one item relative to another (Nelson, 1984). Procedurally, judgments of relative accuracy are obtained by asking readers subsequent to reading a text and prior to being tested for comprehension of the text, "How confident are you that you can answer questions about the text?" Participants provide their confidence judgment from 0% confident to 100% confident. The confident judgments are then correlated with performance. The statistic commonly

used to correlate confidence with performance is the Goodman-Kruskal gamma coefficient (*G*), with 0 indicating chance, -1 indicating poor accuracy, and +1 indicating perfect accuracy. Perfect accuracy means that for every time a person gave high confidence judgment in knowing a question, he or she knew the answer, and for every time a person gave a low confidence judgment, he or she did not know the answer.

Although the use of G in studies of metacomprehension has typically not been sensitive to individual differences, such as verbal ability, and has not shown strong reliability (Masson & Rotello, 2009), it has remained the measure of choice for relative accuracy because it has been shown to be a meaningful measure of accuracy by cross-validating it with other measures of accuracy. In contrast, measures of absolute accuracy have shown sensitivity to individual differences and stronger reliability. In some instances, the reliability of absolute accuracy has exceeded the reliability of test performance (Hacker et al., 2000).

People can provide accurate measures of absolute accuracy but not relative accuracy. For example, a person might make a judgment that he or she will get 80% of the items on a test correct and get 80% correct (i.e., perfect calibration accuracy), but do very poorly at judging exactly which items are correct or incorrect (i.e., a person does not discriminate well between what is known or not known). Maki et al. (2005) have shown that there are weak relations between absolute and relative accuracy, leading these researchers to conclude that "these two types of metacognition tap different processes" (p. 728).

Metacomprehension judgments, whether relative or absolute, can be helpful in improving comprehension. As a predictive judgment, when students judge that they have

not understood a text prior to a test, they can then reread for improved understanding. As a postdictive judgment following a test, when students judge that they have not performed well on particular test items, they can emphasize those particular topics in future study or revisit the test question. The judgment of not understanding is the result of metacognitive monitoring, and rereading is the result of metacognitive control (Nelson & Narens, 1990), both aspects of metacognition being essential ingredients of self-regulation of study (Thiede & Dunlosky, 1999). Effective self-regulation of study requires that predictive judgments of comprehension are accurate.

Unfortunately, studies of both absolute and relative measures of metacomprehension have shown only moderate levels of accuracy. However, researchers are exploring interventions using strategies that appear to improve readers' relative accuracy by encouraging deeper engagement with the text; for example, accuracy was significantly improved by summarizing (Thiede & Anderson, 2003), applying self-explanation strategies while reading (Griffin, Wiley, & Thiede, 2008), or simply rereading texts (Rawson, Dunlosky, & Thiede, 2000). To match judgments more closely to the content with the goal of improving predictive power, researchers have used term-specific judgments regarding a few main concepts that appear in the texts, which helps to specify the nature of the prediction (Dunlosky & Lipko, 2007).

Similarly, other researchers have been exploring interventions to improve absolute accuracy. Schraw, Potenza, and Nebelsick-Gullet (1993) found that incentives for calibration accuracy increased calibration accuracy and performance. Hacker et al. (2008) investigated the impact of extrinsic incentives and reflection on college students' calibration judgments and accuracy on exam performance, as well as relationships

between explanatory style and calibration. Their results indicated that achievement level was associated with calibration accuracy. Higher-achieving students were very accurate in their calibrations and were less affected by incentives and reflection. Lower-achieving students were less accurate in calibration but showed significant improvement in accuracy when incentives were offered. Also, Koku and Qureshi's (2004) research further supports the findings related to calibration accuracy and achievement. They assert that high-performing students are more likely to recognize the extent and limitation of their knowledge, while low-performing students have limited insight into their performance. They suggest using an intervention that requires students to respond to and justify each question response, theorizing that this will increase student metacognitive processes and thereby result in increased calibration accuracy and improved performance.

Comprehension is an important part of metacomprehension, and the construction-integration model is relevant to both because interpreting the meaning of any metacomprehension judgment arguably depends on the level of comprehension being tested (Wiley et al., 2005). Our knowledge of metacomprehension will be better informed by knowing more about the influence of levels of comprehension, especially the inferences that contribute to the level of the situation model, which may provide the best representation of text comprehension (Graesser et al., 1997; Kintsch, 1994). There has been some related research examining accuracy of metacomprehension with tests or texts of varying difficulty. Using relative accuracy of predictive judgments, Weaver and Bryant (1995) examined the effect of text difficulty on judgments of comprehension using narrative and expository types of texts, and they found with both types of texts that subjects were most accurate (using *G* scores) with texts of moderate difficulty, with lower

accuracy for both easy and difficult texts. They termed this the *optimum effort hypothesis* because it may indicate that readers choose a moderate level of effort required for a task, or set of related tasks, and this is generally the most efficient approach for maximum overall accuracy. These findings are similar to the anchoring-and-adjustment heuristic (Epley & Gilovich, 2006; Tversky & Kahneman, 1974) in which people tend to use some value as a starting point, or anchor, in making estimates over a set of related tasks, and then tend to make insufficient adjustments from this anchor in making successive estimates. In other words, the optimum effort hypothesis may imply that readers choose a moderate estimate as the initial anchor, at least with predictive judgments, and then insufficiently adjust from this anchor when judging easier and more difficult texts.

Scheck, Meeter, and Nelson (2004) examined anchoring with the absolute accuracy of immediate versus delayed judgments of learning. They tested the results against three hypotheses: (a) the anchoring hypothesis, proposing an extreme anchor that does not change; (b) the monitoring hypothesis, proposing that monitoring is solely responsible for judgments and there will be no anchor effect; (c) the dual-factors hypothesis, proposing that both anchoring and monitoring will occur to some degree, meaning, there will be evidence of an anchor although subjects will make (imperfect) adjustments to their judgments due to monitoring—a proposal that is very similar to the anchoring-and-adjustment heuristic. The dual-factors hypothesis was consistent with their results.

Also using relative accuracy of predictive judgments, Zaromb, Karpicke, and Roediger (2010) examined the effect of sentence difficulty (hard or easy) with judgments of either comprehension or learning and found in either case that subjects' confidence

was greatest with easy sentences, regardless of experimental conditions that manipulated the use of *effort after meaning* (i.e., completing the tasks with or without clues). Zaromb et al. also manipulated recall performance with two types of clues (embedded or delayed) and a control group who received no clues. Based on levels of confidence, subjects appeared to be unaware of the manipulation, which significantly enhanced recall performance. Therefore, predictive relative accuracy as measured by correlations between judgments and recall performance was greatest in the control condition, with lower accuracy in the embedded-clue condition and very low accuracy (no significant correlations) in the delayed-clue condition. Nevertheless, despite the subjects' inability to detect the benefits of the manipulation, the findings suggest that the subjects were able to monitor the differences in text difficulty and their own related level of comprehension to a significant degree. It may be that the subjects had chosen a more typical anchor in the context of reading for making their judgments (i.e., the control condition with no clues), and the subjects' predictive accuracy in the experimental conditions appeared to suffer in comparison. These findings as well as the implications of the optimum effort hypothesis (Weaver & Bryant, 1995) suggest that using relative accuracy, readers make judgments of metacomprehension with anchors that reflect typical expectations or moderate levels of difficulty related to texts—and a moderate level of difficulty may be a typical expectation, depending on the circumstances and information available.

Using the absolute accuracy of posttest judgments, Schraw et al. (1993) found that subjects were more accurate in their calibration accuracy on easy tests compared with difficult tests and they suggested that subjects performed more consistently on easy items. Therefore, there was a stronger correspondence with those judgments. Another

interpretation could be the finding that people often use 75% as a self-chosen anchor to estimate a moderate probability of success (Epley & Gilovich, 2006; Hacker & Bol, 2011), because mean performance on the easy tests in the study by Schraw et al. was about 75% and mean performance on the difficult tests was about 46%. Scraw et al. (1993) also found that subjects were generally overconfident on difficult tests and underconfident on easy tests, replicating the hard-easy effect. Because of this study's use of posttest judgments, the findings from it were especially relevant to the present study in which I sought to use the levels of comprehension discussed previously as a way to investigate readers' ability to monitor their comprehension at each of these levels using posttest judgments.

Expository and Narrative Texts

Different genres of texts present differing characteristics and instructional implications. Although there are many approaches in the literature to defining the construct of "genre" (Johns, 2002), for the present study, it is sufficient to consider the two "macro-genres" of narrative and expository texts (Grabe, 2002). Expository texts may be broadly defined here as informational in nature, where the author intends to explain or define something to the reader using some type of prose. Narrative texts may be broadly defined as storytelling in nature, where the author intends to describe a sequence of events that may include fiction and/or nonfiction using a story structure that includes a protagonist, goals, and outcomes.

Comparing these genres, narrative discourse in general appears to be more common at a younger age than expository discourse (Berman & Katzenberger, 2004). Following from the informational nature of expository texts, specific prior knowledge

influences the comprehension of expository texts, whereas narrative text is influenced more by general world knowledge that often relates in some way to common experiences (Best, Floyd, & McNamara, 2008; Wolfe & Mienko, 2007). Consistent with these findings, narrative texts appear to be generally easier to comprehend than expository texts (Best, Floyd, & McNamara, 2008; Weaver & Bryant, 1995).

Also, Weaver and Bryant (1995) found that predictive metacomprehension accuracy differed by type of text and question. For expository texts, accuracy was better for detail-oriented questions as compared with thematic questions; for narrative texts, accuracy was better for the thematic questions. According to the construction-integration model (Kintsch, 1998), the detail-oriented questions should apply most to textbase comprehension, whereas thematic questions are types of inferences that should apply to a situation model level of comprehension. In other words, predictive metacomprehension accuracy for questions related to a situation model level of comprehension was better for the narrative texts than expository texts.

Despite the challenges presented by expository texts, current metacomprehension research has focused on them because of their important role in educational contexts (Wiley et al., 2005). However, because relatively little research has been conducted using narrative texts and because of their unique strengths over expository texts (e.g., easier comprehension), the present study uses them to explore comprehension and metacomprehension. Moreover, the narrative texts that have been used in previous research (a) have been short, consisting of about 100 to 200 words, (b) have been contrived in their construction so that propositions needed to generate inferences, whether automatic or generative, were placed in contiguous sentences, with little or no intervening

propositions, and (c) presented a limited number of antagonists or protagonists who were involved in very simple themes. The narrative texts used in the present research are longer, each being 600 words long, have propositions necessary for making inferences spaced out across the text with one or more sentences not directly relevant to the inference intervening between the propositions, and have several characters that are engaged in more complex and realistic themes.

Pupil Size and Cognitive Effort

To examine further subjects' difficulties in answering the three types of question, I introduced into the main study an examination of pupil size of each subject over the course of the experiment based on research that pupil size positively correlates with cognitive effort or demands (e.g., Ahern, 1978; Ahern & Beatty, 1979; Beatty, 1982). Although the method of examining pupil size as an indicator of cognitive effort over the course of a process such as problem-solving is still a matter of discussion (Cook, Zheng, & Blaz, 2009; Just & Carpenter, 1993; Xie & Salvendy, 2000), there is evidence provided by Just and Carpenter (1993) that peak amplitude for each task is a primary correlate of cognitive effort, and therefore I measured the peak amplitude for each task.

