Reconceptualizing Metacomprehension Calibration Accuracy

by

Rylan Graham Egan

M.Ed. (Curriculum Studies), Queen's University, 2007
B.B.A., Brock University, 2001

Dissertation Submitted in Partial Fulfillment
of the Requirements for the Degree of
Doctor of Philosophy

in the
Educational Psychology Program
Faculty of Education

SIMON FRASER UNIVERSITY
Fall 2012

Approval

Name: **Rylan Graham Egan**

Degree: **Doctor of Philosophy**

Title of Thesis: ***An Investigation into Learners' Metacomprehension Accuracy in a Within Text Design***

Examining Committee:

Chair: First name Surname, Position

Philip H. Winne
Senior Supervisor
Professor

First name Surname
Internal Examiner
Assistant/Associate/Professor
School/Department or Faculty

First name Surname
External Examiner
Assistant/Associate/Professor, Department
University

Date Defended/Approved: September 27, 2012

Abstract

Accurate judgment of text comprehension is compulsory for learners to effectively self-regulate learning from text. Unfortunately, until relatively recently the literature on text comprehension judgment, termed metacomprehension, has shown learners to be inaccurate in their judgments. Over the last decade researchers have discovered that when learners use delayed summaries of text to make judgments metacomprehension accuracy increases. In contrast, when learners use individual differences (e.g., knowledge and interest) to make judgments they are less accurate. Traditionally metacomprehension accuracy has been construed as the average correlation of judgments and comprehension assessments across multiple texts.

In the current study multiple alterations to the delayed summarization paradigm were evaluated. Specifically, the difference between learners' comprehension assessments and assessment scores were calculated within text to assess text specific differences in judgment cue use and accuracy. Second, pre-reading prompts were provided to focus learners on connections within the text (graphic organizer) and specific factual information (a list of facts). Third, the relative influence of individual differences (i.e., interest, knowledge, GPA, understanding of university text and Need-For-Cognition), text, pre-reading prompts, and summary delay on comprehension judgments were evaluated. Finally, experimental influences on judgment/assessment score differentials were considered.

Results indicated that the two obscure texts in this experiment, with similar structures, and different levels of interest resulted in statistically detectably different judgments, scores, and accuracy. Delayed summarization did not improve metacomprehension accuracy for either text. This is a departure from the current literature, and may indicate that within-text measures of metacomprehension accuracy react differently to delayed summarization. Individual differences also affected judgments differently across texts and groups. Finally, participants used different cues to make judgments at the detailed and explanation levels of understanding.

Keywords: Metacomprehension; Delayed Summarization Effect; Graphic Organizers; Metacomprehension Accuracy; Judgments of Learning; Elaboration

To Asha and Anya,
may you always disregard artificial barriers.
To Sujji, my wife, friend, and refuge.

Acknowledgements

I would like to begin by acknowledging the ongoing support of my supervisor Dr. Philip Winne. Dr. Winne has provided academic guidance and financial support throughout my studies. He has taught me that hard work, academic rigour, and good planning can seed a long and prosperous career. I would also like to thank my committee member Dr. John Nesbit for his help in the development and refinement of this dissertation. Finally, I would like to thank my mentor Dr. Denise Stockley. Dr. Stockley has been my steady guide, confidant, and support throughout my graduate studies. This was written, and I am here, because she was.

My greatest gratitude, appreciation, and love for my wife, Sujji Murthy, who has shared this goal and made many sacrifices to see it accomplished. To my children, Asha and Anya, who have inspired me to press forward in my career. And to my Mother, Father, and Sister, who have supported my graduate work and educational development from the beginning.

Finally, I would like to thank my co-workers and collaborators who have inspired me and supported my progress. In particular, Dr. Robert McTavish who has kept me laughing, provided valuable advice, and ensured that I didn’t take this process too seriously. Dr. Mingming Zhou, who has been a co-author and guiding figure throughout my doctoral studies. Wenting Ma who provided technical support and a listening ear. And all others who have supported me in my studies.

Thank you.

Table of Contents

List of Tables

List of Figures

1. Rationale

As evidenced by its focus within scholastic entry exams (e.g., SAT, GRE, GMAT, etc.), reading comprehension is an integral skill requisite to success at the university level (e.g., http://www.ets.org/gre/). However, for text comprehension to be successful, learners must also be able to judge qualities of their comprehension (Dunlosky & Metcalfe, 2009) commonly referred to as *metacomprehension* (Maki & Berry, 1984). Accurate metacomprehension allows learners to assess discrepancies between desired and actual states of comprehension, and then employ strategies to mitigate differences. The concept of *metacomprehension accuracy* can be referred to in absolute or relative terms. In a relative sense, metacomprehension accuracy refers to learners' ability to rank texts by the extent to which they are judged to be learned (relative accuracy). In the absolute sense, metacomprehension refers to a judgment of the extent to which actual learning from text (performance) corresponds to their learning goals (absolute accuracy). This study considers the latter, absolute accuracy but is novel in investigating an intervention found to influence the former, relative accuracy.

To be successful in academic settings, learners must accurately predict future comprehension in both relative and absolute terms. However, this process is complicated by differential rates of memory decay over time. Specifically, rote memory for facts, figures, and details tends to be less stable than inferences and connections that learners tie to personal affect and prior knowledge (Anderson & Thiede, 2008).

A large base of experimental literature has developed over the past four decades with the goal of increasing learners' metacomprehension accuracy (for a review see

Thiede, Wiley, Griffin, & Redford, 2009). This literature has largely considered metacomprehension in terms of relative accuracy. Although initial attempts to understand and increase relative metacomprehension accuracy had limited success (e.g., Glenberg, & Epstein,1987; Glenberg, Sanocki, Epstein, & Morris, 1987; Glenberg, Wilkinson, & Epstein, 1982), attempts over the last decade have been much more fruitful (e.g., Dunlosky & Lipko, 2007; Griffin, Wiley, & Thiede, 2008; Thiede, Wiley, & Griffin, 2010). In recent research, interventions designed to refocus participants' attention on particular kinds of judgment cues have increased accuracy. The two cues most often theorized to increase learners' judgment accuracy are: (a) the extent of learners' network of interconnected text events, characters and other causal agents, and prior knowledge (often referred to as the learners situation model of a text, Anderson & Thiede, 2008), and (b) the learners' accessibility to information (amount and ease of access) during recall and/or reading (Baker & Dunlosky, 2006). Unfortunately, when learners judge comprehension, they have tended to focus on less predictive cues such as prior knowledge, interest, and perceived ability (Thiede, Griffin, Wiley, & Anderson, 2010).

The current study investigates learners' use of individual differences and situation model cues for judging comprehension. Prior research suggests this intervention influences learners' to consider the situation model and thus increases metacomprehension accuracy. Individual differences in this study refer to variables that differ among participants including perceived ability to comprehend university text, predicted interest and knowledge about the topic studied, enjoyment of cognitively challenging tasks (i.e., *Need-For-Cognition*), and general academic proficiency as measured by GPA. In this study the transferability of findings from the literature on relative accuracy will be assessed in terms of absolute accuracy. Participants read texts and then wrote summaries after a delay. A unique intervention in the current study is the use of pre-reading prompts intended to shape participants' depth of text processing.

In this experiment I also altered conventional methodology to measure a different element of metacomprehension accuracy than has been measured in prior research. In previous studies, metacomprehension accuracy has been expressed as the correlation of comprehension judgments and assessment outcomes calculated within persons over multiple texts (usually 4 -6). This computational method implicitly assumes that characteristics of individual texts (in the sample) do not influence judgments. This present study tests the effect of interventions when metacomprehension is computed separately for each of two texts. In summary, this study investigated three main questions related to text specific influences: 1) *Do experimental interventions influence judgments, assessment scores and metacomprehension accuracy for detailed and explanation questions when calculated within texts? 2) Are statistically detectable differences in judgments, scores, and metacomprehension accuracy found between texts? 3) Do individual differences - interest, knowledge, GPA, understanding of university text and Need-For-Cognition - influence judgments differently as a function of text studied.*

Writing text summaries has consistently been shown to increase metacomprehension accuracy (for a review see Thiede, Wiley, Griffin, & Redford, 2009). However, increases only appear to occur when there is a delay between reading texts and writing summaries. This contingency is theorized to result from different kinds of information learners use to make judgments based on the summaries they write immediately after studying or after a delay. During immediate summarization, the extent of memory for facts and details is theorized to be the cue on which learners judge their comprehension. Since information at this level of complexity (a) degrades rapidly between writing the summary and a delayed test and (b) may be of little use in tests of comprehension, judgments are inaccurate. In contrast, when participants write summaries at some delay after studying a text, they are theorized to rely on understanding at the level of the situation model to make judgments. Information at this level tends to be more stable over time, more useful on tests of comprehension, and therefore is a more accurate cue on which to base judgments of comprehension.

Previous researchers have not yet considered interactions between depth of text processing during reading, and delayed summary effects. It is plausible that readers' initial focus on text interconnections may influence summary depth in addition to immediate and delayed summarization. Providing advanced organizers prior to reading can increase students attention to interconnections within texts (Ausubel, 1978, Mayer, 1979). To test for differences in learners' focus during text reading either advanced organizers or a list of facts from the text were provided to two of the four experimental groups. In this study two questions regarding immediate and delayed summarization were investigated: 4) *Do experimental interventions influence the number of recalled Specific Details, Facts, and Connections in summaries* 5) *Do the number of Specific Details, Facts, and Connections in summaries differ between texts 6) Do associations between judgments and characteristics of summaries that learners write - the number of Specific Details, Facts, Connections and summary writing time - differ as a function of characteristics of texts learners study and experimental interventions?*

2. Literature Review

2.1. Introduction

In the following review of the metacomprehension literature theoretical foundations required for the current study are provided in two parts. In part I, the measurement challenges and practical benefits of metacomprehension accuracy are reviewed. Specific attention is paid to issues in indexes of metacomprehension accuracy and implications for interpreting an index aggregated over individual vs. multiple texts. Finally, in part I the importance of text characteristics and methods of assessing comprehension are considered. In part II, overarching influences on metacomprehension accuracy with a focus on individual differences and writing text summaries either immediately after studying or at a delay, will be reviewed. Part I and II provide a foundation for the analyses of experimental interventions in this study, and the resulting effect of learners' judgments, comprehension scores, and metacomprehension accuracy. All the studies reviewed here refer to relative metacomprehension accuracy, most commonly measured by the gamma correlation.

2.2. Part I: Challenges and Benefits of Accurate Measures of Metacomprehension Accuracy

Part I begins with a short rationale for, and explanation of metacomprehension accuracy. Next, the implication of measurement on interpretations of metacomprehension accuracy are explored. Finally, characteristics of text required to validly measure metacomprehension, and methods of assessing reading comprehension are reviewed. Implications for this study are discussed in the *relevance to current study* sections.

2.2.1. Nature of metacomprehension measurement.

Metacognition refers to "the awareness learners have about their general academic strengths and weaknesses, cognitive resources they can apply to meet the demands of particular tasks, and their knowledge about how to regulate engagement in tasks to optimize learning processes and outcomes" (Winne & Perry, 2000, p. 532). Metacognition can also be further classified into *monitoring* and *control* functions. Metacognitive monitoring refers to the ability to qualitatively evaluate cognition in reference to pre-established criteria. Metacognitive control refers to the ability to choose and apply cognitive strategies to ameliorate discrepancies between evaluated cognition and individuals' criteria for success (Greene & Azevedo, 2007).

Accurate metacomprehension requires a reader to effectively metacognitively monitor actual and/or expected comprehension before, during, and after reading text (Dunlosky, 2005; Maki & Serra, 1992; for review see Thiede, Griffin, Wiley, & Anderson, 2010). Manipulations targeting metacomprehension as measured by correlations between judgments and assessment scores, have been demonstrated to affect participants' ability to monitor and regulate text learning (Nietfeld, Cao, & Osborne, 2005). Since metacognitive monitoring has not traditionally been observed directly (although educational neuroscience is beginning to make headway on this front; see, e.g., Rieger, Reichert, Gegenfurtner, Noesselt, Braun, Heinze et al., 2008), studies reviewed here are limited to those using behavioural correlates of accuracy.

2.2.2. Influences of metacomprehension accuracy on learning.

Relative metacognitive monitoring accuracy is important for self-regulating one's study of text. Inaccurate monitoring may cause learners to spend time unnecessarily on materials that are actually understood, or fail to employ strategies required to adequately comprehend text (Thiede, Anderson, & Therriault, 2003). Empirically, there have been two main theoretical camps recommending slightly different explanations for how metacognitive monitoring may influence study behaviour.

In one camp, Thiede and colleagues (Dunlosky & Thiede, 1998; Thiede, 1999; Thiede & Dunlosky, 1999) have proposed the d*iscrepancy reduction model*. It contends learners continue studying until "the error between the perceived state of learning and the amount of learning reaches zero" (Thiede & Dunlosky, 1999, p.38). In the other camp, Metcalfe and colleagues (Kornell & Metcalfe, 2006; Metcalf, 2002) propose the *region of proximal learning* hypothesis. It describes that participants choose materials to restudy for which comprehension can be gained with the least effort. A commonality between these theories is that accurate metacomprehension is required for effective regulation of studying from text. As such, the value of accurate metacomprehension will be taken as given, and this review will focus on mechanisms affecting accuracy and not on the importance of metacomprehension accuracy *per se*.

2.2.3. Measurements of metacomprehension accuracy.

To better understand the relevance of measuring metacomprehension on a *per text* basis, it is important to first review how it has been measured across multiple texts. Once traditional measures of metacomprehension have been discussed, methods for evaluating this construct are further explained in the *relevance to current study* section.

Relative metacomprehension accuracy is best measured using the gamma correlation coefficient (Maki, Shields, Wheeler, Zacchilli, & 2005; Thiede, Griffin, Wiley &

Redford, 2009), and this has been the primary measure of metacomprehension accuracy over the past 30 years. Gamma is commonly referred to as *resolution* or *relative accuracy* as it gauges the accuracy of each judgment about how well a text is comprehended relative to other judgments about other texts (Goodman & Kruskal, 1954; Nelson, 1984).

A second measure of metacomprehension accuracy, and the measure used in the current study, is called absolute accuracy. Absolute accuracy is used to assess the magnitude of judgment accuracy across texts. After studying each text participants make a judgment of comprehension, they subsequently take one (or more) test(s) of comprehension for each text, and the squared difference between judgments and the test (or average of multiple tests) of comprehension is divided by 1. This procedure is also referred to as *calibration* as it measures learners' ability to calibrate (or bring into sequence) comprehension or memory judgments and performance on corresponding texts. Absolute accuracy is calculated as shown in Equation 1; where J = Judgment, T = Test Score, and *i* = a judgment/test trial such as a text (Schraw, 2009). Perfect absolute accuracy occurs when the coefficient is 1. Lower scores indicate less accuracy.

Equation 1: Absolute Accuracy Coefficient

$$\frac{1}{N}\sum_{i=1}^{N}\left(J_i - T_i\right)^2$$

A variation of absolute accuracy is called bias. Bias is calculated by not squaring the product of J_i - T_i in Equation 1. As a result, the index calculated for each person is either positive (overconfident) or negative (underconfident, Schraw, 2009). A participant who is neither over- nor underconfident will have a bias score of 0. Other indexes of absolute accuracy include the Prediction Accuracy Quotient (Maki & Swett, 1987), and the Hamann Coefficient (Nietfeld, Enders, & Schraw, 2006). However, as these indexes are far less prevalent in the literature they will not be explored further.

An example will clarify the implications of resolution and calibration for a student. Suppose a student has an exam covering 5 textbook chapters. If resolution (gamma) is high but calibration (absolute accuracy) is close to zero, the student will be able to choose chapters for restudying that are least understood - the chapters are judged relative to one another regardless of how much is understood about each. In contrast if a student had a high absolute accuracy index but a relatively low gamma coefficient, she would be able to effectively judge her average comprehension of texts but would have difficulty determining which chapters were understood better (or worse) than others.

In the most common paradigm, participants read a number of texts, make comprehension judgments about each text, and are then tested for each text (for a review see Lin and Zabrucky,1998). A gamma coefficient is then derived from the non-parametric gamma correlation between judgments and tests across all texts. An absolute accuracy score is calculated by dividing the sum of accuracy indexes for each text by the number of judgment/test trials as demonstrated in Equation 1 (Maki, Shields, Wheeler, & Zacchilli, 2005; Thiede, Griffin, Wiley, & Redford, 2009). When an intervention intended to affect metacomprehension is researched, a group's accuracy is construed as the median of gamma or absolute accuracy scores across participants in each group. In the vast majority of studies, coefficients describe metacomprehension for each student while ignoring potential differences between texts; i.e., it is assumed that characteristics of a text do not influence either a reader's judgment of comprehension or their actual comprehension.

2.2.3.1. Relevance to current study.

In the current study, the potential interaction between text, participant, and experimental manipulation is the focus of investigation. As such, to evaluate text specific influences on metacomprehension judgment accuracy, absolute coefficients were calculated and evaluated for each text. In this study absolute accuracy was measured instead of bias as my intent was to evaluate the alignment of judgments and assessment scores, not whether participants' judgements were over or under confident.

This departs from common practice where, because accuracy (both relative and absolute) is calculated across a number of texts (usually $\geq$4), measures are insensitive to text properties (e.g., personal relevance, interest, etc.) that may mediate learners' metacomprehension accuracy. It is important to consider possible influences of text characteristics on metacomprehension accuracy as learners' are often required to make accurate *per text* judgments (e.g., chapter tests, reading response activities, article reviews, etc.). To gain a single within-text and individual metacomprehension index of absolute accuracy the formula in Equation 2 was used. Since bias was not the focus, the square root of the squared product was calculated, and that quantity was subtracted from 1.0 so that higher scores reflect greater accuracy.

Equation 2: Accuracy Coefficient Used in this Study

$$1-\sqrt{(J_i - T_i)^2}$$

Finally, participants were asked to make two judgments and take two assessments for each text. The first judgment of comprehension was one concerning memory of facts, and the second concerned the more complex comprehension/situation model level understanding. In previous studies participants have been asked for judgments about one or the other type of understanding rather than both. Judgments were analyzed to better understand if/how participants a) use different cues to make judgments at these two levels, and b) if yes, what cues are used.

2.2.4. Metacomprehension, text, and assessment.

Assessment of metacomprehension accuracy requires that a) texts provide a sufficiently complex situation model that requires learners to move beyond mere memorization, b) participants' and researchers' share an understanding of how comprehension is assessed, and c) assessments effectively measure learners' comprehension at the level of the situation model. In this section findings related to these aspects of metacomprehension are reviewed, and a rationale for texts and assessments used in this study is provided in the *relevance to current study* section.

Wiley, Griffin, and Thiede (2005) categorized expository text complexity based on Meyer and Freedle's (1984) categorizations: *collections, comparisons, causations,* and *problem-solution*. Collections are texts that compile facts and figures as may be typically presented in encyclopaedia entries (Mayer & Freedle, 1984; Wiley, Griffin, & Thiede, 2005). Comparison texts are more complex and integrated than collection texts. In comparison texts, two or more propositions, characters, events, etc. are compared and contrasted with each other (Meyer & Freedle, 1984). Wiley et al. (2005) identified a number of texts in a series of studies by Arthur Glenberg and colleagues (e.g., Glenberg and Epstein, 1985, e.g., *A Good Hanging Never Hurt Anyone*) as comparison texts. In these cases, one perspective is favoured over another by the author. Texts such as these may confound measures of comprehension with a participant's disagreement or agreement with the author's and/or researcher's point of view.

Finally, Wiley, Griffin, and Thiede (2005) categorized causal and/or problem-solution texts (e.g., Ice Age Passage, Thiede, Wiley, & Griffin, 2010) into a single *explanatory* category. Comprehending texts in this category requires readers to make multiple kinds of inferences: surface, text base, and situation model based on Kintsch's (1988, 1994, 1998) model of reading comprehension. The lexical or surface level of comprehension concerns properties of text and the meaning of specific words and concepts. The text-based level refers to comprehending propositions that emerge from interpretations at the lexical level. Finally, the situation model emerges as a "high-level" or "deep" form of comprehension when the reader forms meaningful (and accurate) structures of meaning forged from propositions in the text, previous knowledge, and logical inferences grounded in these two sources of information.

Causal texts typically require readers to infer causation from associations implicit to the text. Problem-solution texts typically require readers to infer causes for problems, or determine how and why solutions may (or may not) be effective. The main differences between explanatory, collection and comparison texts are the number and complexity of connections and inferences involving lexical, text based and situation model levels and prior knowledge. Wiley et al. suggest that readers will be able to comprehend collection and comparison texts by memorizing propositions. Thus, tests for collections and comparison texts will be similar in depth and complexity to memory tests. In contrast, tests for explanatory texts will typically evaluate participants' situation model level understanding.

To accurately determine learners' ability to judge their comprehension, it is crucial to first ensure that participants' and researchers agree on what constitutes comprehension. Unlike memory for facts and details, judgment of comprehension cannot be measured dichotomously through recall or recall failure. Rather, metacomprehension is measured as the difference between participants' subjective judgment of personal comprehension (comprehension judgment), and performance on tests (comprehension assessment, Wiley, Griffin, & Thiede, 2005). Clearly, discrepancies between a learner's and an experimenter's criteria as to what constitutes comprehension will influence judgment accuracy, independently from any experimental manipulation (Thiede, Wiley, & Griffin, 2010). To mitigate such discrepancies, researchers have attempted to establish mutual understanding by providing descriptions of assessments, practice trials, and even providing correct responses for comparison (Dunlosky, Hartwig, Rawson, & Lipko, 2011; Thiede, Wiley, & Griffin, 2010).

Given the importance of inference generation in comprehension at the level of the situation model, a logical extension in the comprehension and metacomprehension literatures has been to ask participants to answer inference questions to assess situation model understanding (for reviews see Graesser, Millis & Zwaan, 1997; Thiede, Griffin, Wiley & Redford, 2009). Early metacomprehension researchers asked participants to make logical inferences based (to some extent) on opinions presented by the author and researcher (e.g., Glenberg, Wilkinson, & Epstein, 1982; Glenberg & Epstein, 1985). In these experiments participants were asked to answer single true-or-false questions. Using a Monte Carlo simulation Weaver (1990) demonstrated the importance of using multiple inference questions. Recent studies of metacomprehension at the situation model level have also used inference questions, usually between 5 and 8 four-distracter multiple choice questions for each text (Thiede, Griffin, Wiley & Redford, 2009).

