ABSTRACT

Title of Dissertation: READING COMPREHENSION AND ITS ASSESSMENT: ALIGNING OPERATIONALIZATION WITH CONCEPTUALIZATION OF THE CONSTRUCT

Taslima Rahman, Doctor of Philosophy, 2013

Directed by: Professor Patricia A. Alexander
Department of Human Development

The current study explored ways to improve reading comprehension assessments. Available assessments appeared misaligned with views of comprehension that are emerging in the reading research literature. Further, the measurement models as currently applied to comprehension assessment do not take into account the cognitive perspective of the construct when estimating proficiency. It has been argued that an assessment, an evidentiary argument, when based on a theory of the construct can offer more informative estimates of proficiency and improve validity of the inferences drawn from those estimates (Mislevy, Steinberg, & Almond, 2003).

For this study, the design and the analytic approach for an assessment of comprehension were grounded on a premise that comprehension is influenced by task attributes (e.g., text type or target mental representation) as well as reader attributes (e.g., prior knowledge or interest). Construction of the comprehension measure and the ensuing psychometric analyses were framed following Kintsch's (1998) Construction-Integration model and Alexander's (1997) Model of Domain Learning.

The resulting measure was administered to 160 eighth-grade students with no known status of receiving services for special education. In completing the comprehension task, the students read four text passages and answered a set of text passage-related questions, eight per passage. Those passages varied by text type and text topic, and questions varied by the target mental representations of a text and relations among the events of a situation described in the text. In addition, participants answered a set of questions for self-reporting about their familiarity with and interest in the topic of a text passage that they had read.

In synthesizing the data, a particular form of the Linear Logistic Test Model introduced by Fischer (1973) was applied within a Bayesian framework. When the attributes were incorporated in the measurement model, the comprehension proficiency estimates changed in a way that reflected positive effects of topic familiarity and topic interest. Further, the task and reader attributes considered in the study contributed to estimates of item difficulties. Thus, the study, based on empirical evidence, suggests that developing a comprehension assessment more aligned with views of the construct offered in the literature is indeed viable.

READING COMPREHENSION AND ITS ASSESSMENT: ALIGNING
OPERATIONALIZATION WITH CONCEPTUALIZATION OF THE CONSTRUCT

By

Taslima Rahman

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2013

Advisory Committee:

  Professor Patricia A. Alexander, Chair
  Professor Mariam J. Dreher
  Professor Jonna M. Kulikowich
  Professor Robert J. Mislevy
  Professor Kathryn R. Wentzel

DEDICATION

I dedicate this work to my departed father who believed that education is the only means to eradicate a society's poverty.

I would like to add that I believe reading comprehension is the key for education.

ACKNOWLEDGEMENTS

By completing this dissertation, I fulfilled a dream passed on to me by my father. Over the years while striving to attain my dream, I met many scholars and found many friends. I would like to remember them here because without their support my dream would not have come true.

First and foremost, I would like to express special gratitude to my esteemed advisors for their guidance during my conceptualization to completion of this study. I felt blessed when Dr. Patricia A. Alexander accepted me as her student. You taught me even many basics—the importance of tracing back a thought, juxtaposing thoughts, and interlacing sentences and paragraphs. I felt honored when I received permission from Dr. Robert J. Mislevy to enroll in his class. You taught me the art of visualizing relations among elements of a "black box" to make it transparent to others.

I am deeply grateful to Dr. Mariam J. Dreher, Dr. Jonna M. Kulikowich, and Dr. Kathryn R. Wentzel for graciously accepting my request to be in my advisory committee and for enriching this dissertation with their thoughtful comments and suggestions.

I am indebted to all who reviewed the measures developed for this study. A special thanks goes to Dr. Eileen Kintsch and Dr. Joseph Magliano for their guidance in constructing the comprehension items included in the study. With your input, it was possible for me to ensure that the comprehension items followed the framework adopted in the study.

I appreciate members of the Disciplined Reading and Learning Research Laboratory (DRLR) for their comments and suggestions on my presentations prepared at various stages of the study. More importantly, I would like to thank the children who participated in the study. Without them and cooperation from their parents, teachers, principals, and school districts, this study would not have been possible.

I am also grateful to my colleagues, friends, and family. I am grateful to my supervisors at work for their belief in me and inspiration in pursuing my goal.

I cannot stop here without remembering my friends who made themselves available when I was trying to respond to my need that seemed endless. You were there when I was verbalizing my thoughts, checking grammar, resolving Microsoft issues, seeking cooperation from schools and parents, making sure that I completed requirements for graduation, and when I simply needed a shoulder to lean on.

Last but not least, I would like to thank my husband, as he has been supportive throughout this long endeavor. You urged me to go on while I was in despair, encouraged me while I had my shining moments, and were always so accommodating. I want to extend my love to my mother for her prayers and emotional support. Finally, I thank almighty God for giving me the strength to overcome hurdles and the opportunity to enjoy success.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Reading comprehension, a precondition to learning from text, has long been a

subject of educational assessment in the United States (Pearson & Hamm, 2005; Segel,

1935; Smith, 1965). Reading comprehension assessment, nonetheless, has been a source

of dissatisfaction throughout its history. In 1921, Arthur Gates stated, "The present tests

are useful but not perfect instruments. We need tests constructed with such care that the

numerous defects found in the tests now in existence shall be avoided" (p. 464). About

80 years later, the RAND Reading Study Group ([RRSG]; 2002) panel stated, "currently

used comprehension instruments are all unsatisfactory in various ways" (p. 112).

Discontentment over tests administered to assess the comprehension ability of U.S.

students has also been voiced by many researchers (e.g., Farr & Carey, 1986; Grabe,

1999; Johnston, 1984; Keenan, Betjemann, & Olson, 2008; Magliano, Millis, Ozuru, &

McNamara, 2007; Sarroub & Pearson, 1998; Sweet, 2005; Valencia & Pearson, 1987;

Valencia, Pearson, Peters, & Wixson, 1989). A theme that resonated in those expressions

of discontentment is that comprehension tests are not grounded in a clearly articulated

conceptualization of the construct consistent with the reading research literature.

Further, the interpretations and implications of reports published about results

obtained from reading comprehension assessments are troublesome. The National

Assessment of Educational Progress (NAEP), a congressionally mandated assessment

program, reported that only about one third of the nation's 4th- and 8th-grade students are

proficient in reading (Lee, Grigg, & Donahue, 2007). Experts were not sure where to

point as to the nature and causes of, or prescriptions for students' difficulties, because

reporting of the comprehension results does not delineate what students read, what they were asked to do with what they read, or the factors that might have influenced comprehension of what was read (e.g., Baltimore Sun, July 2008; Education Week, April 2001). In addition, there are inconsistencies in the results reported from tests developed by different education agencies. For instance, the percentages of the students in grades 4 and 8 reported to be read well or proficient in the academic year 2006-07 based on different state assessments, ranged approximately from 25% to 91% (Bandeira deMello, Blakenship, & McLaughlin, 2009). These problems in interpretations or inconsistencies in the results occur in part because the different comprehension tests that emerged do not necessarily reflect a view of comprehension common among the test specialists and/or a view of comprehension supported by empirical studies or theoretical treaties in the reading literature.

Despite these problems, the importance of comprehension testing has been recognized for almost a century. Brown (1914), deputy superintendent of New Hampshire's Department of Education, stated: "There is no subject in the school curriculum in which there is a greater need that tests be devised with which to measure the accomplishment of children than reading" (p. 477). In addition, students in the U.S. have been assessed for comprehension ability since 1916 when Kelly published the Kansas Silent Reading Test, considered the first standardized reading comprehension test (Farr & Carey, 1986; Smith, 1965). Since 2000, the assessment of comprehension ability has become a regular practice by law. Public school students in grades 3 through 8 are required to be assessed annually by all U.S. state education agencies per the federal law, the Elementary and Secondary Education Act (ESEA, reauthorized in 2002). In addition,

samples of students at selected grades attending the public and nonpublic schools across the nation are assessed every two years by NAEP. The students in U.S. also participate every four to six years in comprehension assessments administered in many other countries, for example, Progress in International Reading Literacy Study (PIRLS) or Program for International Student Assessment (PISA).

Information derived from reading testing is useful for a variety of reasons. For one, these data can be used in appraising students' strengths and weaknesses in order to determine the levels of support students require for reading (Caldwell, 2002; Traxler, 1970); supplementing a teacher's judgment of individual students' performance in reading (Judd, 1914), determining efficiency of instruction (Brown, 1914), and directing the planning of reading instruction (Traxler, 1970; Valencia et al., 1989). In current educational practices, information from testing is applied not only to evaluate what individual students know and can do, but is also considered when determining the success of a state, district, or a school (Afflerbach, 2007; Tyler, 1983). Further, comprehension testing is not only necessary to support student learning (Brown, 1914; Durkin, 1993; Pearson & Johnson, 1978; RRSG, 2002), but it also serves as a tool for research that can advance our understanding of this construct (RRSG, 2002; Robinson, Farone, Hittleman, & Unruh, 1990; Smith, 1965).

The importance of comprehension testing is predicated on the importance of the domain of reading. Reading has been identified as "the greatest tool for thought-getting which an individual may have at disposal" (Brown, 1914, p. 478). According to educator and reading researcher Nila Smith (1965): "Reading was the most important subject in…schools and it has continued to be the most important subject all through the years"

(p. vii). Many others agree with Smith's proclamation (e.g., Carroll, 1972; Davis, 1964; Gibson & Levin, 1975; Goodykoontz, 1937; Lorch & van den Broek, 1997). The lack of reading skills is considered by many teachers to be "the most common handicap to learning" (Beckner & Cornett, 1972, p. 313). Reading is considered a medium through which one is initiated into the language of a community (Hampton & Resnick, 2009). More importantly, "the ability to read allows one to navigate a world in which so much of interest and importance is conveyed through written language" (Alexander, 2005, p. 414).

In the domain of reading, comprehension is regarded as the "essence of reading" (Durkin, 1993, p. 12; also see National Reading Panel report, 2000). The view of comprehension as the core of reading has been longstanding. For instance, John Russell Webb, a 19th century school teacher who has been credited with introducing the concept of comprehension into the U.S. classrooms (Robinson et al., 1990; Smith, 1965), said that children should be reading for meaning, and he claimed, "a full comprehension of the matter to be read" as a criterion for determining if one is reading well (1856, p. 5). The ability to make sense of texts in its many forms is essential to acquiring the knowledge and skills in a number of domains and this belief holds as much weight today as it did during the two previous centuries (e.g., Hampton & Resnick, 2009; Webb, 1858; Whipple, 1925). Pearson and Johnson (1978) stated that helping a student to comprehend the written words is like giving a gift that "lasts a lifetime and makes that lifetime more worth living" (p. 2).

As we approach the 100th birthday of the first U.S. publication of the standardized reading comprehension test, the mode of test delivery has already entered a new era. Frequent use of computers in classrooms, without a doubt, may well make the

paper and pencil mode of assessment delivery passé. Clear direction is necessary for the construction of future comprehension tests for delivery via computers. Moreover, tests need to assess comprehension the way research has outlined; otherwise tests may misguide instruction, as argued by Valencia et al. (1989):

> Thus, if we equate high scores on existing tests with good reading, we may be led to a false sense of security. Conversely, low scores may lead us to believe that students are not reading well when, by a more valid set of criteria, they are. (p. 57)

Statement of the Problem

One may ponder the source of discontentment with comprehension tests. Based on the extant literature, comprehension testing is considered a case of misalignment. That is, tests have been regarded as inconsistent in the representation of the construct and they are considered inadequate in the representation of the complexity of the construct affirmed by reading research (e.g., RRSG, 2002; Sarroub & Pearson, 1998; Valencia et al., 1989).

Reading comprehension has many components (e.g., reader, text, activity, and a situation); each component has a role in comprehension of what is being read; and every component entails many attributes (RRSG, 2002). For example, a reader brings his/her experiences and knowledge when asked to read a text. A text passage or what a reader reads is characterized by its content, structure, the language, and level of coherence. Likewise, an activity, or what a reader is asked to do to demonstrate comprehension, can be a simple recognition of what is explicitly stated in a text or entail answering a question that requires reasoning through the given information and integration of information

given in sentences and paragraphs and with what the reader already knows. The

attributes of each of the comprehension components, individually and collectively, are

found to make a difference in the comprehension of a text (Alexander & Fox, 2004;

Duke, 2005; Francis, Fletcher, Catts, & Tomblin, 2005; Graesser, 1981; van den Broek,

Young, & Tzeng, 1999). If a comprehension assessment does not take into consideration

the complexity of the construct, the results obtained by the assessment may not be a valid

representation of what students know and can do.

Achieving a better alignment between reading comprehension tests and a viable

theory of reading comprehension requires a framework to relate the design and analysis

of tests to the theory. An assessment is an evidentiary argument because the claims we

make about what students know and can do are based on observations of a sample

performance in a time-constrained situation (Mislevy et al., 2003). An evidentiary

argument about comprehension manifests misalignment when decisions about the major

components of the measure and the assessment claims are not explicitly articulated and

guided by a valid view of the construct. For instance, misalignment can occur if testing

views comprehension as a set of discrete skills and assuming that those skills when

aggregated would represent comprehension, as opposed to the leading view that considers

comprehension as a complex cognitive process that involves integration of various skills.

To state the problem more specifically, comprehension tests are not being

designed to reflect a view of the construct that is grounded in the empirical and

theoretical research (e.g., Farr & Carey, 1986; Grabe, 1999; Johnston, 1984; Sarroub &

Pearson, 1998; Sweet, 2005). Further, little effort has been made to integrate the

cognitive theory of comprehension into the measurement model applied to synthesize the

comprehension performance observed in an assessment. In other words, the attributes of the components that are said to influence comprehension are not taken into consideration when designing a test and when defining the psychometric approach applied to analyze the evidences provided by the comprehension test. The psychometric approaches, Classical True Score Theory (CTT) and standard Item Response Theory (IRT), currently applied in most cases for summarizing comprehension performances do not take into account the cognitive complexity of the construct when estimating comprehension proficiency. If the influences of the attributes of comprehension components are not considered, the validity of the inferences derived about a student's comprehension ability is in question.