Research Questions

The purpose of the present research was to further integrate the investigation of comprehension with metacomprehension by examining how accurately readers monitor their comprehension at varying levels of difficulty. To provide converging evidence of cognitive effort, response time (or latency) and pupil width were measured during the assessment. To measure comprehension, questions were presented that assess comprehension at three levels: (a) textbase, (b) situation model, temporal ordering (low

difficulty inferences), and (c) situation model, propositional logic (high difficulty inferences). To measure metacomprehension, subjects provided prospective confidence judgments regarding their performance on each section of text and retrospective confidence judgments regarding their performance on each recently-answered question (i.e., immediate posttest judgments). The main research questions related to metacomprehension were limited to the posttest judgments because of their paired relationship to the individual questions.

My research questions addressed both comprehension and metacomprehension. The questions related to comprehension were: (a) Will performance in percent correct vary by question type with literal questions being the easiest and propositional logic questions being the most difficult?; (b) Will response time in milliseconds per character vary by question type with literal questions having the shortest and propositional logic questions having the longest?; and (c) Will maximum amplitude of pupil size vary by question type with propositional logic questions having the largest and literal questions the smallest?

I expected to find that inferential questions would be more difficult to answer than literal questions, and of the two types of inferential questions, the propositional logic questions will be the most difficult. In other words, literal questions should be the easiest, followed by temporal ordering questions, and propositional logic questions should be the most difficult. This ordering by level of difficulty should be evidenced by a decrease in performance with increased difficulty. In addition, this ordering by level of difficulty should be evidenced by increased response time and increased pupil size. The questions related to comprehension served as a manipulation check to verify that the

design of the question types produced the expected results.

The questions related to metacomprehension were: (a) Will the magnitude of metacognitive posttest confidence judgments vary by question type with confidence being greatest for judgments related to literal questions and lowest for judgments related to propositional logic questions?; (b) Will posttest measures of calibration accuracy vary by question type with greatest accuracy for the question type that was answered closest to a moderate anchor of success and with decreasing accuracy for question types that move away from this anchor?; (c) Will bias vary by question type with the greatest overconfidence related to difficult questions (propositional logic) and the least overconfidence (or underconfidence) related to easy questions (literal)?; and (d) Will levels of relative accuracy (*G*) remain consistent across question types?

Confidence in judging comprehension should vary with question difficulty because there already exists evidence that readers can monitor their level of comprehension to a significant degree (Zaromb et al., 2010). Thus, the magnitudes of posttest confidence judgments will be greatest for literal questions and less for inferential questions, with confidence lowest for propositional logic questions.

Subjects were also expected to anchor on an estimate of moderate success that may correspond roughly with 75%—or its equivalent (Hacker & Bol, 2011). The subjects in this study did not use numbers to make confidence judgments. Rather, they made judgments on a continuous scale between no confidence and total confidence, and these judgments were later transformed into values between 25 and 100% because no confidence, or guessing behavior, should produce 25% accuracy by chance given a four-choice format. Therefore, a spatial estimate of 75% on the scale in this study was defined

in numeric terms as 81.25—or about 81%, given that these are rough estimates. Calibration accuracy was predicted to be most accurate for the question type that was answered closest to 81%, with decreasing accuracy for question types that move away from this anchor.

Based on the hard-easy effect observed with posttest judgments, bias was predicted to show the greatest overconfidence with difficult questions (propositional logic) and the least overconfidence (or underconfidence) with easy questions (literal). Overconfidence, in general, was expected because this is a prevalent finding in studies of metacomprehension and bias.

Finally, it was already predicted that subjects should adjust their confidence judgments to compensate for varying levels of question difficulty and that these adjustments should be inadequate in terms of scale. Therefore, it was predicted that relative accuracy (*G*) should not vary by question type because the insufficiency of adjustments in terms of scale will not affect relative accuracy as long as the adjustments move in the right direction (see the first prediction related to metacomprehension). In other words, subjects should monitor to some degree whether one question is more or less difficult than other questions, regardless of its question type, and adjust their confidence judgments accordingly (i.e., relative accuracy), because relative accuracy is only concerned with whether questions are more or less difficult. The size of the adjustments made between confidence judgments only affects absolute accuracy.

CHAPTER II

METHOD

Design

The present study used a mixed design to investigate text comprehension and metacomprehension. Specifically, this was a (6 x 3) within-subject design with six trial positions and three question types, both repeated measures.

Subjects

One hundred and twenty-seven subjects from the Educational Psychology subject pool participated to satisfy 1 hour of their research participation requirement. Of these, 7 subjects did not speak English as a first language. Therefore, their data were not included in the analysis, leaving a sample of 120 subjects. The subjects were college students with a mean age of 24.4 years (*SD* = 6.43). Of the 120 subjects, 95 were female and 108 were Caucasian/White, 6 were Hispanic/Latina, 3 were African American/Black, 1 was Asian, 1 was American Indian, and 1 was self-identified as "other").

Hardware and Software

Personal Computer (PC) with Windows

A PC with Windows XP as the operating system was used to run the eyetracking hardware and software necessary for the experiment. The PC had a standard keyboard, mouse, and an SVGA monitor running at a resolution of 1280x1024.

Eye Tracker

An Arrington eye tracker attached to the PC with an internal PCI card measured pupil size during the assessment. Other pupillometric data such as eye position was recorded for later reference but was not used in this study. A version of ViewPoint software (developed by Arrington Research) recorded pupillometric data that were coordinated with the automated assessment program designed to display the materials and record data related to performance, judgments, response time, response time in milliseconds per character when applicable, and other details.

Automated Assessment

The Automated Assessment of Reading Comprehension and Metacomprehension Using Narratives (AARCMUN) program was designed by the author (with programming assistance from Dr. John Kircher) to present the assessment to the subjects. The narrative stories as well as metacomprehension judgments and comprehension questions were developed during Pilot Studies 1 and 2.

Log Analysis

A software program was designed in FORTRAN by the author to read the log files created by the AARCMUN program, do the necessary calculations, and provide the dependent variables for each subject in comma-delimited files for analysis, as well as providing a file designed to be read by the MCT Analysis program (see below).

MCT Analysis

A software program was designed in FORTRAN by the author to read text files containing multiple-choice test (MCT) data and then conduct an analysis of items. This

analysis included all items and trial positions, providing performance for each subject on particular items and a discrimination score for each item to check for problematic ones (i.e., items with a negative discrimination value would indicate that high-performing subjects did more poorly on those items than low-performing subjects).

Pilot Study 1

To develop an assessment that would measure textbase and situation model levels of comprehension using narrative texts, I conducted a pilot study to test literal and inferential levels of comprehension, using propositional logic questions (i.e., inferences that may best indicate situation model development; Kintsch, 1998). However, due to the difficulty in constructing a sufficient number of propositional logic questions, temporal ordering questions were added to balance the number of literal and inferential questions and significant differences between the two types of inferential questions became apparent in the results.

I developed relatively authentic texts by employing a top-down approach to writing them; therefore, the focus was placed on narrative qualities such as cohesive themes and meaningful beginnings and endings within each text and each section of text. The texts are provided in the Appendix, although the questions there were changed for the dissertation study (see Revisions to the Pilot Assessment). There were no controls for factors such as the types and number of specific propositions or sentential structures. Each text was divided into three sections (200 words each) and shown to subjects one at a time on a computer monitor. Subjects could navigate back and forth through the three sections of text. Using entire sections of text allowed for more natural reading compared with more strictly controlled methods, such as presenting the text to the reader one

sentence at a time.

Assumptions underlying the selections of narrative texts were twofold: Popular books that were thought to be more engaging by virtue of their popularity, and unfamiliar texts were expected to provide a better opportunity for new learning at the level of the situation model than familiar ones. As a compromise between these two competing assumptions, the least-known titles from a list of popular books were chosen. A list of popular literature titles (based on titles available as Cliff notes) was developed into a questionnaire to which a person could indicate his or her level of familiarity on a six-point likert scale, ranging from no familiarity to total familiarity. Any titles associated with television programs or films were not included because of the increased chance of familiarity. These familiarity questionnaires were completed by a convenience sample of undergraduate students (n = 74) enrolled in an educational psychology course as part of a teacher education program who volunteered to complete them.

Based on the results of the questionnaires, six books were chosen: *Black Boy* by Richard Wright; *Black Elk Speaks* by John G. Niehardt; *Bless Me, Ultima* by Rudolfo Anaya; *Death Comes for the Archbishop* by Willa Cather; *Night* by Elie Wiesel; and *Steppenwolf* by Hermann Hesse. The six texts developed for the assessment were based on portions of these books and provided a simplified synopsis of the events. In doing so, liberties were taken with respect to details and presentation.

For assessing comprehension of the texts at the levels of textbase and situation model comprehension, six literal and six inferential questions were developed for each text. For clarity, the title of the prior story was provided above each question in the assessment window; for example, “Based on ‘Richard and His Education,’ choose the

best answer (a or b) to complete the following statement." The questions were in two-choice format and related to content distributed evenly across the three sections. The literal questions included information found directly in the text. For example, "daily life for Richard was made _____ in some very significant ways as a result of his self confidence"; and the choices were "easier" or "more difficult."

The six inferential questions consisted equally of two types: temporal ordering and propositional logic. The temporal ordering questions were inferences of "time" that are believed to be generated automatically while reading narrative texts (Zwaan & Radvansky, 1998). For example, "Richard moved in 1925 to Memphis, and this led eventually to a time when he _____"; and the choices were "met a man named Shorty" or "began to experience persecution."

The propositional logic questions were modeled after the general form of modus ponens logic (or occasional instances of its negation) because this is a common logical form that has been previously applied by Franks (1997) and Lea (1995). Modus ponens makes use of a rule of inference, such that: If P, then Q. P. Therefore, Q. An example is: If today is Saturday, then I will sleep late. Today is Saturday. Therefore, I will sleep late. For example, "according to the advice of Richard's family, Shorty was doing the _____ thing by accepting his mistreatment"; and the choices were "right" and "wrong."

Given the results of other studies using different methods that examined online processing in a similar context, I predicted that the literal questions would be easiest to answer, followed first by the temporal ordering questions and then by the propositional logic questions in order of increasing difficulty. Subject response times were recorded as

a rough measure of cognitive effort, and response times were calculated in milliseconds per character due to variability in question length (M = 130.1 and SD = 16.2). To make the different types of questions sufficiently similar to the reader with respect to length, mean question lengths within each text were made similar across all question types.

The comprehension assessment was piloted using 39 undergraduate students who participated to satisfy a research requirement for an educational psychology course. Thirty-five were female, and 34 were white with 5 being Asian. The mean age was 25.3 years (SD = 8.79). Subjects were told in the consent form and verbally after signing it that they would be reimbursed 10 cents for every question answered correctly (with a max of 72 questions, or $7.20). This was done in order to motivate task engagement. The test was administered by a computer program developed in FORTRAN by the author. Text order was randomized by subject, question order was randomized for each text, and the position of correct responses (i.e., choice a or b) was also randomized. To clear the subject's working memory of textual information and promote assessment of subject's long-term memory of the text, an intervening number Stroop task lasting about two minutes, M = 1.9, SD = .17, was completed between each text and the questions following it.

Results of the pilot study confirmed my predictions about the difficulty of the questions. I conducted a repeated measures analysis of variance (RM-ANOVA) to examine effects related to question type and trial position and including percent correct and response time in milliseconds per character as dependent variables. Greenhouse-Geisser was reported for univariate tests (with one exception described below), and all analyses were conducted with an alpha level of .05. The first trial was treated as a

practice trial and omitted from the following analysis.