2.2.4.1. Relevance to current study.

The two texts used in the current study satisfy Wiley, Griffin, and Thiede's (2005) requirements for an explanatory text. They are based on historical topics and provide a range of interconnections among people and events. In this way, these texts differ substantially from typical encyclopaedia entries that are referred to by Wiley, Griffin, and Thiede (2005) as a collection of facts and events.

To decrease discrepancies between participants' and this researcher's perceptions of comprehension, a new method for soliciting judgments of comprehension, referred to as *explanation tests*, was developed. To make judgments of text comprehension participants were asked "On a scale from 0-10, how well do you think you will be able to discuss the full meaning and implications of *Wind Blown Transportation/Nuclear Dumping in Russian Lake?*" Correspondingly, explanations tests asked participants to "Explain all major events in the history of pneumatics (i.e., Wind Blown Transportation) from its invention to the present day." and "Explain each major nuclear disaster, its causes, its effect, and the Russian reaction as described in Nuclear Dumping in Russian Lake." In this way the task requested at judgment mirrored the assessment task, and abstract terms were avoided (e.g., *comprehend*, *understand*, *know)* that may be differentially interpreted by researchers and participants.

Texts were revised from their original source (damninteresting.com) such that each text presented 16 *main events* each comprised of multiple text propositions. Participants' test scores were calculated based on the number of *main events* provided. As such, participants could provide 16 main events for each text. However, each event required integration of lexical, text based and situation model propositions.

The assessment method for comprehension in this study was chosen as it provides greater coverage of text materials, gives learners the opportunity to use their entire memory/comprehension to rebuild their situation model at the time of assessment, and may increase the calibration between experimental and individualized judgment criterion. Another benefit of this method is that participants' answers are constructed from their situation model understanding without a prompt provided by question stems and options as in a multiple-choice test. Finally, this method does not allow for guessing, a factor that could bias results.

2.3. Part II: Individual Differences and Experimental Influences on Metacomprehension Accuracy

This review first provides justification for examining the potential influence of individual differences on metacomprehension accuracy. Second, experiments that manipulated factors such as summarizing, the timing of summaries in relation to studying a text and rereading are discussed. This provides background and justification for experimental manipulations used in the current study. As in Part I, relevance to the current study are discussed throughout.

2.3.1. Individual differences and metacomprehension accuracy.

Although individual differences have not been a focus of the metacomprehension literature (Maki, 1998b), a number of studies have indicated that these factors may affect metacomprehension judgments (Linderholm, Zhao, Therriault, & Cordell-McNulty, 2008; Moore, Lin-Agler, & Zabrucky, 2005; Zhao & Linderholm, 2008). Individual differences to be reviewed here include expectations of success, prior knowledge, interest, and working memory span (Chiang, Therriault, & Franks, 2010; Griffin, Jee, & Wiley, 2009; Linderholm, Zhao, Therriault, & Cordell-McNulty, 2008; Maki & Berry, 1984; Zhao and Linderholm, 2011). Finally, Need-For-Cognition (NFC), a previously underexplored factor, will be discussed.

Linderholm and colleagues (Study 3; Zhao, Linderholm, & Therriault, 2006 as cited in Linderholm, Zhao, Therriault, and Cordell-McNulty, 2008) investigated self-reported cues used as bases for comprehension judgments. In these studies, participants read either one (Zhao et al., 2006) or two (Linderholm et al., 2008) unrelated texts with Flesch-Kincaid reading grades ranging between 10-11. After reading, participants were asked to describe cues they used for comprehension judgments. In both studies, participants reported rating their comprehension based on multiple individual differences including ability, prior knowledge and topic interest. Thiede, Wiley, and Griffin (2010) also found that, in addition to rote memory, participants relied most heavily on prior knowledge and interest as bases for making comprehension judgments. Interestingly, although very few students reported using comprehension *per se* as a basis for judgment, those who did made the most accurate judgments (Thiede, Griffin, Wiley, & Anderson, 2010).

Zhao and Linderholm (2011) demonstrated the importance of preconceived expectations for success on the magnitude of comprehension judgments. These investigators enhanced or depressed students' expectations for success by providing fake statistics. A high expectation group was told that previous cohorts received test scores of 85% (Exp. 1) and 95% (Exp. 2), while the low expectation group was told that previous cohorts had received test scores of 55% (Exp. 1 & 2). A control group was included that received no false information. Zhao and Linderholm found that average judgments across the three texts were statistically detectably higher for the high expectation compared to the low expectation group. No statistically detectable difference was found between the high expectation and control group, potentially reflecting persistent "natural" overconfidence found in the metacomprehension literature (Thiede, Griffin, and Wiley & Redford, 2009).

Self-reports of perceived interest and knowledge have also been found to affect variations in comprehension judgments (Linderholm, Zhao, Therriault & Cordell-McNulty, 2008; Thiede, Griffin Wiley, & Anderson 2010). Thiede, Griffin, Wiley, and Anderson (2010) found that participants relied on perceptions of both interest and perceived knowledge when making comprehension judgments. Lin, Zabrucky and Moore (1996) tested the effect of students' self-reported topic interest on judgment magnitude, performance, and judgment accuracy. Text judgments correlated with interest ratings computed across texts (collected after reading, *r* = .62), and interest for specific topics (collected before reading, *r* = .70).

Moore, Lin-Algler and Zabrucky (2005) found that expectations formed through prior testing experiences can influence subsequent text judgments. Students were allowed to read and judge their understanding of 12 texts. Using path analysis, Moore et al. found that judgments are based on accumulated prior judgments. This outcome corresponds to similar findings in the metamemory literature (Metcalfe & Finn, 2008). Koriat, Sheffer and Ma'ayan, (2002) also found that participants overconfidence was replaced with underconfidence after experiencing failure. Although confidence was regained with mastery experiences, increases did not compensate for underconfidence, leading Koriat et al. to coin the *underconfidence with practice effect*. Thus, a side effect of testing participants across multiple texts may be to bias judgments of metacomprehension due to perceptions about success or failure with prior reading tasks.

The impact of prior knowledge on metacomprehension judgments is debated. Glenberg and Epstein (1987) proposed the *domain familiarity hypothesis* that participants use perceived domain knowledge as a basis for future judgments. To test this hypothesis Glenberg and Epstein recruited students with extensive experience in one of two distinct fields, physics and music. Students read texts from both fields and made judgments. Participants more accurately predicted comprehension outside their domain of expertise. Other researchers have tested this effect but failed to find similar results (e.g., Lin, Zabrucky, & Moore, 1996; Maki & Serra, 1992).

In a recent study Griffin, Jee and Wiley (2009) tested the effect of baseball knowledge on metacomprehension accuracy for baseball texts. Participants were identified as expert or novice after completing a 45-item test of baseball knowledge. No statistically detectable difference was found for relative accuracy as a function of participants' knowledge of baseball. Moreover, expert participants showed better absolute calibration. Through a rigorous review of the literature, Griffin, Jee, and Wiley contend that there is little evidence that topic expertise negatively influenced metacomprehension accuracy. However, their findings require further verification in other domains.

Need-For-Cognition (NFC) was originally conceptualized as an "individual's tendency to organize his experience meaningfully" (Cohen, Stotland, & Wolfe, 1955). Cohen et al. rated participants' NFC based on their responses to hypothetical situations requiring thinking at various "depths." Next participants were asked to read texts with varying degrees of cohesion and ambiguity. They found that participants high in NFC were more likely to be frustrated and less interested in ambiguous texts. This original conceptualization of NFC assumed those high in NFC required clarity and ease of heuristic processing (Cacioppo, Petty, Feinstein, & Jarvis, 1996). In contrast, Cacioppo and Petty (1982) and Cacioppo, Petty and Kao (1984) construed low NFC as "the relative absence of the motivation for effortful cognitive activities that defines high need for cognition" (p. 198). Generalizing to research on comprehension of texts, those high in NFC are more likely to seek, enjoy and profit from effortful thought in ambiguous and/or complex texts.

Cacioppo and Petty (1982) created the first published measure of NFC. Initially the NFC instrument was investigated using populations assumed to have dichotomously high (university professors) or low (assembly line workers) NFC. Results from these populations, as well as other populations (university students) showed that NFC is negatively related to closed mindedness and positively related to general intelligence. Cacioppo and colleagues (Cacioppo & Petty, 1982; Cacioppo, Petty, & Kao, 1984) found that factor analyses of the full and a shorter ("efficient") version of the NFC instrument returned strong single factors accounting for 27% (full) and 37% (efficient) of overall variance. Cacioppo et al. (1996) noted that the efficient version was developed as the 34-item reached an asymptote for both variance explained and reliability at a length of 18 items.

I speculate NFC should be an integral factor in metacomprehension. According to the dual-processing hypothesis (Nelson & Narens, 1990), readers must comprehend at lexical, text based and situation model levels while monitoring cues indicative of comprehension. Both functions are likely to be more successful as learners are more willing to process text. Moreover, the intensity of engagement when participants' experience interventions such as rereading, self-explanation and concept mapping may increase with NFC, and (as reviewed next) may activate additional metacomprehension cues. In short, NFC may influence readers' comprehension goals, intensity of processing they undertake to meet their goals, and potentially the number and saliency of comprehension judgment cues. For this study the efficient measure of NFC was chosen as a) it limited time requirements for participants, b) Cacioppo, Petty, and Kao found loadings and 1st factor variance to be similar between the normal and short forms, and c) "little [was] sacrificed in terms of reliability" by using the short form (Cacioppo, Petty, and Kao, 1984, p. 306).

Although no direct investigation of NFC and metacomprehension was found in this review, participants with high NFC may be more likely to access feedback (Coutinho, Wiemer-Hastings, Skowronski, & Britt, 2005). Coutinho et. al. (2005) had participants complete problems from the Graduate Records Examination (GRE) and judge their success. After responding to the problem participants had the opportunity to either receive the correct answer with an explanation or just the correct answer. Participants in the high NFC group were more likely to request the explanation in addition to the correct answer.

2.3.1.1. Relevance to current study.

With the limited number of studies, it is unclear how person/text interactions may affect absolute accuracy. One of the main goals of this study is to statistically evaluate the effect of individual differences on metacomprehension accuracy in a between participant and within text design. By conducting identical analyses for each of two texts, a comparison of text specific determinants of accuracy will be evaluated. Moreover, this study investigates how individual differences may influence judgments at the memory and situation model levels of judgment. Finally, this study will investigate the influence of participants' NFC scores on metacomprehension accuracy.

2.3.2. Judgment cues used in metacomprehension.

The metacomprehension literature in the last decade has largely rejected the notion that learners have direct access to the extent of their comprehension (Maki, 1998a). Instead, the cue-utilization theory suggests that learners make inferences about their comprehension based on cues available prior to, during, or after studying (Koriat, 1997; Lin & Zabrucky, 1998). Koriat (1997) proposed three types of cues that can be used to make comprehension judgments: intrinsic, extrinsic, and mnemonic.

Intrinsic cues refer to properties intrinsic to materials that may influence participants' judgments. These include perceived text difficulty, interest, prior knowledge, complexity of the text (e.g., readability), and genre (Lin & Zabrucky, 1998; Thiede, Griffin, Wiley, & Anderson, 2010; Thiede, Wiley, & Griffin, 2010; Weaver & Bryant, 1995).

External cues are cues inferred from the perceived effectiveness of text processing evidenced in behaviour. Examples include generating self-explanations while reading, constructing concept maps, rereading text, answering adjunct questions, and spacing recall trials (e.g., Haenggi & Perfetti, 1992; Nesbit & Adesope, 2006; Rawson, Dunlosky, & Thiede, 2000; Thiede, Griffin, Wiley, & Anderson, 2010). These activities are similar in that they increase cognitive processing. This is commonly referred to as elaboration in cognitive psychology and that label will be used here as an encompassing term (Klein & Kihlstrom, 1986).

Finally, mnemonic cues refer to cues "that may signal for the participant the extent to which an item has been learned and will be recalled in the future" (Koriat, 1997, p. 351). This may include (often unconscious) inferences derived from the speed of text processing, the perceived accessibility of information at recall, and one's memory for previous recall attempts (e.g., Baker & Dunlosky, 2006; Morris, 1990; Rawson, Dunlosky, & Thiede, 2000; Zhao & Linderholm, 2011; Thiede, Griffin, Wiley, & Anderson, 2010.).

Often it is difficult to separate cue categories because interventions expose participants to multiple cues. For example, the second reading in a rereading intervention may make content seem more familiar, allow the text to be processed with less effort, and suggest to the learner that the intervention *per se* should increase comprehension (Dunlosky, 2005; Rawson, Dunlosky, & Thiede, 2000).

2.3.2.1. The influence of elaboration on metacomprehension.

Findings in the metacomprehension literature show that elaborative exercises, such as answering questions about a text while reading, increase monitoring. Put another way, effective elaborative exercises may be partially responsible for detecting discrepancies between current and desired states of comprehension, often termed the "norm of study" (Chiang, Therriault, & Franks, 2010; Griffin, Wiley & Thiede, 2010; Zhao & Linderholm, 2011). An interesting cycle may form whereby learners use elaborative processing to detect comprehension deficits, then correct deficits by employing elaborative processing techniques, which in turn expose further discrepancies and so on. For a similar account see Koriat, Ma'ayan, and Nussinson, 2006.

Various techniques that could be categorized as elaborative exercises have been studied as potential methods for increasing metacomprehension judgment accuracy. These include providing adjunct questions (Pressley, Snyder, Levin, Murray, & Ghatala, 1987; Walczyk & Hall, 1989), delaying judgment (Maki, 1998a), decreasing text coherence (Rawson, Dunlosky, & Thiede, 2000), and requiring information to be inserted or arranged (Thomas & McDaniel, 2007). Unfortunately, most such interventions have been relatively unsuccessful (see Maki, 1998a; Thiede, Griffin, Wiley, & Redford, 2009, cf. Thomas & McDaniel, 2007). This review will concentrate on two elaborative exercises that are most relevant to the current study and have been found to consistently increase metacomprehension accuracy: rereading and delayed-summarization.

2.3.2.2. Metacomprehension and the rereading effect.

The rereading effect predicts that fluency at the lexical and text base levels increases during rereading. Rawson, Dunlosky, and Thiede (2000) suggested a *levels-of-disruption hypothesis* to account for this effect. Disruptions refer to occasions when access to information and information processing are hampered due to deficits of either the reader (e.g., insufficient prior knowledge, poor reading ability) or the author, (e.g., text incoherence). Disruptions are expected to provide mnemonic cues to the learner that text has not been effectively understood, and thus, judgments should be lowered (Dunlosky & Thiede, 1998). On the first reading, decoding information at the lexical and text base levels are predicted to consume readers' attention and provide judgment cues.

Because accurate assessments of comprehension rely on the situation model (for a review see Wiley, Griffin, & Thiede, 2005), cues about lexical and text-based processing are expected to mislead the learner. On the other hand, during the second reading, lexical and text based processing should be more fluent. Therefore, disruptions should take place at the level of the situation model to provide more valid cues about comprehension. Dunlosky (2005) tested this hypothesis by asking participants to delay rereading by one week. The delay was expected to increase cognitive processing because elements of the text base were forgotten from the first reading, thus making it harder to build a complete situation model which, in turn, would provide more invalid cues about comprehension. Data supported the levels-of-disruption hypothesis.

An assumption of the levels-of-disruptions hypothesis is that rereading increases fluency so that learners' are able to attend to disruptions at the situation model level. Recently, researchers have suggested that rereading frees cognitive resources so that learners are able to attend to disruptions *in the first place*, not necessarily at the situation model level *per se*. A lack of cognitive resources during the first reading may constrain learners' abilities to metacognitively monitor comprehension cues. This interpretation of the rereading effect fits within Nelson and Narnes' (1990) *dual-processing hypothesis*, which states that when fewer resources are required at the cognitive level of processing, more resources can be allocated to the metacognitive level (Chiang, Therriault, & Franks, 2010; Griffin, Wiley, & Thiede, 2008; Thiede, Griffin, Wiley, & Anderson, 2010).

Chiang, Therriault and Franks (2010) recently found that rereading and self-explanation strategies increased overall metacomprehension accuracy. More importantly for the dual-processing hypothesis, they found that students with lower working memory span (WMS) showed greater metacomprehension accuracy. Chiang et al. explained this effect by noting that readers with lower WMS require more cognitive effort to process text at lexical and text base levels. As such, fewer resources are available to monitor text comprehension at the meta-level. In contrast, higher WMS learners (or low WMS learners after a first reading) have additional resources available during reading that can be allocated to comprehension monitoring. Similarly, Griffin, Wiley, & Thiede (2008) found that in contrast to the levels-of-disruption hypothesis rereading did not effectively benefit better readers, who may already have resources available for meta-processing. Rather, it provided struggling readers with the resources needed to metacognitively monitor during reading, and as a result increased overall group accuracy.

In short, implications from the dual-processing hypothesis suggest that, after rereading, lower WMS students are better able to think about their thinking. It is not clear however that participants will choose to monitor cues more predictive about comprehension at the level of the situation model. In support of this assumption, Thiede, Griffin, Wiley, and Anderson (2010) recently reported that only a small minority of participants used gist cues (situation model level information) as their primary source of judgment inference.

2.3.2.3. Metacomprehension and the delayed summary effect.

Ebbinghaus (1913) was the first to report that delayed learning over multiple trials improves memory. Since this time delay has been one of the most robust and potent interventions in the cognitive and educational psychology literatures (Carpenter & DeLosh, 2006; Cepeda, Pashler, Vul, Wixten, & Rohrer, 2006; Cull, 2000; Glover, 1989). Two main theories have been proposed to account for improved recall due to spacing; *encoding variability* (Dempster, 1989) and *elaborative retrieval* (Carpenter & DeLosh, 2006). Described simply, the encoding variability account predicts that retrieval of materials in different contexts increases the variety of contextual cues one can use to retrieve a target. The elaborative retrieval hypothesis suggests that after a delay a greater number associations are made with the recalled information, and therefore the strength between cues and responses is greater.

Considerations of cue type and strength have spilled into the metacognition literature. Nelson and Dunlosky (1991) were the first to illustrate the dramatic effect of delayed recall on judgment accuracy for rotely memorized material. In this study, word-pairs were presented to participants who were required to judge future recall either immediately or after a delay. By separating encoding and retrieval attempts by a delay, Nelson and Dunlosky found that participants had on average exceptionally high gamma correlations of +.90. Nelson and Dunlosky initially explained these results with the *Monitoring-Dual-Memories* (MDM) hypothesis: after a delay, participants make judgments by attempting to recall a paired associate from long-term memory. In contrast, when making immediate judgments, participants use short term memory and often make positively biased judgments due to underestimation of information decay over time (Anderson & Thiede, 2008). Since long term memory is required for recall after a delay, judgments that occur after a delay were more accurate. Spellman and Bjork (1992) questioned this explanation contending that items retrieved after a delay increase memory (not metamemory) due to the elaboration required for delayed recall, and are therefore better recalled (also see Kimball & Metcalfe, 2003). The actual explanation for the delayed judgment effect is contentious, but the intervention is one of the most robust in the metamemory and metacomprehension literatures.

The *delayed summary effect* refers to increased metacomprehension accuracy resulting when participants summarize text after a delay between studying the text and generating a summary of it (Thiede & Anderson, 2003). A similar effect has been found when learners generate keywords at a delay; see Thiede, Griffin, Wiley, & Reford, (2009). The delayed summary effect is hypothesized to focus participants' summaries at the situation model since the text base and lexical levels of understanding have degraded (Kintsch,1994; Thiede, Griffin, Wiley, & Redford, 2009). Arguably, this effect has been the most effective way to enhance relative metacomprehension accuracy to date. This intervention has improved gamma correlations from .25 (for a review see Maki, 1998b) to between .60 - .75 across numerous experiments (Thiede, Griffin, Wiley, Redford, 2009).

Thiede and Anderson (2003) were the first to explain increased metacomprehension accuracy resulting from delayed summarization. Previous studies had used delayed judgment with no positive effect on gamma correlations (Maki, 1998a). As explained by the MDM hypothesis, generating delayed summaries before making judgments requires participants to monitor long term memory using cues from the situation model (referred to as *gist cues)* because these cues are more robust to decay than details (i.e., memory level understanding, Thiede, et al., 2009). Since the texts' situation model and learners' comprehension are both defined by connections between text content (and prior knowledge in the case of comprehension), delayed summarization cues were expected to reflect the extent of learners' comprehension and thus increase metacomprehension accuracy. Thiede and colleagues have labelled this effect the *situation model hypothesis.*

Thiede, Anderson, and Therriault (2003) provided the first and most direct evidence of the benefits of delayed summarization on judgment accuracy of comprehension. In this experiment, a control group read six texts then took a test on each text. The immediate summary group wrote a summary of the text immediately after reading. The delayed summary group read each text in order, and then proceeded to summarize each text in the same order. Results showed that participants in the delayed summary group, in comparison to the no summary and immediate summary groups, chose to restudy texts they comprehended less well, and avoided restudying texts that were better comprehended. As a result, participants in the delayed summary group performed better on the second test. In contrast, the no summary and immediate summary groups improved less, restudied texts they didn't need to restudy, and failed to study texts that were poorly comprehended.

An alternative to Thiede and Anderson's (2003) situation model hypothesis is the accessibility hypothesis. Anderson & Thiede (2008) explain that the accessibility hypothesis "states that metacognitive judgments are based on the amount of information accessed from memory" (p. 111). Similar to the levels-of-disruption hypothesis during reading, the accessibility hypothesis assumes that, at a delay, participants will retrieve less information and, therefore, confidence based on amount of recall will more closely approximate recall on the test. Using the definition of accessibility above Anderson & Thiede (2008) tested the accessibility hypothesis by correlating metacomprehension judgments with idea units included in immediate and delayed summaries.