In sum, the fact that reading comprehension tests are not being grounded in a viable model of comprehension is problematic for several reasons. If tests do not assess what has been defined as comprehension, then the test results cannot validly be used to determine students' ability to comprehend, detect the problems in comprehension failure of the students, or determine any progress those students might have made toward the goals set for them (Valencia et al., 1989).

Purpose of the Study

The current study was undertaken to investigate the viability of developing an assessment of reading comprehension following a conceptualization of the construct posited in the reading research literature. More specifically, the intent was to operationalize a theoretically grounded view of comprehension when conducting an assessment. This operationalization entails framing the design and the analytic approach for an assessment with a theory of the construct.

To this end, the current study was guided by the following two questions:

1. How can a reading comprehension measure be designed that supports a view that comprehension is influenced by the attributes of task and reader, the two key components of comprehension?

2. How can a psychometric model for such a comprehension measure be defined that can take into account the task and reader attributes when estimating item difficulties and reader comprehension proficiencies?

In other words, the purpose of the study was to develop an exemplar for reading comprehension assessment by building jointly on reading research and psychometric modeling. Thus, the work demonstrates methodological as well as reading-research aspects.

To achieve this overarching goal, four major tasks were undertaken in conducting the study. The first task involved reviewing a set of comprehension assessments and compiling information on how the construct, comprehension, was operationalized in those assessments. This task also involved reviewing the literature on reading comprehension to compile information on how this construct has been conceptualized.

The second task entailed constructing an assessment measure that was framed by conceptualization of comprehension offered in research, specifically Kintsch's (1998) Construction-Integration model and Alexander's (1997) Model of Domain Learning. Drawing from the CI model and MDL, it was posited that what a reader brings in terms of prior knowledge and interest to a reading of a text passage that can vary in type and topic influences the reader's comprehension of the text. Since a reader's mental representation of a text is not readily evident to observers, the reader's text representation

generally is gauged through what understanding of the text he or she demonstrates. This conceptual model of comprehension, as shown in Figure 1, was applied in developing the assessment measure. Accordingly, the second study task entailed two subtasks. They were: (a) constructing a comprehension task where a set of text passages was selected and sets of questions were prepared for the text passages that shared the same attributes across the selected texts; and (b) constructing a measure to gather information on reader attributes considered pertinent to the comprehension of a text passage, specifically, topic familiarity and topic interest.

The third task entailed determining a procedure for gathering information on the comprehension of the target population, 8th-grade students. This task also involved gathering comprehension performance data from a sample of the target population.

The final task entailed defining the psychometric approach for synthesizing the resulting data. This task also involved applying a more advanced psychometric approach known as the Linear Logistic Test Model, a cognitive-based IRT model introduced by Gerhard Fischer (1973), in synthesizing the obtained data. In this psychometric model, text passage-related reader attributes were incorporated along with task attributes. These attributes were taken into consideration when estimating student comprehension proficiencies.

In order to demonstrate the use of the psychometric modeling approach, the analyses of the data addressed the following two research questions:

1. In what ways do the attributes of task and reader influence the comprehension of the 8th-grade students who participated in the study?

2. Do the task and reader attributes considered in the study contribute to the item difficulties and reader comprehension proficiencies?

The first question, addressed by means of a statistical approach, was a prerequisite for the second question, which was addressed by means of a psychometric approach. Viability is warranted when the patterns of effect of the attributes, as revealed in the psychometric approach, are in accord with those seen by the statistical approach, and are in line with the reading literature, which is the foundation of theoretically grounded comprehension assessment.
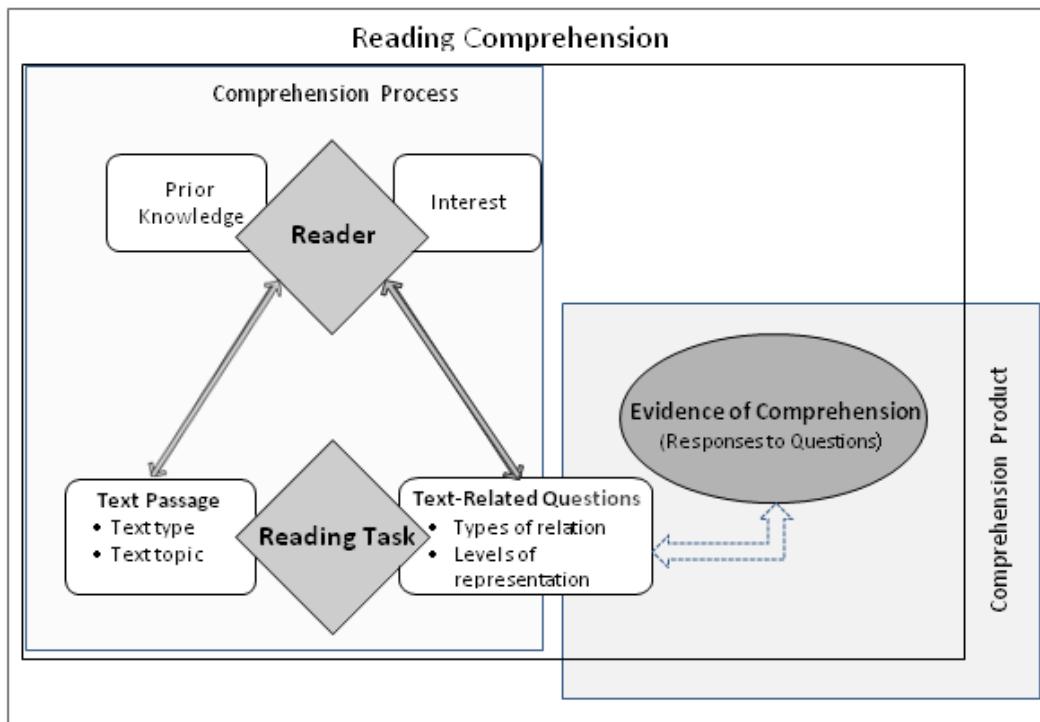
Figure 1. A Conceptual Model of Reading Comprehension Posited in the Study

Definitions of Terms

In order to discuss and answer the aforementioned research questions, it is necessary that the following constructs are broadly defined.

An assessment, according to the Evidence-Centered-Design (ECD) principle, is viewed as an evidentiary argument—"an argument from what we observe students say, do, or make in a few particular circumstances, to infer what the students know, can do, or have accomplished more generally" (Mislevy, Steinberg, & Almond, 2003, p. 1). The terms, assessment and testing were used interchangeably in this document to refer to the device and the relevant procedures applied in obtaining information in order to draw inferences about one's learning.

The term *comprehension* concerns a reader's representation of a text. The term can be used to refer to both the cognitive process by which comprehension occurs and the product of this process. According to Kintsch (1998), comprehension involves a cyclical process of construction and integration of text elements. Since a reader's representation of a text is not obvious, the mental representation of a text generally is gauged through a variety of activities that the reader is asked to perform, such as recall of what is read or answers to questions about what is read. Attributes of texts (Duke, 2005; Grasser, 1981), as well as readers (Alexander & Fox, 2004; van den Broek et al. 1999) are claimed to make a difference in the meaning a reader derives from a text. In this study, the cognitive items that were answered correctly served as evidence of a reader's comprehension.

A *causal relation* is a two-part relation between events. In this two-part relation, one event, a cause or an antecedent that brings about the other event, an effect or a consequence. Events can be causally connected in a variety of ways (Mislevy & Rahman, 2009). A non-causal relation could be a spatial relation or a temporal relation. A spatial relation indicates the spatial setting or context where an event, action or a state took place. Temporal relation indicates the time interval when an event, state, or action

took place. Causal, spatial, and temporal relations are represented in situations portrayed in narratives (Rinck, Hahnel, Bower, & Glowalla, 1997; Zwaan, Magliano, & Graesser, 1995), as well as in expository texts (Graesser, McNamara, & Louwerse, 2003; Millis & Cohen, 1994). The comprehension questions about text passages constructed for the study's measure asked for these relations.

Expository text and *narrative text* are two general types of text referred to in the reading research and comprehension assessment. According to Brewer (1980), these two broad categories vary in organization and in the intent of the information presented in texts. In general, an expository text explains, informs, or describes an idea, event, or a problem, whereas a narrative text, using fictional or non-fictional information tells a story or recounts an event or feeling. Texts of these two types were included in the measure. Many varieties of text fit into these types. In the study, the texts included as expository type were of informational in nature and as narrative type, were recount of events.

The term *prior knowledge* refers to "readers' familiarity with the content" (Alexander & Jetton, 1996, p. 99). It has also been referred to "the sum of what an individual knows" (Alexander, Schallert, & Hare, 1991, p. 333). Topic familiarity is a form of prior knowledge (Spyridakis & Wenger, 1991). Topic familiarity represents a student's reported familiarity with the topic discussed in a particular text (Ainley, Hidi, & Berndorff, 2002; Graesser & Bertus, 1998; Schraw, 1997). Topic familiarity reported by study participants was used as a reader attribute when analyzing comprehension performance.

Textbase and *situation model* are two levels of text representation identified by Kintsch (1998). Textbase, a lower level of representation, is a text-driven mental

representation of the text, and situation model, a higher level of representation, is a blend of text-driven and knowledge-driven representation of the text (Kintsch, 1998). A reader's understanding of a text—the situation model—is formed when the reader integrates information from a given text with his or her knowledge. The comprehension questions about text passages constructed for the study measure probed these two forms of representation of a text.

A *theta* represented by the Greek letter, θ in Item Response Theory (IRT) "characterizes an examinee's propensity to answer an item correctly rather than incorrectly" (Mislevy & Huang, 2006, p.7). Theta represents a value on a scale referred to as the ability scale. An IRT model posits that an examinee has an underlying "ability" that determines his or her probability of answering an item correctly. An examinee is assigned a θ using the examinee's responses to a set of items. However, the term ability does not connote something that is "fixed, innate, or immutable" (Hambleton, Swaminathan, & Roger, 1991, p. 79). Thus, the IRT ability can change over time. The term ability is used to describe what it is that the test items measure. An examinee's ability score, θ and true score are monotonically related.

Further, IRT ability score θ is not dependent on the items that the examinee has answered or the group who answered those items (See Hambleton et al., 1991 for a more detailed description of the ability score). Sometimes, the IRT ability score is referred to as a proficiency score. However, the term proficiency in IRT may not have the same connotation as the term used by others. In the context of reading comprehension, an IRT θ is operationally defined as the propensity to correctly answer the particular kinds of items in the test, under the conditions of testing, without accounting for particular

influences of a reader's background knowledge, interest, attention, and so on. It is a summary measure of patterns in the item responses at hand.

Topic interest is a form of interest referred to in the context of text reading (Hidi, 2000; Renninger, 2000; Schiefele, 1996). It represents a student's reported interest in the topic discussed in a particular text (Alexander & Jetton, 1996). Topic interest reported by study participants was used as another reader attribute when analyzing comprehension performance.

Significance of the Study

The current study is expected to improve the assessment of comprehension ability and advance our understanding of the construct, reading comprehension. With respect to the comprehension assessment, the study is expected to contribute in the following ways.

This study is expected to show how to design and construct an assessment measure that integrates a conceptualization of the construct that is consistent with the reading literature. This contrasts with current practices in which task development is based on a test specification table that typically concerns coverage of a content area in a test and surface features of tasks included in the test.

Further, the current study is expected to suggest ways to define the psychometric model in order to integrate the conceptualization of the construct. In current practices, where standard IRT is used, the psychometric models do not associate reader or text attributes with comprehension ability. Thus, the current study is expected to put forward a new way to design comprehension tests in order to improve test-construct alignment.

With respect to reading research, this study is expected to enrich our understanding of the construct of comprehension in the following ways.

The study is expected to enhance our understanding of the effects of multiple attributes of the key comprehension components, that is, task and reader. More specifically, with an application of two analytic approaches, statistical and psychometric, the study is expected to enhance our understanding of (a) how reader attributes—a reader's topic familiarity and topic interest—influence comprehension; (b) how the task attributes—text type, text topic—influence comprehension; and (c) whether students can attain different levels of comprehension and whether they can identify different relations among events of a situation, namely, temporal, spatial, and causal which are necessary to derive meaning of what is read. This is unlike current practices where the various comprehension components and their attributes are often explored in isolation.

Limitations of the Study

While the present study attempts to overcome shortcomings involved in comprehension assessment, there are some limitations. For instance, the student sample was comprised of a relatively small voluntary sample. Further, the number of text passages and number of cognitive questions included in the measure were limited. This limitation was set to minimize the burden on students. In addition, the information on readers' attributes of topic familiarity and topic interest was based on the students' self-report and self-estimation and no other information were available to validate what students reported. Thus, the results are not meant to be representative, but rather suggestive.

CHAPTER II

REVIEW OF RELEVANT LITERATURE

The purpose of this chapter is to provide a conceptual framework for each of the two research questions identified in Chapter I. The literature review consists of three major sections. The first section presents a snapshot of reading comprehension assessment. The second section presents an overview of conceptualization of comprehension. The last section presents a description of the measurement methodology that was applied to integrate the view of comprehension into the measurement model when estimating item difficulty and comprehension proficiency.

Reading Comprehension Assessment

The review of reading comprehension assessment presented in this section is based on a set of tests. The tests were selected from two eras: when testing began (1916) and when testing practice was mandated in the U.S. educational system per a federal law (2001). The review of tests from the two eras provides a broader picture on the general trend in the operationalization of comprehension. This literature review concludes with a discussion on why tests should be examined vis-à-vis theoretical models of the construct, as opposed to the traditional approach in which the focus is more on the correlates of test scores when evaluating the quality of a test.

For the earlier era, all tests that emerged before 1920 and used silent reading to measure comprehension were considered for selection. For the contemporary era, the selection aimed at breadth of coverage of various comprehension tests. Tests that have been used after 2000 for measuring comprehension, sponsored by different education agencies, devised for reporting performance at various levels (e.g., individual, state, or

national), and meant to address various purposes (e.g., determining progress, making policies, resolving research issues, or counseling) were considered for the selection. Multiple tests sponsored by similar agencies, for example, state education agencies, were also selected. These states, located in different regions of the country, had different scores on the NAEP reading tests. The tests reviewed from those two eras are listed in tables in the relevant sections.