The univariate tests showed a significant main effect of question type with percent correct, $F(1.8, 3798.0) = 9.84$, $p = .001.$, $\eta^2 = .21$, and response time, $F(1.8, 3798.0) = 58.86$, $p = .001.$, $\eta^2 = .61$, with response time having a much larger effect than percent correct. Pairwise comparisons showed that the propositional logic inferences were significantly more difficult than temporal ordering as measured by a decrease in percent correct: propositional logic $M = 75.73$ and $SD = 9.97$, temporal ordering $M = 80.51$ and $SD = 14.38$, $p = .034$, and they required more cognitive processing as measured by an increase in response time in ms/character: propositional logic $M = 96.54$ and $SD = 21.33$, temporal $M = 86.57$ and $SD = 18.38$, $p = .001$. The temporal ordering questions were more difficult than literal questions as measured by a decrease in percent correct, but this difference was only approaching significance: temporal ordering $M = 78.1$ and $SD = 10.34$, literal $M = 84.0$ and $SD = 9.65$, $p = .056$. And as expected given the larger effect size associated with response time, temporal ordering questions required significantly more cognitive processing than literal questions as measured by an increase in response time in ms/character: temporal ordering $M = 91.06$ and $SD = 18.62$, literal $M = 73.65$ and $SD = 16.11$, $p = .001$. These results confirmed most expectations and this was considered encouraging given a small sample size; however, some revisions were needed, and these revisions served as the basis for the second pilot study.

Revisions to the Pilot Assessment

Based on the results of the pilot study, I made several changes to the assessment in preparation for the main study. First, to improve the assessment of comprehension, the three question types were balanced by generating three questions of each type for each of

the three sections of text. Second, the two-choice (true-false) format of the comprehension questions was increased to a four-choice multiple-choice format to reduce the effects of chance from 50% to 25%. Three, prompts for the metacomprehension judgments were added. Four, the number Stroop task was removed because with the addition of the metacomprehension task, the many tasks became an impractical burden for the subject who in addition to reading six 600-word texts, also made 18 or more predictive comprehension judgments, and answered 54 comprehension questions with 54 posttest confidence judgments in one session.

Other revisions were made to the questions in order to address the possibility that differences between question types in the pilot study were due to either the extra retrieval necessary to generate the inferences or to differences in question content. The generation of some propositional logic inferences in the pilot study required the retrieval of two propositions in addition to generating the inference (either while reading the text or answering the question, the timing or necessity of generation was not established), whereas the other questions each required the retrieval of one basic proposition. To control for this in the main study, the propositional logic questions were revised if necessary to have enough of the information necessary to the inference given in the question so that the retrieval of one proposition would be sufficient to successfully answer the question. For example, in the story Black Elk and His Visions, the original question was, “Black Elk began to hear voices in a new and extraordinary way for the first time around the year _____”; and the choices were “1867” and “1875.” This required the reader to know as stated in the text (a) that Black Elk was born in 1863 and (b) that the visions began when he was 4 years old. The question was revised to read,

"Black Elk began to hear supernatural voices for the first time at the age of four, sometime around the year _____," so that remembering the year of his birth should be sufficient to answer the question.

With these revisions, it may be assumed that differences between literal and inferential types of questions in the main study will be due to the extra cognitive processing required for the generation of inferences rather than simply extra retrieval, although the specific timing of inference generation remains unclear. If some inferences are made online and others offline, the more difficult inferences are more likely to be made offline (Graesser et al., 2007) when answering the related question, adding to the reader's response time following that question and appropriately facilitating the detection of difficult questions with the measure of response time.

Finally, to examine the effects of question content, a literal-only version of the revised assessment was created in which the general content of the inferential questions was transformed into literal questions. This version was tested in the second pilot study.

Pilot Study 2

I conducted the second pilot study to check for potential differences in difficulty between question types due to question content. Because each of the questions made reference to different parts of the relevant story, content that was difficult to grasp or unfamiliar to the subject could produce differences in cognitive processing. The questions now consisted of three types of literal questions: literal (the same literal questions from the first pilot study), literal questions derived from the temporal-ordering questions, and literal questions derived from the propositional logic questions. Any differences in difficulty among the questions should be due to question content and not to

the added effects from inferential processing. For example, the propositional logic question discussed earlier that requires knowledge of Black Elk's year of birth was revised to read, "According to the story, Black Elk was an Oglala Lakota who was born sometime during the year _____"; and the choices were "1853," "1857," "1863," or "1867."

To expedite data collection, this test was administered to a convenience sample in the same manner as was done with the topic familiarity questionnaires in the first pilot study: Undergraduate students (n = 37) in an educational psychology course volunteered to participate in the group task after class. Subjects volunteered with no compensation and no demographic or personal information were collected. The test was administered in pen-and-paper format. Therefore, response times for questions were not available and only the dependent variable of performance was examined. However, the order of texts and questions was balanced with Latin squares using six versions of the test.

Results showed that percent correct for each group of questions was similar: literal (M = .67, SD = .47), literal from temporal ordering (M = .71, SD = .45), and literal from propositional logic (M = .70, SD = .46). I conducted a RM-ANOVA to examine the effects of question type using percent correct as the dependent variable and text order as a between-subjects factor. Within-subject contrasts showed no significant effect for question type, $F(2.0, 60.5) = .91$, $p = .41$, $\eta^2 = .03$. Given these results, it seems unlikely that differences observed in measures of difficulty among question types would be due to question content; in other words, differences among question types in the main study should be due to the type of cognitive processing required to answer the question.

Narrative Stories and Questions

The protocol included six stories that were used in both pilot studies. Each story was separated into three sections, with each section 200 words in length. The three question types relating to the stories were developed in pilot studies 1 and 2: (a) literal, (b) temporal ordering, and (c) propositional logic. Each story had three of each question type for each section of text for a total of 9 questions per story, and a total of 54 for the entire protocol. The order of the stories was counterbalanced between subjects using balanced Latin squares. The questions and question choices were randomized except that the first question shown after a story was never from the third (and most-recently read) section of text. The texts and questions have been provided in the Appendix.

Measures

The dependent variables used in the study were: As a measure of reading comprehension for literal, temporal ordering, and propositional logic questions, percent of correct responses were reported as a function of trial position; cognitive effort was measured using response time and peak amplitude of pupil size for each task; and metacomprehension was measured using confidence judgments and performance to calculate both absolute and relative accuracy.

Judgments of confidence in performance were made on a continuous scale from "no confidence (or guessing)" to "total confidence" with tick marks indicating spatial distance at repeated intervals like a ruler, except with no numbers given. The scale was assumed to present interval properties due to the spatial intervals presented, however, this assumption of interval properties for the scale may be mistaken. Later, these spatial choices were converted to numbers representing probabilities from 25% (chance) to

100%.

Absolute metacognitive accuracy was measured in two ways: bias and calibration accuracy. A bias score was calculated for each subject by subtracting actual performance from the pretest or posttest confidence judgments, summing these differences, and then taking the mean of the summed differences. The bias score (or bias index) provides a signed difference of accuracy, with negative values indicating underconfidence, positive values indicating overconfidence, and 0 as perfect absolute accuracy. The equation for bias is:

$$\text{Bias} = 1/n \sum_{i=1}^{n} (c_i - p_i)$$

Absolute metacognitive accuracy, or calibration accuracy, was calculated as a magnitude of accuracy for each subject in a similar fashion, but taking the mean of the absolute value of each of the differences between confidence and performance. With the exception that this produces all positive values, this measure of absolute accuracy can be interpreted similarly to bias, with 0 indicating perfect accuracy and increasing values indicating increasing inaccuracy. The equation for calibration accuracy is:

$$\text{Calibration accuracy} = 1/n \sum_{i=1}^{n} \left| c_i - p_i \right|$$

Relative accuracy was measured for each subject using the Goodman-Kruskal (1954) gamma correlation. The gamma correlation is a nonparametric statistic that has been widely used to calculate relative metacognitive accuracy (Masson & Rotello, 2009).

Gamma provides an ordinal association between two measures. In the present study, gamma was used to measure the extent to which an individual's high or low confidence across items was associated with his or her high or low performance, respectively. Gamma correlations range from +1.0 to -1.0, with a correlation of 0 representing an association that is no better than chance.

For each subject, cognitive effort was measured using the mean response time in milliseconds per character for every task requiring a response; likewise, the peak amplitude of pupil size was calculated for each subject and task. Noise was filtered from the pupil data in the following way. A minimal number of samples identified by the ViewPoint program as problematic were removed. Samples where the pupil aspect was less than or equal to .5 were removed as blinks. Samples greater than three standard deviations from the mean of the remaining data were removed as noise. Using this method, the percentage of data removed for all subjects was considered acceptable, M = 4.66, SD = 4.81.

Procedure

There was one session for each subject lasting about an hour. Following the process of informed consent, the computerized assessment collected demographic information. Then, before beginning the main assessment, subjects read the following instructions:

> You will be reading six short stories and each story will be shown in three sections. You can return to each story section as many times as you wish except you cannot return to a story after you begin answering the questions. Before answering the questions, you will judge your confidence in being able to correctly answer questions related to each section of the story on a continuous scale from TOTAL CONFIDENCE to NO CONFIDENCE (or guessing) in the following manner:

> After reading each section of the story, you will judge your confidence related to that section: "How confident are you that you can correctly answer the questions related to this section of the story?"
>
> Note: if rereading causes you to make more than one judgment for a section or story, your most recent judgment will be used.
>
> After reading a story and making your predictions, you will be shown nine incomplete statements about the story and you will choose the best of four responses to complete them. You can change your answer after clicking on (A), (B), (C), or (D), but once you click on the submit button, you cannot change it.
>
> After you submit an answer, you will judge your confidence in the chosen response; for example, "How confident are you that you just answered correctly?" These confidence judgments will use the same scale as the others.

Following the instructions, the subjects were shown a screen asking them to wait for nine seconds, with the number counting down as the seconds progressed. Then the experiment began with the first section of text and proceeded as described in the instructions (see Table 2) and with a 9-second wait screen before viewing a new text. When the subject finished the assessment, the final screen provided a summary of test performance, general accuracy of metacomprehension judgments, an acknowledgment for the authors of the original books used to develop these stories, and the statement, "Thank you for participating in this research!"

Table 2

Procedure Followed by Each Subject

Instructions to Experiment	Wait Screen	Text 1									Wait Screen	Begin Text 2
		Read Section 1	Predictive Judgment for Section 1	Read Section 2	Predictive Judgment for Section 2	Read Section 3	Predictive Judgment for Section 3	Proceed Screen	Answer 9 Questions in Random Order	Posttest Judgment after Answering Each of the Questions		
			Prompt: *How confident are you that you can correctly answer the questions related to this section of the story?*	Subject may choose to navigate back to section one from here	Prompt: *How confident are you that you can correctly answer the questions related to this section of the story?*	Subject may choose to navigate back to section two from here	Prompt: *How confident are you that you can correctly answer the questions related to this section of the story?*	Prompt: *Press NEXT to proceed to the questions or PRIOR to return to the story*	Set consists of 3 literal, 3 temporal ordering, and 3 propositional logic questions	Prompt: *How confident are you that you just answered correctly?*		

CHAPTER III

RESULTS

The first trial was considered a practice trial and omitted from the following analyses. To test for carryover effects with the six text orders (i.e., did a particular text order have either subtractive or additive effects on performance on subsequent texts), one-way ANOVAs were conducted to examine the effects of text order as the independent variable with each dependent variable (DV). The results showed no significance with text order for any of the DVs (smallest $p = .34$).