Anderson and Thiede (2008) coded summaries for detail, gist, and total idea units. They conjectured that if comprehension judgments and performance scores were both correlated with total idea units then the amount of information retrieved, or accessibility, could account for increased metacomprehension accuracy. In contrast, if comprehension judgments correlated with gist units after delayed summarization, but detailed units after immediate summarization, the researchers argued there would be support for the situation model hypothesis. Indeed gist units were more highly correlated with comprehension judgments after a delay, and idea units were more highly correlated with comprehension judgments after immediate summaries. These findings provided support for the situation model hypothesis. In addition, correlations between comprehension performance scores and gist units, in both groups, highlighted a potential source of metacomprehension inaccuracy for the immediate summary group.

Koriat (1993) originally presented the accessibility hypothesis as a means by which learners make meta-memory judgments. To some extent Anderson and Thiede's (2008) account of the accessibility hypotheses reflects Koriat's definition in that he noted that participants may be "relying mostly on the amount of relevant information that is recruited" (p. 610) when using mnemonic accessibility cues. However, in contrast to Anderson and Thiede's (2008) definition of accessibility, Koriat (1993) made clear that accessibility has two components "the sheer amount of information accessible and its intensity" (p. 613).

Researchers in the social psychology literature on beliefs have found robust interactions between the amount of information retrieved and the cognitive effort *(intensity)* of retrieval (e.g., Schwarz, 1998, in press; Schwarz, Bless, Strack, Klumpp, Rittenauer-Schatka, & Simons, 1991). In a common experimental paradigm participants are asked to provide a more or a less extensive account of a typical behaviour. For example, Aarts and Dijksterhuis (2000) asked participants to generate either 3 or 8 locations where they had travelled by bicycle in the past month. When participants provide a more extensive account of their behaviour (i.e., more examples), recall becomes more effortful or intense and, as a result, ratings of frequency of behaviour described by the recalled examples decrease. Participants in Aarts and Dijksterhuis' study who generated 3 locations (low intensity recall group) rated their bicycle use higher than the 8 locations (high intensity recall group).

Morris (1990) and Baker and Dunlosky (2006) investigated accessibility viewed as intensity of recall. Morris (1990) asked participants to read texts, and after a 24hr delay recall as many words, concepts, or ideas as they could within 15 seconds. Three measurements were taken: 1) recall latency, the time between the onset of the recall trial and the first utterance; 2) recall production, the number of content words produced; and 3) post access retrieval rate, the number of words produce divided by 15 seconds (recall time limit) subtracted from recall latency. Morris (1990) found that judgments correlated with all three measures and that recall latency and post-access retrieval rate were largely independent. Baker and Dunlosky (2006) replicated and expanded this experiment by including an immediate recall group. Morris' general findings were duplicated and results indicated that the immediate recall group was also influenced by all three accessibility factors. However, judgment variance and accessibility correlation coefficients were far less pronounced in the immediate group. Thus, the saliency of accessibility cues, including intensity of recall, may be more impactful after a delay.

2.3.2.4. Relevance to the current study.

In the current study summary delay is manipulated between participants for each of two texts. As a result, the impact of summarization delay can be investigated both within and between texts. Using regression analyses the relative impact of the amount, type, and latency of Specific Details, Facts, and Connections will be considered. In addition, a new intervention was explored whereby immediately summarized texts were preceded by pre-reading graphic organizers, and texts summarized after a delay were preceded by presentation of facts. The intent of this additional intervention was to test the anchoring and adjustment judgment hypotheses. Specifically, by focusing the reader at the text base (facts) or situation model (graphic organizer) judgments may reflect the anchor rather than (or in addition to) the level of processing predicted by the situation model hypothesis.

A second deviation from traditional metacomprehension research in this study was the requirement for participants to make judgments for comprehension differentiated from judgments for their memory for facts. In the currently available literature it is unclear if learners are able to select judgment cues relevant to a given purpose. More specifically, it is untested if learners' are able to make comprehension judgments based on situation model level recall, and judgments based on rote memory recall for the same text. Moreover, the influence of delay of summarizing on level of judgment has also been previously untested.

Finally, a hybrid of the accessibility and situation model hypotheses was tested in the current study. I reason that, as with findings in the social psychology literature, as more information is recalled intensity of recall will increase; and, as a result, judgments of comprehension decrease. This should result in an interaction where the speed of recall x the amount of information (in this study the number of Specific Details, Facts, or Connections) is negatively related to judgment magnitude.

In the current study effects of delayed summarization cues were investigated by coding summaries into three levels of response: *Specific Details, Facts, and Connections.* Each of these levels were referred to generically as a *summary characteristic*. Meaningful segments of text were categorized as Facts if they were disconnected with other summary segments, or as Connections if they were either implicitly or explicitly connected to another segment(s). After all Facts and Connections were counted a separate count was made of Specific Details referring to names, dates, or measurements. Finally, all summary characteristics that were relevant to the reading materials were counted regardless of their accuracy. Since all characteristics were expected to be valid to participants, they were expected to potentially influence judgments. These three summary levels were analyzed in terms of their associations with experimental interventions, and their differentiated role within each text.

2.4. Conclusion

The complexity and diversity of methods used to conduct and evaluate experiments of metacomprehension accuracy has led to a large diversity of experimentation. However, there is agreement in the literature that metacomprehension accuracy is influenced by individual differences and elaboration experiences. Moreover, texts and assessments must focus on the situation model to truly measure metacomprehension as opposed to metamemory.

In the current study, metacomprehension accuracy is evaluated separately for each text. As such, the relationship between experimental interventions and text characteristics can be evaluated. Moreover, a new method of assessing comprehension is employed. Finally, participants are asked to evaluate their memory and comprehension of a single text after immediate or delayed summarization. This affords the opportunity for measuring judgment cues used for different texts at different levels of cognition, under different experimental conditions. As well, interactions between text, judgment level, and experimental condition can be evaluated.

Finally, in the current study the delayed summary intervention is evaluated based on metacomprehension as measured by absolute accuracy. The delayed summary literature prior to this study has been based on metacomprehension as measured by relative accuracy. The current study evaluates the transferability of findings about delayed summarization when metacomprehension is indexed by absolute accuracy.

3. Methods

3.1. Participants

Participants (n=116) in this study were members of the Simon Fraser University (SFU) community. Their ages ranged from 18 to 49 (M = 24.1 SD = 6.16). Forty-three percent of participants were male, and participants identified highest level of education and/or degree in progress as undergraduate (84%), master's (5.2%), doctorate (1.7%), or "other" (8.6%). Participants were enrolled in various degree programs including Arts (37.9%), Education (18.1%), Science (24.1%), Health Science (6.9%), Business Administration (5.2%), and "other" (7.7%). In accordance with the diverse cultural representation at SFU, 55.2% of participants rated English as their second language. This reflects SFU's high English language proficiency standards for acceptance. A large majority (88.7%) of participants reported their grades averaged between "A+ and B-", with only 2.6% of participants reporting grades below a "C". Participants rated their understanding of university text as "Excellent" (37.1%), "Very Good" (34.5%), "Good" (19.8%), or "Fair" (8.6%).

3.2. Treatments

All participants read two texts: "Nuclear dumping in Russian Lake" (NDRL) and "Wind Blown Transportation" (WBT). For each text participants wrote a summary, answered 10 detail questions and 1 explanation question. Differences between the experiences of participants in the four groups in this study are shown in Table 1.

Table 1: Experimental Procedure by Group

Delayed NDRL Summary	Delayed WBT Summary	Prompted Delayed NDRL Summary	Prompted Delayed WBT Summary
Read NDRL	Read WBT	Study NDRL Facts	Study WBT Facts
Read WBT	Read NDRL	Read NDRL	Read WBT
Summarize WBT	Summarize NDRL	Study WBT Graphic Organizer	Study NDRL Graphic Organizer
Summarize NDRL	Summarize WBT	Read WBT	Read NDRL
Confidence Judgment	Confidence Judgment	Summarize WBT	Summarize NDRL
Detailed NDRL Questions	Detailed WBT Questions	Summarize NDRL	Summarize WBT
Explanation NDRL Questions	Explanation WBT Questions	Confidence Judgment	Confidence Judgment
Detailed WBT Questions	Detailed NDRL Questions	Detailed NDRL Questions	Detailed WBT Questions
Explanation WBT Questions	Explanation NDRL Questions	Explanation NDRL Questions	Explanation WBT Questions
		Detailed WBT Questions	Detailed NDRL Questions
		Explanation WBT Questions	Explanation NDRL Questions

3.3. Assignment to Treatments

Participants were assigned to one of the four previously discussed groups. As participants entered the computer laboratory they were presented with four square pieces of paper placed on a table. Each piece of paper contained a number (1-4) on the concealed side. After the first participant had selected a piece of paper it was removed from the pile. The next participant therefore had 3 pieces of paper from which to choose. This continued until all four pieces of paper had been selected. Once all groups were chosen the next participant once again was able to choose from all four pieces of paper.

Unfortunately, it became clear that 4 participants cheated by returning to the readings when answering questions. Their data were eliminated. In addition, 4 participants were removed due to failure to participate in all experimental requirements. Some of these difficulties were caught in the lab. On these occasions participants were excused and new participants were assigned specifically to groups in an attempt to create equal group numbers. Other difficulties were caught *post hoc* and therefore group samples are not exactly equal.

3.4. Materials

The two texts in this study were adapted with permission from Alan Bellows of the website damninteresting.com (see Appendices A, B, and C). Texts were chronologically sequential expository explanations involving people and events. They were chosen for their complex and integrated explanations and descriptions of obscure historical situations. Both readings required inferences to be made within and between characters, geographic locations, and socio-political motives. As such, comprehension required participants to move beyond understanding at the level of the text base to process texts at the situation model level. Texts were specifically chosen to meet the requirements of "higher order" comprehension (as described by Wiley, Griffin, & Thiede, 2005). Passages focused on topics outside the world purview of most people and, in this way, were unlikely to be biased by extensive prior knowledge. To test this assumption, a list of 20 relatively obscure reading topics, including the experimental texts, were provided and participants were asked to "Rate [their] knowledge of each topic as accurately as possible" on a scale from "1 (none) to 5 (Very Strong)".

Texts were edited from their original in an attempt to ensure structural homogeneity. Moreover, texts were edited to equalize the number of "main events" described in each passage. Both WBT and NDRL had 16 main events described with approximately equal detail. Both texts had Flesch-Kincaid Grade levels of 12 (see Table 2). Although nearly identical in a structural sense, the texts were expected to differ in regards to reader interest. The subject matter of NDRL was expected to draw more interest given historical and modern day concerns about nuclear disaster. Moreover, NDRL chronicled the inhumane suffering of people at the hands of a dictatorial government. This was expected to create empathy in participants and therefore promote interest. Although some students may have found the historic and rather technical description of pneumatic transportation interesting in WBT, it was not expected to engage the participants on an equally emotional level. Using an identical rating methodology as for assessing topic knowledge, participants were asked to "Rate [their] interest in each topic as accurately as possible".

Table 2: Readability Statistics for Wind Blown Transportation and Nuclear Dumping in Russian Lake

Reading Statistics	WBT	NDRL
Words	1544	1397
Characters	8125	7740
Paragraphs	14	12
Sentences	61	63
Sentences per Paragraph	5.1	5.7
Words per Sentence	25	22.1
Characters per Word	5.2	5.4
Passive Sentences	27%	30%
Flesch Reading Ease	37.6	25.3
Flesch-Kincaid Grade Level	12	12

3.5. Participant Consent

All participants were given the opportunity to provide (and retract) consent to participati in this study in accordance to Simon Fraser University's ethical guidelines (see Appendix D). Although specific experimental intentions were not disclosed to participants, main requirements were clearly explained.

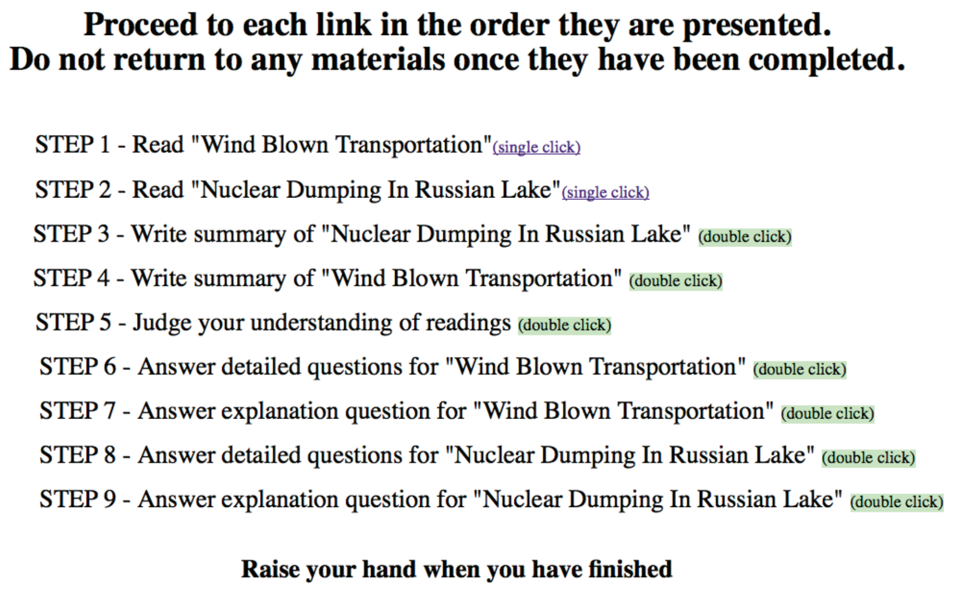
3.6. Pre-Study Questionnaire

The pre-study questionnaire had four key purposes. First, it was used to determine participants' demographics. Factors including gender, age, and faculty enrolment were recorded. Second, indicators of reading comprehension and GPA were collected by asking participants to report their judgement of reading ability and their GPA. Third, as previously described participants' interest and knowledge of experimental texts were examined. Finally, *The Efficient Assessment of Need for Cognition* was administered with permission (see Appendix E) to evaluate participants' propensity to enjoy effortful consideration of material they study (Cacioppo, Petty, and Kao, 1984).

3.7. nStudy

Data for this study were collected on nStudy software. nStudy records fine grained time-stamped behavioural data recorded as participants interact with the user interface. In this study, summaries and questions were answered in editable text fields called "Notes". Texts, *graphic organizers*, and *fact lists* were presented in the nStudy browser. Participants accessed the browser by single clicking hyperlinks, and returned to the experiment instruction homepage by clicking a second link (see Figure 3).

Figure 3: Experimental nStudy Homepage



Proceed to each link in the order they are presented.
Do not return to any materials once they have been completed.

STEP 1 - Read "Wind Blown Transportation"(single click)

STEP 2 - Read "Nuclear Dumping In Russian Lake"(single click)

STEP 3 - Write summary of "Nuclear Dumping In Russian Lake" (double click)

STEP 4 - Write summary of "Wind Blown Transportation" (double click)

STEP 5 - Judge your understanding of readings (double click)

STEP 6 - Answer detailed questions for "Wind Blown Transportation" (double click)

STEP 7 - Answer explanation question for "Wind Blown Transportation" (double click)

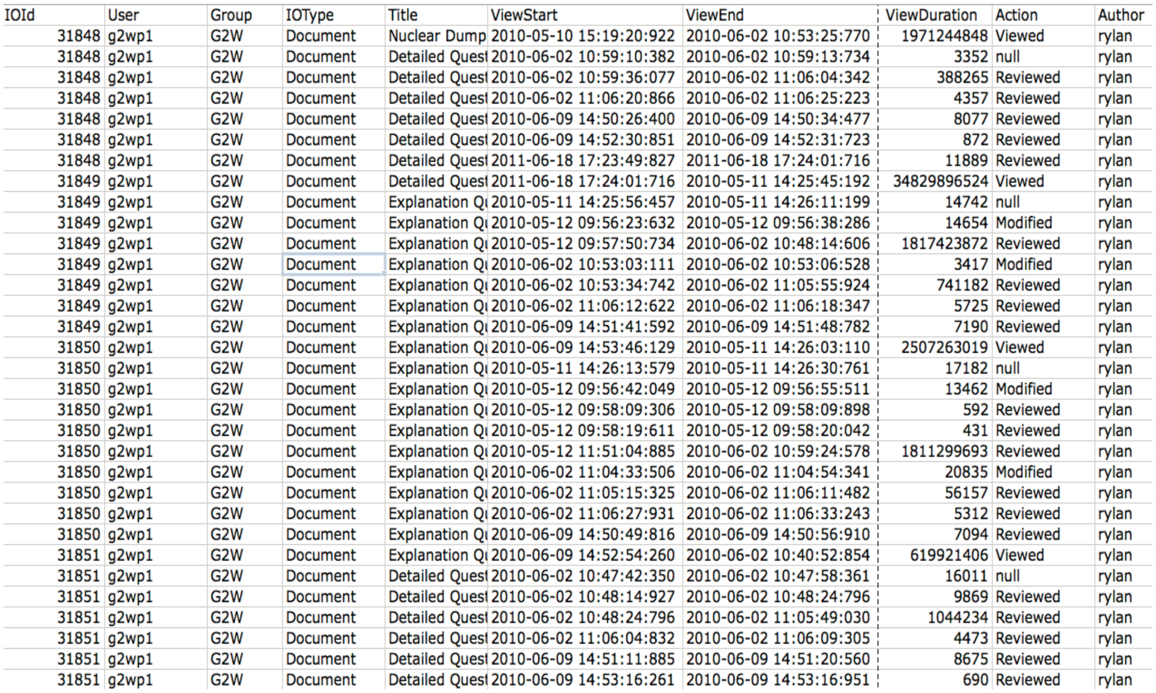
STEP 8 - Answer detailed questions for "Nuclear Dumping In Russian Lake" (double click)

STEP 9 - Answer explanation question for "Nuclear Dumping In Russian Lake" (double click)

Raise your hand when you have finished

Using nStudy allowed me to review participants' actions as they progressed through stages in the study. Specifically, instances of cheating, skipping experimental procedures, and *clicking through* links without adequate attention were caught and removed from data set. In addition, summary writing, reading, and judgment times could be precisely calculated as a result of precise time stamps in nStudy's log files (see Figure 4).

Figure 4: Sample nStudy Log File



IOId	User	Group	IOType	Title	ViewStart	ViewEnd	ViewDuration	Action	Author
31848	g2wp1	G2W	Document	Nuclear Dump	2010-05-10 15:19:20:922	2010-06-02 10:53:25:770	1971244848	Viewed	rylan
31848	g2wp1	G2W	Document	Detailed Quest	2010-06-02 10:59:10:382	2010-06-02 10:59:13:734	3352	null	rylan
31848	g2wp1	G2W	Document	Detailed Quest	2010-06-02 10:59:36:077	2010-06-02 11:06:04:342	388265	Reviewed	rylan
31848	g2wp1	G2W	Document	Detailed Quest	2010-06-02 11:06:20:866	2010-06-02 11:06:25:223	4357	Reviewed	rylan
31848	g2wp1	G2W	Document	Detailed Quest	2010-06-09 14:50:26:400	2010-06-09 14:50:34:477	8077	Reviewed	rylan
31848	g2wp1	G2W	Document	Detailed Quest	2010-06-09 14:52:30:851	2010-06-09 14:52:31:723	872	Reviewed	rylan
31848	g2wp1	G2W	Document	Detailed Quest	2011-06-18 17:23:49:827	2011-06-18 17:24:01:716	11889	Reviewed	rylan
31849	g2wp1	G2W	Document	Detailed Quest	2011-06-18 17:24:01:716	2010-05-11 14:25:45:192	34829896524	Viewed	rylan
31849	g2wp1	G2W	Document	Explanation Qu	2010-05-11 14:25:56:457	2010-05-11 14:26:11:199	14742	null	rylan
31849	g2wp1	G2W	Document	Explanation Qu	2010-05-12 09:56:23:632	2010-05-12 09:56:38:286	14654	Modified	rylan
31849	g2wp1	G2W	Document	Explanation Qu	2010-05-12 09:57:50:734	2010-06-02 10:48:14:606	1817423872	Reviewed	rylan
31849	g2wp1	G2W	Document	Explanation Qu	2010-06-02 10:53:03:111	2010-06-02 10:53:06:528	3417	Modified	rylan
31849	g2wp1	G2W	Document	Explanation Qu	2010-06-02 10:53:34:742	2010-06-02 11:05:55:924	741182	Reviewed	rylan
31849	g2wp1	G2W	Document	Explanation Qu	2010-06-02 11:06:12:622	2010-06-02 11:06:18:347	5725	Reviewed	rylan
31849	g2wp1	G2W	Document	Explanation Qu	2010-06-09 14:51:41:592	2010-06-09 14:51:48:782	7190	Reviewed	rylan
31850	g2wp1	G2W	Document	Explanation Qu	2010-06-09 14:53:46:129	2010-05-11 14:26:03:110	2507263019	Viewed	rylan
31850	g2wp1	G2W	Document	Explanation Qu	2010-05-11 14:26:13:579	2010-05-11 14:26:30:761	17182	null	rylan
31850	g2wp1	G2W	Document	Explanation Qu	2010-05-12 09:56:42:049	2010-05-12 09:56:55:511	13462	Modified	rylan
31850	g2wp1	G2W	Document	Explanation Qu	2010-05-12 09:58:09:306	2010-05-12 09:58:09:898	592	Reviewed	rylan
31850	g2wp1	G2W	Document	Explanation Qu	2010-05-12 09:58:19:611	2010-05-12 09:58:20:042	431	Reviewed	rylan
31850	g2wp1	G2W	Document	Explanation Qu	2010-05-12 11:51:04:885	2010-06-02 10:59:24:578	1811299693	Reviewed	rylan
31850	g2wp1	G2W	Document	Explanation Qu	2010-06-02 11:04:33:506	2010-06-02 11:04:54:341	20835	Modified	rylan
31850	g2wp1	G2W	Document	Explanation Qu	2010-06-02 11:05:15:325	2010-06-02 11:06:11:482	56157	Reviewed	rylan
31850	g2wp1	G2W	Document	Explanation Qu	2010-06-02 11:06:27:931	2010-06-02 11:06:33:243	5312	Reviewed	rylan
31850	g2wp1	G2W	Document	Explanation Qu	2010-06-09 14:50:49:816	2010-06-09 14:50:56:910	7094	Reviewed	rylan
31851	g2wp1	G2W	Document	Explanation Qu	2010-06-09 14:52:54:260	2010-06-02 10:40:52:854	619921406	Viewed	rylan
31851	g2wp1	G2W	Document	Detailed Quest	2010-06-02 10:47:42:350	2010-06-02 10:47:58:361	16011	null	rylan
31851	g2wp1	G2W	Document	Detailed Quest	2010-06-02 10:48:14:927	2010-06-02 10:48:24:796	9869	Reviewed	rylan
31851	g2wp1	G2W	Document	Detailed Quest	2010-06-02 10:48:24:796	2010-06-02 11:05:49:030	1044234	Reviewed	rylan
31851	g2wp1	G2W	Document	Detailed Quest	2010-06-02 11:06:04:832	2010-06-02 11:06:09:305	4473	Reviewed	rylan
31851	g2wp1	G2W	Document	Detailed Quest	2010-06-09 14:51:11:885	2010-06-09 14:51:20:560	8675	Reviewed	rylan
31851	g2wp1	G2W	Document	Detailed Quest	2010-06-09 14:53:16:261	2010-06-09 14:53:16:951	690	Reviewed	rylan

3.8. Detail Questions

Detail and explanation questions were used to assess comprehension of both texts in all groups. Detail questions tested participants' recall of details at the level of the text base (see Appendices F and G). An effort was made to ensure that questions targeted details that were relevant to text comprehension and evenly distributed across both texts. Item's did not require inference at the level of the situation model. Reliability was approaching acceptable for both NDRL (α = .67) and WBT (α = .69).