When It Began

Reading comprehension testing had its start as a component of intelligence testing. George Romanes, a Canadian biologist became interested in reading comprehension as a way to understand individual differences in intelligence. His measure of adults' silent reading rate, published in 1884, is considered as the first comprehension test by some (e.g., Smith, 1965; Venezky, 1977; Willis, 2008). A few years later, in 1895, psychologist Alfred Binet developed an intelligence test, considered the first such test in the world, in which he incorporated items assessing reading comprehension (Freeman, 1939). Since the time of Binet, many intelligence tests (e.g., WAIS-R, K-ABC; see Ebel, 1972 for a description) have been developed in the U.S. that included tasks that asked examinees to complete a sentence, interpret a sentence, or answer questions after reading a short text passage. However, the comprehension measures in intelligence tests have been combined into a category labeled, in general, verbal ability or verbal reasoning (Ebel, 1972).

Comprehension testing as such began in the U.S. when the mode of reading instruction shifted from oral to silent reading in the early 20th century (Pearson, 2000; Smith, 1965). Such testing came into being much later than testing of intelligence or

testing of other learning domains. This delay, some believed, was an outcome of uncertainty among test developers in not knowing how to represent comprehension (e.g., Gates, 1921; Smith, 1965).

The comprehension tests that emerged in the early 20th century, as shown in Table 1, addressed different goals and reflected different points of view regarding comprehension. It was the purpose underlying the test development that determined the conceptual framework for operationalizing the construct. For example, Kelly (1916) developed Kansas Silent Reading Test to address the needs of teachers who wanted to know how successful they were in fostering their students' ability to get meaning from texts. In designing the test, Kelly stressed the administration and scoring procedures in order to measure comprehension in an objective manner, so that any trained teacher could administer the test. He asked test-takers to read only sentences and select a response from the given choices.

At the same time, Starch (1915) developed a test to provide a growth profile of reading capacity. He asked test takers to read their grade appropriate text passage and a passage appropriate for a grade below or above theirs so he could make comparisons across grades using the same text passages. The examinees taking Starch's test were asked to reproduce thoughts in writing immediately after reading. Similar to Kelly, Starch did not explicitly define comprehension.

Table 1

Reading Comprehension Tests That Emerged Before 1920

Test by Author	Test Description
Brown Silent Reading Test Brown (1916)	**Goal:** Examine comprehension development in students in grades three through high school. **Content:** Used connected paragraphs (e.g., short story, a description or an exposition), connected passage (passage length extended up to 800 words). **Task:** Students were asked to reproduce in writing what was read. **Index of comprehension:** Comprehension score represented quantity (percentage of total number) and quality (percentage that were entirely correct) of thoughts reproduced. In addition, considered the rate at which the students read silently. In evaluating the reproductions, used a key enumerating the essential ideas. **Point of view:** Comprehension is the ability to interpret and remember the thoughts conveyed in a text.
Kansas Silent Reading Test Kelly (1916)	**Goal:** Develop a scale for measuring comprehension in an objective manner in order to address teachers' need to know how effective they were in their teaching. **Content:** Used sentences. **Task:** Students were asked to complete 16 items in five minutes; Examples of item: • Draw a line around the name of each animal that is useful on the farm: Cow Tiger Rat Wolf [where cow is the only correct answer], • Circle to indicate: Mary is older than Nellie, Nellie is older that Kate. Which girl is older, Mary or Kate? **Index of comprehension:** Comprehension score represented an aggregate of the weighted correct answers. Each item was assigned a value indicating the time length an average student in a given grade needed to answer the item correctly. For example, a value of 2.1 was assigned to the above-described second item for students in grades 9 through 12. Each answer was considered as either wholly correct or wholly incorrect. **Point of view:** Comprehension is the ability to get meaning without any confounding from the ability to reproduce, vocabulary knowledge, and lack of information available to the examinees.

Table 1 (*continued*)

Test by Author	Test Description
Monroe Silent Reading Test Monroe (1918)	**Goal:** Identify differences in comprehension of students attending different grades; compared 3rd-grade students with those of 4th grade. **Content**: Used paragraphs and passages, selected from *Readers*. **Task**: Students were asked to answer questions based on the text read. **Index of comprehension:** The total number of errors made by a student, the time the student used to reproduce, and the number of questions answered correctly. **Point of view:** Comprehension is the ability to understand because comprehension depends in part upon one's store of information.
Pintner Silent Reading Test Pintner (1913)	**Goal:** Identify differences between silent and oral reading. **Content:** Used passages from supplementary *Reader* appropriate for 4th-grader. **Task:** Students were asked to read as much as possible during a two-minute period and reproduce in writing upon completion of reading. **Index of comprehension:** A total number of thoughts reproduced. **Point of view:** Comprehension is an ability to reproduce thoughts conveyed in a text passage.
Starch Silent Reading Test Starch (1915)	**Goal:** Identify a growth profile in students' reading capacity. **Content:** Used graded series of passages selected from *Readers*. **Task:** Students were asked to reproduce in writing after reading for thirty seconds. This two-day testing presented text passage identified as appropriate for the regular grade on day one, and text identified appropriate for grade immediately below on day two. **Index of comprehension**: The average number of written words describing the essential thoughts. **Point of view**: Starch believed students in any given grade individually and collectively are superior to the average or better readers compared to the students in the next lower grade; further, he believed writing ability is not essential for comprehension, and a short interval could give as reliable measure of reading ability as a long interval.

Table 1 (*continued*)

Test by Author	Test Description
Alpha Scale Thorndike (1917)	**Goal:** Develop a scale of reading ability to understand mechanism in comprehension.
	Content: Used sentence and text paragraph of 300-400 words.
	Task: Students were asked literal and interpretational questions upon completion of reading. For example: "Tom gave a gray cat to Mary. She gave him a black dog, followed by question: What did Tom gave to the girl?" Or "Was the dog of same color as the cat?"
	Index of comprehension: Mistakes in students' answers; identified three forms of mistakes in failure to understand: (a) the author's intended word associations, (b) the correct emphasis on words (over and under potency), (c) the overall text.
	Point of view: Reading comprehension is a reasoning process because understanding of texts is like solving mathematics problem.

Note. Smith (1922) was also consulted for information on the tests included in the table.

In contrast, Thorndike (1917a) developed his test to understand the mechanisms underlying reading comprehension. Although Thorndike included some tasks similar to Kelly (e.g., reading a couple of sentences), he also required the test takers to read text passages of about 300-400 words and answer literal and interpretational questions. Further, Thorndike (1917b) explicitly conceptualized comprehension as reasoning because understanding a paragraph to him was like "solving a problem in mathematics" (p. 329).

Further, the process of developing the tests varied. For example, two features, length and difficulty, were considered by those who included text passages in their tests. The process of text selection, nonetheless, varied. For example, Pintner (1913) selected text passages from the supplementary Reader because the materials were already determined to be well within students' comprehension ability, while Monroe (1918) selected texts with topics that were familiar to students at the target grade.

In addition, the index of comprehension varied. For example, Kelly counted the number of items answered correctly in the given time. Thorndike analyzed errors made in the items answered. The others who asked their test takers to write down as much as they could recall evaluated the reproductions differently. For example, Pintner (1913) counted the number of thoughts reproduced, Brown (1916) checked the quality of the thoughts reproduced, and Monroe (1918) counted the number of errors made in the reproduction

Across the tests, speed was consistently considered in measuring comprehension. In all tests, the number of words read per minute, or number of questions answered correctly in a given time was emphasized. The underlying assumption was that speedier readers, compared to those reading at a slower rate, have more cognitive resources for getting the thoughts from print (Gates, 1921; Gray, 1917). Further, none of the earlier test developers took the nature of the tasks included in the tests into consideration when they summarized comprehension performance.

The discussion that follows presents a snapshot of comprehension tests administered in recent years. However, a brief description of the key developments that took place in testing in general since the early years is presented first to understand what shaped contemporary comprehension tests.

Key Developments in Testing

Reading comprehension testing since the 1920s has become an integral part of educational programs instituted by state education agencies. By 1935, more than 20 state mandated testing programs incorporated comprehension testing (Segel, 1935). Education agencies used comprehension tests primarily for student promotion and placement.

Those tests included sentences, paragraphs, or short text passages to ensure that the tests have, at least, face validity (Segel, 1935).

One reform that shaped the comprehension testing, similar to the other tests, is the application of IRT, introduced in modern form by Lord in 1952 (Hambleton & Swaminathan, 1985; Yen & Fitzpatrick, 2006). In contrast to the CTT, introduced in the work of Charles Spearman in 1904, IRT considers the pattern of a student's responses in addition to his/her total number of correct responses when estimating proficiency. Two students with the same total scores might receive different estimates of proficiency by IRT, if their response patterns differ. By taking into consideration the patterns in responses, IRT provides finer performance distinctions among students who took the test. IRT provides a scale along which both the ability of examinees and the difficulty of test items can be located.

Another reform that shaped comprehension testing is Criterion-Referenced Testing (CRT), a concept introduced by Robert Glaser in 1963 that tied testing to the curriculum (Carver, 1974). The CRT went further than the prevailing Norm-Referenced Testing (NRT), a concept introduced by Francis Galton and Spearman in the late 1800 (Spearman, 1904a; 1904b) that describes the ability of a student with respect to his/her peers and provides an indication of learning simply through the contents of the test. In addition to gauging the differences between individual students, CRT explicates the differences within an individual student by indicating his/her mastery of the learning objectives. The advent of CRT led to the creation of different forms of testing, such as large-scale, high-stakes, and minimum competency tests. These different forms of testing

differ in the configuration of skills and knowledge and approaches applied in the test construction (Pearson & Johnson, 1978).

Test construction procedure also received attention as a result of the publication of standards for best testing practices, first issued by the American Psychological Association in 1966. The standards required test developers to develop a specification or framework for the test. The use of test specifications was stressed so that the content of a test and the process of testing would be transparent to the various players involved in test development and to promote greater uniformity in testing practices (Crocker & Algina, 2006). This influenced the item writing process. Item-writing practices for comprehension usually proceeded with minimal instruction (Rinsland, 1935). Item writers for comprehension testing were generally asked to get a graded series of text passages and to generate items with a different number of answer choices for different grades. A lack of clear and detailed guidelines, in many cases, resulted in a subjective definition of the construct, omission of areas relevant to the construct, and variability in the item writing (Bormuth, 1970; Crocker & Algina, 2006). To overcome these problems, many suggestions were offered in order to provide guidelines for constructing items for comprehension tests. Bloom's (1956) taxonomy—a classification scheme of educational objectives—was one of them. This suggestion has been implemented in limited occasions for various reasons (Anderson, 1982; Davis, 1972).

These reforms in test developments, however, did not mandate that comprehension testing should be grounded in a theoretical framework of the construct. The problems of just what was being measured that characterized the earlier testing

continued to prevail because the focus of reforms primarily was on how to analyze, how to expedite the reporting process, and how to use the test results.

Current Comprehension Testing Practices

Similar to the earlier tests, contemporary tests pursue different goals. However, their test design is guided by one of two overarching reporting goals: Criterion-Referenced-Testing (CRT) and Norm-Referenced-Testing (NRT). These two reporting goals determine the form adopted for a test. A test can be large-scale and high-stakes or minimum competency testing. Among the tests described in Table 2, NAEP, PIRLS, and state tests are examples of CRT; Gates-MacGinitie and ITBS are examples of NRT. Further, NAEP and PIRLS represent low-stakes, large-scale testing, whereas state tests generally represent high-stakes, large-scale testing.

Contemporary comprehension tests, produced either under the auspices of state or federal education agencies, or by test publishers for commercial consumption, usually ask examinees to read texts and answer questions about those texts in a time-constrained session. When a complete text passage is included, it could fall under two broad text types: narrative and expository. The questions typically ask an examinee to identify the meaning of words or phrases, locate information, ascertain the gist, or draw inferences. The comprehension questions are predominantly formatted as Multiple-Choice-Question (MCQ) that requires the selection of answers. Questions are also formatted that require examinees construct their own answers. The responses of the comprehension test-takers are synthesized, in most cases, either by adding up scores or by applying different IRT models. As the goal and form of the test permits, results are reported for an individual student or a group of students to indicate comprehension proficiency.

Table 2

Reading Comprehension Tests Administered in The 2000s

Test Name	Target Population & Format	Aspects of Comprehension & Item Expectations
National Assessment of Educational Progress (NAEP, 2003)	**Goal**: Reporting progress in performance at the national and state levels. **Time**: A 50-minute test for students at grades 4th-, 8th-, and 12th. **Content**: Two passages of different genres per student. Passage length is 100-800 word for 4th, 400-1,000 words for 8th, 500-1500 words for 12th grade. **Task:** Asked answer questions where MCQ (50% with 4-options), and CRQ (short and extended). About 22 to 26 items for 8th grade. **Reporting:** A *Criterion-Reference-Testing*	Test focuses on: • Basic comprehension where item requires to: identify, and/or locate important information in a given sentence, paragraph, or adjacent paragraphs; • Integration of information in various portions, or whole text where item requires to: integrate and interpret information, compare and contrast ideas, or actions of character, and identify alternative suggestions or explanations; • Important issue where item requires to: identify likelihood of an event, adequacy of an argument, or metaphors.
Progress in International Reading Literacy study (PIRLS , 2006)	**Goal:** Comparing national level comprehension performance across different countries. **Time:** A 80-minute test for students who are in their 4th year of schooling. **Content**: Narrative and expository text passages (biographies, direction, leaflets etc.) and maximum length of each passage is 1,000 words. **Task:** Asked answer questions where MCQ (50% with 4-options) and CRQ. **Reporting:** A *Criterion-Reference-Testing*	Test focuses on: • Literary experience and use of information (i.e., acquiring and using information). Items require to: • Retrieve explicitly stated information, • Make straightforward inferences, • Interpret and integrate information, • Examine and evaluate content, language, and textual elements.