Reliability of the DVs was calculated by estimating the consistency of performance across trials 2 through 6. The reliability was moderate to high for each of the DVs related to comprehension: percent correct, Cronbach's Alpha (α) = .75; response time, $\alpha = .87$; maximum pupil size, $\alpha = .99$. The reliability was also high for posttest confidence judgments, $\alpha = .91$, and for each of the DVs related to absolute accuracy of metacomprehension for those judgments: bias scores, $\alpha = .82$; calibration accuracy, $\alpha = .81$.

However, as a measure of the relative accuracy of posttest judgments, the reliability of *G* scores was low, $\alpha = .15$. The low reliability may have been due, in part, to the elimination of 48 cases (40%) from the analysis when analyzing *G* by trial position. Comparing pretest and posttest judgments avoided missing cases, but the reliability of *G* scores was still low, $\alpha = .14$. The low reliability of *G* scores has been a typical problem

in metacomprehension research. Nevertheless, as argued by Maki et al. (2005), it would be difficult to explain consistently positive *G* scores that imply some degree of successful monitoring by being significantly above zero as arising from random noise. A one-sample *t*-test showed that *G* scores for both pretest and posttest judgments in this study were significantly greater than zero: pretest $M = .11$ and $SD = .27$, $t(119) = 4.59$, $p = .001$, $\eta^2 = .15$; posttest $M = .52$ and $SD = .19$, $t(119) = 29.43$, $p = .001$, $\eta^2 = .88$. As typically seen in studies of metacomprehension that have shown the testing effect (Pressley & Ghatala, 1990), posttest accuracy was greater than pretest accuracy and a RM-ANOVA to examine the effect of pretest/posttest condition on relative accuracy (*G*), reporting Greenhouse-Geisser for the univariate test, showed the difference to be significant, $F(1.0, 119.0) = 195.10$, $p = .001$, $\eta^2 = .62$. Furthermore, comparing posttest scores with zero produced an impressive effect size ($\eta^2 = .88$). Therefore, the relative accuracy (*G*) of posttest judgments that were focused on in this study presented compelling evidence that *G* may provide a meaningful measure despite its low reliability.

In doing an item discrimination analysis, the higher- and lower-performing subjects were each defined as closely as possible to the higher and lower quartiles, respectively, resulting in 33 subjects in the upper quartile and 29 subjects in the lower quartile. For each question, the mean percent correct of the lower performers was subtracted from the mean percent correct of the higher performers, producing a discrimination value between -100 and 100, with positive values indicating that the question properly discriminated between the higher and lower performers. All discrimination values were positive, discrimination $M = 28.05$ and $SD = 12.7$.

Comprehension

My research questions related to comprehension were: (a) Will performance in percent correct vary by question type with literal questions being the easiest and propositional logic questions being the most difficult?; (b) Will response time in milliseconds per character vary by question type with literal questions having the shortest and propositional logic questions having the longest?; and (c) Will maximum amplitude of pupil size vary by question type with propositional logic questions having the largest and literal questions the smallest?

To answer these questions, I conducted RM-ANOVAs to examine effects related to the repeated measures of question type and trial position (omitting the first trial), including percent correct, response time in milliseconds per character, and maximum amplitude of pupil size as dependent variables. Greenhouse-Geisser statistics were reported for univariate tests, and all analyses were conducted with an alpha level of .05.

The univariate tests showed a significant main effect of question type with percent correct, $F(2.0, 235.5) = 16.13$, $p = .001$., $\eta^2 = .12$, response time, $F(1.9, 231.0) = 104.00$, $p = .001$, $\eta^2 = .47$, and pupil size, $F(1.6, 192.9) = 3.84$, $p = .031$, $\eta^2 = .03$. As observed in the first pilot study, there was a larger effect size related to response time as compared with percent correct, and the measure of pupil size introduced for the main study presented the smallest effect size.

Examining the measures of percent correct and response time by question type (see Table 3) with pairwise comparisons showed significance for all comparisons in the expected directions. Propositional logic inferences were significantly more difficult than temporal ordering as measured by a decrease in percent correct, $p = .01$, and they

Table 3

Means and Standard Deviations for Percent Correct, Response Time in ms/character, and Maximum Amplitude of Pupil Size by Question Type (n = 120)

Question Type	Percent Correct *M* (*SD*)	Response Time in ms/character *M* (*SD*)	Maximum Amplitude of Pupil Size *M* (*SD*)
Literal	74.94 (14.99)	96.40 (24.54)	.1581 (.0409)
Temporal Ordering	70.83 (18.15)	109.70 (26.56)	.1583 (.0409)
Propositional logic	66.94 (15.20)	118.05 (26.78)	.1592 (.0402)

required more cognitive processing as measured by an increase in response time in ms/character, p = .001. Temporal ordering questions were significantly more difficult than literal questions as measured by a decrease in percent correct, p = .003, and they required more cognitive processing as measured by an increase in response time in ms/character, p = .001. Examining the maximum amplitude of pupil size, propositional logic questions required more cognitive effort than temporal ordering questions as indicated by an increase in pupil size, p = .01. However, there was no significant difference between temporal ordering and literal questions, and this may have been related to the small effect size of this measure.

Overall, the results related to comprehension supported predictions that propositional logic questions should be the most difficult, temporal ordering questions should be of moderate difficulty, and literal questions should be the easiest. The only exception was that maximum amplitude of pupil size did not discriminate between literal and temporal ordering questions. Nevertheless, the other two measures presented

significant differences as predicted and therefore, these results were considered sufficient to establish the relative difficulty of question types and proceed to examine their influence on metacomprehension.

Metacomprehension

The questions related to metacomprehension were: (a) Will the magnitude of metacognitive posttest confidence judgments vary by question type with confidence being greatest for judgments related to literal questions and lowest for judgments related to propositional logic questions?; (b) Will posttest measures of calibration accuracy vary by question type with greatest accuracy for the question type that was answered closest to a moderate anchor of success and with decreasing accuracy for question types that move away from this anchor?; (c) Will bias vary by question type with the greatest overconfidence related to difficult questions (propositional logic) and the least overconfidence (or underconfidence) related to easy questions (literal)?; and (d) Will levels of relative accuracy (*G*) remain consistent across question types?

To test these hypotheses, RM-ANOVAs were conducted to examine effects related to question type with the DVs of posttest confidence, *G* as a measure of relative accuracy, and bias or calibration accuracy scores as measures of absolute accuracy. The first trial was omitted. Nine cases were not analyzed due to missing data; therefore, the sample size for this analysis was 111. The Greenhouse-Geisser statistic was reported for univariate tests and all analyses were conducted with an alpha level of .05. Trial position was not included as a factor because calculating *G* scores by question type and trial position produced too many missing values that there were no cases (subjects) remaining for analysis, furthermore, there were no significant effects for trial position with percent

correct after removing the first trial, and comprehension as measured by correct responses serves as the basis for metacomprehension accuracy. The missing values with *G* scores were due to the fact that the calculation of *G* ignores comparisons of equal values, therefore, small sets of judgments will sometimes present no variation and no *G* score can be calculated from them.

It was predicted that the magnitude of posttest confidence judgments should be greatest for literal questions and less for inferential questions, with confidence lowest for propositional logic questions, and this was the case (see Table 4). The univariate test for question type with magnitude of posttest confidence was significant, $F(2.0, 215.9) = 53.95$, $p = .001$, $\eta^2 = .33$, and pairwise comparisons showed that these differences were significant between propositional logic and temporal ordering questions, $p = .001$, and between temporal ordering questions and literal questions, $p = .005$.

Table 4

Means and Standard Deviations for Magnitude of Posttest Confidence Judgments, Calibration Accuracy, Bias, and Relative Accuracy—*G* of Posttest Confidence Judgments by Question Type ($n = 111$)

Question Type	Magnitude of Posttest Confidence Judgments *M* (*SD*)	Posttest Calibration Accuracy *M* (*SD*)	Posttest Bias *M* (*SD*)	Posttest Relative Accuracy—*G* *M* (*SD*)
Literal	76.68 (11.56)	.31 (.11)	.03 (.15)	.54 (.34)
Temporal Ordering	75.23 (11.91)	.34 (.12)	.06 (.17)	.53 (.37)
Propositional Logic	71.66 (11.77)	.37 (.10)	.05 (.17)	.47 (.32)

As one measure of absolute accuracy, posttest calibration accuracy was predicted to be the most accurate for questions that subjects answered most closely to 81% of the time because this was a good numeric estimate of 75% of the spatial scale. This was the case (see Tables 3 and 4). Performance on literal questions was closest, followed by temporal ordering questions, and propositional logic questions were the farthest from this anchor, and then as predicted, calibration accuracy was greatest for judgments related to literal questions, followed by judgments related to temporal ordering and propositional logic questions. The univariate test for question type with calibration accuracy was significant, $F(1.9, 213.83) = 28.01$, $p = .001$, $\eta^2 = .20$. Pairwise comparisons showed that the difference between propositional logic and temporal ordering questions was significant, $p = .002$, and the difference between temporal ordering and literal questions was significant, $p = .005$.

As another measure of absolute accuracy, posttest bias was predicted to vary by question type, showing the greatest overconfidence with difficult questions (propositional logic) and either the least overconfidence or underconfidence with easy questions (literal). The univariate test for question type with bias scores was significant, $F(2.0, 219.03) = 3.83$, $p = .02$, $\eta^2 = .03$. The difference between temporal ordering and literal questions was in the correct direction and pairwise comparisons showed it to be significant, $p = .007$. However, the difference between propositional logic and temporal ordering questions was not significant, $p = .46$ (see Table 4). In other words, literal questions presented the least overconfidence, as predicted, but propositional logic and temporal ordering questions presented the most overconfidence in equal amounts. With bias scores closer to zero indicating greater accuracy, a one-sample *t*-test was conducted

to compare these scores to zero and check for significant differences. There were significant differences with small effect sizes for scores related to temporal ordering questions, $t(119) = 3.04$, $p = .003$, $\eta^2 = .07$, and propositional logic questions, $t(119) = 3.52$, $p = .001$, $\eta^2 = .09$, meaning the subjects were slightly overconfident with these types of questions. However, there was no significant difference for scores related to literal questions, $t(119) = 1.65$, $p = .10$, $\eta^2 = .02$, meaning the subjects showed virtually no bias (i.e., over- or underconfidence) with judgments related to literal questions.

Finally, it was predicted that posttest relative accuracy as measured by *G* scores would not vary by question type because item-to-item performance at various levels of question difficulty would be consistent with item-to-item confidence. This was the case (see Table 4). The univariate test for question type with *G* scores was not significant, $F(1.9, 211.47) = 1.37$, $p = .26$, $\eta^2 = .01$. Therefore, relative accuracy did not vary by question type. In other words, as predicted, the subjects were moderately successful at monitoring the relative difficulty of items across varied levels of difficulty.