3.9. Explanation Questions

Explanation questions were asked to reveal participants' understanding of text events and their interrelations at the level of the situation model. Specifically, participants were asked to *Explain all major events in the history of pneumatics from its* *invention to the present day.* After reading NDRL, participants were asked to *Explain each major nuclear disaster, its causes, its effect, and the Russian reaction as described in "Nuclear Dumping in Russian Lake".* In both cases, participants were reminded to *Make sure you describe events in the order they occurred from oldest to most recent*. As previously described, texts had been edited to have an equal number of main story events. Responses were given 1 point for every main event implicated in the situation model of the text. If participants' answers were described out of order, .5 points were rewarded. Main points were coded by the investigator and a graduate student collaborator. Spearman's rho, a non-parametric correlation was calculated for an inter-rater reliability of .86. Discrepancies were resolved through discussion.

3.10. Graphic Organizers

Graphic organizers were developed for both WBT and NDRL texts (see Appendices H and I). Brief descriptions, vivid images, and directional lines were used to describe the progression of events in the text. The graphic organizers were not intended to identify all 16 main points used to evaluate explanation question responses but to facilitate organizing information of the text. Graphic organizers have a mixed record as a substitute for studying text, however a recent meta-analysis indicates that graphic organizers can "assist[s] in recall of both central ideas and detail ideas, but the effect may be stronger for central ideas" (Nesbit & Adesope, 2006, p. 434, also see Kiewra, Mayer, & Dubois, 1996).

This advanced description of text organization, was expected to promote higher level coordination of text content that would support both text processing and summary writing. As such, participants were expected to judge their ability to *discuss the full meaning and implications* from the text more favourably. If the graphic organizers served this function they may also counteract and potentially mitigate reliance on details expected to result from immediate summarization (Anderson & Thiede, 2008).

3.11. Facts

For each text 15 detailed questions were developed. Ten of these questions were used for the detail tests, information corresponding to the remaining 5 questions was presented in the list of facts. The intent of providing the list of 5 facts was to increase participants' recognition and potential recall of text level information (see Appendices J and K) by prompting readers to concentrate on the text based and/or lexical level(s) during reading and summary writing. As a result, participants may judge future recall of detailed information more favourably. If facts served this function they may counteract (or supplement) the tendency for participants to concentrate on situation model level cues at recall (Anderson & Thiede, 2008).

3.12. Detailed Judgments

Participants' judgments of memory of specific text facts and details were prompted on a web page by two questions that read: *On a scale from 1-10 how well do you think you will remember specific facts from "Nuclear Dumping in Russian Lake"/"Wind Blown Transportation"*. A slider was provided with the toggle located at the far left, and a 0 in the adjacent display window. As the slider was moved from left to right numbers increased from 0-10 in equal intervals. Judgments were made prior to testing, but subsequent to all experimental interventions including summary writing, reading delays, and pre-reading prompts. It should be noted that time intervals between reading, interventions, and judgment varied (see Table 1). However, these were not expected to influence judgments. Thiede, Dunlosky, Griffin and Wiley (2005) altered lags between judgments, and readings and summaries and found no statistically detectable influence on judgments.

3.13. Explanation Judgments

Judgments about comprehension were made using an identical scale system by responding to the question, "*On a scale from 0-10 how well do you think you will be able to discuss the full meaning and implications of "Nuclear Dumping in Russian Lake"/Wind Blown Transportation"?*

3.14. Study Environment

The experiment was carried out in Dr. Philip H. Winne's lab in the Education building at Simon Fraser University. The lab consists of a large room with 8 computers oriented in a semi-circle around the parameter of the room. Each participant worked at a personal computer facing the wall. In the centre of the room there is a large table used to orient new participants.

Because two experiments were conducted in tandem and were facilitated by the author or a colleague who was a fellow PhD student, my colleague and I took turns recruiting students and facilitating experiments. Participants were offered the opportunity to choose either research study, or if they desired they were allowed to complete both studies. It was made clear to participants choosing both studies that a one hour break was required between experiments. This measure was taken in an attempt to limit participant fatigue.

The term *investigator* will be used to describe the individual who facilitated the experiment at a given time. My PhD student colleague was fully trained by me to carry out tasks required for this study.

All efforts were made to ensure participants in the same group did not sit together to limit any opportunity for collaboration. Although both experiments focused on learning, the associated texts, methods, and questions were dissimilar and should not have influenced each other.

3.15. Log Data

nStudy allows experimenters to retrieve fine grained behavioural data after learners study the materials presented in the nStudy environment. A vast array of data is available (e.g., reading time, typed text, Likert scale ratings, highlighting activity, and glossary items created). For the present experiment, data included writing time, word counts, typed text, and responses to explanation questions and details questions. In addition, time stamps for all hyperlink clicking actions were recorded. Using this data, the precise time when links were opened and closed could be obtained. As such, the reading time for experimental texts was recorded to the millisecond. Finally, by tracking link clicking it became clear when participants failed to observe the required experimental protocol.

3.16. Summary Analysis

Participants summaries of the texts were analyzed to assign three types of scores. Points for *Specific Details* corresponded to dates, names, and measurements of time, weight, percentage etc. Points for *Facts* referred to isolated information provided without a plausible connection to other aspects of the text. Finally, points for *Connections* were assigned to Facts that were causally related to one another. Connections were identified whether derived from the text or otherwise, e.g., from prior knowledge.

Accuracy of Specific Details, Facts or Connections was not a requirement to receive points. Since participants are often unable to determine the validity of summary assertions (Dunlosky, Hartwig, Rawson, & Lipko, 2011), all assertions were assumed to be *true for the participant.* As such, points for correct and incorrect content were assumed to have equal potency in contributing to a participant's judgment of comprehension. A minority of Facts and causal assertions were a) not related to the text, b) repeated, or c) non-comprehensible. In these cases the fact or connection was not counted. The author trained a PhD student in education to rate summaries from 20% of participants. Spearman's Rho was calculated to determine inter-rater reliability of the summary characterstics measure. Results showed acceptable reliability for Facts (.78), Connections (.88), and Specific Details (.93). Discrepancies were resolved through discussion.

3.17. Pilot Study

Pilot studies were conducted with 3 graduate students, one undergraduate student, 2 professionals, and 2 adult participants who had not completed university. Verbal and written ratings of prior topic knowledge was low for both texts. Prior topic interest of NDRL was rated higher except for one mechanically inclined individual who rated WBT higher. The pilot study indicated that completion time would be around 1 hour. Initially, graduate student colleagues reported that the graphic organizers were somewhat difficult to follow. Revisions were made and presented to these students until they confirmed that graphic organizers were understandable. A subsequent convenience sample of students working in the lab confirmed these opinions.

3.18. Procedure

3.18.1. Recruitment.

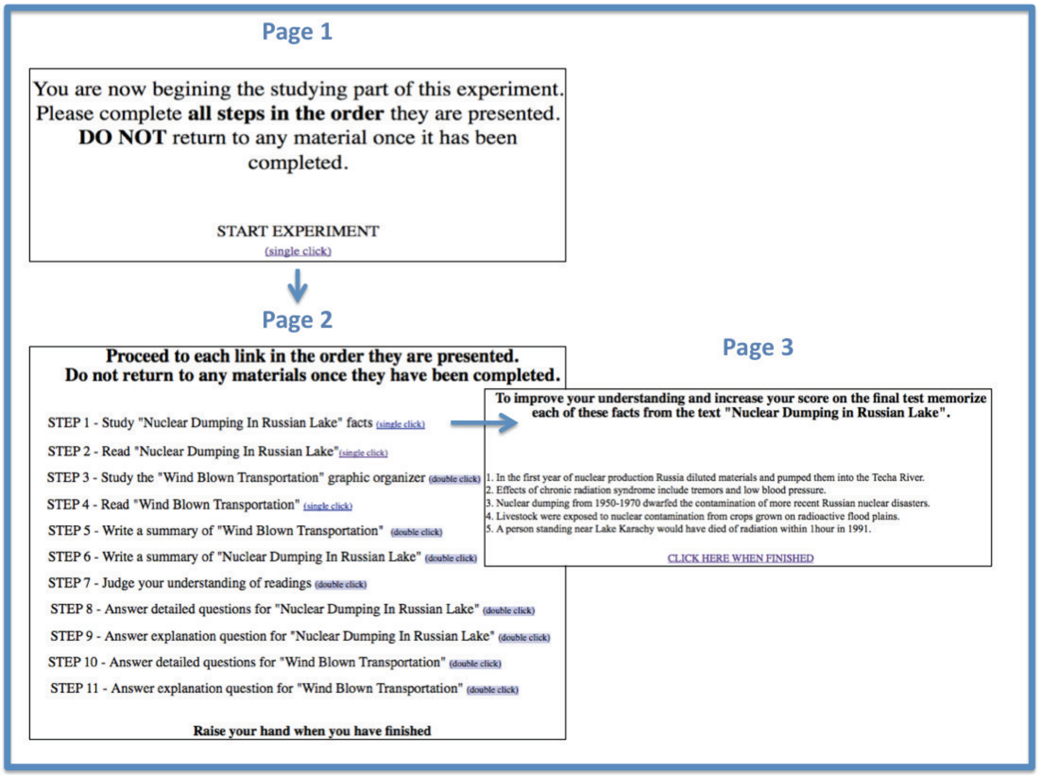
Participants were recruited from the SFU community. The first attempt at recruitment consisted of making announcements in summer session 2010 courses. In addition, posters were displayed in high traffic areas to introduce the study and ask potential participants to contact the author via an email address created for the purpose. Participants were compensated $20 for their time with the promise of a $35 reward for the highest score. Given the sub-optimal success of these two strategies, a more direct approach was adopted. A desk was placed in a busy thoroughfare near the laboratory. The thoroughfare connects multiple buildings and was therefore expected to contain students from diverse disciplines. My colleague and I took turns, one running experiments while the other recruited new participants.

3.18.2. Participant orientation.

Experimental instructions were provided verbally and in writing. Participants were first asked to read and, upon agreement, to sign the consent form for the study. Participants were informed they would be required to *study, read, and summarize two separate texts.* They were also informed that they would be required to judge their understanding of each text and take a test of memory and understanding. Finally, it was made clear that demographic information would be recorded along with various measures of scholastic ability, data regarding nourishment and sleep, and inclination to think deeply.

After being briefed about the experiment's requirements, participants were logged into a previously created nStudy account. An investigator oriented participants to the nStudy interface. The interface provided clear step-by-step procedural instructions (see Figure 5). An investigator made it clear that each step should be completed in the sequence indicated. Participants were told that a single click was sufficient to open hyperlinks but to open "highlights" they would need to double click. For clarity, this difference was also explicitly written on the instruction homepage. Most importantly, an investigator made clear to each participant that each "page" was to be visited and completed only once. Going back to a previous page would render the experiment unusable and, if backtracking was observed, the experiment would be terminated. Participants were told that there was no time limit and that they should put up their hand if they had any questions.

Figure 5: Sequence of Experimental Instructions



You are now begining the studying part of this experiment. Please complete **all steps in the order** they are presented. **DO NOT** return to any material once it has been completed.

START EXPERIMENT
(single click)

Proceed to each link in the order they are presented.
Do not return to any materials once they have been completed.

STEP 1 - Study "Nuclear Dumping In Russian Lake" facts (single click)
STEP 2 - Read "Nuclear Dumping In Russian Lake"(single click)
STEP 3 - Study the "Wind Blown Transportation" graphic organizer (double click)
STEP 4 - Read "Wind Blown Transportation" (single click)
STEP 5 - Write a summary of "Wind Blown Transportation" (double click)
STEP 6 - Write a summary of "Nuclear Dumping In Russian Lake" (double click)
STEP 7 - Judge your understanding of readings (double click)
STEP 8 - Answer detailed questions for "Nuclear Dumping In Russian Lake" (double click)
STEP 9 - Answer explanation question for "Nuclear Dumping In Russian Lake" (double click)
STEP 10 - Answer detailed questions for "Wind Blown Transportation" (double click)
STEP 11 - Answer explanation question for "Wind Blown Transportation" (double click)

Raise your hand when you have finished

To improve your understanding and increase your score on the final test memorize each of these facts from the text "Nuclear Dumping in Russian Lake".

1. In the first year of nuclear production Russia diluted materials and pumped them into the Techa River.
2. Effects of chronic radiation syndrome include tremors and low blood pressure.
3. Nuclear dumping from 1950-1970 dwarfed the contamination of more recent Russian nuclear disasters.
4. Livestock were exposed to nuclear contamination from crops grown on radioactive flood plains.
5. A person standing near Lake Karachy would have died of radiation within 1hour in 1991.

CLICK HERE WHEN FINISHED

On a few occasions computer malfunction required moving participants to different computers where they resumed activities at the point of the malfunction. Most participants easily navigated the interface and did not require assistance. One participant was caught with two windows open and answering questions using text from the open window. This participant was thanked for his time, paid, asked to leave, and his data was removed from the study. On two occasions participants required the use of the washroom and were permitted to leave, but were asked not to access any materials external to the experiment.

4. Results

4.1. Overview

This study investigates five specific questions. First, the effect of delaying summary writing and pre-reading prompts on judgments, assessment scores, and metacomprehension accuracy was evaluated. Since each group was subjected to a unique experimental program (see Table 1), a multivariate analysis of variance (MANOVA) was conducted with groups as independent variables and measures as outcome variables. Before this analysis could be conducted each individual difference measure was evaluated for group homogeneity. If groups were heterogeneous it would be unclear if experimental effects resulted from the influence of individual differences or experimental interventions *per se*.

Second, repeated measures analyses of variance (ANOVA) were conducted to investigate differences in judgments, assessment scores, and metacomprehension accuracy between texts.

Third, the influence of individual differences on judgments were examined. Correlations between judgments and individual differences were examined to identify associations. Next, canonical correlation analyses were used to identify specific judgment variance accounted for by individual differences.

Fourth, the influence of experimental intervention on the number of summary characteristics (i.e., Specific Details, Facts, and Connections) was examined. As illustrated in Chapter 2, findings in the literature indicate that the number of details and gist idea units (analogous to Specific Details/Facts and Connections in this study) are greater in immediate summaries. To test this finding, a MANOVA was conducted with groups as independent variables and summary characteristics as outcome variables. Fifth, the effects of text on the number of summary characteristics was tested using repeated measures ANOVA.

Finally, associations between judgments and summary characteristics as a function of experimental texts and groups were examined. Correlation analyses was used to explore independent shared variance between individual summary characteristics and judgments. Next, regression analyses was used to parse out unique shared variance between judgments and summary characteristics both within and between groups. A total summary characteristics x summary time interaction term was introduced into the regression analyses to test for associations between cognitive intensity and judgment magnitude as described in section 2.3.2.3.

4.2. Individual, Group and Text Differences

To investigate the relative influence of delayed summarization and pre-reading prompts on metacomprehension accuracy, individual differences needed to be homogeneous between groups. The following set of analyses tested for group differences in ratings of prior topic knowledge and interest, NFC, GPA, and ability to read university text. In addition to comparing NFC between groups, a principle components analysis was used to examine NFC. A profile analysis was used to analyze group and text differences in prior topic knowledge and interest. This analysis tested both for group rating differences and the methodological expectation that NDRL would be more interesting.

4.2.1. *Group and text differences in prior topic knowledge and prior topic interest.*

A profile analysis was computed with text (2 levels) and topic perception (interest and knowledge, 2 levels) as within-subject variables, and group as a between subject factor. Main effects for both Text F(1, 109) = 43.30, $p < .001$, $\eta = .28$, $\beta = 1.0$ and perceived topic knowledge and interest F(1, 109) = 195.59, $p < .001$, $\eta = .64$, $\beta = 1.0$ were statistically detectable. Specifically, participants rated their interest (M = 2.8, SE = .08) statistically detectably higher than their knowledge (M = 1.6, SE = .06), and NDRL (M = 2.5, SE = 1.92) statistically detectably higher than WBT (M = 1.92, SE = .06) on combined ratings of topic knowledge and prior interest. No group or interaction effect were statistically detectable (p > .7). However, since differences in interest ratings between text was of particular relevance to this study the estimated marginal means were computed. A statistically detectable difference was found between interest ratings in NDRL (M = 3.04, SE = .10) and WBT (M = 2.51, SE = .10) F(1, 109) = 21.70, $p < .001$, $\eta = .17$, $\beta = 1.0$. Similarly, statically detectable differences were found for knowledge ratings in NDRL (M = 1.91, SE = .09) and WBT (M = 1.34, SE = .06) F(1, 109) = 44.07, $p < .001$, $\eta = .28$, $\beta = 1.0$.

4.2.2. *Group difference in Need-For-Cognition.*

After observing that the efficient measure of NFC had acceptable reliability in my sample ($\propto = .865$), a principle components analysis (PCA) was calculated without rotation, as Cacioppo, Petty, and Kao (1984) did. The Kaiser-Meyer-Olkin measure of sampling adequacy was adequate, KMO = .85 (Field, 2009). In addition, correlations among items were sufficient for analysis according to Bartlett's test of Sphericity χ^2 (136) = 609.83, $p < .001$. Three components had eigenvalues greater than 1, which is often used as a requisite for extraction (Field, 2009). In this analysis only the first factor was retained for analysis. Cacioppo, Petty, and Kao (1984) argue for four constraints that must be satisfied if a single factor is to be extracted from a truncated psychometric measure, and used to represent a construct. First, the first factor (in a shortened inventory) should account for a similar amount of overall variance as the more extensive instrument from which it is derived. The variance accounted for by a first factor in the long form (34-item) NFC instrument was 27% (Cacioppo & Petty, 1982). In the efficient version tested by Cacioppo, Petty, and Kao (1984) a comparatively large 37% of variance was found. In the current study, 33.5% of overall variance was accounted for by the first factor.

Second, "subsequent factors [should] explain fairly equal (though, of course, decreasing) proportions of the remaining variance" (Cacioppo, Petty, & Kao, 1984, p. 3). In the current study, factors 2-17 explained between 9.1% - 1.6% of total variance. The range between factors with eigenvalues greater than 1 was 9.1% - 6.1%. Third, most items should have substantial loadings on the first factor. In the current study, variables' loadings on the first factor ranged from .44 - .71 for all but two items. The loading for *I would prefer complex to simple problems* and *I usually end up deliberating about issues even when they do not affect me personally*, loaded on the first factor at .28 and .31, respectively. This closely mirrors findings of Cacioppo, Petty, and Kao (1984). Lastly, variables should have higher factor loadings on the first factor than subsequent factors. In the current study, as was the case in Cacioppo, Petty, and Kao (1994), only one item had a higher loading on the second component. For these reasons, and because the NFC has almost exclusively been studied using only the first extracted factor (Cacioppo, Petty, Feinstein, & Jarvis, 1996), only the first factor (calculated as all the items weighted by loadings) was retained and used as a predictor in the current study.

To test for group differences in NFC an ANOVA was conducted with groups as independent variables and the sum of raw item scores identified by the first NFC factor as the outcome variable. Main effects showed no statistically detectable NFC differences between groups F(3, 110) = .18, *p* = .91. Unfortunately, a post experimental review found that questionnaire item number 12 which stated "Learning new ways to think doesn't excite me very much" (p. 307), was mistakenly omitted from the questionnaire. Given the statistical reliability of the collected data set, and the replication of Cacioppo, Petty, and Kao's (1984) results, this error was not expected to negatively impact the usefulness of the NFC score.

4.2.3. Group differences in Grade Point Average and perceived ability to read university level text.

To test for *a priori* group differences between GPA and perceived ability to read university text, a MANOVA was computed with perceived GPA and ability to read university text as outcome variables, and group as the independent variable. No statistically detectable main effect for group was found F(6, 220) = 1.39, *p* < .22, η = .04, β = .537.

4.3. Group Differences in Judgments, Assessment Scores, and Metacomprehension Accuracy

To examine if experimental interventions influenced judgments, assessment scores and metacomprehension accuracy, a MANVOA was conducted with judgments, assessment scores, and metacomprehenesion accuracy for both texts at the detailed and explanation levels as dependent variables and groups as independent variables.

A 2 (Text) x 2 (Level: explanation, detail) x 3 (Measure: judgment, assessment score, metacomprehension accuracy) MANOVA was conducted. Levene's tests for equality of variance were not statistically detectable for most dependent variables, however statistically detectable difference in variance was found for knowledge judgments for NDRL explanation questions (*p* = .05), and metacomprehension accuracy for NDRL explanation questions (*p* = .04). In addition, Box's test of variance-covariance between groups was significant (*p* < .001). Hotelling's Trace statistic was used to test for statistical significance as it is more robust against between group variance-covariance (Field 2009). No statistically detectable differences were found for the main effect of group F(36, 291) = .964, p = .89, η = .11, β = .89.