Table 2 (*continued*)

Test Name	Target Population & Format	Aspects of Comprehension & Item Expectations
Illinois Standard Achievement Test (2006)	**Goal:** Comparison in annual status **Time**: A 125-145 minute test for students in grades 3 to 8 (145 minutes for 8th grade) **Content**: Narratives and others that include maps, graphs, diagrams, instruction. (Includes 10 passages for 8th grade and length of the passages varies from 50 to 1000 words). **Task**: Asked questions that are formatted entirely in MCQ (70-75 items at 8th grade). **Reporting**: *Criterion-Referenced-Testing*	Test focuses on meaning where item asks to • Summarize (i.e., identify the main idea not explicitly stated, for example, choose a title, and distinguish between main ideas and supporting details); • Sequence (i.e., order events in a story or non-fiction text; • Draw conclusion (i.e., make inferences, conclusions, make generalizations, differentiate fact from opinion, or interpret an image); • Interpret (i.e., determine whether a complex instruction or procedure is complete and clear); • Identify author's purpose in fiction or non-fiction, and explain the points used to illustrate a mood, or imagery.
New-England Tri-state (2003)	**Goal:** Statewide accountability **Time:** A two-hour testing for students in grades 3 to 8 that includes comprehension. **Content**: Text passage: Four passages (literary and informational) per student **Task:** Asked questions that are MCQ (4-options) and CRQ. **Reporting**: *Criterion-Referenced-Testing*	Test focuses on: • Initial understanding of literary text where item asks to identify information explicitly stated; • Initial understanding of informational text where item asks to use text feature to locate information, identify specific information, word's definition; • Analysis and interpretation where item asks to explain, generalize, or connect ideas and support responses with information given in a text.

Table 2 (*continued*)

Test Name	Target Population & Format	Aspects of Comprehension & Item Expectations
North Carolina State Assessment (2004)	**Goal:** Estimate level of knowledge and skills to determine growth. **Time**: A 115-minute test administered over three weeks to students at 6th grade. **Contents**: A total of 9 text passages (6 literary, 3 expository) per student. **Task**: Asked questions (3 to 9 items per passage) that are MCQ (100%). **Reporting**: *Criterion-Referenced-Testing*	Test focuses on: • Cognition where item asks to use context clues to determine meaning or summarize main points, (29% of the total items); • Interpretation where item asks to clarify, explain the significance, extend ideas, (40% of the total items); • Critical Stance where item asks to compare and contrast, identify impact of literary elements, (25% of the total items); • Connections where item asks to relate events beyond the selection and make association between two selections, (6% of the total items).
Washington State Student Learning (WASL, 2006)	**Goal**: Determine whether students (in grades 3 to 8) have met the learning standards **Time**: 75-90 minutes depending on grade level (It is 90 minutes for 8th grade) **Contents**: Six passages (50% literary and 50% expository) and one of them is a paired passage per student. (4 to 6 passages for 8th grade with total of 2500 to 4000 words in total) **Tasks**: MCQ (3 to 6 options depending on grade level), CRQ (short and extended); (35 items in 8th grade) **Reporting**: *Criterion-Referenced-Testing*	Test focuses on: • Understanding of theme supporting with details where item asks to summarize with evidence; • Making inference or prediction where item asks to interpret vocabulary, analyze author's purposes, evaluate information for different audiences, extend information beyond text, compare and contrast elements, make connections (e.g., cause and effect); • Application of information from four literary and informational texts where item asks to sequence events or processes, understand literary elements (e.g., plot, character, setting, and stylistic devices), and graphic elements. The description of the test indicates that all items are tied to passages in order to minimize impact of prior knowledge.

Table 2 (*continued*)

Test Name	Target Population & Format	Aspects of Comprehension & Item Expectations
Gates-MacGinitie (2008)	**Goal**: Identify status of Reading ability. **Time**: A 35-minute test assess students at levels 3 to 10/12 **Content**: A total of 11 passages that includes narrative and expository, and selected from published documents. **Tasks**: Asks questions that are entirely MCQ **Reporting**: A *Norm-Referenced-Testing*	Test focuses on: • Literal comprehension where items require understanding of the information stated explicitly; • Inferential comprehension where items require understanding of the information stated implicitly; • Vocabulary where items require identifying meaning of words.
Iowa Test of Basic Skills (ITBS, 2008)	**Goal:** Identify **s**tatus of reading ability. **Time**: Two separate periods of 25- and 30- minutes testing for students in Levels 9 through 14. **Contents**: Three passages (narrative, a poem, and a prose about science or social studies), passage length extends from a few lines to a full page. **Tasks**: Asks questions that are entirely MCQ. **Reporting:** A *Norm-Referenced-Testing*	Test focuses on: • Factual or literal understanding where items require understanding of what is stated; • Making inference or interpretation where items require understanding of what is implied; • Analysis and generalization where item asks for the main points or ideas, or author's viewpoints and use of language.

Note. Sources of information on the tests included in the table are provided in the reference.

Similar to the earlier tests, the contemporary tests vary in the points of view applied in constructing the test. They also vary from each other in many ways. As shown in Table 2, tests vary in the time allocated for testing, the length of the texts given to read, and the number of passages given for each text type. Further, the number of questions asked and the nature of questions asked across the passages are not consistent.

In addition, the proportion of questions formatted as MCQ varies. For instance, the commercially produced tests, ITBS and Gates-MacGinitie use the format for the entire test, while two federally sponsored tests, NAEP and PIRLS, assigned about 50% of the testing time to the format, and state sponsored tests format 80% to 100% of the items as MCQ (Bandeira de Mello et al., 2009). The number of response options offered in MCQ also varies (e.g., 3 to 5 options).

In addition, the process applied in selecting the text passages is not consistent. An example of variations observed in test specifications is presented in Table 3. Accordingly, the specification of state A's test stated that approximately 80% of the passages should be at or near grade level, and 20% should be up to two years below grade level for testing students at grades 3 through 5. The percentage of passages below grade level increases for students at higher grades. The specification of state B's test stated that the selection must use three readability formulas and different formulas for different grades. State C specification stated that the readability formula should be used to ensure reading difficulty, and teacher judgment should be used to ensure that the included text passages meet the respective state's target of reading complexity. The test D specification simply stated that two types of readability formulas should be applied in selection of text passages. Further, when judgment of a panel comprising education specialists is applied in the selection, the criteria the panel applied, however, were not described in the specifications.

The conceptualization of comprehension used in assessment has not changed much over time. Contemporary tests place less stress on retention and more on answering questions that mostly focus on explicitly stated information. Testing generally

views comprehension as a set of discrete skills as opposed to the view that considers

comprehension as a complex cognitive process where various skills need to be integrated

to uncover the meaning of a text.

Table 3

An Example of Variations in Text Selection Process

Tests	Description
Test A	• Specification states that approximately 80% of the passages should be at or near grade level, and 20% should be up to two years below grade level for testing students at grades 3 through 5. The percentage of passages below grade level increases for testing students at higher grades.
Test B	• Specification states that the text selection uses three readability formulas and different formulas for different grades (e.g., Powers-Summer-Kearl, Flesch-Kincaid, Spache or any other deemed appropriate for grades 3 and 4 materials, Dale-Chall, Flesch-Kincaid, Smog or Lexile for grades 5 through 8).
Test C	• Specification states that a readability index is used to ensure reading difficulty, and teacher judgment is used to ensure that the included text passages meet the respective state's target of reading complexity.
Test D	• Specification states that selection of text passages uses two to three readability formulas (two or three depending on grade level) and test authors' evaluation to ensure appropriate level of tone, topic, and familiarity.

Note. It is not customary for tests to reveal details on test selection process, only a few tests provided such information. The tests included in the table represent states: Illinois, Oklahoma, and Washington. Also included is Gates-MacGinitie.

As with this review, Johnstone et al. (2007) reviewed comprehension tests

developed for 49 states. They found that those tests assess different types of

comprehension, which the reviewers termed as forming a general understanding or

forming a higher-level understanding. The questions that students were asked are

grouped under two subthemes distinguishable by text passage type or the learning

objectives set in the state's curriculum standards (i.e., a state's expectations of what

students should know and learn). In another review, Keenan et al. (2008) examined a set

of tests used for research or counseling. They found low correlations among the tests. Further, they found the same test developer conceptualizes comprehension differently for students of different age and ability.

These variations, perhaps, are necessary to maintain the specifications. The rationales for these variations, however, are not transparent to the consumers of the tests. Most importantly, currently available tests did not emerge from a view of reading comprehension common among test specialists or shared by reading researchers. The questions asked to probe comprehension are primarily selected using psychometric properties, which, as currently used, do not explain the sources of difficulty. Further, when reported, comprehension performance is not explained in terms of the tasks students performed. For instance, results are not reported by the type of texts students read or the type of questions students answer. Moreover, the tests do not reveal the relation between comprehension performance and student attributes. The design and construction of comprehension tests appears to be carried out, in most cases, without a model of reading comprehension. If a model is applied as a foundation, the model is seldom explicitly articulated or the model does not represent what the reading researchers defined as comprehension.

It should be noted that a full picture of comprehension tests is difficult to portray. An assessment is a black box. Testing involves a multitude of steps and people, and current practices surrounding testing do not require that all information be compiled in one accessible document. Further, testing information is seldom disclosed or is disclosed on a limited basis to the public in order to maintain confidentiality. This inherent nature

of testing practices precludes a full overview of what the various parts of a test are and how those parts are assembled in testing machinery.

Role of Construct Definition in Assessment

As laid out in the Evidence-Centered-Design framework, the view of assessment as an argument is considered the cornerstone of test validation (Kane, 1992, 2006; Messick, 1989). To ensure validity, as argued in the ECD assessment principle (Mislevy et al., 2003), claims and decisions about major components of a test should be grounded in a theory of the construct targeted for testing. When the major components of an assessment and the claims made from the assessment are explicitly articulated and backed by research, the entire design process can be made more transparent. A design process that is transparent allows test developers and users of the test results to determine if the inferences drawn from the test are warranted (Mislevy, 2006). To the contrary, if not guided by a valid view of comprehension, the evidentiary argument about comprehension manifests misalignment. Inferences drawn about students' ability to comprehend cannot be sound if a test is misaligned (National Research Council [NRC] report, 2001; Valencia et al., 1989).

The standard approaches applied in evaluating the quality of an educational test stress the reliability and validity of the test. Sheehan and Mislevy (1990) contend, "It is increasingly recognized that mere high reliability coefficients do not guarantee a 'good' test, nor do high predictive relationships guarantee a 'valid' one" (p. 270). The reliability of a test, as explained by Carpenter and Paris (2005), can be low as a result of using "a highly skewed distribution rather than a poor assessment;" similarly, the reliability of a test can be high as a result of "assessing discrete skills" (p. 303). Likewise, ensuring

content relevance—that is, comprehension tests requiring reading of text passages and using criteria established for a classroom practices—may not be enough. A clear conceptualization of validity encompasses more than content relevance (Cronbach & Meehl, 1955; Messick, 1995).

Further, Embretson (1983, 1994, 1998) explained that traditional reliability and validity typically look for correlates of test scores when the test has already been constructed; in other words, quality is checked after the fact. According to her, a cognitive theory arises from the empirical findings. Therefore, a theory of the construct should play a key role in designing and structuring a test. As Embretson explained, designing a test is similar to designing a research study where a theory plays a key role. She also contends that a cognitive theory can make the item development become a scientific process rather than an artistic process, and is critical for making sound inferences and understanding the correlations of test scores. Developing test items by specifications based on a theory is expected to help explain psychometric parameters, for example, difficulty and discrimination, and provide more details on the strengths and needs of those who took the test. This notion stresses the understanding of the processes, knowledge structures, and strategies involved in responding to an item and the relations between test performances and other measures of individual differences that can strengthen the validity of a test. Embretson (1998) said, "tests should be validated from cognitive theory" (p. 380).

As described in the NRC Report (2001), constructing an educational test without a theory of learning and an understanding of the construct cannot support the collection of evidence to draw reasonable conclusions about what students know and can do. As

stated in the report, the purpose of an assessment is not simply to determine what students

know. Therefore, a more coherent picture of comprehension ability seems likely if tests

are built by incorporating a model of the construct. In this way, the cognitive processes

and products of the learning domain can be explicated (NRC Report, 2001). According

to Kintsch and Kintsch (2005), an assessment when conceptualized within a framework

of a theory of comprehension can be valuable in many ways including enhancing our

understanding of that theory.

Conceptualization of Comprehension

Regarded as the "essence of reading" (Durkin, 1993, p. 12), comprehension is an

internal activity carried out in deriving, constructing, or generating meaning[1] from one's

engagement with a text. The goal of this internal activity is to understand what is

conveyed through a text by an author. The understanding of how this internal activity

occurs and what would be the manifestation of this process is still evolving.

Since 1980, much reading research has been framed following a model of

comprehension within the Cognitive Constructive perspective. The discussion in this

section traces the way to the model to show how the conceptualization of comprehension

evolved. The conceptual model of comprehension as depicted in the Figure 1 and

described here served as the theoretical framework for the current study.

---

[1] One might consider adding "extracting" meaning to this list, reasoning that an author encodes a particular meaning in a text and a reader decodes this meaning. Introspection may lead a reader to this interpretation, especially when a writer and a reader share many experiences and views. However, the research indicates that comprehension is always a construction, even in these cases.

A Model of Comprehension

Edmund Huey (1908/1968), an early psychologist attempted to understand comprehension through his experiments on reading. He emphasized the importance of background knowledge and text content for a reader's ability to derive meaning from a text. He contended that two readers—one knowledgeable and one not-so-knowledgeable—can derive different meanings from reading the same text. Further, a reader's success in deriving meaning also depends on what the reader is reading. This conceptualization, however, did not explain the process involved in the comprehension of a text.

Another notable work on comprehension was by Fredrick Davis. In his effort to understand what is involved in comprehension, Davis (1944, 1968) identified a set of comprehension skills by means of statistical processes and concluded that two skills—knowledge of word meaning and ability to see the relation among ideas—can adequately explain comprehension. This conceptualization seems to suggest that aggregation of certain skills can ensure adequate representation of what is stated in a text.