CHAPTER IV

DISCUSSION

The purpose of the present research was to integrate the investigation of comprehension with metacomprehension by examining subjects' comprehension to establish that the question types varied in difficulty as expected, and then examining how accurately they monitored comprehension at each level of difficulty in their posttest judgments with both absolute and relative measures. The results supported many of the expectations for comprehension and metacomprehension. By integrating an examination of metacomprehension with varied levels of comprehension, this research has added to the literature by providing support to the premise that readers' metacomprehension judgments can be accurate (Nelson & Narens, 1990) and that metacomprehension related to retrospective (posttest) judgments reflects appropriate differences in comprehension. In addition, the results provided evidence that changes in absolute accuracy by level of question difficulty may be attributed to the anchoring-and-adjustment heuristic (Epley & Gilovich, 2006), as well as anchoring on a moderate estimate of success specifically (Hacker & Bol, 2011).

For comprehension, I expected to find that inferential questions would be more difficult to answer than literal questions, and of the two types of inferential questions, the propositional logic questions would be the most difficult. For the measures of percent correct and response time in milliseconds per character, these predicted relations were

supported. Response time provided a large effect size in discriminating question types, η^2 = .47, and percent correct provided a smaller effect size but still substantial, η^2 = .12.

With the measure of maximum amplitude of pupil size, the propositional logic questions were the most difficult, but there was no difference between temporal ordering and literal questions. One practical implication of these results would be that pupil size data provided little extra to the study while requiring significant effort and expensive eyetracking equipment. Future studies with this assessment may not benefit much from eyetracking technology, unless better ways were found to analyze pupil size or the method was expanded to include other pupillometric measures, such as time spent reading in particular areas of interest. These additional measures may present stronger effect sizes.

Once the level of difficulty for the three question types had been addressed, question about metacomprehension could be addressed. I expected confidence in judging comprehension should vary with question difficulty, with the magnitudes of posttest confidence judgments being greatest for literal questions and less for inferential questions, with confidence lowest for propositional logic questions. Results supported this prediction. With the large effect size of η^2 = .33, this indicates that the subjects were very successful at monitoring the relative difficulty of the questions.

I had predicted that calibration accuracy, measured by the absolute value of the difference between judgments and performance, would be most accurate for the question type that was answered closest to 81%, with decreasing accuracy for question types that move from this anchor. Results supported this prediction and with a large effect size, η^2 = .20. Performance on literal questions was closest to 81% correct, followed by temporal

ordering questions, and propositional logic questions were the farthest from this anchor. As predicted, calibration accuracy was greatest for judgments related to literal questions, followed by judgments related to temporal ordering and propositional logic questions.

Based on the hard-easy effect, I had predicted that bias would show the greatest overconfidence with the most difficult questions (propositional logic) and the least overconfidence with easy questions (literal). Although the results showed a small effect size, $\eta^2 = .03$, the prediction had support. As predicted, literal questions (i.e., the easiest questions) presented the least overconfidence—in fact, subjects showed virtually no bias with literal questions, and propositional logic (the most difficult questions) showed the greatest overconfidence. Counter to the prediction, there was no significant difference in bias between temporal ordering and propositional logic questions. Subjects showed greater overconfidence with temporal ordering questions in comparison to literal questions. However, they did not show greater overconfidence with propositional logic questions in comparison to temporal ordering questions. In other words, for some reason, there appeared to be similar levels of overconfidence with both types of inferences, although these question types appeared to differ in difficulty using multiple measures. It is possible that subjects recognized the inferential nature of both temporal ordering and propositional logic questions as being difficult because they required more than simple memory (i.e., literal information) to answer, and that this recognition produced similar amounts of overconfidence for posttest judgments related to both question types, despite the fact that they were also related to significant differences in confidence judgments, calibration accuracy, and measures of comprehension.

Finally, I had predicted that relative accuracy should not vary by question type

because the insufficiency of adjustments in terms of scale will not affect relative accuracy as long as the adjustments move in the right direction, i.e., appropriately indicating a greater or lesser probability of success. This prediction also was supported. There were no significant differences for relative accuracy among the three question types. Moreover, subjects showed relatively high accuracy for each question type (.54 for literal, .53 for temporal ordering, and .47 for propositional logic).

Subjects in the current study adjusted the magnitudes of confidence in knowing questions directed at different levels of comprehension, with greater confidence given to literal questions that probe textbase levels of comprehension, less confidence to temporal ordering questions that probe inferences at a situation model level of comprehension, and even less confidence to propositional logic questions that probe inferences at a deeper level of the situation model. In addition, absolute accuracy (i.e., the absolute value) was greatest for the literal questions, less so for temporal ordering questions, and the least for propositional logic questions. In general, these results indicate that subjects were sensitive to question difficulty, but when calibrating their predictions to actual performance (i.e., absolute accuracy), their predictions appeared to be anchoring somewhere around a moderate estimate of success.

The current study did provide some support for the hard-easy effect, which (using bias scores) has shown that readers are overconfident in answering difficult questions about text information and underconfident with easy questions. Although Juslin, Winman, and Olsson (2000) have provided some rationale for this effect, it still remains an elusive finding. Some research has shown the effect, whereas others have not. The small effect size shown in the present study does not provide strong support of this

curious finding about metacomprehension, and suggests that more research is needed to clarify this relation between confidence and question difficulty.

Results also showed that relative accuracy did not vary significantly by question type. This was consistent with the predicted nature of relative accuracy, i.e., the sizes of differences between judgments were unimportant and only the relative consistency of choices contributed to the measure. Because of differences in the behavior of relative versus absolute accuracy, as mentioned frequently in the literature, absolute and relative accuracy appear to be measuring different aspects of metacognitive monitoring, for example, relative accuracy as measured by *G* appears to be relatively free from the effects of anchoring. Unfortunately, *G* presented low reliability as often occurs in the literature. Admittedly, a measure that is not reliable may not be valid, but I propose that it is possible for a measure with low reliability to still possess enough validity to be meaningful because there are consistent observations in the literature related to *G* scores, such as positive scores that are significantly different than zero, and posttest scores that are significantly greater than pretest scores. Despite its low reliability, subjects' relative accuracy provided evidence that they were monitoring differences in difficulty between items to a significant degree of success, and this evidence was more compelling with measures of absolute accuracy also presenting evidence of successful monitoring.

Finally, I made a posthoc observation related to bias scores that the choice of scale as an appropriate estimation of probability was crucial for measuring over- and underconfidence. Specifically, if subjects have a greater than zero probability of answering something correctly by chance, scales from 0 to 100 may deflate bias scores and transform small amounts of overconfidence into underconfidence. This should

clearly be avoided. For example, judgments in this study were measured on a scale from 25 to 100 because there was a 25% chance on a four-choice test that a subject with no confidence would answer a question correctly by guessing. In other studies, similar judgments have sometimes been placed on a scale from 0 to 100, but this will result in the deflation of bias scores. Transforming the current data for purposes of illustration to a scale from 0 to 100 would make it appear that the subjects as a whole were underconfident with literal and propositional logic questions and almost perfectly accurate with temporal ordering questions (see Table 5). Using a one-sample *t*-test to compare the bias scores to zero, there were significant differences for scores related to literal questions, $t(119) = -3.55$, $p = .001$, $\eta^2 = .10$, and propositional logic questions, $t(119) = -2.44$, $p = .02$, $\eta^2 = .05$. However, there was no significant difference for scores related to temporal ordering questions, $t(119) = -1.80$, $p = .074$, $\eta^2 = .03$. Given prevalent findings of overconfidence on tests as a general rule in metacomprehension, the deflated scale produces strange results here that would be difficult to explain.

Table 5

Means and Standard Deviations for Absolute Accuracy (Bias Score)

With Posttest Judgments on a Scale of 0 to 100 by Question Type ($n = 120$)

Question Type	Posttest Bias Scores *M* (*SD*)
Literal	-.05 (.17)
Temporal Ordering	-.03 (.19)
Propositional Logic	-.04 (.18)

Limitations and Implications for Further Research

This research was limited in several ways: (a) the uncertainty of measuring online or offline comprehension, (b) the use of narrative texts, (c) factors uncontrolled in the comprehension test, (d) limited analysis of eyetracking data, (e) questions related to the scale of measurement for confidence judgments, and (f) the exploratory nature of the study.

First, the study design required that subjects were asked questions about their comprehension of the text after reading the text and without the text available. The answers to the questions may have been formulated online as the subjects were reading the texts or they may have been formulated offline only as a consequence of being probed by the question. This has important implications for theories of comprehension, particularly pertinent to understanding inferential processing. When inferences occur has been the focus a great deal of research, and unfortunately, the present study does not add to this research. In future research, probing comprehension more proximally to reading will be necessary to shed light on this "when" question.

Second, expository texts are more common in educational settings than narratives; therefore, it would be useful to repeat this method of research with expository texts to see whether subjects appear to monitor their comprehension in a similar way to that observed with narrative texts, for example, anchoring their calibration accuracy for posttest questions on moderate expectations of success.

Third, the quality of the comprehension test could be improved. Although the questions in the comprehension test were controlled in some important ways, other aspects remain to be considered. Further testing could examine whether the overall test

might present enough reliability with less texts and questions to reduce the demands of the test. Furthermore, each of the test questions could be reexamined more closely to control for issues related to its salience to the main character or the general plot, using one or more norming studies to estimate the relative levels of salience for each question.

Fourth, only the maximum amplitude of pupil size was examined in this study but a wide range of pupillometric data can be examined with eyetracking technology. Other measures of pupil size and other pupillometric measures (e.g., reading times with specific areas of interest) could be tested for ability to discriminate differing question types that have shown significant differences with larger effect sizes using other measures (i.e., percent correct, response time in milliseconds per character).

Fifth, as a measurement issue, I have concluded that researchers should avoid deflating scores with scales of probability that set a minimum at zero—unless zero would be the expected probability of success with guessing, for example, a fill-in-the-blank assessment with no choices given. However, it remains unclear whether a scale between no confidence and total confidence that does not show numbers to the subject but translates later to a probability between 25 and 100 would present similar results as a scale marked explicitly from 25 to 100. Therefore, this point may be especially relevant to continuous scales that do not present explicit numbers, and future research could examine the effects of explicitly- and implicitly-numbered scales that predict probability of success.

Sixth, although the study provided information about how levels of comprehension may influence different measures of metacomprehension, it does not say anything about the specific processes that may be used to make judgments of

metacomprehension. It is likely with metacomprehension that one or more heuristics may be used to facilitate the making of confidence judgments (Tversky & Kahneman, 1974), such as the heuristics of availability (Gilovich, Griffin, & Kahneman, 2002), retrieval fluency (Benjamin & Bjork, 1996), ease-of-processing (Begg, Duft, Lalonde, Melnick, & Sanvito, 1989), levels-of-disruption (Dunlosky, Rawson, & Hacker, 2002), or cue utilization (Koriat, 1997). Perhaps methods can be developed to test predictions implied by these heuristics and how these heuristics might interact with differing levels of comprehension.

Other research could examine the effectiveness of specific interventions to improve accuracy (including the reduction of the anchoring effect apparent in this study), delayed assessments of comprehension and metacomprehension as indicators of long-term learning, or the effect of incentives on comprehension and metacomprehension.

Conclusions

This research provided a major contribution by showing that metacomprehension should not be treated as some monolithic process in which a reader either does or does not engage. In the earliest studies of metacomprehension, researchers had measured metacomprehension by asking a single question about a text (see Glenberg, Wilkinson, & Epstein, 1982). Although this was modified in later research to include multiple questions (see Maki & Berry, 1984), there was no consideration given to varying kinds of comprehension, such as comprehension at a textbase level or a situation model level. Most of the research on metacomprehension since 1985 has shown that readers are not very adept at monitoring their comprehension (for a review, see Maki, 1998). One reason proposed for these poor results was that metacomprehension needs to be tied more

directly to comprehension to be able to fairly assess readers' monitoring ability. There have been recent attempts to measure metacomprehension at various levels of comprehension (e.g., Dunlosky & Lipko, 2007; Dunlosky et al., 2002; Salmen, 2004); however, results from this research have not provided a clear picture.