4.4. Judgment, Assessment Score, and Metacomprehension Accuracy Differences Between Texts

To examine differences between judgments, assessment scores, and metacomprehension accuracy as a function of text repeated measures ANOVA were conducted. It would have been possible to conduct a larger profile analysis to test for group and text differences in 4.3. However, since within and between group analyses were conducted in the service of different experimental questions, and interactions available for the more extensive test were not required, the choice was made to conduct these analyses separately.

4.4.1. Detailed and explanation judgment differences between text.

A 2 (Text) x 2(Question Type: Detail, Explanation) repeated measures ANOVA was used to investigate text differences for detailed and explanation judgments. Statistically detectable main effects were found for judgments between WBT and NDRL texts F(1,109) = 38.79, *p* <.001, η^2 = .262, β = 1.0, explanation and detailed questions

F(1,109) = 18.26, p < .001, η^2 = .143, β = .99, and the interaction between texts and questions F(1,109) = 6.67, p = .011, η^2 = .058, β = .73.

Simple effects indicated that participants judged their knowledge of NDRL text (M = 5.66, SE = .18) higher than WBT (M = 4.58, SE = .18) text. The simple effects analysis on test question type indicated that participants judged their knowledge of explanation questions (M = 5.50, SE = .19) higher than detailed questions (M = 4.75, SE = .17). The pattern of higher judgments for explanation questions persisted across both texts. Judgments for NDRL explanation (M = 6.14, SD = 2.17) questions exceeded NDRL detailed (M = 5.19, SD = 2.05, p =.003), WBT explanation (M = 4.86, SD = 2.28, p <.001), and WBT detailed (M = 4.30, SD = 1.96, p < .001) questions. Judgments of NDRL detailed questions exceeded WBT detailed (p < .001) questions, and WBT explanation questions (p = .047). Finally, WBT explanation judgments exceeded WBT detailed questions (p< .001).

4.4.2. Detailed and explanation assessment score differences between text.

A 2 (Text) x 2 (Assessment Scores: Detail, Explanation) repeated measures ANOVA was conducted. A statistically detectable main effect was not found between WBT and NDRL text F(1,113) = 1.23, p = .257, η^2 = .011, β = .204. A statistically detectable main effect was found for assessment scores between explanation and detailed questions F(1,113) = 37.95, p < .001, η^2 = .251, β = 1.0, and the interaction between texts and question type F(1,113) = 37.96, p < .001, η^2 = .251, β = 1.0.

Simple effects indicated, in contrast to their expectations, participants scored higher on detailed assessments (M = .33, SE = .02) than explanation assessments (M = .26, SE = .02). There were no statistically detectable differences between NDRL detailed (M = .30, SE = .02) and explanation (M = .30, SE = .021, p = .40) assessment scores. However, there was a statistically detectable difference between WBT detailed (M = .36, SE = .02, *p* < .001) and explanation (M = .22, SE = .02) assessment scores. Moreover, detailed assessment scores were statistically detectably higher in WBT than NDRL (*p* <.001), and explanation scores were statistically detectably higher for NDRL than WBT (*p* = .003).

4.4.3. Detailed and explanation metacomprehension accuracy differences between text.

To explore differences in metacomprehension accuracy between texts, a 2 (Text) x 2 (Test: Explanation, Detailed) repeated measures ANOVA was conducted. Main effects indicated statistically detectable differences between NDRL (M = .70, SE = .02) and WBT (M = .74, SE = .01) texts F(1,109) = 9.59, *p* = .002, η^2 = .08, β = 0.87. In addition, there were statistically detectable differences discovered between metacomprehension accuracy scores for explanation (M = .69, SE = .02) and detailed (M = .76, SE = .01) questions types F(1,109) = 14.65, *p* <.001, η^2 = .12, β = 0.97. No statistically detectable main effect was found for the interaction between texts and levels (detailed, explanation) F(1,109) = 2.24, *p* = .14, η^2 = .02, β = 0.32. However, further analyses of the estimated marginal means showed statistically detectable differences between metacomprehension accuracy for WBT (M = .79, SE = .02) and NDRL (M = .72, SD = .02) detailed questions F(1, 109) = 11.38, *p* = .001*,* η^2 = .10, β = 0.92. Statistically detectable mean differences were also found between NDRL explanation (M = .67, SE = .02) and detailed (M = .72, SE = .02) questions F(1, 109) = 5.66, *p* = .019*,* η^2 = .05, β = 0.65. Similarly, statistically detectable mean differences were found between WBT metacomprehension accuracy scores for detailed (M = .79, SE = .02) and explanation (M = .70, SE = .02) questions F(1, 109) = 15.24, *p* < .001*,* η^2 = .12, β = 0.97.

4.5. Individual Differences, and Judgment Differences Between Texts

To examine the effect of individual differences on judgments between texts, a correlation analyses was conducted. Next, canonical correlation was used to isolate unique influences of individual differences.

4.5.1. Correlations between individual differences, and judgments.

Correlations in Table 3 between individual differences and judgments describe associations between these variables.

Table 3: Summary of Intercorrelations, Means, and Standard Deviations for Individual Differences and as a Function of Text Type

Measure	1	2	3	4	5	6	7	M	SD
1. JExp	—	.52**	.45**	.14	.14	.38**	.16	4.86	2.27
2. JDet	.59**	—	.33**	.12	.18	.14	-.13	4.31	1.96
3. NFC	.32**	.25*	—	.13	.07	.37**	.22*	.006	.99
4. Prior Topic Interest	.26**	.06	.26**	—	.51**	-.06	-.03	2.5	1.06
5. Prior Topic Knowledge	.04	.01	-.09	.21*	—	-.09	-.03	1.34	.62
6. Understanding University Text	.24*	.04	.37**	.12	-.17	—	.17	4.0	.96
7. GPA	.05	-.11	.22*	.18	-.06	.17	—	4.6	.92
M	6.14	5.19	.008	3.03	1.93	.40	4.6	—	
SD	2.17	2.05	1.0	1.08	.92	.96	.92		—

Note: Correlations for Nuclear Dumping in Russian Lake (n = 115) text are presented above the diagonal, and correlations for Wind Blown Transportation (n = 115) are presented below the diagonal. Means and standard deviations for Nuclear Dumping in Russian Lake text are presented in vertical columns, and means and standard deviations for Wind Blown Transportation are presented in the horizontal rows. JExp = Judgments for explanation questions, JDet = Judgments for detailed questions, NFC = Need For Cognition, and GPA = Reported Grade Point Average. **p < .05. **p <.01.*

These correlations indicate that participants' interest in the operationally defined less interesting WBT text was positively correlated with explanation question judgments. In contrast, NDRL judgments were not statistically detectably correlated with prior topic interest. For both texts, understanding of university text was positively correlated with explanation judgments. NFC was statistically detectably correlated with judgments across texts and test types. As predicted by the literature NFC was also statistically detectably correlated with GPA and understanding of university text.

4.5.2. Canonical analysis of individual differences in relation to explanation and detailed judgments.

Canonical correlation analyses were conducted to examine associations between WBT and NDRL explanation and detailed judgments and individual differences. Analyses on WBT text indicated that the two canonical roots accounted for 23.6% of the shared variance between judgments and individual differences. However, only the first root was statistically detectable, F(10, 204) = 2.75, *p* = .003. The second root will not be further analyzed as it was not deemed statistically detectable (*p* = .08) and accounted for relatively little variance.

The standardized canonical coefficients for dependent (judgment) variables and the canonical variate representing individual differences was 1.07 for explanation judgments and .14 for detailed judgments. Therefore, the canonical variate can be interpreted as describing explanation judgment variance. Explanation judgments were most strongly influenced by topic interest (.54), NFC (.53), and understanding text at the university level (.38).

Analyses on the NDRL text indicated that the two canonical roots accounted for 38% of the shared variance between judgments and individual differences. Canonical root 1 and 2 accounted for 28% and 10%, respectively. Both roots one F(10, 206) =

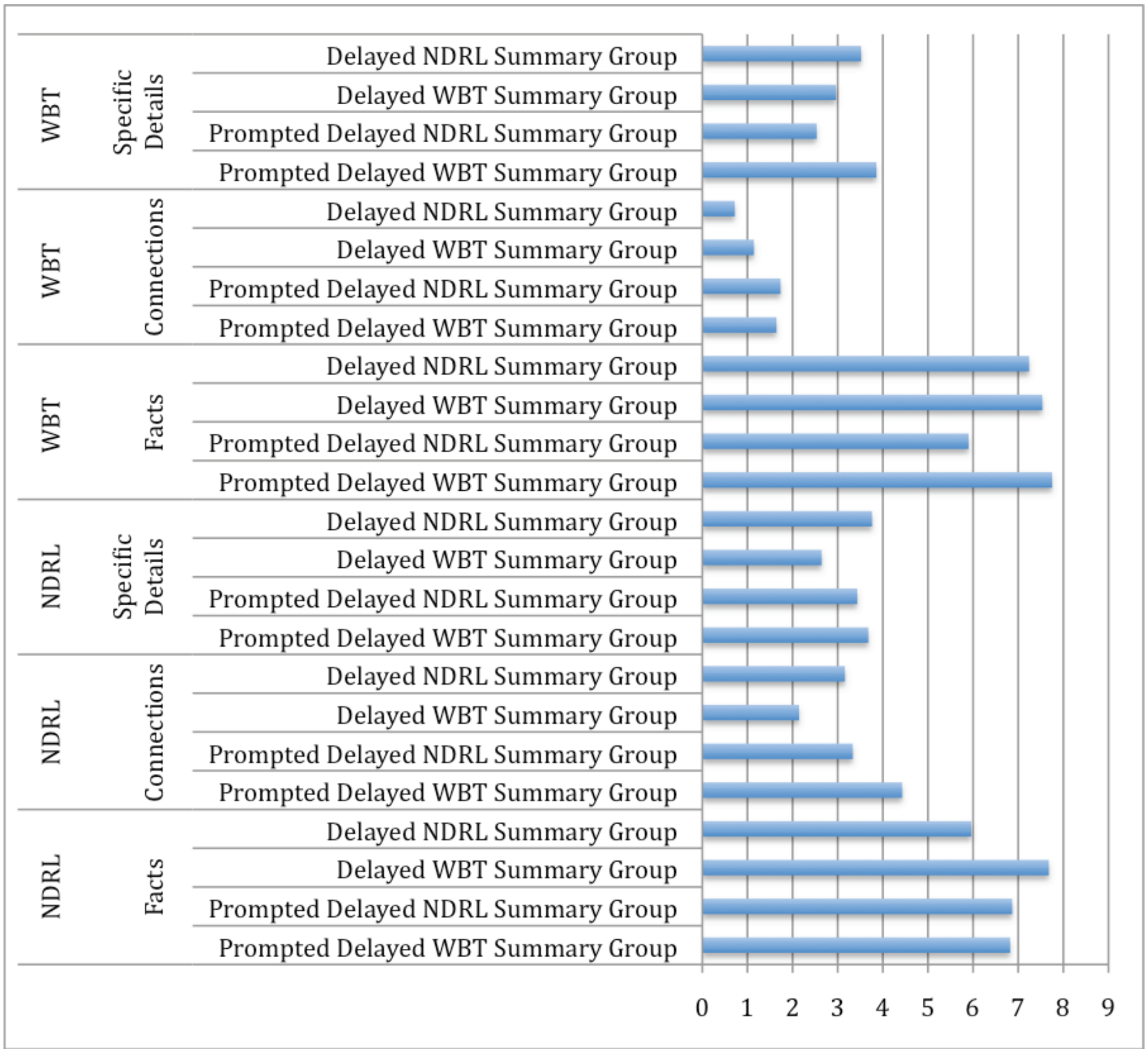
2.76, *p* = .01, and two F(4, 104) = 3.01, *p* = .02 were statistically detectable at traditional levels. The standardized canonical coefficients for judgments were .88 for explanation judgments, and .20 for detailed judgments. The standardized canonical coefficients for dimension 2 were .76 and 1.2 for explanation and detailed judgments, respectively. Thus, dimension one can be considered a strong representation of explanation judgment variance, whereas dimension 2 can be considered to represent judgment variance less likely to be influenced by judgment level (i.e., detailed or explanation). Dimension 1 was most strongly influenced by understanding of university text (.46) and NFC (.67). Dimension 2 was most strongly influenced by GPA (.86), NFC (.45), and understanding university text (.38).

4.6. Summary Characteristic Differences as a Function of Group

To test for between group differences in summary characteristics, a MANOVA with summary characteristics as outcome variables and groups as independent variables was conducted. As was the case in analyses 4.3, a profile analysis could have been conducted to test for statistically detectable mean differences between texts and groups separately analyzed in 4.6 and 4.7. However, these tests were also conducted in the service of different experimental questions. Thus, the choice was again made to conduct these analyses separately.

A statistically detectable main effect was found for Group F (18, 312) = 3.6, *p* = .001, η = .171, β = 1.0. A post hoc comparison indicated that statistically detectable main effects for both texts resulted from the number of Connections. Specifically, for the WBT text, the Delayed NDRL Summary group participants recorded fewer Connections than Prompted Delayed NDRL Summary group or Prompted Delayed WBT Summary group F(3, 107) = 5.45, *p* = .004, η = .116, β = .88. Similarly, for the NDRL text, Delayed WBT Summary group recorded fewer Connections than the Prompted Delayed WBT Summary group F(3, 107) = 4.63, *p* = .004, η = .12, β = .88 (see Figure 6).

Figure 6: Summary Characteristics by Text and Group



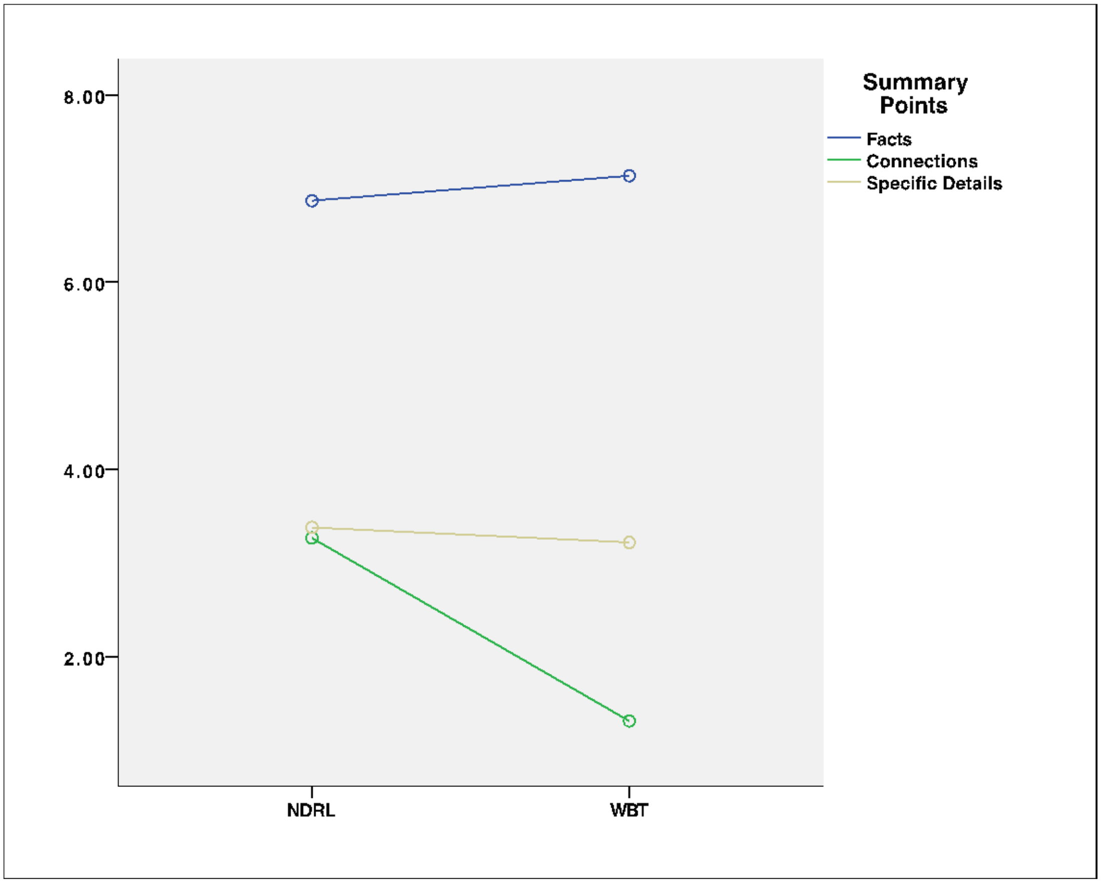
Note: Units are whole numbers representing the number of characteristics encoded in summaries.

4.7. Summary Characteristic Differences as a Function of Text

To examine differences between the number of Specific Details, Facts, and Connections as a function of text, a 2 (Text) x 3 (Summary Characteristics) repeated measures ANOVA was conducted. There were statistically detectable main effects for Text F(1, 110) = 13.33, *p* < .001, η = .11, β = .95; Summary Characteristics F(2, 109) = 275.09, *p* < .001, η = .84, β = 1.0; and Text x Summary Characteristics F(2, 109) = 17.69, *p* = .001, η = .245, β = 1.0. Further analysis of the estimated marginal means for text showed statistically detectably more summary characteristics in NDRL (M = 4.5, SE = .21) than in WBT (M = 3.9, SE = .17). Marginal means for summary characteristics showed more Facts (M = 7.0, SE = .25, *p* < .001) than Connections (M = 2.3, SE = .13, *p* < .001) or Specific Details (M = 3.3, SE = .21, *p* < .001). Additionally, there were statistically detectably more Specific Details than Connections (*p* < .001).

Interaction effects for Text x Summary Characteristics showed, after Bonferroni adjustment, statistically detectably more Connections for NDRL than WBT F(1, 110) = 69.56, *p* < .001, η = .39, β = 1.0 (see Figure 7). Text differences between NDRL and WBT Facts (*p* = .46) and Specific Details (*p* = .55) did not reach traditional levels of statistical detection (see Figure 7).

Figure 7: Marginal Means Interaction Between Text and Summary Characteristics



4.8. Associations Between Summary Characteristics and Judgments

In this section, associations between judgments and characteristics of summaries (Specific Details, Facts, and Connections) are examined. First, Pearson *c*orrelations between summary characteristics and judgments (detailed and explanation) provides an overview of relationships.

4.8.1. Correlations between summary characteristics and judgments.

A review of Table 4 shows that WBT judgments (at both the explanation and detail levels) shared variance with Facts and Specific Details. However, WBT judgments did not share statistically detectable variance with Connections. NDRL explanation judgments shared statistically detectable variance with all three summary characteristics. However, detailed judgments only shared statistically detectable variance with Specific Details and Connections.

Table 4: Correlations Between Summary Characteristics and Judgments

Measure	1	2	3	4	5	M	SD
1. JExp	—	—	.36**	.36**	.19*	4.86	2.27
2. JDet	—	—	.14	.30**	.26**	4.31	1.96
3. Facts	.47**	.35**	—	.36**	.48**	7.1	3.17
4. Connections	.14	.04	.28**	—	.42**	1.35	1.13
5. Specific Details	.27**	.23**	.68**	.11	—	3.21	2.28
M	6.14	5.19	6.8	3.3	3.4	—	
SD	2.17	2.05	3.0	2.5	3.0		—

Note: Intercorrelations for Nuclear Dumping in Russian Lake (n = 115) text are presented above the diagonal, and intercorrelations for Wind Blown Transportation (n = 115) are presented below the diagonal. Means and standard deviations for Nuclear Dumping in Russian Lake text are presented in vertical columns, and means and standard deviations for Wind Blown Transportation are presented in the horizontal rows. **p < .05. **p <.01.*

4.9. Summary Characteristic and Judgment Regression Analysis

Correlation analysis provides a good overview of relationships between summary characteristics and judgments. However, bivariate correlation cannot extract unique variance shared between judgments and characteristics. Unique variance is an important consideration as ignoring overlapping variance may result in overestimating the effect of any one summary characteristic on explanation or detailed judgements. In the following analysis unique influences of summary characteristics and the interaction between summary writing time and total summary characteristics is tested. A stepwise regression entry method was used. The possibility of spurious inclusion of variables due to small differences in statistical variance (a potential problem with the stepwise approach) was checked by comparing stepwise results to (an unreported) single block, direct entry analysis.

4.9.1. Regression of summary characteristics on explanation judgments.

The number of Facts recalled in both texts was positively associated with judgment magnitude (see Table 5). In addition, the interaction of summary seconds and Specific Details was negatively associated with judgment magnitude for the WBT text. A positive relationship was found between Connections and NDRL explanation judgments.

Table 5: Stepwise Regression Analysis for Summary Characteristics and Explanation Judgments

	WBT		NDRL	
Variable	ß	95% CI	ß	95% CI
Facts	.518	[.25, .49]	.266	[.01, .05]
Total Summary Characteristics x Summarization Time	-.199	[-.001, -.00]	-	-
Connections	-	-	.243	[.04, .37]
Adj R^2	.238		.170	
F	17.74**		11.93**	

**p* < .05. ***p* <.01

4.9.2. Regressions of summary characteristics on detail judgments.

An identical analysis to 4.9.1 was conducted on detailed judgments. Interestingly Connections remained a statistically detectable predictor of NDRL detail judgments (see

Table 6). A positive relationship was also found between Facts and WBT detailed judgments.

Table 6: Stepwise Regression Analysis for Summary Characteristics and Detailed Judgments

	WBT		NDRL	
Variable	ß	95% CI	ß	95% CI
Facts	.356	[.11, .33]	-	-
Connections	-	-	.292	[.09, .38]
Adj R^2	.119		.077	
F	15.43**		9.88**	

**p* < .05. ***p* <.01

4.9.3. Regression of WBT summary characteristics on explanation judgments between groups.

Anderson and Thiede (2008) found a larger association between detailed idea units and judgment magnitude when summaries were written immediately, and between gist idea units and judgment magnitude when summaries were written after a delay. To test for a similar relationship, regression analyses were conducted within each group predicting judgments using summary characteristics. Facts consistently predicted explanation judgments for WBT text across Delayed WBT Summary, Prompted Delayed NDRL Summary and Prompted Delayed WBT Summary groups (see Table 7). The effect size for fact recall between groups was relatively consistent. The interaction between summary time and total characteristics was a negative predictor of judgment in the Delayed WBT Summary group.