During these early years, many researchers who studied comprehension and who were closely involved in testing viewed comprehension as a mode of thought. Gray (1937a) classified these reading-as-thought perspectives into three categories depending on the degrees of emphasis placed on the readers' thinking process. Some considered comprehension as a supplement to the thinking process. Others considered comprehension as a thought-getting process, whereby various elements of meaning were fused or organized into a chain of ideas, or an integrated system of thought. In some

more inclusive views, comprehension was considered as a process of recognizing and discovering the significance of facts and ideas presented in a text.

In addition to offering very little insight into how comprehension could occur or what would be its appropriate manifestations, these early conceptualizations offered no direction for explaining the comprehension differences observed within and between readers. These early attempts to understand comprehension, nevertheless, underscored the need for an in-depth understanding of this complex construct.

Since 1970, reading research has focused more on how readers process comprehension. Two approaches—bottom-up and top-down—were identified within the framework of information processing. These two approaches, that have many metaphors in literature, suggest different mechanisms underlying comprehension. The bottom-up approach suggests that readers comprehend text passages through organizing the meanings of words (e.g., Gough, 1972; LaBerge & Samuels, 1974). Alternatively, the top-down approach suggests that readers encode and decode information conveyed in a text using knowledge structures to uncover meaning (Adams & Collins, 1985; Anderson & Pearson, 1984; Rumelhart, 1977/1980; Wilson & Anderson, 1986).

The tenet espoused in the top-down approach, that a reader's knowledge plays the key role in comprehension, received further attention when schema theory came into prominence. Proponents of schema theory contended that reading comprehension is principally controlled by what a reader brings to a reading task (e.g., Anderson, 1977; Ausubel, 1963; Minsky, 1975; Rumelhart, 1975; Schank & Abelson, 1977; van Dijk & Kintsch, 1983). In explicating this view, Rumelhart (1980) stated that a reader proceeds with some hypotheses about what the writer may convey based on his/her knowledge of

the world when reading a text. The reader is claimed to comprehend a text when he/she is "able to find a configuration of those hypotheses that offers a coherent account for the various aspects of the text" (p. 38). In other words, a reader without a schema—a knowledge structure developed by organizing prior knowledge—can encounter difficulty in deriving meaning from reading a text passage.

This exclusive emphasis on readers shifted when researchers recognized that comprehension depends not only on the text or the reader alone, but, rather, on an interaction between the two components (e.g., Kintsch & van Dijk, 1978; Thorndyke, 1977; van den Broek et al., 1999). A change in the attributes of a text passage (e.g., text content, text type, text length, word frequency) or in the attributes of a reader (e.g., a change in a reader's attention, or prior knowledge) was found to make a difference in the meaning derived from reading a text. This premise that comprehension involves reader-and-text interaction is a building block of some comprehension models currently being applied by reading researchers. Two notable models that describe the process underlying comprehension are the Construction-Integration model offered by Walter Kintsch and the Landscape model offered by van den Broek et al. (1999). The Landscape model emphasizes what readers do during reading and in constructing a representation in memory. The discussion that follows focuses on Kintsch's model, as this conceptualization was used in framing the current investigation.

In his landmark publication, Comprehension: A paradigm for cognition, Kintsch (1998) described his Construction and Integration or CI model which he refined over many years in collaboration with others (e.g., Kintsch & van Dijk, 1978, van Dijk & Kintsch, 1983). In describing the model, Kintsch enumerated the processes posited to be

involved in comprehension, identified different levels of comprehension, and explained the possible roles that readers' and texts' attributes can play in comprehension. For Kintsch, a text usually describes a real or an imaginary situation in the world, and the description is seldom fully coherent. A writer relies on the reader to fill the gaps, which the reader does with his/her knowledge—about language, about the world, and about the specific communicative situation.

As Kintsch explained, text representation or comprehension is built up sequentially through a process of construction and integration of text elements (e.g., word, sentence, proposition). In the construction phase, features of the text activate associations in a reader's memory, whether or not those features are relevant to the current circumstance. In the integration phase, only certain aspects of the activated knowledge—the contextual input and memory that are mutually associated—are carried forward. This cyclical process of CI continues through all segments of the text until a coherent representation is realized.

Kintsch identified two levels of comprehension: textbase and situation model. The textbase results when a reader's comprehension, or representation of a text passage, is dominated by the propositions in the text. The situation model is a blend of text-driven and knowledge-driven representation of the text; this higher level of comprehension results to the extent that rich and relevant knowledge is available to a reader. A reader's understanding of a text (i.e., the situation model) is formed when the reader integrates information from a given text with his or her knowledge.

Textbase corresponds to the propositional representation of a text, both at the micro- and macrostructure levels, whereas the situation model corresponds to a

representation of the text that is integrated with other knowledge. When irrelevant and inconsistent information is deactivated and important knowledge is integrated, the textbase represents a form of the situation model (Kintsch, 1988). In a totally coherent or explicit text, a complete and adequate textbase represents the situation model. Thus, a textbase that tells it all is a good situation model. However, texts are generally incomplete. In addition to integrating information given in a text, a reader is required to integrate information from the text with prior knowledge.

The textbase and situation models constructed from a text passage by different readers are not always exactly the same. As Kintsch noted, a reader's knowledge enables formation of a mental model of a text even when the text is disorganized or poorly written. In general, poor comprehension or a poor textbase results when a reader fails to properly encode what is in the text or when the text is poorly written. A poor situation model is the consequence of the reader lacking or failing to apply the required knowledge to attain comprehension. Textbase and situation model representations together play roles in comprehension of a text.

Further, according to Kintsch (1998), the comprehension of a literary text is different from that of a nonliterary text. Each type of text presents a set of constraints and can demand specific encoding strategies and specific knowledge. For example, a scientific text can demand logical argument, knowledge of scientific theory and methodology, whereas a literary text can emphasize images, styles, and mood. However, Kintsch points out that the psychological processes involved in the comprehension of different types of texts are the same with respect to the cyclical process of construction and integration. That is, the role of knowledge and experiences, and production rules do

not change. However the various types of texts will vary in the "what" (i.e., the product) but not in the "how" (i.e., the process) of the comprehension (p. 205).

In describing the model, Kintsch also offered guidelines on the nature of questions to pose when probing one's comprehension. The lower level of comprehension—textbase—can be probed by asking questions based on information explicitly stated in the text. The higher level of comprehension—the situation model—can be probed by asking questions based on the information that goes beyond what is explicitly stated. For example, if a comprehension task requires a reader to make inferences, or reason through the situation depicted in the text, this task will be a measure of comprehension at the situation model level.

The conceptualization of the comprehension process in the CI model is unlike the top-down or schema theory. The rules of processing, identified in the CI model imply that, underlying comprehension is neither robust to all situations, nor is one exclusive of the other. The rules for the CI are context sensitive, as the model assumes that text comprehension can vary from one context to another as readers can vary in their extent of knowledge, experience, and their goals. Likewise, a situation model as conceptualized in the CI model is unlike a schema. Situation model represents the specifics of a situation described in a text as opposed to a schema that represents a notion based on a set of similar situations.

Many researchers have applied the conceptualization of comprehension offered in the CI model. Zwaan, Langston, and Graesser (1995) stated, "there is empirical support for the claim that readers construct situation models" (p. 292). According to Graesser (2007) "Kintsch's CI model is currently regarded as the most comprehensive model of

reading comprehension" (p. 11). Further, "the Construction-Integration (CI) model …has proven to be a useful frame for investigating reading comprehension," (Rupp, Ferne, & Choi, 2008, p. 445). Despite these claims, this model is evolving and can be further enriched. For instance, readers' attributes that other reading researchers consider important for comprehension of a text are left unspecified in the CI model.

For example, Alexander (1997a, 1997b) identified three reader attributes in her Model of Domain Learning (MDL) that are considered relevant to comprehension. These reader-attributes are: prior knowledge, interest, and strategic processing. In addition, a text that is to be comprehended has attributes that researchers claim make differences in comprehension. Any change in those attributes can significantly alter the representations formed of a text (Alexander & Fox, 2004; Duke, 2005; Fox & Alexander, 2009; RRSG, 2002; Samuels & Kamil, 1984). The RRSG panel (2002) contended that each key comprehension element—a reader, a text, and an activity—possesses some attributes and it is critical to consider those attributes when studying comprehension.

The discussion that follows provides a description of certain attributes of readers as well as tasks considered pertinent when designing a study to explore comprehension proficiency. The discussion is based on a set of studies that were found from a search in the empirical literature that began by constructing search terms representative of the key variables. The search focused on the key terms—prior knowledge, interest, text-type, levels of comprehension, and dimensions of a situation presented in a text. The search focused on regular students, measures of comprehension, and methodology applied to ascertain reader attributes. To obtain articles meeting the criteria, two approaches were followed. First, articles were searched for in major literature databases (PsycINFO,

Academic Search Premier, and ERIC) using the search terms. Second, the "snowball method" was used to find articles by looking through articles' references.

Reader Attributes

A reader is the primary component in the comprehension of what is being read. Reading research has established that what a reader brings in terms of knowledge and interest to a reading situation influences his or her comprehension (RRSG, 2002). According to the RRSG panel (2002), a reader must have a wide variety of cognitive capacities to comprehend a text. In addition, the reader needs to have various forms of knowledge—vocabulary, domain and topic, discourse, linguistics (Alexander, 1997b; Kintsch, 1998; RRSG, 2002). The reader also needs to have an interest in what is read (Alexander, 2005). The importance of these two reader attributes—prior knowledge and interest—in comprehension of a text was underscored even in early writings (Dewey, 1913, 1915; Huey, 1908). These two reader attributes, individually or collectively, have been recognized as making a difference in the meaning a reader derives from reading a text (Alexander, 2005).

As mentioned, various test results indicate that a substantial percentage of our nation's students are not reading as well as expected, and there is an inadequate explanation for this failure. Guthrie and Wigfield (2005) speculated that students who fail or perform poorly on reading comprehension tests might lack various reader attributes that extend beyond basic linguistic skills. Expanded information on the relation between reader attributes and reading comprehension can offer suggestions to test developers for designing more informative comprehension tests and can help explain the results to the test users (Bray & Barron, 2003). By understanding the relation between

comprehension and reader attributes, classroom instructors might be able to foster children's continued engagement in reading in order to promote learning from texts (Schiefele, 1999; Wigfield, 1997).

The discussion that follows focuses on two reader-attributes—prior knowledge and interest—in relation to comprehension. The discussion provides a definition of each construct, explains its importance to comprehension, and includes a description of a set of studies that explored effects of them on comprehension. Certain studies explored one attribute at a time, while others concurrently investigated the influence of both attributes. These studies are discussed in order to understand how the effects of reader attributes are examined in general. The description of the studies focuses on the methodology applied to measure these attributes, the population employed, and the results obtained. The ultimate goal is to inform the design of the current study.

Prior knowledge and comprehension. The concept of "prior knowledge" has a multitude of synonyms in the literature including background knowledge, experiential knowledge, world knowledge, pre-existing knowledge, and personal knowledge, and generally refers to "the sum of what an individual knows" (Alexander et al., 1991, p. 333). Alexander and Jetton (1996) refer to prior knowledge as "readers' familiarity with the content" (p. 99). Such knowledge comes in many forms such as declarative, procedural, or conditional knowledge (Dochy, 1992; Dochy & Alexander, 1995; Pearson, 1985), can vary in breadth or depth, and can be tacit or explicit (Alexander, 2003; Murphy & Alexander, 2002). Prior knowledge can be acquired in a deliberate manner as well as in an unintended manner (Ruddell & Unrau, 2004). This complex collage of a reader's knowledge develops over the years (Dochy & Alexander, 1995).

Two types of knowledge, varying in depth and breadth, have frequently been referred to when the relation between comprehension and prior knowledge is explored. These knowledge types are: domain knowledge, or knowledge of how much one knows about reading, and topic knowledge, or knowledge of a specific topic referenced in a text (Alexander, 1997b). Accordingly, domain knowledge is the knowledge that a reader has of a subject area (e.g., history or reading) and topic knowledge is the knowledge that a reader has related to the content of what is read (Alexander, Kulikowich, & Schulze, 1994). The difference can be illustrated with this example: a reader's general knowledge about geology represents domain knowledge, whereas knowledge about dunes located in Jockey's Ridge as described in a text passage, represents topic knowledge.

Prior knowledge is considered pivotal in comprehension of what is read (e.g., Alexander, 2003; Fletcher & Chrysler, 1990; Kintsch, 1988; Johnston, 1984; Millis & Cohen, 1994; Pearson & Johnson, 1978; RRSG, 2002; Rumelhart, 1980; Schank, 1982; van den Broek et al., 1999). Prior knowledge is important because texts are seldom fully coherent (Anderson, Reynolds, Schallert, & Goetz, 1977; Kintsch, 1998). Readers fill in those gaps, or figure out what remains implicit, with prior knowledge (Johnston, 1984; King, 2007; Kintsch, 1998). Prior knowledge is considered an integral factor in various comprehension models (e.g., schema theory, Landscape, and CI model).

According to Kintsch (1998), prior knowledge provides a context. As he explains, such context can be thought of as a kind of filter through which a reader perceives the world and interprets a text. In other words, prior knowledge enables a reader to form a mental model of a text. Alexander and Jetton (2000) further explained the role. According to them, a reader needs prior knowledge because it provides a

perspective of the text content, helps the reader allocate attention, direct interest, and make a judgment of the importance of information, all that are necessary to deriving meaning. The advantage of the readers who have high knowledge compared to those who have low knowledge is more in the higher level of comprehension (Kintsch, 1998). However, the relation between prior knowledge and comprehension might not be linear, especially when the existing knowledge does not correspond to the information presented in a passage (Alvermann, Smith, & Readence, 1985; Kendeou & van den Broek, 2007).