The current study has shown that readers are sensitive to different levels of comprehension in making retrospective (posttest) confidence judgments of performance. In addition, as evidenced with immediate and delayed judgments of learning (Scheck et al., 2004), absolute accuracy appears to be susceptible to the effects of anchoring-and-adjustment (Epley & Gilovich, 2006) when making posttest confidence judgments across question types of varying difficulty. Relative accuracy (*G*) did not show any effects related to anchoring, therefore, as with previous studies, these results suggest that absolute and relative accuracy are measuring different components of metacomprehension.

APPENDIX

TEST MATERIALS

Note: The stories in these test materials were adapted from portions of the following books: *Black Boy* by Richard Wright; *Black Elk Speaks* by John G. Niehardt; *Bless Me, Ultima* by Rudolfo Anaya; *Death Comes for the Archbishop* by Willa Cather; *Night* by Elie Wiesel; and *Steppenwolf* by Hermann Hesse. Liberties were taken with respect to details and presentation. To provide proper credit to the originals, the subjects were shown the names of these books and their authors after they had finished the assessment.

Richard and His Education

Richard was a black American living in the Southern United States in the early 1900s—also known as the Jim Crow South. Racial segregation was common there and black people were often considered to be a lower class of citizen. Richard had serious difficulties because of his high self confidence. He was not willing to play the passive role that many white people expected of blacks. He suffered abuse from whites as well as being shunned by his black friends and family who told him that it is better for black people to behave and simply stay in their place. In 1925, he moved to Memphis and found a job with an optical company that employed blacks and whites. Some of his coworkers were members of the Ku Klux Klan and the black employees were subjected to regular abuse. For example, the elevator operator was a black man named Shorty who had been there for years. He endured kicking and insults by white people in the company. Shorty appeared to have no hope for change or future escape and he made a joke of being kicked in order to receive better tips from his abusers. Richard felt contempt for Shorty.

Richard enjoyed reading and writing. He read an article about H. L. Mencken once and wanted to read more. However, as a black man, he did not have access to the public library. Richard worked with a Catholic man who had known some persecution against his religion. And as a result, the two of them shared a common understanding about prejudice. Richard was then able to borrow the man's library card to continue his own education. He read Mencken's essays and discovered that writing could be a powerful weapon. He also learned the names of other American authors. He read the books of Theodore Dreiser and Sherwood Anderson. It was a revelation when Richard discovered that there were others who felt alienated from the American way of life. Richard also read some of the major European authors. Through reading, he began to understand himself better. Richard also developed a better understanding of white people. They no longer seemed so strange to him; however, he felt compelled to hide everything that he was learning. He worked, read, and played dumb for the sake of himself, his mother and brother. He wanted to be able to bring them to Memphis with him.

For Richard, his books and learning carried a burden. He was isolated in his own secret world with no one to share his dreams and insights. Conscious of the many forces in life that have helped to make him who he is, he knew that he must continue onward to become an American writer. He also realized that black Americans lived as outsiders in their own country. Richard came to accept these things. He also learned to play the role expected of him by his family and by white people. Richard did this because it could bring eventual escape by helping to provide the means to travel north and become a writer. It was only a matter of time and he did so. Unfortunately, Richard remained isolated and shared his intentions with no one when he left. He lied about his reasons for leaving. Traveling north, he did not feel a sense of promise or joy; rather, he felt the usual tension and fear. Nevertheless, he had valuable lessons to share. Richard was no victim. He also realized then that his revenge would be to succeed despite those who tried to destroy him. And his success might make them change.

Life for Richard was made significantly _____ on a regular basis as a result of his high self confidence.

a. more challenging
b. easier
c. less predictable
d. more social

Some of the books that Richard borrowed from the public library and read were written by _____ European authors.

a. famous
b. obscure
c. controversial
d. academic

Richard was feeling _____ as he traveled north from Memphis to pursue his dream to become a writer.

a. the familiar anxiety
b. a new sense of happiness
c. less isolated
d. sick

Richard moved to the city of Memphis and this led to a time when he _____.

a. met Shorty
b. first experienced persecution
c. first disagreed with friends
d. learned to write

Richard worked with a Catholic who knew persecution, and this led to Richard _____.

a. reading more books
b. having self confidence
c. moving to Memphis
d. becoming Catholic

Playing the role expected of a black person in some way led to Richard _____.

a. moving north
b. borrowing a library card
c. losing his friends
d. moving south

According to the advice of Richard's family, Shorty was doing a _____ thing by generally accepting his mistreatment.

a. praiseworthy
b. contemptible
c. shameful
d. curious

Following from Richard's discovery about writing when reading Mencken, his future writing could likely be _____.

a. threatening
b. ignored
c. popular
d. unclear

After a period of reading and learning, Richard realized that he _____ within his own country.

a. was alienated
b. belonged
c. had many opportunities
d. felt trapped

Black Elk and His Visions

Black Elk was an Oglala Lakota of the Sioux Nation. He was born in 1863 and began hearing voices when he was four years old. This frightened him a little at first. When he was five years old, he had a vision of two men in the sky singing a sacred song. Then one day, at the age of nine, Black Elk was eating when a voice told him to hurry because his Grandfathers were waiting for him. He became very ill. His arms, legs, and face became swollen and he could not walk. Lying in his parents' tipi, looking through the opening in the top, he saw the same men in the sky who sang the sacred song years ago. The men in the sky called to him, saying that his Grandfathers were waiting for him. Then Black Elk was transported in a cloud to a place made entirely of clouds, where he saw an incredible vision. A horse greeted Black Elk and said that the horse would tell the story of himself and the others there. The horse turned toward the four directions of the compass, and Black Elk saw that there were twelve horses in each direction.

The horses in each direction were matching in color. The horses in the north were white, the southern horses were yellowish gray, the eastern horses were light brown, and the western horses were black. The horse beside him was a bay horse, a reddish brown color. The bay horse told Black Elk that the horses will take him to his Grandfathers. As they moved along, the horses from the four directions followed in a formation and the sky was filled with dancing horses that changed into different animals. They arrived at a cloud. It

became a tipi whose door was a rainbow. Inside, there were six Grandfathers. Black Elk recognized them as the Powers of the World. The Grandfather of the West told Black Elk that all of his Grandfathers around the world were having a council, and they would teach him. Grandfather of the West gave Black Elk a cup of water containing the sky, which was the power to live, and also a bow, which was the power to destroy. He told Black Elk that his spirit was Eagle Wing Stretches. And then Grandfather of the West ran toward the west and changed into a starving, black horse.

Grandfather of the North gave Black Elk an herb that strengthened the black horse. Grandfather of the North told Black Elk that he will create a nation and have the power of the white giant's wing; then he ran toward the north, changing into a white goose. Grandfather of the East gave Black Elk a peace pipe with a spotted eagle on its stem, telling him that he will use this to heal the sick. Grandfather of the South gave Black Elk a stick. It sprouted branches. Then birds were singing in them. He told him that he would brace himself upon this cane, and his nation would brace itself upon it. Grandfather of the South also told Black Elk these powers would exist for four generations. The fifth Grandfather was Great Spirit Above who stretched out his hands and became a spotted eagle, telling Black Elk that the birds will come to you, and the stars will be like brothers. The sixth Grandfather was an old man who told him to have courage returning to earth. Then the sixth Grandfather left the tipi through the rainbow door, becoming younger until he was Black Elk at nine years of age.

At the age of nine, Black Elk became very sick and eventually his illness grew so serious that he could not even _____.

a. walk
b. speak
c. see
d. eat

Grandfather of the West gave a cup of water to Black Elk and said that this provided him with the power to _____.

a. live
b. destroy
c. heal
d. satisfy

Black Elk received a _____ from Grandfather of the East that was decorated with the image of a spotted eagle.

a. pipe
b. bow
c. stick
d. picture

Black Elk saw men in the sky when he was nine years old and this led to him _____.

a. having a vision
b. hearing a song
c. getting sick
d. singing

Black Elk saw a host of horses dancing and this led to him seeing _____.

a. a tipi with a rainbow door
b. a place made of clouds
c. two men
d. his parents

Black Elk's meeting with Grandfather of the South led to him _____.

a. hearing birds sing
b. being given a pipe
c. seeing a goose
d. speaking with a bird

Black Elk began to hear supernatural voices for the first time at the age of four, sometime around the year _____.

a. 1867
b. 1871
c. 1860
d. 1850

Black Elk _____ the Powers of the World in a tipi that had a rainbow for a door.

a. was given gifts from
b. did NOT see
c. briefly saw
d. heard about

In Black Elk's vision, the spotted eagle was associated in an important way with _____.

a. spirituality
b. music
c. destruction
d. plants

Antonio and His Sorrow

Antonio's sleep was troubled. One night, Antonio dreamed of three young friends who had died. In the dream, his three friends fought each other with sticks and knives. Cico was also there and he had a spear. He killed the golden carp with it. And this turned the waters red. Then Florence told him that the old gods were dying and pointed to the hills. There, Tenorio had killed the night-spirit of Ultima. And now Ultima was dying. Antonio was overwhelmed and he asked why God has forsaken him. He awoke from the dream, crying. Ultima comforted him. She suggested that he has known too much death in his life. She also said that becoming a man always involves great sadness. He should go work with his uncles in El Puerto to learn more about "growing life." Maria Luna y Márez, Antonio's mother, asked Ultima to bless them both. Ultima did so. Then Antonio drove away with his father, Gabriel Márez. He was traveling with his father to El Puerto to learn the art of being a good sheepherder. Gabriel told his son that he became a man while learning the art of sheepherding. Now it was time for Antonio.

Speaking with his father on the journey, Antonio learned some things. The opposing things in his life, such as the plains and valley, moon and sea, even God and the golden carp must merge and change. The same thing must happen with his settled Luna heritage and his sense of Márez freedom. In becoming a man, Antonio must combine his opposing backgrounds to create something new. His father also told him that "evil" is only something that people do not understand. He said true understanding can require a lifetime to achieve. It was not as simple as being in one communion ritual. And so Antonio spent the summer of 1947 working with his uncles. In doing so, he learned to respect and care for the earth. Then he gained strength from this experience. As a result, he slept peacefully. He realized that they were working in harmony with the lunar cycles. And at one point, Pedro told Antonio that he and the other Lunas were proud of what he'd learned. However, their talk was interrupted when Juan arrived to speak with Pedro in private. But Antonio overheard something from their conversation: A man named Tenorio had vowed to kill Ultima.

Pedro decided to drive to town in order to help Ultima. Antonio understood and headed into town on foot. He met Tenorio riding on horseback. Cursing Antonio, Tenorio tried to trample him but failed. Then Tenorio vowed to kill the owl that was the spirit of the old witch, and he rode away. Antonio remembered something else that his father had told him while traveling: True understanding requires sympathy for others. He said that Ultima's sympathy was so complete that she can heal others. Antonio came to the home where Ultima was staying. Pedro's truck arrived. Antonio saw Tenorio standing near a juniper tree with a rifle and cried out. Tenorio turned his rifle on Antonio, so Ultima quickly called to her owl. The owl flew above Tenorio, startling him. Tenorio fired the rifle upwards and then picked up the dead owl in triumph. Then Tenorio aimed again at Antonio. Another shot was heard. Pedro had fired a pistol. Tenorio was hit in the stomach and he fell. Antonio wrapped the owl in a blanket and rushed to Ultima. She said her owl was simply flying to a new place, and she was now preparing to do the same thing.