Table 7: Stepwise Regression for WBT Summary Characteristics and Explanation Judgments Between Groups

Variable	**Delayed NDRL Summary**		**Delayed WBT Summary**		**Prompted Delayed NDRL Summary**		**Prompted Delayed WBT Summary**	
	ß	95% CI	ß	95% CI	ß	95% CI	ß	95% CI
Facts	-	-	.587	[.26, .81]	.466	[.15, 1.11]	.439	[.04, .44]
Total Summary Characteristics x Summary Time	-	-	-.326	[-.002, .00]	-	-	-	-
Adj R^2	-		.521		.188		.161	
F	-		14.58**		7.48*		6.19*	

**p < .05. **p <.01*

4.9.4. Regression of NDRL summary characteristics on explanation judgments between groups.

The influence of fact recall on explanation judgment was less pronounced for NDRL judgments. Only the Prompted Delayed NDRL Summary group showed a statistically detectable association with fact recall (see Table 8). However, Connections predicted explanation judgments for Delayed WBT Summary and Prompted Delayed WBT Summary groups.

Table 8: Stepwise Regression for NDRL Summary Characteristics and Explanation Judgments Between Groups

Variable	Delayed NDRL Summary		Delayed WBT Summary		Prompted Delayed NDRL Summary		Prompted Delayed WBT Summary	
	ß	95% CI	ß	95% CI	ß	95% CI	ß	95% CI
Facts	-	-	-	-	.611	[.21, .65]	-	-
Connections	-	-	.394	[.01, 1.24]	-	-	.570	[.13, .50]
Adj R^2	-		.120		.350		.298	
F	-		4.41*		16.06**		12.03*	

*p < .05. **p <.01

4.9.5. Regression of WBT summary characteristics on detailed judgments between groups.

Summary predictors for detailed WBT judgments differed from explanation judgments (see Table 9). Specifically, Facts in the Delayed NDRL Summary group were positively associated with detailed judgments but not with explanation judgments. Delayed WBT Summary and Prompted Delayed NDRL Summary groups did not have statistically detectable associations between Facts and detailed judgments, but they did with explanation judgments. Interestingly, a negative interaction between summary time and number of total characteristics was found in the Delayed WBT Summary group for both detailed and explanation judgments.

Table 9: Stepwise Regression for WBT Summary Characteristics and Detailed Judgments Between Groups

Variable	**Delayed NDRL Summary**		**Delayed WBT Summary**		**Prompted Delayed NDRL Summary**		**Prompted Delayed WBT Summary**	
	ß	95% CI	ß	95% CI	ß	95% CI	ß	95% CI
Facts	.439	[.02, .38]	-	-	-	-	.465	[.06, .42]
Total Summary Characteristics x Summary Seconds			-.418	[.002, .00]	-	-	-	-
Adj R^2	.257		.140		-		1.86	
F	5.48*		5.07*		-		7.18*	

*p < .05. **p <.01

4.9.6. Regression of NDRL summary characteristics on detailed judgments between groups.

Whereas Facts and Connections were used to predict NDRL explanation questions in the Delayed WBT Summary, Prompted Delayed NDRL Summary, and Prompted Delayed WBT Summary groups, for detailed judgments only Prompted Delayed NDRL Summary group Connections were statistically detectably uniquely associated with detailed judgments (see Table 10).

Table 10: Stepwise Regression for NDRL Summary Characteristics and Detailed Judgments Between Groups

Variable	**Delayed NDRL Summary**		**Delayed WBT Summary**		**Prompted Delayed NDRL Summary**		**Prompted Delayed WBT Summary**	
	ß	95% CI	ß	95% CI	ß	95% CI	ß	95% CI
Connections	-	-	-	-	.559	[.21, .81]	-	-
Adj R^2	-		-		.287		-	
F	-		-		12.28*		-	

*p < .05. **p <.01

5. Discussion

5.1. Introduction

Over the past decade experimental interventions have produced substantial increases to relative metacomprehesion accuracy. One particularly successful method evaluated in the current study is delayed summarization. In addition, effects of individual differences and pre-reading prompts on metacomprehension accuracy were tested. Experimental interventions and factors will be discussed in order of the six questions posed in chapter one.

5.2. Differences in Judgments, Assessment Scores and Metacomprehension Accuracy as a Function of Experimental Intervention

Thiede and colleagues (e.g., Thiede, Griffin, Wiley, & Redford, 2009) proposed that delayed summarization could increase metacomprehension accuracy. I tested this effect, and the influence of a new pre-reading prompting intervention within texts. In contrast to findings by Thiede and Anderson (2003) and Anderson and Thiede (2008), no statistically detectable differences were recorded as a function of summarization or prompting for judgments, assessment scores, or metacomprehension accuracy. The average Pearson correlations between judgment and assessment scores for detailed questions was .29, this is similar to the average of .27 found by Maki (1998b) across 25 studies in her lab, using a similar level of question complexity (Wiley, Griffin, & Thiede, 2005). In this study the average correlation between assessment scores and judgments for explanation questions was higher (.43), but did not reach the level (>.60) of correlation found in most delayed summarization studies (Thiede, Griffin, Redford, & Wiley, 2009). However, comparisons between these studies should be interpreted with caution since methodological differences between gamma and Pearson correlations may result in different interpretations (Anderson & Theide, 2008).

For the purposes of the current study, the most significant finding is that delayed summarization does not seem to positively benefit judgment accuracy within texts. This is important for practice as students are required to determine the extent of per text comprehension in preparation for chapter tests, article reviews, discussions, or topic specific components of larger assessments. In these situations learners are given only one opportunity to accurately judge their understanding.

5.3. Differences in Judgments, Assessment Scores, and Metacomprehension Accuracy Between Texts

The following analyses were conducted to investigate metacomprehension measures between texts. Text specific differences in metacomprehension measures indicate that the character of studied texts *per se* may impact metacomprehension accuracy, and differences between texts need to be considered in addition to general interventions and individual differences.

5.3.1. Judgments and scores.

Participants' judgments differed between and within texts. Specifically, participants uniformly judged NDRL higher than WBT for detailed and explanation questions. Within texts, explanation judgments exceeded detailed judgments for both texts. In contrast, explanation assessment scores for the WBT text were statistically detected to be lower (by 14%) than detailed scores. Statistically detected differences were not found between detailed and explanation scores for the NDRL text. Thus, although participants' judged explanation questions higher than detailed questions, the reverse effect was found in assessment scores. Features of specific texts may add to this effect, as participants were generally less confident in WBT judgments.

In the literature reviewed here, no other study has asked for both explanation and detailed level judgments. Although in most studies students are asked to judge comprehension or understanding, it is not clear if participants uniformly attribute the same depth of learning to these terms. For example, students with more epistemologically naïve perspectives may judge *comprehension* to be a rather surface level of learning (Hofer and Pintrich, 1997). Although no systematic studies have looked into participants' perceptions of learning depth based on judgment question wording, this study suggests that any differences between participants may confound results as participants tend to judge memory for details lower than general comprehension.

5.3.2. Metacomprehension.

Metacomprehension accuracy was higher for the WBT than the NDRL text. Moreover, participants were statistically detected to be more accurate on detailed compared to explanation questions. In accordance with findings in the literature, participants' confidence was inversely proportional to judgment accuracy. Moreover, WBT detail questions had the lowest judgments and highest judgment accuracy. Thus, from this analyses it would seem that characteristics of WBT that were associated with lower judgments may have increased detailed and explanation level metacomprehension accuracy.

5.4. Effects of Individual Differences on Judgments, Assessment Scores, and Metcomprehension Accuracy

The influence of text characteristics, and the lack of influence of experimental interventions on judgments, assessment scores, and metacomprehension accuracy has been demonstrated. In the remaining sections influences of individual differences and summary characteristics on judgments will be explored. To investigate the effect of individual differences on judgment, Pearson and canonical correlation analyses were conducted between individual differences and judgments for both texts.

5.4.1. Pearson correlations between individual differences and judgments.

Correlations among individual differences and judgments suggest that explanation and detailed judgments are influenced by multiple factors. Specifically, for both texts NFC shared variance with detailed and explanation judgments, and understanding of university text shared variance with explanation questions. All correlations were positive. Thus, it seems that participants who were more willing to engage in intensive cognitive activity were more likely to make more positive judgments. As may be expected, participants who perceived they had better understanding of university text were more confident in their comprehension of texts they read in this study. For the less interesting WBT topic, there was a statistically detectable correlation between topic interest and explanation judgments. There was also a statistically detectable correlation between topic interest and NFC. Neither of these correlations was statistically detectable at traditional levels for the NDRL text. Thus, for the WBT topic that was rated *a priori* to be less interesting, a statistically detectable association was found between interest and judgment magnitude.

Because individual differences appear to influence participants' text judgments, I sought to isolate variables that share unique variance with judgment using canonical correlation, as discussed in the next section.

5.4.2. Canonical correlations between individual differences and judgments.

The canonical correlation between individual differences and WBT judgments supports findings from Pearson's correlations that Interest and NFC share variance with WBT explanation level judgments. Although it is impossible to determine a causal relationship between factors through correlation, it is unlikely that NFC was influenced by interest in a specific text topic. Therefore, two more likely alternatives remain: 1) a *third variable* may be responsible for the shared variance between WBT explanation judgment magnitude and topic interest and NFC, or 2) higher NFC may increase the likelihood that participants will have interest for less interesting topics, and will therefore positively influenced WBT explanation judgments. In either case, it is interesting that statistically detectable associations between interest and text in this study were specific to WBT.

The canonical correlation between individual differences and judgments for NDRL found two canonical roots accounting for 38% of variance. Judgment loadings on these roots indicated that root one represented explanation judgments, and root two represented general judgment tendencies that were less influenced by judgment level. NFC loaded on both variables at .67 and .45 for root one and two, respectively. Thus, participants' willingness to engage with the text from both self-reported motivational and aptitude perspectives positively correlated with explanation judgment magnitude. Loadings on root two with GPA (.86) indicated that general academic aptitude and motivation to engage in cognitive activity influenced both detailed and explanation judgments.

Two important inferences can be drawn from these analyses. First, topic interest may impact judgment for explanation questions when the text topic is identified as marginally interesting (2.5/5 for WBT). When the topic is slightly (and statistically detectably) more interesting (3/5 for NDRL), and the text explores topics of general human intrigue (death, corruption, and catastrophe) participants' general ability to comprehend text positively correlates with explanation judgment magnitude. Moreover, general academic proficiency as measured by GPA was associated with overall judgment magnitude for NDRL.

Second, NFC was identified as influential variable for judgments across texts and levels (i.e., detailed and explanation). NFC has been a previously understudied variable in the metacomprehension literature. Results in the current study indicate participants with higher NFC may also have higher judgments. NFC should be considered in future investigations into metacomprehension accuracy.

5.5. Summary Characteristic Differences as a Function of Experimental Intervention

Anderson and Thiede (2008) found that the number of *gist* and *detail* summary idea units were greater after immediate summarization. In the current analysis, when participants produced WBT summaries immediately after reading (without prompting) they produced fewer Connections than groups that were provided a graphical organizer prior to immediate summary or which were provided pre-reading facts prior to writing delayed summaries. In NDRL summaries there was no statistically detectable difference between immediate and delayed summary groups. However, when a graphic organizer was provided prior to the immediate summary group, more Connections were produced than when immediate summarization occurred without pre-reading prompts. No other within text summary characteristic group differences was found.

These findings indicate that texts *per se* may influence the impact of experimental intervention on text summaries. Specifically, when the WBT text was delayed and pre-reading facts were provided participants produced more Connections than when WBT text was summarized immediately without prompting. It is unclear if the pre-prompting with facts or the delayed summarization influenced recorded Connections. However, the prompted delayed NDRL group did not show a similar increase in the number of Connections recorded when provided with pre-reading facts. The number of Connections increased in both texts when pre-reading graphic organizers were provided. This may indicate that graphic organizers not only increase participants' attention to Connections (Anderson & Thiede, 2008) but also influence recall during summarization. However, as previously noted and in contrast to findings by Anderson and Thiede (2008), increased summary Connections did not increase metacomprehension accuracy between groups.

5.6. Summary Characteristic Differences as a Function of Text

For both texts, participants recalled more Facts than Specific Details, and more Specific Details than Connections. Thus, participants seem to be better able to recall lexical and text based ideas from text. Results also indicated that in all groups participants recalled more NDRL than WBT characteristics (Facts, Specific Details, and Connections). Thus, it seems that, *ceteris paribus*, participants were better able to recall the more challenging situation model idea units, and more simplified lexical and text based units from NDRL than WBT text.

5.7. Associations Between Judgments and Summary Characteristics as a Function of Texts and Groups

To provide an overview of associations between summary characteristics and judgments, Pearson correlations were calculated and will be discussed first. Next, unique shared variance between detailed and explanation judgments and summary characteristics will be discussed between texts and groups.

5.7.1. Pearson correlations between summary characterstics and judgments.

Correlation analysis indicated that Facts and Specific Details shared statistically detectable variance with all WBT judgments and were proportional to those judgments. In contrast, Facts did not share statistically detectable variance with NDRL detailed judgments. Participants' detailed and explanation judgments shared statistically detectable positive variance with Connections in NDRL, but no statistically detectable correlation was found in WBT. Thus, Connections were seemingly an untapped resource for making WBT judgments. Potentially participants relatively low judgment about success on WBT explanation questions (64% NDRL vs. 48% WBT) diminished their attention to summary Connections.

5.7.2. Regression analyses between summary characteristics and judgments between texts.

Regression analysis was conducted for NDRL and WBT across texts and groups. The intent of these analyses was to determine a) which summary cues share variance with judgments at different levels within each text, and b) if summary cue variations differed as a function of text. In addition, evidence from the social psychology and metacomprehension literatures indicated a potential interaction between the total number of retrieved characteristics and cognitive intensity as measured by recall latency (see Baker & Dunlosky, 2006). Adding the interaction of summary time x total summary characteristics into the analysis tested this effect.

The predicted negative summarization time x total summary characteristics interaction effect was supported for WBT explanation judgments, as they shared statistically detectably unique negative variance with WBT explanation judgments. This effect was not found for NDRL. Facts shared statistically detectably unique variance with explanation judgments for both texts however beta values were nearly twice as high for the WBT text. As indicated by Pearson correlation analysis, Connections were a statistically detectable unique predictor of NDRL explanation judgments but not WBT judgments. These results indicate that for the NDRL text, participants relied on a combination of Facts and Connections to make explanation judgments, a strategy that is expected to increase metacomprehension accuracy (Anderson & Thiede, 2008). In contrast, WBT participants relied heavily on Facts for explanation judgments. This is expected to be a less effective judgment strategy. Therefore it is interesting that WBT explanation judgments were more accurate than NDRL explanation judgments. Possibly the summary time x summary characteristics interaction, and generally lower interest and knowledge could have suppressed judgments and thus mitigated the negative effect of overconfidence.

Unique variance for detailed judgments was accounted for differently by WBT and NDRL summary characteristics. Specifically, Facts accounted for unique variance with WBT detailed judgments while Connections accounted for unique variance with NDRL detailed judgments. According to Anderson and Thiede (2008), the use of fact based summary cues for detailed judgments should be most effective. In the current study this prediction was supported as WBT detailed judgments were statistically detectably more accurate than NDRL detailed judgments.

Two conclusions can be drawn from the proceeding analysis. First, participants tended to rely more heavily on Connections to make NDRL judgments, and on Facts to make WBT judgments. Explanation judgments for WBT were also negatively affected by the intensity of information recall. Second, text characteristics influence summarization cues used to make both explanation and detailed judgments.

5.7.3. Regression analyses between summary characteristics and judgments between groups.

Regression analyses were conducted between groups to test for effects of delay and pre-reading prompts on judgment cue use. Facts were statistically detectably uniquely associated with WBT detailed but not explanation judgments in the Delayed NDRL Summary group. Although it is unclear if this effect represents an intentional decision, it indicates that participants made a theoretically optimal choice to associate Facts with detailed but not explanation judgments. Optimally, Connections would have been associated with explanation judgments but this was not the case. Potentially, the immediate summary condition precluded the use of Connections for explanation judgments. However, this is likely not the case as there was no statistically detectably unique variance shared between Connections and judgments across groups or levels.

A negative interaction between total summary characteristics and summary time was found in for the WBT text in the Delayed WBT Summary group for both explanation and detailed questions. It is possible that unique individual differences resulted in this group showing a similar effect across question type. However, since individual differences did not statistically detectably differ between groups, and no interaction effects were found for the Delayed WBT Summary group's NDRL explanation judgment, this rationale is not supported by available data. Another plausible explanation is that the summarization delay made it more taxing for participants to recall information during summary. However, there were no interaction effects found for the Prompted Delayed WBT Summary group. It is interesting that when all variables were entered in a single block for each group and for both texts, the interaction effect was predominantly negative, although *p*-values were > .3. This could indicate that an underlying effect may be present but only statistically detectable in certain circumstances. It is possible that WBT text characteristics and delayed summarization resulted in this effect reaching traditional levels of statistical detection. However, further research would be required to verify or dispute such a contention.

Statistically detectable shared variance between judgments and Facts were found for the WBT text in the prompted delayed NDRL and WBT summary groups' for explanation questions. Consistency between prompted delayed groups was not found for NDRL explanation questions. Shared variance between judgments and Facts were found in the Prompted Delayed NDRL Summary group, whereas judgments were predicted by Connections in the Prompted Delayed WBT Summary group. Thus, it seems that for the NDRL text, presenting Facts prior to delayed summarization resulted in participants relying to a greater extent on Facts when making explanation judgments. Conversely, when a graphic organizer was used as a pre-reading prompt, participants used Connections to a greater extent when making judgments. Regardless of prompting condition, participants used Facts to a greater extent when making WBT explanation judgments. Although no specific connection can be made to text characteristics based on these findings, it does suggest that text characteristics may play a role in the influence of pre-reading prompts on judgment cue use.

No statistically detectably unique variance between explanation or detailed judgments and summary characteristics was found for the NDRL text in the Delayed NDRL Summary group. In fact, the only summary group that showed an association between summary characteristics and detailed judgments was the Prompted Delayed NDRL Summary group. In this case Connections were associated with judgments. It is unclear why participants detailed NDRL judgments were largely not associated with summary characteristics. In the case of the Prompted Delayed NDRL Summary group Connections were associated with details after prompting with pre-reading facts. Given that correlations between Connections (.30) and Specific Details (.26) and detailed NDRL judgments were similar, it may be that participants used both of these cues. Potentially, the tendency to rely on Connections, and the cognitive necessity to consider rote memory for detailed judgments may have caused a lack of unique association. Further research would required to support or dispute this suggestion.

5.8. Limitations

Most limitations in this study fall into two categories: statistical and logistical. These and steps taken to mitigate their influence are discussed below. A third potential kind of limitation was the use of perceived GPA and ability to read university text ratings, and a context neutral NFC scale. GPA and ability to read university text were both provided based on participants' perception. If participants were truthful about their perceptions then this would be the most prudent measure, as perceptions will be used when those people make judgments. However, there is the potential that these factors were inflated due to social desirability. NFC is both a dispositional and situational factor (Cacioppo, Petty, Feinstein, Blair, & Jarvis, 1996). However, in this study (like many others) a general measure of NFC was used as a predictor over two different text contexts. Also, as previously mentioned one item of the efficient NFC instrument was missing in this study.

5.8.1. Statistical considerations.

WBT knowledge ratings were positively skewed (3.2). In addition, in accordance with suggestions by Field (2009) two WBT summary time outliers were adjusted to one second above the highest timing. Similarly, one specific detail and one connection score was adjusted to one score above the highest score. These slight adjustment reduced distribution skew to acceptable levels. Another statistical concern in this study is the use of parametric statistical methods with ordinal judgment ratings. For this to permissible there must be an assumption of equal distance between ratings (Field, 2009). Participants in this study provided judgments on a sliding scale between 0-10. To check for adverse effects of this method non-parametric (Spearman's Rho) correlations were calculated in addition to parametric (Pearson's *r*) correlations. Correlations were very similar and therefore use of parametric methods, including regression was considered tenable. A third concern may be that of interrater reliability. Although levels were greater than .70 which is usually considered an acceptable level (Jonsson & Svigby, 2007), interrater reliability for Facts of .78 were somewhat lower than expected. Finally, the decision was made to use stepwise regression to isolate variables sharing unique statistically detectable variance. As discussed within the paper this method may take some power away from the research, but the author judged this was the prudent technique given experimental conditions.

5.8.2. Logistical considerations.

One concern with the current study was its length. Participants were required to spend on average 1hr to complete the entire study. Some participants extended this requirement to as much as 1.5hrs. Given the extent, difficulty, and length of the readings and assessments it is possible that participant fatigue may have affected results. However, since group differences were not discovered, and reading order was alternated between groups this concern may be unfounded. Secondly, on approximately 6 occasions computers in the lab froze. This difficulty was addressed by restarting computers or moving participants to adjoining computers. Since all experimental data was recorded in the cloud, no data was lost. However, it is possible that the disruption and inconvenience of pausing in the middle of the study could have influenced results. All metadata was reviewed to ensure that pauses did not create false traced in study duration logs. Finally, it is possible that participants' misunderstanding of assessment requirements during judgment may have influenced results.

5.9. Conclusion

Findings in this study suggest that metacomprehension accuracy calculated using a within text design may behave differently from relative or absolute accuracy calculated across texts. In this study judgments, scores, and metacomprehension accuracy at the level of explanation and detail showed statistically detectable variance between texts. This suggests that previous findings for the delayed summarization effects using across text judgments, scores, and metacomprehension accuracy measures may have averaged out different text effects. In situations where learners must judge their understanding of multiple diverse texts, and/or relative accuracy between texts, these differences may not be detrimental. However, if learners are attempting to determine comprehension for specific passages, measures may be inaccurate.