The importance of prior knowledge for comprehension testing has been noted by many reading researchers (e.g., Durkin, 1993; Farr & Carey, 1986; Johnston, 1984; Maria, 1990; Pearson & Hamm, 2005; Pearson & Johnson, 1978). For instance, as Johnston (1984) stated:

> The basic assumption, that it is possible to construct reading comprehension tests which will produce scores that are immune to the influence of prior knowledge, is erroneous. …. Consequently, if test constructors managed to produce a test in which performance was indeed unaffected by prior knowledge, whatever is measured, it would not be measuring reading comprehension. (p. 221)

Johnston urged that the sources of failure in comprehension be distinguished in testing, because each source may require different types of support. As Johnston elaborated, poor performance of one student may be due to a lack of prior knowledge, whereas the poor performance of another student who has adequate prior knowledge might be due to a lack in another attribute. Durkin (1993) pointed out that test developers control the variation in knowledge, in most cases, by confining the student evaluation to literal

comprehension. Early literacy researchers (e.g., Dewey, 1935; Gray, 1937b) warned that literal comprehension should not be considered as a true understanding of what is read.

In a review of studies that explored the effects of prior knowledge in performance involving reading of texts, Dochy, Serger, and Buehl (1999) found a number of them claimed a positive effect regardless of how such knowledge was measured and what sample composition was considered. Although the influence of prior knowledge on reading comprehension is no longer a debate among reading researchers, how the information on prior knowledge is obtained seems to be an issue (Dochy et al., 1999; Spyridakis & Wenger, 1991; Valencia, Stallman, Commeyras, Pearson, & Hartman, 1991). Prior knowledge of readers is measured in two broad ways: formal assessment and self-reporting. Within each, a variety of techniques has been applied. The techniques used in a formal assessment include multiple-choice question, open-ended question, free word association test, recognition test, free recall, judgment by an experimenter, whereas the techniques used in self-reporting include questionnaires and checklists (Dochy et al., 1999; Valencia et al., 1991). A set of studies discussed next provides examples of ways this attribute was measured.

In a study with college students, Ozuru, Dempsey, and McNamara (2009) used a combination of different techniques to study the effect of prior knowledge on comprehension of texts. The biology domain knowledge of the participants was measured with 21 multiple-choice questions and topic knowledge was measured with 16 constructed-response questions. The topic knowledge focused on the information not included in the texts but essential in understanding of the texts. In addition, the participants' reading skill and text cohesion of the given reading materials were also

assessed. Alternatively, Graesser and Bertus (1998) determined topic familiarity of adult readers (over 40 years and 60 years) using the self-report technique when they studied the influence of prior knowledge on comprehension of expository texts. To indicate familiarity with a text's content, after reading text passages, participants were asked to make ratings on a 6-point scale with response options: "very unfamiliar," "moderately unfamiliar," "somewhat unfamiliar," "somewhat familiar," "moderately familiar," and "very familiar."

Studies have also been conducted with K-12 students. For example, Langer (1984) selected 6th-grade students who were assigned to two different groups, depending on the activity performed prior to or after reading the texts. Prior knowledge of the students was measured with three content words selected from the passages that were rated by teachers as the key concepts on the passages. The students who had a discussion on those major concepts prior to reading the passages scored higher on a 20-item test designed to assess comprehension. In this study, background knowledge accounted for approximately 12% to 17% of the variance in comprehension performance.

For a study with 8th-grade students, Johnston (1984) constructed 33 content specific vocabulary related questions to assess prior knowledge and 54 textually explicit or textually implicit questions to assess the comprehension of three text passages. The questions were presented in multiple-choice-question format. The participants were assigned to different groups depending on the availability of the text when answering the questions. Prior knowledge accounted for about 3.5% between subject variance, over and above other factors, like IQ. Further, students with high prior knowledge performed

better than students with low prior knowledge did on certain questions (e.g., questions
that ask for causal chains).

Similar to Johnston (1984), Recht and Leslie (1988) compared the comprehension
performance of a group of students selected from grades 7 and 8 based on their reading
ability. The students' prior knowledge of baseball was measured with 42 multiple-choice
questions. The students who scored above the 70th percentile were considered as having
high prior knowledge, and those who scored below the 30th percentile were considered as
having low prior knowledge. Comprehension performance was assessed after reading a
text on baseball using a variety of measures: recall, sentence sorting, enactment, and
summarizing. In addition to the main effects of prior knowledge, the study revealed that
the students with high reading ability but low knowledge of baseball were not at any
advantage over those who were low in reading ability and low in knowledge of baseball.

In a recently conducted study, Tarchi (2010) explored both domain knowledge
and topic knowledge of 7th-grade students by asking the participants to write down all
they know about a given concept and to define key words taken from the text passages
that were being read. Prior knowledge explained a significant proportion of variance
whether the text was of science or history. However, the proportion of variance
accounted for by different forms of prior knowledge varied by text topic. Topic
knowledge accounted more for the history text than for the science text. The researcher
concluded that topic knowledge is more related to comprehension than domain
knowledge.

Referring to the comprehension of middle grade students, Wharton-McDonald
and Swiger (2009) contended that students who know more about a topic or are more

familiar with a particular text structure would be better able to comprehend a text. They further contended that students' knowledge of terms and concepts can be a proxy for what they know about the world, which correlates highly with their comprehension.

Interest and comprehension. Interest, a facet of motivation, is a psychological state in an individual (Alexander, 1997a; Hidi, 2000; Renninger & Hidi, 2002). Such a state is characterized by an affective reaction that results in focused attention to a particular content or an activity, and tends to have a relatively enduring predisposition to engage in certain classes of objects, events, or ideas (Ainley et al., 2002; Alexander, 1997b; Hidi, 2000). Interest is related to a specific object, activity, or subject area (Schiefele, 2009). Interest explains, "how an individual becomes motivated to occupy himself/herself in an activity in a certain way" (Krapp, 2002, p. 384). An interest, in general, emerges with an interaction between a person and his/her environment or a task-at-hand (Hidi & Renninger, 2006; Krapp, Hidi, & Renninger, 1992; Schiefele, 1999).

Two forms of interest are recognized in the reading literature. They generally are referred to as personal interest and situational interest[2] (e.g., Alexander, 1997b; Schiefele, 2009). Personal interest (also referred to individual interest) is conceptualized as a more stable, longer, and deeply seated form of interest as it is motivating from within (Alexander, 1997b). Such interest, it is posited, is accompanied by an accumulation of knowledge about a given domain that a person values. Further, as Alexander and Jetton explained (1996), individual interest "is more likely to result from repeated experiences or exposures to a related topic, subject, or activity" (p. 97). In contrast, a situational

---

[2] Different terms have been used to refer to different forms of interest (Alexander, 1997b; Jetton & Alexander, 2001). Terms are used, for example, text-base, emotional, or cognitive. Those classifications might not be a parallel to the classifications referred to as individual and situational interests (Alexander, 1997b).

interest is a temporary state and is triggered by a situation or moment (Alexander, 1997b; Schiefele, 2009). A situational interest is materialized quickly and can be short lived unless there are factors in the environment to sustain this initial experience (Hidi & Anderson, 1992; Hidi & Renninger, 2006; Murphy & Alexander, 2002). In context of reading, various features of a text can elicit a situational interest. Alexander and Jetton (1996) further explained the distinction using how interest is rated. According to them, "passage level ratings [are] more indicative of individual interest, whereas situational factors may … [influence] paragraph ratings and sentence underlining" (p. 103).

Topic interest is a form of interest referred to in the context of text reading (Ainley et al., 2002a; Ainley, Hillman, & Hidi, 2002; Alexander & Jetton, 1996; Hidi, 2000; Renninger, 2000; Schiefele, 1996). Topic interest is related to interest in a topic presented in a text. As Renninger (2000) explained, a particular topic is interesting because it offers personal meaning and fulfillment in its own right. Such interest is considered as "a specific form of individual interest," (Schiefele, 1996, p. 4). Other researchers claimed topic interest is influenced by both individual and situational interests (Ainley et al., 2002a; Ainley et al., 2002b; Hidi, 2000). According to them, topic interest elicited by a title, word, or a paragraph presented to a reader can be influenced by existing individual interest in the content of a text. Such interest could also be influenced by situational interest because features of text title and/or the theme expressed in the title can trigger topic interest (Ainley et al., 2002a; Ainley et al., 2002b; Hidi, 1990). Hidi (2000) explained that this is because "the arousal of topic interest initiated the process contributing to student persistence with reading and increased learning" (p. 314).

In the MDL, the importance of both forms of interest—personal and situational—is emphasized. About the importance of interest, Alexander (1997b) stated that the existence of personal interest can "result in an active pursuit of experience, knowledge, and skills associated with that interest" (p. 88). Alexander further stated, "schools can do much to nurture emerging competence by allowing students to pursue topics and tasks of interest and by immersing them in meaningful learning experiences that are fertile grounds for the growth of enduring interest" (Alexander, 2003, p. 12). Although the role of interest has not been integrated in the CI model's conceptualization of comprehension, Kintsch (1998) emphasized the importance of interest. Kintsch identified two types of interest—emotional and cognitive—to be pertinent to text comprehension, and he expects the comprehension model to evolve further when enriched with some aspects of motivation.

According to some researchers (e.g., Krapp, 2002; Schiefele, 2009), interest has a facilitative effect on performance. According to Hidi (2001), interest elicited by topics, theme, and/or text segments contributes to comprehension of a text. A text can be interesting if the pattern of events described in a text is intricate, the content is organized in a coherent manner, the events described in the passage have personal relevance to the reader, or the details presented are fascinating (Alexander, 1997b).

However, the relation between interest and performance is claimed to be complex (Bray & Barron, 2004). The relation may depend on other factors. For example, an individual's knowledge can influence the effect of interest on performance (Alexander, 1997b). Further, as Kintsch (1980) stated, a reader's interest increases as more is known and then decreases when the reader's knowledge reaches the point where the text no

longer presents the reader with new or challenging content. The relationship may be reciprocal; in other words, interest may facilitate performance and the success in performance subsequently influences the interest (Bray & Barron, 2004).

Unlike prior knowledge, studies conducted to examine the relation between interest and comprehension predominantly employed the self-reporting technique. However, these studies exploring the relation differed in what was asked, the rating scale that was used, the number of points included in the scale, and the response options offered. The following discussion presents a set of studies as examples.

In a study, Schraw, Bruning, and Svoboda (1995) explored the influence of college bound students' interest on different levels of text processing using a questionnaire. They found that interest influenced comprehension even when students' prior knowledge and general intelligence were controlled. Further, interest was found to be strongly related to deep-level processing. The high interest participating students processed the reading materials more deeply and applied more elaborate learning strategies than their less interested peer.

Schraw (1997) explored the sources of interest using college students. He developed a questionnaire for his investigation. He included six questions in his questionnaire to measure interest. These questions and their associated factor loadings shown in the parenthesis were: "I thought the story was very interesting" (.77), "I'd like to discuss this story with others" (.68), "I would read this story again if I had a chance" (.87), "I liked this story a lot" (.84), "The story was one of the most interesting things I've read in a long time" (.73). As a source of interest, topic familiarity was also measured by three questions: "The story discussed a topic I had read about before" (.83),

"The story covered a topic I knew a lot about" (.63), "The story contained information I already knew" (.36), and "The story was like others I have read" (.35). In addition, Schraw included text coherence, text complexity, and informational completeness. Schraw found the following percentages of variance in interest accounted for by the sources: topic familiarity explained about 10%, text coherence about 15%, complexity about 13%, and completeness accounted for about 10%.

In a recently conducted study, Magliano, Durik, and Holt (2011) examined the influence of topic interest on college students' comprehension of four expository text passages measured by recall of idea units. Interest was probed both before and after reading the text passages on a 7-point scale that ranged from "not at all" to "very much." Pre and post measures of interest were taken to see whether interest in a text passage changed as a consequence of reading the text. Magliano et al. found the results to vary by text passages. Post-interest increased in three of the four text passages. The results also showed that post-interest, knowledge of text topic in addition to the other predictors (e.g., general knowledge, and reading ability measured by a standardized reading test) were positively related to recall of three of the four text passages. Further, post-interest and recall correlated positively among the students who had low general knowledge. These associations were not observed among those who had high general knowledge. Topic interest was credited to facilitate comprehension. The researchers surmised that topic interest could be used as a scaffold to promote comprehension.

In a study using K-12 students, Stevens (1980) examined the relation between topic interest and reading comprehension. The participants, 5th- and 6th- grade students were asked to rate their interest on several topics using verbal and pictorial characteristics

of the text passages three days prior to the comprehension test. The topics that were rated

consistently as high or low on both characteristics were chosen for the comprehension

test. Further, students were grouped based on their reading ability into three categories:

(a) high ability, (b) medium ability, and (c) low ability. The results showed that interest

was only a significant factor in the comprehension of high ability students who

performed better on text passages that were of high interest.

Research that looks at the effect of interest on comprehension in a testing

situation is relatively new. Bray and Barron (2004) examined the relation using data

from a large-scale testing. Overall, a small but statistically significant association

between situational interest and comprehension test performance was claimed for the

students at grades 4 through 8 (r=.06 to .11 after correcting for verbal ability). This

relation was stronger among female students and among higher ability students. Further

analysis showed interest to vary due to the nature (poem vs. other text type) and content

focus (male or female focused) of text passages. Boys had less interest in poems and

girls had more interest in female-focused text passages. However, this study was limited

in some ways. The interest rating was based on only one self-reported question asked

upon completion of all the comprehension tasks that involved reading multiple text

passages. Further, no analysis was conducted to see whether students' knowledge about

the text or the perceived ease of the comprehension tasks influenced the psychological

state underlying student interest.

Prior knowledge and interest in comprehension. Some researchers

concurrently studied effects of prior knowledge and interest on comprehension to see

whether one attribute influences the effect of the other or their effects are independent.

According to Alexander (1997b), an individual's personal interest is expected to be strongly related to his/her knowledge. The following discussion presents a set of studies that explored effect of both attributes. The purpose was to guide the measures of current study that included both prior knowledge and interest as attributes of readers.