Antonio _____ when he eventually awoke from his dream that occurred at the beginning of the story.

a. was crying
b. realized something
c. spoke with his uncles
d. felt rested

After spending time with them, Antonio's uncles became proud of him for his _____.

a. learning
b. Luna heritage
c. Márez heritage
d. lofty dreams

When Antonio first met Tenorio on the way into town, Tenorio was _____.

a. riding a horse
b. driving a truck
c. walking on foot
d. riding a motorcycle

Cico appeared in Antonio's dream with a spear, and this led to _____.

a. water turning red
b. fighting between friends
c. friends dying
d. fighting with Florence

In El Puerto, Antonio learned to respect the earth and this led to him _____.

a. sleeping peacefully
b. working with his uncles
c. being blessed
d. riding a horse

Ultima called to her owl near the end of the story and following this, _____.

a. the owl died
b. Antonio saw Tenorio
c. Pedro arrived
d. Pedro saw Tenorio

If Ultima's advice to Antonio is correct, Antonio will surely have experienced

_____ when he is a man.

a. sadness
b. peace
c. compassion
d. hatred

According to Antonio's father, if people decided that Tenorio was evil, they _____.

a. did not understand him
b. probably knew of him
c. were likely evil
d. were judgmental

According to Antonio's father, if he had true understanding then he would have _____ other people.

a. compassion for
b. sadness for
c. conflicts with
d. power over

Latour and His Final Years

Latour had been a French Jesuit missionary serving in Ohio in the mid-1800s. He was elevated to bishop and sent to Santa Fe in the New Mexico territory. Bishop Latour left for Santa Fe with an old friend, Father Vaillant. Although the region of New Mexico was predominantly Catholic, the local faith had been corrupted through a process of neglect lasting over three centuries and rogue priests had become greedy and abusive. Latour and Vaillant were sent to change things and they had some success over the years. Latour eventually became an archbishop. Years later, Archbishop Latour retired and moved to a small estate four miles north of Santa Fe. He planned to spend the remaining years of his life there. An apricot tree that was about two hundred years old grew on the estate. Even after so much time, the apricot tree continued to bear delicious fruit. The Archbishop cultivated an orchard and a garden there. He spent a lot of time gardening. He also instructed new priests in Spanish and in the local customs of the diocese. He advised the priests to plant fruit trees in their parishes. The fruit would help them to balance their Mexican diet.

Archbishop Latour cultivated wildflowers on the estate that eventually covered the surrounding hillside in many shades of purple. He also received a young priest, Bernard, for training. Bernard was like a son to the Archbishop and helped to care for him. Then, in 1889, Latour was drenched in a January rainstorm which caused a fever. He asked the new Archbishop for permission to return to Santa Fe and die there. Bernard told Latour that he wouldn't die of a cold and Latour responded that he will not die from a cold, he

will die from having lived. And then Latour began speaking only French. This transformation alarmed the household. In February, Latour returned to Santa Fe with Bernard. He scheduled his return to occur with the sunset because that was the same time of day that he had first entered Santa Fe. His family had expected him to return to France, but he preferred to stay in New Mexico because he felt young there and he loved breathing the air. He stopped to admire the cathedral in Santa Fe. He was pleased with how the French architect had built it in a manner that seemed to fit the surrounding landscape.

In Santa Fe, Latour had little time remaining. He dictated a local history of the Catholic Church to Bernard. Latour described how the early Catholic missionaries from Spain entered New Mexico as a hostile territory and made many sacrifices, and as a result of sacrifice, Catholics arriving now were greeted by friendly people. Latour tried to impress the significance of this upon the younger priests. Latour also remembered his decision to leave France for America with Vaillant, who accompanied him on a mission from France to the New World, and later from Ohio to New Mexico. Latour had encouraged Vaillant to pursue his mission in the New World. With this encouragement, Vaillant decided to continue on. And Vaillant accomplished many things despite bad health and other challenges. Latour also realized that he had outlived most of his old acquaintances. During his last days, Latour slept most of the time and ate little. He eventually refused food. The cathedral filled with parishioners who prayed for him and Latour received last rites. On his deathbed, he returned to the day many years ago when he convinced Vaillant to travel with him. His final words were encouraging Vaillant onward to the New World.

Archbishop Latour would often spend his time _____ after he retired to the estate north of Santa Fe.

a. working in the garden
b. visiting other parishes
c. visiting the cathedral
d. sleeping

A priest _____ helped to take care of Archbishop Latour during his retirement on the estate north of Santa Fe.

a. named Bernard
b. named Vaillant
c. whose name was not mentioned
d. named Parish

When Archbishop Latour was finally given last rites on his deathbed, _____ prayed for him within the cathedral in Santa Fe.

a. many people
b. a few people

c. no one
d. one stranger

Latour and Vaillant were sent to New Mexico, and this led to _____.

a. less corruption there
b. New Mexico becoming mostly Catholic
c. Vaillant converting to Catholicism
d. more rogue priests

Latour asked for permission to return to Santa Fe and prepare for death and this led to Latour _____.

a. speaking only French
b. planting many flowers
c. receiving Bernard
d. planting many trees

Latour encouraged Vaillant and this led to Vaillant _____.

a. succeeding more than once
b. becoming Catholic
c. improving his health
d. retiring with Latour in New Mexico

The old apricot tree that grew on the estate where Latour eventually retired in the late 1800s was a young sprout in _____.

a. the late 1600s
b. the late 1500s
c. the early 1500s
d. the early 1400s

If Latour had first entered Santa Fe a few hours later, he would have scheduled his final return to Santa Fe during the _____.

a. night
b. late afternoon
c. early afternoon
d. morning

In Latour's view, people in New Mexico were usually _____ him because of missionaries who came there centuries ago.

a. friendly toward

b. suspicious of
c. hostile toward
d. apathetic toward

Elie and His Internment

In the fall of 1944, the Nazi SS began a process of selection at the concentration camp. They separated the weak prisoners from the strong. The weak were eliminated and Elie was given work dragging heavy blocks of stone. Elie worried most for his aging father, Chlomo. Following the initial selection, Elie's father was called along with nine others from Block 36 for a second examination. Chlomo feared that he would not return to see his son again. Chlomo gave him a knife and a spoon, the only inheritance he had to pass along. At the end of the day, he returned and Elie gave back his inheritance. Unfortunately, others did not return. For example, the concentration camp had weakened Akiba and he knew he would not pass the selection process. Even worse, he once held a strong faith in God and both his faith and body had been broken. Akiba made a last request of his friends before leaving that they recite the Kaddish (a Jewish prayer) in his memory. Over the next three days, conditions at the camp were terrible with exhausting work and cruel punishments. And Akiba's friends forgot their promise to recite the Kaddish for him.

Their captors provided warmer clothes when winter came, but the prisoners still suffered from the icy temperatures at night and the difficult work conditions. Elie entered a hospital in January of 1945 in order to drain pus from his foot. Another inmate there in the ward told him that the sickest patients were selected for death. The Jewish surgeon was nice to Elie, though, and told him that he would recover in a couple of days. Then, in two days, they began to hear the sound of guns in the distance. Rumors spread that the Red Army was approaching. The SS evacuated most of the inmates from the hospital on the following day, taking them to central Germany. Among them, Elie and Chlomo moved through the snow on a dangerous journey. They learned later that the inmates who remained in the hospital were eventually freed by the Russians. The evacuating prisoners were made to run and the SS shot all prisoners who fell behind. Elie almost welcomed death. Only concern for his father kept him going. They kept going through the night and into the morning. Then, at a rest stop, many froze to death beneath a covering of snow.

Even the soldiers of the SS were weary. Chlomo helped to keep Elie awake so that he would not freeze and they kept moving. When they finally arrived at Gleiwitz, the prisoners were assigned into crowded barracks for the night. In the darkness of the building, Elie was almost crushed and had to bite and claw for a breath of air. Somewhere in the crowd, he heard his friend Juliek play a fragment from a Beethoven concerto. Juliek was a Polish musician who had managed to keep his violin with him. He was dead by morning. Juliek and his instrument had been trampled on the floor. The SS kept them in Gleiwitz for three days and they received no food or water. The inmates also began to hear the sounds of gunfire outside the barracks. This revived hopes the Red

Army may be advancing. On the last day, the inmates were marched to the rail lines. They waited to be given one ration of bread and ate snow from each other's backs to quench their thirst. In the evening, a train made of roofless cattle cars arrived. The SS herded a hundred of them into each car before setting out.

The challenges and continual abuse associated with life in the concentration camp had _____ Akiba's faith.

a. weakened
b. strengthened
c. changed nothing about
d. created

The surgeon who treated Elie in the hospital for a problem with his foot and who treated him well was _____.

a. Jewish
b. Polish
c. German
d. Russian

Elie and the other prisoners were kept by the SS in Gleiwitz for _____ days before being marched to the rail lines.

a. three
b. two
c. four
d. five

News of a second exam in the camp led to Elie being given _____.

a. a knife and spoon
b. blocks of stone
c. a summons to be examined again
d. a violin

After the SS evacuated inmates from the hospital, the Red Army _____.

a. freed inmates there
b. approached from far away
c. fired guns nearby
d. departed

Elie's assignment to a barracks in Gleiwitz led to him _____.

a. hearing music

b. meeting a surgeon
c. almost freezing to death
d. being examined by the SS soldiers

It became clear to the SS during the initial selection process that Elie was one of the _____ prisoners in Block 36.

a. stronger
b. weaker
c. less dangerous
d. more dangerous

Before eventually reaching Gleiwitz, Elie and Chlomo _____ the group.

a. stayed with
b. fell behind
c. became separated in
d. slipped away from

Elie's friend Juliek was eventually killed by _____ in a crowded barracks.

a. other prisoners
b. SS soldiers
c. extreme cold
d. lack of water

Steppenwolf and the Galleries

Pablo invited Steppenwolf and Hermine into the Magic Theater. It was the 1920s. In blue light, they drank beverages containing mind-altering drugs. According to Pablo, the elixir allowed someone to see the world of the soul. Pablo held up a small mirror and Steppenwolf saw two different identities in it: a broken man and a beautiful wolf. Pablo told Steppenwolf that he must laugh at himself in the mirror in order to participate in the theater, where there were an endless number of galleries. Steppenwolf did so. Consequently, the small mirror became dark. Then Pablo turned Steppenwolf toward a larger mirror where he saw countless images of himself at all ages. Some images jumped from the mirror into the galleries. A fifteen-year-old boy jumped into the gallery labeled "All Girls Are Yours. One Quarter in the Slot." However, the first gallery that Steppenwolf chose to enter was "Jolly Hunting. Great Hunt in Automobiles." Within, war raged between humans and machines. An old friend to Steppenwolf, Gustav, appeared in the gallery. He was a theologian who left his job to fight in the war. Gustav chose to fight against the machines although he did not especially care which side he chose.