This study also evaluated the delayed summarization effect within text. Findings here showed no statistically detectable increases in metacomprehension accuracy for either text. Moreover, delayed summarization in some instances had effects opposite to predictions in the literature. In some cases judgment and score correlations with summary characteristics supported findings in the literature, however these results were inconsistent. Other consistencies were found within groups and across text and summary delays. Further research is required to validate the use of delayed summarization to increase metacomprehension accuracy within texts, and to more closely examine group effects in within text designs.

Pre-reading prompts were used in addition to delayed summarization groups in this study. These prompts seemed to statistically detectably influence judgments and scores. However, the effects were inconsistent across texts and groups. Moreover, providing facts often impacted both detailed and explanation judgments and scores. Similarly, graphic organizers influenced memory for Specific Details and Facts. These findings are inconsistent with predictions, but potentially useful foundations for future research.

Strong correlations between individual differences and judgments were found in this study. Specifically, NFC and interest were associated with judgments for explanation questions in WBT text. NFC and GPA was associated with judgments across both detailed and explanation judgments, and NFC and ability to read university text was associated with NDRL explanation judgments. Again this points to variance due to text differences that should be considered in future research.

Finally, this study suggests that further investigation may be required to determine the impact of statistical calculations of accuracy outside of the intent of learners' judgments. In traditional studies learners make judgments of comprehension or memory for specific texts. Unlike tests of rote cue-response memory, metacomprehension requires participants to consider question complexity, text structure and individual differences when making judgments. The impact of these factors may influence judgments differently if participants were asked to make normative and or averaged judgments across texts. Future research could ask learners' to make both text specific, averaged, and ranking judgments of comprehension. Comparing metacomprehension based on different methods of quantifying comprehension could help to clarify these questions.

6. References

Aarts, H., & Dijksterhuis, A. (2000). Habits as knowledge structures: Automaticity in goal-directed behavior. *Journal of Personality and Social Psychology, 78,* 53-63.

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

Ausubel, D. P. (1978). In defense of advance organizers: A reply to critics. *Review of Educational Research, 18*, 251-257.

Baker, J. M., & Dunlosky, J. (2006). Does momentary accessibility influence metacomprehension judgments? The influence of study-judgment lags on accessibility effects. *Psychonomic Bulletin & Review, 13*, 60-65.

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology, 42,* 116-131.

Cacioppo, J. T.,Petty, R. E., & Kao, C. F. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment, 48,* 306-307.

Cacioppo, J. T., Petty, R. E., Feinstein, J. A., & Jarvis, W. B. G. (1996). Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. *Psychological Bulletin, 119,* 197-253.

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

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin, 132,* 354-380.

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

Cohen, A. R., Stotland, E., & Wolfe, D. M. (1955). An experimental investigation of need for cognition. *Journal of Abnormal and Social Psychology, 51,* 291-294.

Coutinho, S., Wierner-Hastings, K., Skowronski, J. J., Britt, M. A. (2005). Metacognition, need for cognition and use of explanations during ongoing learning and problem solving. *Learning and Individual Differences, 14*, 321-337.

Cull, W. (2000). Untangling the benefits of multiple study opportunities and repeated testing for cued recall. *Applied Cognitive Psychology, 14*, 215-235.

Dempster, F. N. (1989). Spacing effects and their implications for theory and practice. *Educational Psychology Review, 1,* 309-330.

Dunlosky, J. (2005). Why does rereading improve metacomprehension accuracy? Evaluating the levels-of-disruption hypothesis for rereading effect. *Discourse Processes, 40*, 37-55.

Dunlosky, J., Hartwig, M. K., Rawson, K. A., & Lipko, A. R. (2011). Improving college students' evaluation of text learning using idea-unit standards. *The Quarterly Journal of Experimental Psychology, 64*, 467-484.

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

Dunlosky, J., & Metcalfe, J. (2009). *Metacognition.* Thousand Oaks, CA: Sage Publications, Inc.

Dunlosky, J., & Thiede, K. W. (1998). What makes people study more? An evaluation of factors that affect self-paced study. *Acta Psychologica, 98*, 37-56.

Ebbinghaus, H. (1913). *A contribution to experimental psychology*. New York: Teachers College, Columbia University.

Field, A. P. (2009). *Discovering statistics using SPSS.* London: Sage.

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

Glenberg, A. M., & Epstein, W. (1987). Inexpert calibration of comprehension. *Memory & Cognition, 15*, 84-93.

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

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

Glover, J. A. (1989). The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology, 81*, 392-399.

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

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

Greene, J. A., & Azevedo, R. (2007). A theoretical review of Winne and Hadwin's model of self-regulated learning: New perspectives and directions. *Review of Educational Research, 77*, 334-372.

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

Griffin, T. D., Wiley, J., & Thiede, K. W. (2008). Individual differences, rereading, and self-explanation: Concurrent processing and cue validity as constraints on metacomprehension accuracy. *Memory & Cognition, 26*, 93-103.

Haenggi, D., & Perfetti C. A. (1992). Individual differences in reprocessing of text. *Journal of Educational Psychology, 84*, 182-192.

Hofer, B. K., & Pintrich, P. R. (1997). The development of epistemological theories: Beliefs about knowledge and knowing and their relation to learning. *Review of Educational Research, 67,* 88-140.

Jonsson, A., Svingby, G. (2007). The use of scoring rubrics: Reliability, validity, and educational consequences. *Educational Research Review, 2*, 130-144.

Kiewra, K., Mayer, R., & Dubois, N. (1996). Effects of advance organizers and repeated presentations on students' learning. *The Journal of Experimental Education, 65*, 147-159.

Kimball, D. R., & Metcalfe, J. (2003). Delaying judgments of learning affects memory, not metamemory. *Memory & Cognition, 31*, 918-929.

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

Kintsch, W. (1994). Learning from text. *American Psychologist, 49*, 294-303.

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

Klein, S. B., Kihlstrom, J. F. (1986). Elaboration, organization, and the self-reference effect in memory. *Journal of Experimental Psychology: General, 115*, 26-38.

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychology Review, 100*, 609-639.

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

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

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

Kornell, N., & Metcalfe, J. (2006). Study efficacy and the region of proximal learning framework. Jou*rnal of Experimental Psychology: Learning Memory and Cognition, 32*, 609-622.

Lin, L.-M., & Zabrucky, K. (1998). Calibration of comprehension: Research and implications for education and instruction. *Contemporary Educational Psychology, 23*, 345–391.

Lin, L., Zabrucky, K., & Moore, D. (1996). The relations among interest, self-assessed comprehension, and comprehension performance in young adults. *Reading Research and Instruction, 36,* 127-139.

Linderholm, T., Zhao, Q., Therriault, D., & Cordell-McNulty, K. (2008). Metacomprehension effects situated within an anchoring and adjustment framework. *Metacognition Learning, 3*, 175-188.

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

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

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

Maki, R. H., & Serra, M. (1992). Role of practice tests in the accuracy of test predictions on text material. *Journal of Educational Psychology, 84,* 200-210.

Maki, R. H., & Swett, S. (1987). Metamemory for narrative text. *Memory & Cognition, 15*, 72-83.

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*, 723-731.

Metcalfe, J. (2002). Is study time allocated selectively to a region of proximal learning? *Journal of Experimental Psychology: General, 131*, 349-363.

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

Mayer, R. E. (1979). Twenty years of research on advance organizers: Assimilation theory is still the best predictor of results. *Instructional Science, 8*, 133-167.

Meyer, B. J. F. & Freedle, R. O. (1984) Effects of discourse type on recall. *American Educational Research Journal, 21,* 121-143.

Moore, D., Lin-Agler, L. M., & Zabrucky, K. (2005). A source of metacomprehension inaccuracy. *Reading Psychology, 26*, 251-265.

Morris, C. C. (1990). Retrieval processes underlying confidence in comprehension judgments. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 16*, 223-232.

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

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

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. 125-173). New York, NY: Academic Press.

Nesbit, J. C., & Adesope, O. (2006). Learning with concept and knowledge maps: A meta-analysis. *Review of Educational Research, 76*, 413-448.

Nietfeld, J. L., Cao, L., & Osborne, J. W. (2005). Metacognitive monitoring accuracy and student performance in the postsecondary classroom. Journal of Experimental Education: *Learning and Instruction, 74*, 7–28.

Nietfeld, J. L., Enders, C. K., & Schraw, G. (2006). A Monte Carlo comparison of measures of relative and absolute monitoring accuracy. *Educational and Psychological Measurement, 66*(2), 258–271.

Pressley, M., Snyder, B. L., Levin, J. R., Murray, H. G., & Ghatala, E. S. (1987). Perceived readiness for examination performance (PREP) produced by initial reading of text and text containing adjunct questions. *Reading Research Quarterly, 22,* 219-236.

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

Rieger, J. W., Reichert, C., Gegenfurtner, K. R., Noesselt, T., Braun, C., Heinze, H.-J., Kruse, R., and Hinrichs, H. (2008). Predicting the recognition of natural scenes from single trial MEG recordings of brain activity. *Neuroimage, 42*, 1056-1068.

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

Schwarz, N. (1998). Accessible content and accessibility experiences: The interplay of declarative and experiential information in judgment. *Personality and Social Psychology Review*, 2, 87-99.

Schwarz, N. (in press). Feelings-as-information theory. In P. Van Lange, A. Kruglanski & E. T. Higgins (Eds.), *Handbook of theories of social psychology*. Los Angeles, CA: Sage.

Schwarz, N., Bless, H., Strack, F., Klumpp, G., Rittenauer-Schatka, H., & Simons, A. (1991). Ease of retrieval as information: Another look at the availability heuristic. *Journal of Personality and Social Psychology, 61,* 195-202.

Spellman, B. A., & Bjork, R. A. (1992). When predictions create reality: Judgments of learning may alter what they are intended to assess. *Psychological Science, 5*, 315-316.

Thiede, K. W. (1999). The importance of accurate monitoring and effective self-regulation during multitrial learning. *Psychonomic Bulletin & Review, 6*, 662-667.

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

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

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

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

Thiede, K. W., Griffin, T. D., Wiley, J., & Redford, J. S. (2009). Metacognitive monitoring during and after reading. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), Handbook of metacognition and self-regulated learning (pp. 85-106). New York, NY: Routledge.

Thiede, K. W., Wiley, J., & Griffin, T. D. (2010). Test expectancy affects metacomprehension accuracy. *British Journal of Educational Psychology, 81*, 264-273.

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

Thomas, A. K., & McDaniel, M. A. (2007). The negative cascade of incongruent generative study-test processing in memory and metacomprehension. *Memory & Cognition, 35*, 668-678.

van den Broek, P., Risden, K., Fletcher, C. R., & Thurlow, R. (1996). A "landscape" view of reading: Fluctuating patterns of activation and the construction of a stable memory representation. In B. K. Britton & A. C. Graesser (Eds.), *Models of understanding text* (pp. 165-187). Mahwah, NJ: Erlbaum.

Walczyk, J. J., & Hall, V. C. (1989). Effects of examples and embedded questions on the accuracy of comprehension self-assessments. *Journal of Educational Psychology, 81,* 435-437.

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

Weaver, C. A., III. (1990). Constraining factors in calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 214-222.

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

Winne, P. H., & Perry, N. E. (2000). Measuring self-regulated learning. In M. Boekaerts, P. Pintrich and M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 531-566). Orlando, FL: Academic Press.

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

Zhao, Q., & Linderholm, T. (2011). Anchoring effects on prospective and retrospective metacomprehension judgments as a function of peer performance information. *Metacognition Learning, 6*, 25-43.

7. Appendices

Appendix A.

Nuclear Dumping in Russian Lake

In late 1945, along the banks of the Techa River in the Soviet Union, 70,000 inmates were sent from a dozen labor camps to begin construction of a secret city. Mere months earlier, the United States' atomic bombs had flattened Hiroshima and Nagasaki, leaving Soviet leaders envious of the massive power of that weapon. In a rush to close the gap in weapons technology, the USSR commissioned a sprawling plutonium-production complex in the southern Ural mountains. The secretive military-industrial community was to be operated by Russia's Mayak Chemical Combine, and it would come to be known as Chelyabinsk-40.

Within a few years new nuclear reactors were producing plutonium to fuel the Soviet Union's first atomic weapons. Chelyabinsk-40 was absent from all official maps, and it would be over forty years before the Soviet government would even acknowledge its existence. Nevertheless, the small city became a hidden danger in the Soviet Union, producing nuclear contamination that dwarfed the devastation of more recent nuclear disasters. By June 1948, after 31 months of brisk construction, the first of the Chelyabinsk-40 "breeder" reactors was brought online. Soon bricks of common uranium-238 were being bombarded with neutrons, resulting in weapons-grade plutonium. In their haste to begin production, Soviet engineers lacked time to establish proper waste-handling procedures, so most of the byproducts were dealt with by diluting them in water and squirting nuclear waste into the Techa River. The watered-down waste was a cocktail of "hot" elements including strontium-90 and cesium-137. To decompose only 100 grams of these isotopes approximately 500 years would be required.

In 1951, after about three years of operations at Chelyabinsk-40, Soviet scientists conducted a survey of the Techa River to determine whether radioactive contamination was becoming a problem. In the village of Metlino, just over 6 kilometers downriver from the plutonium plant investigators used sensors to track radioactivity. Rather than the typical gamma radiation in the environment of about 0.21 Röntgens per year, the edge of the Techa River was emanating 5 Röntgens per hour. Such elevated levels were rather distressing since the river was the primary source of water for the village's 1,200 residents. Subsequent measurements found extensive contamination in 38 other villages. Over 128,000 people were either exposed directly to contaminated water or elevated-but-not-quite-as-deadly doses of gamma radiation from floodplains where crops and livestock were raised.

In an effort to avoid serious radiological health effects among the populace, the Soviet government relocated about 7,500 villagers from the most heavily contaminated areas, fenced off the floodplain, and dug wells to provide an alternate water source for the remaining villages. Engineers were brought in to erect earthen dams along the Techa River to prevent radioactive sediments from migrating further downstream. The Soviet scientists at Chelyabinsk-40 also revised their waste disposal strategy, halting the practice of dumping effluent directly into the river. Instead, they constructed a set of waste vats where waste water could spend some time bleeding off radioactivity. After lingering in these vats for a few months, diluted sediment was periodically piped to the new long-term storage location: a ten-foot-deep, 45 hectare lake called Karachay. For a while these measures spared people living near the Techa River from further increases in exposure.

By the mid 1950s, workers at the plutonium production plant began to complain of classic symptoms of chronic radiation syndrome such as low blood pressure, loss of coordination, and tremors. The facility itself was also beginning to display chronic complications, particularly in the waste vat storage system. The row of waste vats now sat in a concrete canal a few kilometers outside the main complex, submerged in a constant flow of water to carry away the heat generated by radioactive decay. Soon the technicians discovered that the hot isotopes in the waste water tended to cause a bit of evaporation inside the tanks, resulting in more buoyancy than had been anticipated. This upward pressure put stress on the inlet pipes, eventually compromising the seals and allowing raw radioactive waste to seep into the canal’s coolant water. To make matters worse, several of the tanks’ heat exchangers failed, crippling their cooling capacity.

Unable to shed enough heat, the concentrated radioactive slurry continued to increase in temperature in the defective 352,000 liter containers. On 29 September 1957, one tank reached an estimated 350 degrees Celsius. At 4:20 pm local time, the explosive salt deposits in the bottom of the vat detonated. The blast ignited the contents of the other dried-out tanks, producing a combined explosive force equivalent to about 85,000 kilograms of TNT. The thick concrete lid which covered the cooling trench was hurled 35 meters away, and 70,000 kilograms of highly radioactive fission products were ejected into the open atmosphere. The buildings at Chelyabinsk-40 shuddered as they were rocked by the shockwave. While investigators probed the blast site in protective suits, a 1.5 kilometer-high column of radionuclides dragged across the landscape. The gamma-emitting dust cloud spread hazardous isotopes of cesium and strontium over 14,500 square kilometers, affecting some 270,000 Soviet citizens and their food supplies.

The facilities at Chelyabinsk-40 were swiftly decontaminated with hoses, mops, and squeegees, and soon plutonium production was underway again. The intermediate storage system had been partially compromised by the accident, but the factory was still able to squirt its constant flow of radioactive waste into Lake Karachay. The lake lacked any surface outlets, so optimistic engineers reasoned that anything dumped into it would remain entombed there indefinitely.

Ten years later, in 1967, a severe drought struck Chelyabinsk Province. Much to Russian scientists’ alarm, shallow Lake Karachay gradually began to shrink from its shores. Over several months the water dwindled considerably, leaving the lake about half-empty. This exposed the radioactive sediment in the lake basin, and fifteen years’ worth of radionuclides took to the breeze. About 2330 square kilometers of land was peppered with Strontium-90, Cesium-137, and other unhealthy elements. Almost half a million residents were in the path of this latest dust cloud, many of them the same people who had been affected by the 1957 waste-tank explosion. Soviet engineers hastily enacted a program to help prevent further sediment from leaving Lake Karachay. For a dozen or so years they dumped rocks, soil, and large concrete blocks into the tainted basin. The Mayak Chemical Combine conceded that the lake was an inadequate long-term storage system, and ordered that Lake Karachay be slowly sealed in a shell of earth and concrete.

Thirty-nine years of nuclear waste had saturated the lake with nasty isotopes, including an estimated 120 megacuries of long-lived radiation. In contrast, the Chernobyl incident released roughly 100 megacuries of radiation into the environment. A delegation who visited Lake Karachay in 1990 measured the radiation at the point where the effluent entered the water, and the needles of their nuclear sensors danced at about 600 Röntgens per hour – enough to provide a lethal dose in one hour.

A report compiled in 1991 found that the incidence of leukemia in the region had increased by 41% since Chelyabinsk-40 opened for business, and that during the 1980s cancers had increased by 21% and circulatory disorders rose by 31%. It is probable, however, that the true numbers are much higher since doctors were required to limit the number diagnoses issued for cancer and other radiation-related illnesses. In the village of Muslyumovo, a local physician’s personal records from 1993 indicated an average male lifespan of 45 years compared to 69 years in the rest of the country. Birth defects, sterility, and chronic disease also increased dramatically. In all, over a million Russian citizens were directly affected by the misadventures of the Mayak Chemical Combine from 1948 to 1990, including approximately 28,000 people classified as “seriously irradiated.”

Today, there are huge tracts of Chelyabinsk land still uninhabitable due to the radionuclides from the river contamination, the 1957 blast, and the 1967 drought. The surface of Lake Karachay is now made up of more concrete than water, however the lake’s payload of fission products is not completely captive. Recent surveys have detected gamma-emitting elements in nearby rivers, indicating that undesirable isotopes have been seeping into the water table. Estimates suggest that approximately 4.4 billion liters of groundwater have already been contaminated with 5 megacuries of radionuclides. The neighboring Norwegians are understandably nervous that some of the pollution could find its way into their water supply, or even into the Arctic Ocean.

Appendix B.

Wind Blown Transportation

Among the proponents of subterranean transportation was Alfred Beach, well-known inventor of a typewriter for the blind, founder of a school for freed slaves, and editor of a new publication known as *Scientific American*. In 1849 he wrote an article in his magazine proposing a network of underground tunnels for horse-drawn trolleys, but that fancy passed once he discovered the great strides being made in England in the field of pneumatics. Pneumatics refers to the study of air and gas movement and, more specifically, using wind as a transportation device.

Although the basic principle of pneumatic tubes was first explored in ancient times, it was not until the early 1800's that practical applications began to appear. It was around that time that the Scottish inventor William Murdoch demonstrated his pneumatic apparatus, a device which used compressed air to whisk notes through a length of pipe to distant recipients. Among the first to appreciate the potential of such systems was a London tinkerer named George Medhurst, who described some practical large-scale applications in his 1812 pamphlet.

"…an hollow tube or archway must be constructed the whole distance, or iron, brick, timber, or any other material that will confine the air, and of such dimensions as to admit a four-wheeled carriage to run through it … The tube must be made air tight, and of the same form and dimensions throughout, having a pair of cast iron wheel-tracks securely laid all along the bottom … and the carriage must be nearly the size and form of the tube, so as to prevent any considerable quantity of Air from passing by it."

Medhurst went on to describe how a large, stationary steam powerplant could produce enough pressure to propel a carriage to an average speed of 80 kilometers per hour, with a fuel efficiency of 6.8 kilometers per coal-bushel. At a time when the most common form of propulsion was feet - either human or horse - it was exciting to consider the prospect of a feasible high-speed transportation system using combinations of existing technologies. Medhurst was aware that travelers might be reluctant to spend long journeys sealed within dark tunnels, so he also described a claustrophobia-friendly alternative which later came to be known as the atmospheric railway. He proposed a system that would use a twelve-inch-wide iron pipe laid between two rails with a sealable slot along its length. Trains would then be connected to an arm protruding from the slot that was connected to a piston within the tube. Several such atmospheric railways were constructed in Europe during the 1840s, most notably by the innovative British engineer Isambard Brunel. One of these peculiar trains achieved an unheard-of 113 kilometers per hour during trial runs. However, decreased pressure due to weakened leather seals caused the technology to be quickly abandoned in favor of steam locomotives.

In the meantime, smaller pneumatic tubes proved their usefulness in shuttling telegraph transcriptions between London's central telegraph offices and the Stock Exchange. The newly-formed London Pneumatic Dispatch Company also began installing iron pipes in the earth to transport postal freight. These pressure tubes carried coffin-sized carts between the post offices of London at speeds up to 96 kilometers per hour.

Eventually a group of investors arranged for a pneumatic-powered passenger carriage tunnel to be installed as a demonstration at the Crystal Palace Exhibition of 1864 at Sydenham, south London. This working prototype aroused much public interest, but its promoters waited too long and failed to bring the technology to fruition.

Upon learning of the strides being made by the London engineers, Alfred Beach became one of pneumatics' most enthusiastic advocates in the United States. To increase public awareness of pneumatic technology, he financed the construction of a 100-foot-long wooden tube to cross the ceiling of the American Institute Fair in 1867. A steam-powered fan installed at the end of this tunnel created enough vacuum pressure to suck carriages of attendees through the tube's length in mere seconds. It then reversed thrust, gently blowing the delighted passengers back to the doorway where they had entered.