For example, Entin and Klare (1985) examined the effect of interest and prior knowledge on comprehension of college students. For interest, participants were asked to rate 12 topics on a 3-point scale with response options: "interesting," "moderately interesting," and "uninteresting." For prior knowledge, the participants rated on a 5-point scale their knowledge of a set of terms. Such terms were not taken from the passages included in the study. Those terms were considered because the knowledge of those terms was assumed necessary to comprehend the passages given to read. The results showed main effects of interest even after the means were adjusted for prior knowledge. Further, prior knowledge was rated high for the topics that were rated as interesting.

In another study, where participants were college students, Alexander, Kulikowich, and Schulz (1994) investigated the influence of situational interest and prior knowledge using a fill-in-the-blank comprehension task with two text passages. The study participants were asked to rate their topic and domain knowledge prior to reading text passages and to rate interest while reading the texts. Their results showed that students were more interested in text passages that were less abstruse and included more personally involved information. Further, interest in addition to the domain knowledge was found to predict a significant amount of variance in the comprehension of a more technical passage. However, domain knowledge was found to be the more potent predictor of reading comprehension.

The influence of prior knowledge and interest on comprehension was also explored among K-12 students. For example, Baldwin, Peleg-Bruckner, and McClintock (1985) asked students selected from grades 7 and 8 to rate their interest on a 10-point scale and answer 100 multiple-choice questions in prior knowledge test for 10 topics in order to group them into four experimental conditions. Those groups were: (a) High Knowledge and High Interest, (b) High Knowledge and Low Interest, (c) Low Knowledge and High Interest, and (d) Low Knowledge and Low Interest. Several days after the grouping, the students' comprehension was assessed with four text passages corresponding to those four conditions and 40 multiple-choice questions. Results showed comprehension was better for passages with higher prior knowledge than for passages with lower prior knowledge. In addition, comprehension was better for passages with higher topic interest than passages with lower topic interest. Further, correlation between topic interest and prior knowledge varied by topic (e.g., ranged from .74 to -.49, average r=.09).

Further, Ainley et al. (2002a) conducted a study with 8th-grade students in order to see whether prior knowledge mediated the effect of interest on comprehension. Topic familiarity was determined by asking the participants to rate how much they knew about each of the four topics described in the texts prior to reading them. The rating was done on a 5-point scale that ranged from "very little" to "a lot." Topic interest was measured by having the students' rate how interesting they expected each of the four titles to be on a 5-point scale that ranged from "a little" to "a lot." The results showed a significant correlation between topic familiarity and topic interest in three out of four text passages. In addition, the degree of correlation varied by topic.

Schiefele (1996) explored the effects of prior knowledge and topic interest on the comprehension of high-school students using sentence verification tasks. Prior knowledge was assessed with 22 questions formatted as MCQ. Interest was measured with two sets of response options. In one set, response options were: "meaningful," "useful," or "worthless" and in another set, the options were: "stimulated," "engaged," or "bored." The results showed that interest accounted for variance for both texts, whereas prior knowledge accounted variance for one text and one form of representations. Schiefele suspected that the small effect of prior knowledge might be due to the low difficulty level of the texts and the nature of the outcome measure (e.g., used recognition task). Further, the correlation between prior knowledge and interest was not significant for any of the texts.

Studying prior knowledge and interest in the context of comprehension testing is infrequent. Artelt, Schiefele, and Schneider (2001) explored the effect of prior knowledge and interest in order to determine the predictors of comprehension using 15-year-old students who participated in a national level test, were asked 18 knowledge questions prior to reading three text passages. Some of the questions were formatted as multiple-choice and others as constructed-response. The participants were also asked to rate their interest in each text topic on a 5-point scale that ranged from "very uninteresting" to "very interesting" before and after working on the respective text. The results showed significant correlation between prior knowledge and comprehension (r=.41), and between interest and comprehension (r=.25). The correlation between prior knowledge and interest was about .20.

In general, the review indicates that prior knowledge as well as interest influences comprehension. Further, the relation between prior knowledge and interest varies by topics. However, there are issues as to the technique used in measuring either of the reader attributes. To date, no technique has been identified as the best technique to determine prior knowledge (Spyridakis & Wenger, 1991). As a result, whether formal assessment or self-report is used, the type of questions asked and the number of questions asked vary from study to study. Dochy et al. (1999) contend that superficial methods like familiarity rating, self-estimation, or experimenter assumption may fail to show a clear relation between prior knowledge and learning outcomes, but they urge that those techniques should not be abandoned. They recommended that the method to apply should be considered in light of a study's goal.

Another issue related to measures of interest raised by some researchers is when to obtain the information. More specifically, the issue is whether the information on interest should be obtained prior to reading text passages, after reading the text passages, or after obtaining the outcome measure. This issue has been raised because it is suspected that the perception of the task or the ability to answer the comprehension questions might taint the self-reported information on interest (Tobias, 1994). To address this issue, the proposed study suggests asking questions at multiple time points in relation to the comprehension task to enrich the obtained information on interest.

Task Attributes

A comprehension task generally is devised to evoke evidence about a reader's ability to derive meaning. One part of such a task involves reading a connected discourse

(i.e., text). The other part of the task involves, in most cases, probing through a variety of activities in order to have the reader demonstrate comprehension.

Each task part can have many attributes. Psychometricians (e.g., Embretson & Wetzel, 1987; Gorin, 2005; Lee, 2002) contended that comprehension performance and the psychometric properties of comprehension tests (e.g., estimates of item difficulty and proficiency) are influenced by task attributes (e.g., texts that are being read, the way comprehension is gauged, and the questions a reader is asked to demonstrate comprehension). The review presented next discusses the components of comprehension tasks and relevant attributes that were considered in designing the measure of comprehension for the current study.

Text passages. A text passage contains words, phrases, and sentences. An author, through selection of words, use of rhetorical style, and methods of organization, conveys a message or presents a circumstance to a reader. According to Kintsch (1998), an author describes in a text passage a situation, real or imaginary. The situation described can be a constellation of events or episodes. Each episode is represented by a network of propositions, also referred to as idea units, meaning units, or predicate-arguments. When comprehending a text, a reader forms a mental representation of what is described in the text.

Text passages can vary in ways that can impact comprehension (Weaver & Kintsch, 1996). According to RRSG report (2002), the ways texts can vary include type (e.g., expository, narrative), content, and sentence difficulty (e.g., vocabulary load, idea units). Variations in texts are found to make differences in a reader's comprehension of

the texts (Johnston, 1983; Pearson, 2000). Meyer and Rice (2002) argued that comprehension of one passage should not be generalized to another passage.

Expository and narrative are two types of text passages frequently referred to in literature and in assessments. Brewer (1980) defined these types by taking into consideration the linguistic form underlying cognitive representation, and the intent of texts. Accordingly, an expository text is one that represents mental processes, such as classification, comparison, induction, illustration, and persuasion, whereas a narrative text is one that represents a series of temporally-occurring episodes perceived as having a causal or thematic coherence. In some cases, a feature of expository texts that distinguishes them from other type of texts is the inclusion of a paragraph convention (Graesser, McNamara, & Louwerse, 2003). In such a convention, headers and sub-headers are used where the first sentence identifies the main theme of the paragraph and the subsequent sentences supply additional details relevant to the theme. A narrative text can also have categorical organization; such organization groups provide information about characters, settings, and particular episodes (Roller & Schreiner, 1985).

Further, the underlying intent of an expository text is generally to inform or to persuade whereas the intent of a narrative is more to entertain or to appreciate literary aesthetics (Brewer, 1980). However, it is not necessary for a text in its entirety to conform to either of these two types. A text can be a mixture of different types (Alexander & Jetton, 2000). An expository text could be an essay, a journal or newspaper article, a brochure, or text passages found in a textbook. A narrative text could be a short story, a novel, or a fable. This general portrayal of categories of texts seems to align with the conceptualizations of different texts held by others (e.g., Millis,

Morgan, & Grasser, 1990; Magliano et al., 2007; Pilonieta, 2011; Weaver & Kintsch, 1996; Wolfe, 2005).

Although assessments of comprehension typically include text of both narrative and/or expository types (Pilonieta, 2011), the results of assessments are not reported by text type. Duke (2005) said that readers' ability to comprehend one type of text does not necessarily correspond to their ability to comprehend another type of text. The panel of RRSG (2002) suggests comprehension assessments to report on how the comprehension difficulty and proficiency of a reader differs across different texts. Such information can help guide instruction offered to students. Otherwise it is not easy to provide support to students who must be equally proficient in comprehending both narrative and expository text passages if they are to perform well on comprehension tests. Pressley and Block (2002) recommended use of both types to understand comprehension as it is a complex process.

The text type is important to consider when exploring comprehension because different texts are claimed to elicit different processing (Alexander & Jetton, 2000; Kintsch, 1998; Weaver & Kintsch, 1996; Wolfe & Woodwyk, 2010; Zwaan, 1994). As Wolfe and Woodwyk (2010) explained, the processing of narrative texts tends to stress the understanding of the organization of events, while the processing of expository texts tends to stress the activation and integration of prior knowledge. Kintsch (1998) stated that the textual schema or rhetorical structure, whether or not it is a narrative, plays a role in the understanding of a text.

In general, expository texts are found more difficult to process than narrative texts. Reading researchers provided explanations for such differences. As explained by

Millis et al. (1990), an expository text seems to be dry in comparison to a narrative text. This is perhaps because the primary function of an expository text is to inform, not to entertain readers. According to Lorch and van den Broek (1997), the structures of expository texts are not as homogeneous as the narrative texts. Further, Graesser, Golding, and Long (1996) suggest that more familiarity with narratives, as we in general enact or experience, makes it easier to follow compared to expository texts. Graesser et al. (2003) contended that it is the unfamiliarity with expository texts and their typical high density of unfamiliar concepts that makes the processing them difficult for readers, particularly young readers. Wolfe and Woodwyk (2010) provided further explanation. According to them, an expository text is more difficult than a narrative to comprehend, partly because readers are more likely to lack relevant prior knowledge, essential to establishing a coherent representation of what the text describes.

Readers in various studies were found to respond differently with respect to text type. For example, in a study, Tun (1989) found readers of different age groups recalled more and answer more comprehension questions (formatted as true/false) from narrative than expository texts (using three of each type). Tun concluded that a lack of familiarity with the expository texts may have led to the differences in performance that were found between two types of text.

Wolfe (2005) with his colleagues (Wolfe & Mienko, 2007), and (Wolfe & Woodwyk, 2010) conducted a series of studies to explore effects of text type using college students. The researchers observed that the participating students recalled more from narrative than from the expository texts. Wolfe and Mienko (2007) found that the college students' memory of narrative was better than their memory of expository texts

when content is not controlled. Wolfe and Woodwyk (2010) also showed that the students performed better on narrative texts compared to expository texts.

Zwaan (1994) reported further differences in performance due to text type. In his study, college students who were assigned to the literary text condition scored higher on surface and textbase levels comprehension (i.e., verbatim and paraphrase), whereas the students who were assigned to the news condition scored higher on situational level comprehension (i.e., comprehension requiring inferences). As Zwaan speculated, the differences in performance might be due to the inability to integrate information. As Kintsch (1998) points out, comprehension of expository texts would entail, in general, more integration in which irrelevant or unimportant information is deactivated quickly. The more integration needed in an expository text results in a stronger situation model.

Studies that were conducted with children also showed the effect of text type in comprehension performance. For example, Best, Ozuru, Floyd, and McNamara (2006) found that 4th-grade students answered more questions from narrative than expository text passages in a study that included two of each text type and a total of 48 questions. Further, those children answered more global questions (i.e., questions that draw upon information located across six or more clauses or ideas that needed to be connected by the reader) from narrative texts, compared to the expository texts, where the children answered more local questions (i.e., questions that draw upon information mostly within two sentences or fewer than five clauses). The performance on global or local questions, however, did not differ between texts within a type. Best, Floyd, and McNamara (2008) were able to replicate those findings with 3rd-grade students.

However, there are some studies where the effects of text type were not observed. For example, Kintsch and Young (1980) found no difference in students' recall of a set of target sentences that were embedded in two types of text. Kintsch and Yarbrough (1982) later explained that the failure to see the effect of text type in such a situation indicates that text type influences macro-processing but not micro-processing of a text. In another study, Roller and Schreiner (1985) trained two groups of 6th graders to identify the main and supporting ideas from texts using narrative and expository texts that were crafted with the same content. The groups in the final assessment did not differ in total number of questions answered. However, the group that was tested with the expository texts took a longer time to read the materials than their peers who were tested with narrative texts. This study suggested that students, perhaps, could perform as well in expository texts as in narrative texts with increased familiarity with such texts.

In addition to type, text content is considered pertinent. The RRSG (2002) panel contended that the content presented in a text has a bearing on comprehension. Magliano et al. (2007) also contended that individual differences due to text type might depend on other text attributes such as content. Duke (2005) stated that we cannot assess "comprehension, but comprehension of …." so we need to take into consideration type and topic when analyzing comprehension (p. 98).

However, a few studies to date were found to explore the joint effect of text type and content on comprehension. In a recent study, Cervetti, Bravo, Heibert, Pearson, & Carolyn (2009) developed different types of text passages for two topics. In other words, the same topics were embedded in two text types, narrative and expository. Students from grades 3 and 4 were asked first to recall and then to respond to a set of

comprehension questions. The study showed a significant effect of text type for only one topic. The researchers speculated that this might be the result of topic familiarity moderating the effect of type. Wolfe and Woodwyk (2010) contended that the comparison of performance differences due to content along with text type is difficult to explore because the same readers cannot be asked to read a common content presented in two different types due to a carry-over effect.

Comprehension activity. Since comprehension is a process not open for external observation, it is in most cases gauged by activities that are assumed to be the manifestations of comprehension (Francis et al. 2005; Pearson, 2000; Readence & Moore, 1983; Simons, 1971; Spache, 1963). Generally, the actions the students are asked to perform in order to obtain evidence of comprehension are observed in two time modes: on-line (i.e., while reading) and off-line (i.e., after reading). The common off-line measure of comprehension is taken via reproduction, which can be free or probed (Kintsch & Rawson, 2005; Palincsar, Magnusson, Pesko, & Hamlin, 2005). The probed reproduction technique in which students are asked questions is "by far the most widely used method in psychology" (Kintsch, 1998, p. 295).