Gustav and Steppenwolf hid in a tree house and fired on passing cars. They killed some

people and spared others; and they were ashamed of themselves when it was done. The second gallery that Steppenwolf entered was labeled "Guidance in the Building Up of the Personality. Success Guaranteed." There, he met a chess player who offered to help Steppenwolf assemble his life. The chess player resembled Pablo. Steppenwolf looked again into a mirror and his image shattered into multiple selves that became chess pieces. These pieces were arranged on the chess board in different ways, with each exploring alternate realities. Steppenwolf was distracted from the chess board by the gallery labeled "Marvelous Taming of the Steppenwolf." Within the gallery, a man commanded a wolf to perform various tricks: kneeling, playing dead, retrieving a whip, etc., and Steppenwolf recognized himself as the man. Steppenwolf felt these menial tasks destroyed the wolf's nature and this horrified him. Then the man and the wolf switched roles. The wolf commanded the man to perform tricks. The man removed his clothes, played dead, etc. Then he killed a rabbit and lamb. He devoured them raw. Steppenwolf was shocked and ran disgusted from the third gallery.

Steppenwolf entered the gallery labeled "All Girls Are Yours. One quarter in the Slot." He repeatedly relived the day when he met Rosa Kreisler and then decided to change the outcome in which no relationship had developed. He proclaimed his love for Rosa and they grew to love each other. Following Rosa, he relived moments with all of his loves, real or imagined, until he came to Hermine. Steppenwolf left the fourth gallery exhausted to find the real Hermine. The next gallery surprised him with the label, "How One Kills for Love." This reminded him of his first dinner with Hermine. She had told him that she would eventually command him to kill her. He became desperate and tried to retrieve his chess pieces to rearrange them. The pieces in his pocket were gone. Instead, he only found a knife. Steppenwolf looked in the mirror and saw the wolf grinning back. Checking the theater, Pablo and Hermine had disappeared. He checked the mirror again and the wolf was gone. Instead, Harry was looking back. Harry was his real self who said that he waited for death and it was approaching. Steppenwolf heard Don Giovanni playing. Then Mozart appeared and laughed.

When Gustav appeared in the first gallery, he had been _____ to Steppenwolf.

a. an old friend
b. a new friend
c. an acquaintance
d. a stranger

The label for the second gallery that Steppenwolf entered contained the words _____.

a. Success Guaranteed
b. How One Kills
c. One Quarter in the Slot
d. The Magic Theater

In the fourth gallery, Steppenwolf relived a prior experience with _____ and then decided to change the outcome.

a. Rosa
b. Hermine
c. Pablo
d. Gustav

Steppenwolf was told to laugh at himself and this led to the _____.

a. small mirror darkening
b. drinking of an elixir
c. appearance of blue light
d. small mirror showing two images

Steppenwolf met a chess player and this led to Steppenwolf seeing _____.

a. his image shatter
b. Gustav
c. himself at the age of fifteen
d. blue light

Steppenwolf found the chess pieces in his pocket were gone and this led to him seeing _____.

a. a grinning wolf
b. Hermine
c. Pablo
d. a chess player

Steppenwolf saw many things when he first looked in the large mirror, including an image of _____.

a. himself at the age of four
b. a wolf
c. Hermine
d. Mozart

In the gallery labeled “Marvelous Taming of the Steppenwolf,” the wolf commanded _____ to do various tricks.

a. Steppenwolf
b. Pablo
c. Gustav
d. a stranger

If the words that Hermine spoke to Steppenwolf came true, Steppenwolf would surely meet _____ again.

a. Hermine
b. Pablo
c. the wolf
d. Mozart

REFERENCES

Ahern, S. K. (1978). Activation and intelligence: Pupillometric correlates of individual differences in cognitive abilities. Unpublished doctoral dissertation, University of California, Los Angeles, CA.

Ahern, S. K., & Beatty, J. (1979). Pupillary responses during information processing vary with Scholastic Aptitude Test scores. *Science, 205*, 1289-1292.

Alter, A. L., Oppenheimer, D. M., Epley, N., & Eyre, R. N. (2007). Overcoming intuition: Metacognitive difficulty activates analytic reasoning. *Journal of Experimental Psychology, 136*(4), 569-576.

Beatty, J. (1982). Task-evoked papillary responses, processing load, and the structure of processing resources. *Psychological Bulletin, 91,* 276-292.

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

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

Berman, R. A., & Katzenberger, I. (2004). Form and function in introducing narrative and expository texts: A developmental perspective. *Discourse Processes, 38*(1), 57-94.

Best, R. M., Floyd, R. G., & McNamara, D. S. (2008). Differential competencies contributing to children's comprehension of narrative and expository texts. *Reading Psychology, 29*, 137-164.

Blattenberger, G., & Lad, F. (1985). Separating the Brier score into calibration and refinement components: A graphical exposition. *The American Statistician, 39,* 26-32.

Braine, M. D. S., O'Brien, D. P., Noveck, I. A., Samuels, M. C., Lea, R. B., Fisch, S. M., Yang, Y. (1995). Predicting intermediate and multiple conclusions in propositional logic inference problems: Further evidence for a mental logic. *Journal of Experimental Psychology, 124*(3), 263-292.

Braine, M. D. S., Reiser, B. J., Rumain, B. (1998). Evidence for the theory: Predicting the difficulty of propositional logic inference problems. In M. D. S. Braine & D. P. O'Brien (Eds.), *Mental logic* (pp. 91-144). Mahwah, NJ: Lawrence Erlbaum Associates.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probabilities. *Monthly Weather Review, 78,* 1-3.

Cook, A., Zheng, R. & Blaz, J. W. (2009). Measurement of cognitive load during multimedia learning activities. In R. Zheng (Ed.), *Cognitive effectives of multimedia learning* (pp. 34-50). Hershey, PA: Information Science Reference/IGI Global Publishing.

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

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

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.

Efklides, A. (2008). Metacognition: defining its facets and levels of functioning in relation to self-regulation and co-regulation. *European Psychologist, 13*(4), 277-287.

Epley, N., & Gilovich, T. (2006). The anchoring-and-adjustment heuristic: Why the adjustments are insufficient. *Psychological Science, 17*(4), 311-318.

Fajardo, D. M., & Schaeffer, B. (1982). Temporal inferences by young children. *Developmental Psychology, 18*(4), 600-607.

Franks, B. A. (1997). Deductive reasoning with prose passages: Effects of age, inference form, prior knowledge, and reading skill. *International Journal of Behavioral Development, 31*(3), 501-535.

Gerrig, R. J., & O'Brien, E. J. (2005). The scope of memory-based processing. *Discourse Processes, 39,* 225-242.

Gilovich, T., Griffin, D., & Kahneman, D. (Eds.). (2002). *Heuristics and biases: The psychology of intuitive judgment*. New York: Cambridge University Press.

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

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

Grabe, W. (2002). Narrative and expository macro-genres. In A. M. Johns (Ed.), *Genres in the classroom: Multiple perspectives* (pp. 249-267). Mahwah, NJ: Lawrence Erlbaum Associates.

Graesser, A. C., Louwerse, M. M., McNamara, D. S., Olney, A., Cai, Z., & Mitchell, H. H. (2007). Inference generation and cohesion in the construction of situation. In F. Schmalhofer & C. Perfetti (Eds), *Higher level language processes in the brain: Inference and comprehension processes* (pp. 289-310). Mahwah, NJ: Lawrence Erlbaum Associates.

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

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. (submitted). Comparing absolute and relative accuracy in a classroom context. *Contemporary educational psychology.*

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*, 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. 429-456). New York: Psychology Press.

Johns, A. M. (2002) (Ed.). *Genre in the classroom: Multiple perspectives.* Mahwah, NJ: Lawrence Erlbaum Associates.

Juslin, P., Winman, A., & Olsson, H. (2000). Naïve empiricism and dogmatism in confidence research: A critical examination of the hard-easy effect. *Psychological Review, 107*(2), 384-396.

Just, M. A., & Carpenter, P. A. (1993). The intensity dimension of thought: Pupillometric indices of sentence processing. *Canadian Journal of Experimental Psychology, 47*(2), 310-339.

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

Kintsch, W. (1993). Information accretion and reduction in text processing: Inferences. *Discourse Processes, 16*, 193-202.

Kintsch, W. (1994). Text comprehension, memory, and learning. *American Psychologist, 49*, 294-303.

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

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85*, 363-394.

Kleiter, G. D., Doherty, M. E., & Brake, G. L. (2002). The psychophysics metaphor in calibration research. In P. Sedlmeier & T. Betsch (Eds.), *ETC.: Frequency processing and cognition* (pp. 239-255).

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

Koku, P. S., & Qureshi, A. A. (2004). Overconfidence and the performance of business students on examinations. *Journal of Education for Business, 79,* 217-224.

Lea, R. B. (1995). On-line evidence for elaborative logical inferences in text. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(6), 1469-1482.

Lea, R. B., Mulligan, E. J., & Walton, J. L. (2005). Accessing distant premise information: How memory feeds reasoning. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 31*, 387-395.

Maki, R. H. (1998). 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: Erlbaum.

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

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.

Masson, M. E. J., & Rotello, C. M. (2009). Sources of bias in the Goodman-Kruskal Gamma coefficient measure of association: Implications for studies of metacognitive processes. *Journal of Experimental Psychology, 35*(2), 509-527.

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

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.), *The psychology of learning and motivation, vol. 26* (pp. 124-141). New York: Academy Press.

Pressley, M., & Ghatala, E. S. (1990). Self-regulated learning: Monitoring learning from text. *Educational Psychologist, 25*, 19-33.

Rader, A. W., & Sloutsky, V. M. (2002). Processing of logically valid and logically invalid conditional inferences in discourse comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(1), 59-68.

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

Salmen, D. (2004). *Differences in reading comprehension post acquired brain injury: Real or imagined?* Unpublished doctoral dissertation, University of Utah, Salt Lake City, Utah.

Scheck, P., Meeter, M., & Nelson, T. O. (2004). Anchoring effects in the absolute accuracy of immediate versus delayed judgments of learning. *Journal of Memory and Language, 51*, 71-79.

Schraw, G. (2009). Measuring metacognitive judgments. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of metacognition in education* (pp. 415-429). New York: Routledge.

Schraw, G., Potenza, M. T., & Nebelsick-Gullet, L. (1993). Constraints on the calibration of performance. *Contemporary Educational Psychology, 18*, 455-463.

Suantak, L., Bolger, F., & Ferrell, W. R. (1996). The hard-easy effect in subjective probability calibration. *Organizational Behavior & Human Performance, 67*(2), 201-221.

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

Thiede, K. W., & Dunlosky, J. (1999). Toward a general model of self-regulated study: An analysis of selection of items for study and self-paced study time. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 1024-1037.

Tversky, A. & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185* (4157), 1124–1131.

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

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

Wolfe, M. B. W., & Mienko, J. A. (2007). Learning and memory of factual content from narrative and expository text. *British Journal of Educational Psychology, 77*, 541-564.

Xie, B., & Salvendy, G. (2000). Prediction of mental workload in single and multiple tasks environments. *International Journal of Cognitive Ergonomics, 4*(3), 213-242.

Yates, J. F. (1990). *Judgment and decision making.* Englewood Cliffs, NJ: Prentice Hall.

Zaromb, F. M., Karpicke, J. D., & Roediger, H. L. (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.

Zwaan, R. A., Langston, M. C., & Graesser, A. C. (1995). The construction of situation models in narrative comprehension: An event-indexing model. *Psychological Science, 6*, 292-297.

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin, 123*, 162-185.