In late 1868, the Beach Pneumatic Transit Company acquired a five-year lease on the basement of Devlin's clothing store on Broadway, and began their conspicuous construction. The details of the endeavor were kept quite secret, but the scale of the operation was evident from the large equipment outside of the building and the parade of horse carts hauling away mounds of dirt each night. Six meters beneath Broadway, a unique machine designed by Alfred Beach himself slowly gnawed a three-meter-wide passage into the Earth. The sharp end of the disk-shaped tunneling shield was arrayed with sharp horizontal shelves which tore through the earth until it fell into the tunnel through a number of openings. Using an collection of eighteen hydraulic rams, workers forced the shield forward sixteen inches, used wheelbarrows to haul away the loosened earth, erected masonry around the newly bored part of the tunnel, and then repeated the process.

On the twenty-sixth of February 1870, Alfred Beach finally exposed his secret tunnel for inspection by the public. For a fare of two cents per passenger, twenty guests at a time could take a ride on the pneumatic carriage. The custom-built, fifty-ton blower was situated in an adjacent chamber, separated from the waiting area by a long corridor. The blower was 7 meters high, 5 meters long, and 4 meters wide. It contained two colossal lengthwise paddles which rotated to draw air in through the rear and thrust it out the front. The magnificent blower was also outfitted with a special set of adjustable fan blades which allowed it to switch from sucking to blowing without reversing rotation. By tapping a telegraph wire, the conductor signaled the boiler engineer to engage the 100 horsepower steam engine. Atmospheric pressure increased by "a few grains per inch," pressing the carriage into the tunnel as the air rushed to escape through the vent at the far end. As quoted in a company booklet, a visitor described her experience on the Pneumatic Transit:

“We took our seats in the pretty car, the gayest company of twenty that ever entered a vehicle; and before we knew it, so gentle was the start, we were in motion, moving from Warren street down Broadway. In a few moments the conductor opened the door, and called out, Murray street with a business-like air that made us all shout with laughter.”

Unfortunately, excessive speculation in post-Civil-War railroads created an investment bubble which burst in 1873, triggering a severe economic depression in the US. In the wake of this calamity, investors in rapid-transit projects were nowhere to be found. In the years that followed, Beach Pneumatic Transit lost its lease on the Devlin building basement, and the tunnel’s entrance was sealed with a wall of brick. In September 1878, Alfred Beach resigned as president of the company and moved on to other endeavors, having invested over $200,000 of his own money in the ill-fated project. He died in 1896. When the building was rebuilt in 1900, the hastily-assembled brick wall was replaced with one of concrete, leaving the ventilation shaft in City Hall Park as the only way to enter the prototype pneumatic tunnel. Beach’s experimental subway lay virtually forgotten beneath the busy street until officials from the Public Service Commission paid it a visit in 1912. Their task was to organize the disassembly of the tunnel to clear the way for a new electric subway line; Beach’s vision for subterranean transit below Broadway was finally becoming a reality. Aside from the rusted rails, the tube was found in excellent repair. Beach’s pneumatic carriages were also found inside and, though they had somewhat disintegrated due to age and neglect, there was still evidence of their once-opulent decor and upholstery. Additionally, at the end of the tunnel, Beach’s innovative tunneling shield remained, its wooden teeth still sunk into the earth.

Beach’s original proposal for a network of pneumatic postal tubes also became a reality after he died. Around the turn of the century, New York City began installing hundreds of miles of medium-sized pneumatic tunnels to ferry freight between post offices. Some of these lines remained in operation until 1953. Ultimately, however, trucks proved more efficient at information-moving than the series of tubes. Many miles of these decommissioned iron transportation tunnels still linger beneath the streets of New York.

The notion of pneumatic transit was revisited in the 1960s by the Lockheed company and Massachusetts Institute of Technology, with the assistance of the United States Department of Commerce. Together these organizations conducted feasibility studies on a system of magnetically levitated tube-trains powered by ambient atmospheric pressure and “gravitational pendulum assist.” Such pneumatic vactrain technology was found to be a superior mode of transportation in many ways, not the least of which was speed - the study indicated a typical line could achieve an average velocity of 630 kilometers per hour. The system was never built due to the enormous expense of such an undertaking, although research into related technologies continues even today. Perhaps in the distant future mankind will traverse the countryside in a network of pneumatic tubes. And if that fine day ever comes, Mr Alfred Beach and his extraordinary 138-year-old experiment will finally be vindicated.

Appendix C.

Allan Bellows Facebook Permission to Adapt WBT and NDRL

Hi Rylan,

My apologies for my lack of replies...I have been stuck in that unproductive pattern of putting important emails into a "reply-to-these" folder and then forgetting that they exist.

The next few weeks are a mess for me, so I am uncertain when I'll have a chance to review the writings. But rather than hold up your progress, I'll just trust that your revisions are acceptable. Please feel free to move forward with your project with my permission...I highly doubt that any of your edits would be sufficient to cause me concern.

Best of luck, and sorry again for the delays,

Alan

Appendix D.

Approved SFU Participant Consent Form

Project: Reconceptualizing Metacomprehension Calibration Accuracy (Study # 2010s0103)

Investigators: Rylan G. Egan, Philip H. Winne

Department: Faculty of Education

The goal of this project is to obtain data on your ability to judge text comprehension after studying. You will be asked to study, read, and summarize two separate texts. Next, you will judge your ability to recall and explain information from the text. Finally, your memory and understanding of each text will be tested. Each text is between 1000-1200 words in length. Texts will be read, summarized, judged, and tested in a web-based learning environment called nStudy. nStudy is learning software that records the duration and content of materials you read and write. Prior to the experiment you will be asked to provide data about your, age, sex, faculty affiliation, level of study (e.g., undergraduate, graduate, etc.), English language proficiency, grade point average, hours since hunger was last satisfied, hours of sleep in previous night, and inclination to think deeply.

Benefits: Participation in this project will contribute to deeper understanding of study and restudy behaviour.

Risks: No risks have been identified.

I agree to the following (check if appropriate):

Release of data to investigators: demographic information, Need-for-Cognition scores, self-reported sleep and hunger measures, test scores, text summaries analyses, duration and content of all writing and reading collected by nStudy, comprehension judgments

Gratuity: $15 cash after completing all activities in the experiment. Highest score on the post study test will receive an additional $30. If there are multiple high scores entries a random draw of those with the high score will be conducted.

To ensure confidentiality, you will be provided a random identity number that will be matched to your student number and your name will not appear on any documentation other than this consent form. Data collected in this study will only be used for research and may be used in presentations and publications resulting from this research. All data will be kept for a period of three years after the completion of the research. Participant consent forms, questionnaires, and participant checklists will be kept in a locked file cabinet. Student’s written summaries, test answers, and traces of study activities (e.g., reading time, words typed, typing time etc.) will be collected on nStudy software, stored on a computer in a locked office with access limited to Philip Winne’s nStudy design/development team. This information will also be password protected.

I understand that I may withdraw my participation at any time and may register any complaint with the Director of Research Ethics, Burnaby, B.C., Canada, V5A 1S6, (Dr. Hal Weinberg,778-782-6593, email, [redacted]). Refusal to participate or withdrawal after agreeing to participate will have no adverse effects on your grades or any evaluation in the classroom or coursework. Upon withdrawal from the study all collected data will be destroyed.

I understand I can obtain copies of the results of this project upon its completion by contacting [redacted].

I certify that I have read this form and I understand the procedures to be used in this project.

Last Name: ____

First Name: _______________________________________

Email Address: ___________________________________

Phone Number (optional)______________________

First Language: __________________________________

Year of Birth: _________________

Credits Completed: _________________

Signature: ____

Date: _________________

Student Number:________________________________

Signature Witness: ______________________________

Name of Witness: __________________________________

Appendix E.

Dr. John Cacioppo (Email) Permission to Use *The Efficient Assessment of Need for Cognition*

Dear Dr. Cacioppo,

I am writing to request the use of the efficient assessment of need for cognition (Cacioppo & Petty, Kao, 1984). I am a Doctoral Candidate at Simon Fraser University in Burnaby, B.C., Canada. I am studying under Dr. Philip Winne in the Educational Psychology program. I would like to use your (and Dr. Petty's) measure for my doctoral dissertation entitled "Re-conceptualizing Metacomprehension Accuracy". The goal of my research is to find text specific influences on individuals' ability to accurately judge their understanding of text. I feel that one's ability and desire to think deeply may be one such variable.

Thank you kindly for considering my request.

All the best,
Rylan Egan

No problem, you can use the scale for research purposes.

All the best,
John
--

John T. Cacioppo, Ph.D.
Tiffany and Margaret Blake Distinguished Service Professor, and
Director, Center for Cognitive and Social Neuroscience
The University of Chicago
5848 S. University Avenue
Chicago, IL 60637
Email: [redacted]
Phone: [redacted]

Administrative Assistant:
Angela McCoy
Email: [redacted]

Phone: (773) 834-7458
Fax: (773) 702-4580

Society for Social Neuroscience (http://S4SN.org)
Center for Cognitive and Social Neuroscience (http://ccsn.uchicago.edu)
Homepage
(http://psychology.uchicago.edu/people/faculty/cacioppo/index.shtml)

Appendix F.

NDRL Detailed Questions

Detailed Questions
(Answer all 10 items)

A specific answer is required for each question below. Your response should be short and provide only information required by the question. If you don’t know the answer to the question make your best guess. Spelling and grammar will not be marked.

Scroll Down To Answer ALL Questions

1. Why were reports of diseases resulting from radiation near Chelyabinsk-40 inaccurately reported in the early 1990s?

Answer:

2. What is the most significant concern at Chelyabinsk-40 today?

Answer:

3. What happens to nuclear waste when it sits in still water for a long time?

Answer:

4. Why were nuclear materials handled unsafely at Chelyabinsk-40?

Answer:

5. Which country is currently concerned about radioactive materials at Chelyabinsk-40?

Answer:

6. What caused the vat inlet pipe seals to break?

Answer:

7. What caused the nuclear waste storage vats to overheat?

Answer:

8. What triggered the explosion at Chelyabinsk-40?

Answer:

9. What materials were used to seal nuclear materials into Lake Karachay?

Answer:

10. Where in Lake Karachay was radiation recorded at 600 Rotgens per hour?

Answer:

When finished press save and then press the red circle in the top left hand corner of this window.

Appendix G.

WBT Detailed Questions

Detailed Questions
(Answer all 10 items)

A specific answer is required for each question below. Your response should be short and provide only information required by the question. If you don’t know the answer to the question make your best guess. Spelling and grammar will not be marked.

Scroll Down To Answer ALL Questions

1. Which materials were recommended for lining the inside of pneumatic train tunnels?

Answer:

2. What caused above ground pneumatic transportation to fail?

Answer:

3. What was special about the fans used at the first demonstration of pneumatic transportation in America?

Answer:

4. Why wasn't the pneumatic tunnel accessible after the Beach Pneumatic Dispatch Company failed?

Answer:

5. How were the Beach Pneumatic Dispatch Company's rail cars described when they were recovered after being abandoned for 34 years?

Answer:

6. The pneumatic-powered prototype at the Crystal Palace Exhibition was not built. Why?

Answer:

7. How would new trains designed by scientists at MIT and Lockheed be lifted off the ground?

Answer:

8. What was the most common form of transportation when George Medhurst wrote his pamphlet on pneumatic transportation?

Answer:

9. Why was the pneumatic prototype in New York abandoned?

Answer:

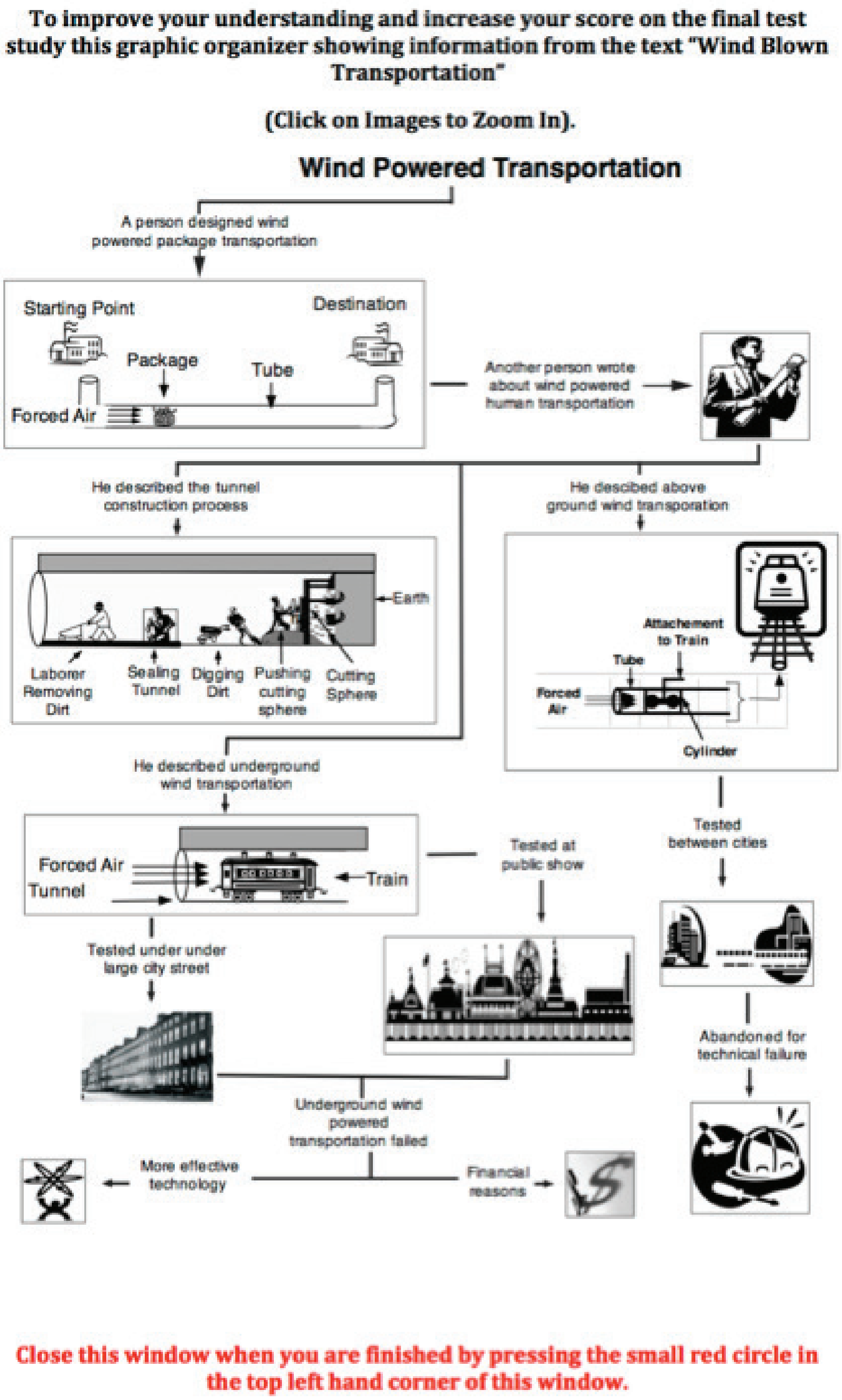
10. Why were New York City’s pneumatic postal tubes retired in 1953?

Answer:

When finished press save and then press the red circle in the top left hand corner of this window.

Appendix H.

WBT Graphic Organizer



To improve your understanding and increase your score on the final test study this graphic organizer showing information from the text "Wind Blown Transportation"

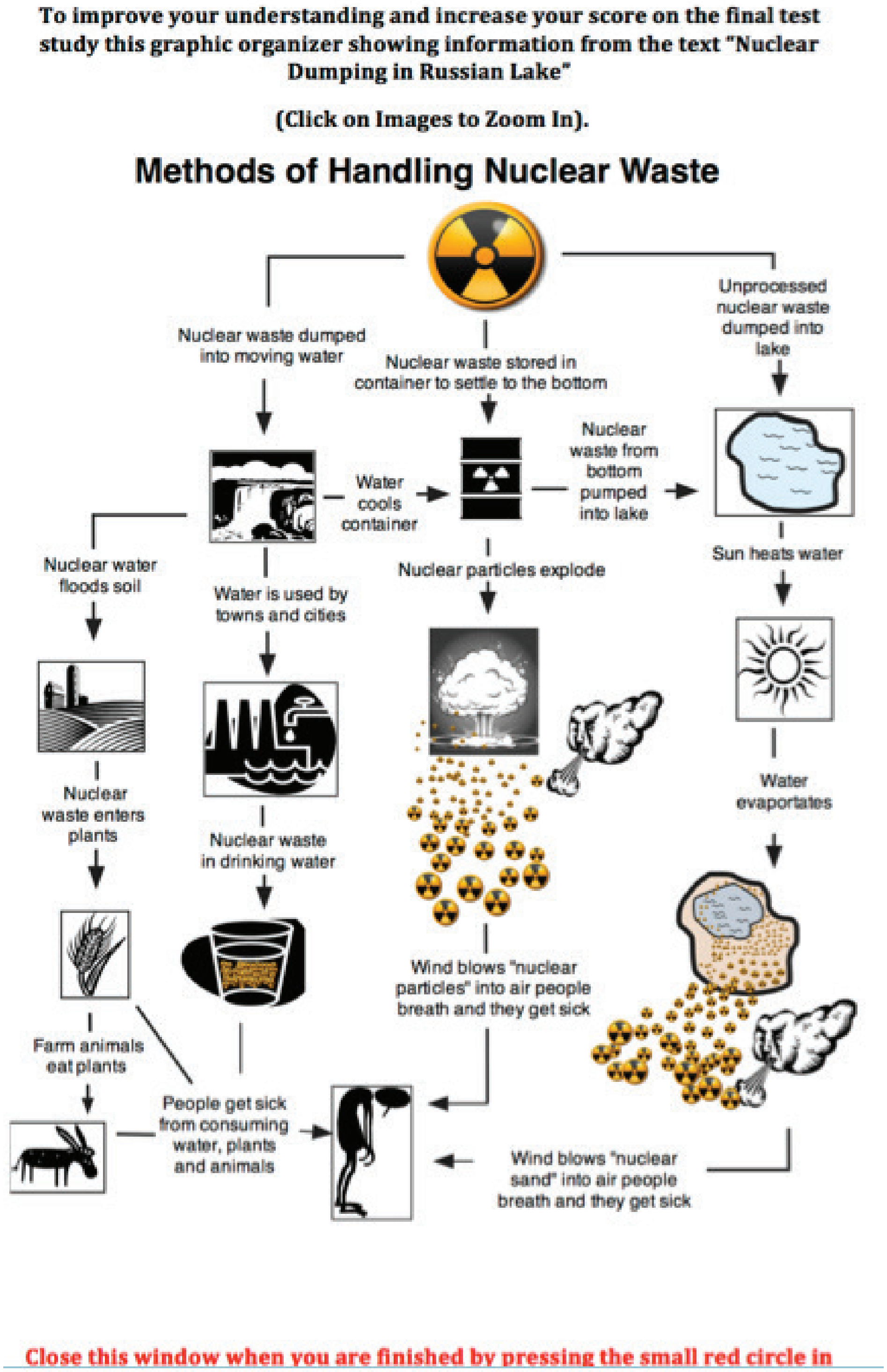
(Click on Images to Zoom In).

Wind Powered Transportation
A person designed wind powered package transportation
Starting Point
Destination
Package
Tube
Forced Air
Another person wrote about wind powered human transportation
He described the tunnel construction process
He descibed above ground wind transporation
Earth
Laborer Removing Dirt
Sealing Tunnel
Digging Dirt
Pushing cutting sphere
Cutting Sphere
Attachement to Train
Tube
Forced Air
Cylinder
He described underground wind transportation
Forced Air
Tunnel
Train
Tested at public show
Tested between cities
Tested under under large city street
Abandoned for technical failure
Underground wind powered transportation failed
More effective technology
Financial reasons

Close this window when you are finished by pressing the small red circle in the top left hand corner of this window.

Appendix I.

NDRL Graphic Organizer



To improve your understanding and increase your score on the final test study this graphic organizer showing information from the text "Nuclear Dumping in Russian Lake"

(Click on Images to Zoom In).

Methods of Handling Nuclear Waste

Nuclear waste dumped into moving water
Nuclear waste stored in container to settle to the bottom
Unprocessed nuclear waste dumped into lake
Water cools container
Nuclear waste from bottom pumped into lake
Nuclear water floods soil
Water is used by towns and cities
Nuclear particles explode
Sun heats water
Nuclear waste enters plants
Nuclear waste in drinking water
Water evaportates
Wind blows "nuclear particles" into air people breath and they get sick
Farm animals eat plants
People get sick from consuming water, plants and animals
Wind blows "nuclear sand" into air people breath and they get sick

Close this window when you are finished by pressing the small red circle in

Appendix J.

WBT Pre-Reading Facts

To improve your understanding and increase your score on the final test memorize each of these facts from the text "Wind Blown Transportation".

1. George Medhurst wrote a pamphlet containing the first description of wind powered transit.
2. A passenger on the first wind powered train in America described the ride as fast and gentle.
3. Engineers tapped a telegraph wire to signal that fans should be started to move wind powered trains.
4. Telegraph transcriptions were shuttled between London offices and the stock exchange.
5. Risky investments after the American civil war caused America's first wind powered transportation company to fail.

CLICK HERE WHEN FINISHED

Appendix K.

NDRL Pre-Reading Facts

To improve your understanding and increase your score on the final test memorize each of these facts from the text "Nuclear Dumping in Russian Lake".

1. In the first year of nuclear production Russia diluted materials and pumped them into the Techa River.
2. Effects of chronic radiation syndrome include tremors and low blood pressure.
3. Nuclear dumping from 1950-1970 dwarfed the contamination of more recent Russian nuclear disasters.
4. Livestock were exposed to nuclear contamination from crops grown on radioactive flood plains.
5. A person standing near Lake Karachy would have died of radiation within 1hour in 1991.

CLICK HERE WHEN FINISHED