A question that is posed to a reader can ask for a variety of information. For example, a question can ask a reader to identify a detail, make a summary, or give the gist of what has been described in a text, or make an inference given the information in a text. The questions included in comprehension tests are constructed in many ways. For example, item writers were found to use their understanding of the construct, various taxonomies, or test specifications (Drum, Calfee, & Cook, 1981; Lucas & McConkie, 1980; Pearson, 2000). However, one of the concerns has been raised about the items

included in tests or studies. There is a lack of information on the types of questions asked

and the process applied in their construction. As Anderson (1982) pointed out, lack of

information prevents one from generalizing the findings of comprehension performance

across assessments or studies. Also, it makes one unable to explain successes and

failures in comprehension assessments (Snow, 2003). Kintsch (1998) suggests that the

process of constructing questions should explain the psychological process and the

knowledge involved in answering a particular question. The discussion that follows

focuses on Kintsch's CI model that was used in developing a framework for the questions

included in the current study's comprehension measure.

In the CI model, it is posited that a reader's comprehension of what he or she

reads can be explained by a combination of the textbase and the situation model

representations. An examinee's textbase comprehension can be assessed with questions

asking for recognition of information included in the text using meaning preserving

paraphrases or sentences (Kintsch, 1998; Fletcher & Chrysler, 1990). For textbase,

readers will connect sentences or integrate information stated explicitly in a few

sentences that are within the proximity of the target proposition. An examinee's situation

model comprehension can be assessed with questions that require making inferences.

Such an inference requires both the integration of the propositions presented in a text and

the integration of what is said in the text with the knowledge that a reader brings to the

reading task. The distinction can be explained with an example texts excerpted from

Kintsch (1998). For the text: Jane could not find the vegetables and the fruit she was

looking for. She became upset. A question can be posed: "Why did Jane become upset?"

This is an example of textbase question. Similarly, for the text: A turtle sat on a log. A

fish swam under the log. A question can be posed: "Where was the fish?" This is an example of a situation model question as the question requires the understanding that the fish was in the water underneath the log where turtle was sitting because the text said the "fish swam."

However, there is not just a single textbase or situation model formed when representing a text (Kintsch, 1998). A full understanding of a text requires the construction of a number of textbase and situation model representations. A situation can be conceptualized as a constellation of episodes or events. In order to comprehend a text, a reader needs to identify those events and various relations, also referred to as dimensions or facets of a situation represented in those events. The relations embedded among events can be varied because, as van den Broek (1990) explained, events described in a text are not perceived as an isolated case. In a narrative, these facets include a protagonist or an agent, the goals or motivation of the agent, and different relations like temporal, spatial, and causal (Rinck et al., 1997; Therriault & Raney, 2007; Zwaan et al., 1995; Zwaan & Radvansky, 1998). Temporal, spatial, and causal relations are also represented in the situations depicted in expository texts (Graesser et al., 2003; Millis & Cohen, 1994).

An understanding of a temporal relation involves keeping track of a sequence of events, states, or actions occurring in a time interval and changes in time or time interval referred to or described in a text. Temporal information is crucial to the construction of situation models. It is important because, as Zwaan and Radvansky (1998) explained, a reader needs to keep track of when an event took place, when the event took place in relation to other events and in relation to the time referred in the text in order to achieve a

coherent understanding of the situation described by a text. Further, temporal relation is crucial for identifying causal links among events, as an effect can never happen before its cause; therefore, readers need to keep track of the temporal relation in order to generate a causal connection. In addition, temporal information is claimed to provide a cue to the unknown events. Ohtsuka and Brewer (1992), similar to Mandler (1986), using canonical and non-canonical texts and flash backward technique, explored whether readers were aware of temporal order. They observed changes in reading time when there was a discontinuity.

Understanding spatial relations involves keeping track of the location or the context and any change in the context of an event described in the text. It is an understanding of where the events occurred, and the characteristics of the location, imaginary or real. A spatial setting can be a room, a region, or a scenario that has distinctive features and is discernible from an alternative spatial setting. The existence of spatial representation is supported by research (Rinck, Williams, Bower, & Becker, 1996). If a reader does not use spatial information, whether the text is narrative or expository, then, "it is likely that the comprehension of the situation will be less than optimal" (Millis & Cohen, 1994, p. 369). Construction of spatial relations is found more difficult for readers than temporal relations. As Zwaan and Radvansky (1998) explained, there are fewer verbal cues for spatial representation compared to temporal relation (e.g., then, after, next, since, yet). Further, spatial information is not as closely intertwined as temporal information with other relations. Readers are more likely to encode information on spatial relation when that information is causally relevant or specifically instructed (Zwaan & Oostendorp, 1993).

Causality is a two-part relation between events. An event, or an antecedent event is a cause and another event is the effect or the consequence. The understanding of causality involves how one event, the cause, brings about through some mechanism another event, the effect. This cause-and-effect relation is considered crucial for interpreting texts, both narrative and expository (Graesser, 1981; Mandler & Johnson, 1977; Norman & Rumelhart, 1975; Schank & Abelson, 1977; Warren, Nicholas, & Trabasso, 1979; van den Broek, 1990; Zwaan & Radvansky, 1998). This is because readers need to develop coherence in understanding the situation. In addition, a reader, in general, would like to know why an event happened (Zwaan & Radvansky, 1998). Two aspects of these relations could be psychological or physical. The psychological cause refers to the perceived relations between/among events. Some causal relations could be more complex than others, for example, a single event can have multiple causes and a single cause can lead to multiple consequences (Mislevy & Rahman, 2009).

Singer (1994) stated, "the research has been spurred by the consensus that causal relations are fundamental to narratives, expository texts, and many other types of discourse" (p. 488). An example that the researcher used to explain this relation is the following: The spy threw his report in the fire, the ashes floated up the chimney; the question about consequence: What happened to the report? Answer: The report was burned to ashes. Lorch and van den Broek (1997) reported that causal relations received a "lion's share of research attention" in discussion on inference generation (p. 220). Further, one aspect of this relation, for example, antecedent causal relation was found easier to make than the causal consequence relation (Magliano, Baggett, Johnson, & Graesser, 1993).

Events in memory are organized based on a series of specific relations embedded in the events (Magliano et al., 1993). The existence of these relations has been supported by research (Magliano et al., 1993; van den Broek, Kendeou, et al., 2005). The results showed readers taking longer when there were discontinuities in these relations described in texts. The elevated time was more for temporality and causality, and less so for spatiality. Results also showed less coherent recalls when there was a gap or ambiguity in those relations presented in the texts. Further, understanding of some of these relations took longer time than others.

These relations—temporal, spatial, and causal—are not independent; rather they are inextricably linked (Zwaan et al., 1995). As these researchers explained: "movement through space is not possible without time, and an effect never occurs before its cause and things happen in a context" (p. 386). The weight of each dimension, however, can vary by texts or the focus of what an author wants to convey in a text. Further, readers need to identify, or figure them out, and organize them when constructing their situation models of what is described because these relations are not always expressed in a linear fashion or stated explicitly (Zwaan & Radvansky, 1998).

Temporal, spatial, and causal relations were considered when developing a framework of comprehension items for the study at hand. These relations were probed at different levels of comprehension.

Alexander and Fox (2004) mentioned that although our understanding of text comprehension has been enriched over the years, we have not yet seen tests that were developed by incorporating the advances in our knowledge about reading comprehension. Further, Kintsch and Rawson (2005) stated:

Text comprehension is a complex process, requiring the involvement of many

different components, relying upon many different kinds of information, and

yielding complex mental representations…. Almost all of the extant research on

text comprehension has focused on identifying and examining the various

components in isolation, which has gone a long way toward furthering our

understanding of how comprehension works…. An important direction for future

research will be to examine the interaction of the various components to

understand how they work together as a system to give rise to comprehension. (p.

225)

It seems assessment as a context can be used to explore comprehension in a

comprehensive way. Constructing questions that ask for information on the various

relations embedded in a situation, in order to determine different levels of comprehension

for text passages that vary in type and content can allow one to take into account the

greater complexity that is represented in comprehension of a text.

Multiple-choice-question format. Multiple-Choice-Question (MCQ) format is

one of the ways questions are presented in off-line measures of comprehension. Such

format consists of a stem and a set of response options from which examinees select the

best answer. This format was used in tests developed in early years (Kelly, 1916;

Thorndike, 1917) and is ubiquitous in contemporary tests (Campbell, 2005). All

questions included in the current study's comprehension measure were formatted as

MCQ.

MCQ format meets the criteria suggested for comprehension questions by Gates

(1921). He suggested that students should be directed to select correct answers or cross

out wrong answers when answering comprehension question so the students'

understanding of a text is not confounded with their writing ability. The format is found

to have many advantages over other formats (Campbell, 2005). MCQ format allows a

broader coverage of a domain being assessed within a time constraint and scores are

considered more objective. Further, the psychometric procedure, IRT, can conveniently

be applied to the responses obtained using this format.

Despite these advantages, reading researchers have raised concerns about the use

of the format (Farr & Carey, 1986; Kintsch & Kintsch, 2005; Pearson & Hamm, 2005).

However, researchers argued in favor of multiple-choice questions stating that its use can

be as informative as other formats when formulated with more thought and care (e.g.,

Kintsch, E. 2005; Kintsch & Kintsch, 2005; Pearson & Hamm, 2005). In keeping with

this argument, Campbell (2005) said that relying on MCQ may not be considered an

underrepresentation of the construct if the rules for writing such items are derived from

theory and empirical research. Durkin (1993) proclaimed that the MCQ format will

persist because it easily meets educators and policy makers' demands, and is considered

most viable for probing students' comprehension on a large scale.

Measurement Methodology

This section presents a discussion on a psychometric approach that was applied in

the study in order to integrate the conceptualization of comprehension offered in

cognitive psychology with the measurement model. The discussion is pertinent to the

research question: "Do the task and reader attributes considered in the study contribute to

the item difficulties and reader comprehension proficiencies?"

Item Response Theory

In educational measurement, IRT has been found to overcome many inadequacies of Classical Test Theory. IRT, a set of procedures for studying characteristics of items relative to an ability scale, emerged with the work of many. The modern form of IRT emerged with the work of Frederick Lord in 1950s (For more on IRT, see Hambleton & Swaminathan, 1985; Yen & Fitzpatrick, 2006). The focus of IRT models is in the estimation of proficiency expressed in terms of the probability that students at different points on the ability scale will answer an assessment item correctly. The probability of a given response x_{ij}, is a function of ability of an examinee, θ_i, and parameters of an item j.

In 1960, George Rasch offered an IRT model in which each item has only one parameter—difficulty. In this one-parameter model, all items are assumed to have same discrimination power and guessing is assumed to be minimal. The probability of a correct response by an examinee to an item j according to the one-parameter model is:

$$P_j(\theta_i) = P(x_{ij} = 1 \mid \theta_i, \beta_j) = \frac{\exp(\theta_i - \beta_j)}{1 + \exp(\theta_i - \beta_j)}$$

where

x_{ij} = denotes the response of an examinee i to item j; a response is 1 if correct and 0 if incorrect
β_j = represents difficulty of the item
θ_i = represents the ability of an examinee, in our case the reader.

In the above equation, it is not the individual item probabilities that are of primary interest, but rather the item difficulties and the person proficiencies (Embretson, 1983).

The estimation procedures underlying the Rasch model generally anchor the item difficulties to a mean of zero $(i.e., \bar{\beta} = 0.0)$. An examinee, at the level of the item set, would have a probability of .50 for passing an item on the average. For example, if an

average item is selected, that is, $\beta = 0.00$ and an examinee who is at the level of the

selected item (i.e., has an ability of zero) will have this probability of correct response:

$$P(x_{ij} = 1) = \frac{exp(0 - 0)}{1 + exp(0 - 0)} = .50 .$$

Likewise, if a difficult item is selected, say $\beta_i = 2.0$, an examinee at the level of

the item needs to have a high ability to compensate for the high difficulty of the item. In

this case, θ_j must equal to 2.0. Thus:

$$P(x_{ij} = 1) = \frac{exp(2.0 - 2.0)}{1 + exp(2.0 - 2.0)} = .50 .$$

IRT models are found to be more viable than the CTT models for various reasons.

With IRT, finer distinctions among the examinees are possible. The patterns in the

examinees' responses and the total number of correct responses are taken into

consideration when estimating student proficiencies. In addition, item difficulty is taken

into consideration when calculating the test scores. Further, comparable estimates of

ability are possible even from item subsets that differ vastly in difficulty. There are many

IRT models, some of which are more complex than the Rasch model.

IRT models are based on two assumptions: (a) local independence (i.e., a

response to an item is independent of responses to other items) and (b) unidimensionality

(i.e., a test measures a single latent trait). When data do not fit the IRT models, this is

considered an indication that those assumptions are violated. Under the assumption of

local independence, the probability of a vector pattern, $x_i = (x_{i1},...,x_{in})'$ represents

responses to n items is: $P(X_i \mid \theta_i, \beta) = \prod_j P_j(\theta_i)^{x_{ij}} (1 - P_j(\theta_i))^{1-x_{ij}}$ where $\beta = (\beta_1,...,\beta_n)'$.

Assuming the independence of responses over examinees, the probability of the data

matrix, $X = (x_1..., x_N)'$ for N examinees is the product of expressions like

$P(X \mid \theta, \beta) = \prod_j P(x_i \mid \theta_i, \beta)$. Once **X** has been observed, $P(X \mid \theta, \beta) = \prod_j P(x_i \mid \theta_i, \beta)$ is

interpreted as likelihood and provides a basis for estimating parameters.

Linear Logistic Test Model

In standard IRT model, such as the Rasch, the features of a task that can make the

items difficult or the cognitive processes that an examinee may employ to have a high

probability of making a correct response are not taken into consideration (Sheehan &

Mislevy, 1990). Gerhard Fischer (1973) offered a special model, which he called Linear

Logistic Test Model, or LLTM. This special model is also referred to as cognitively

based IRT as it incorporates features of tasks in estimating difficulties of items. This

incorporation of item features allows the LLTM model to shed more light into the pattern

of responses and offers insight into the question: "What aspects of a task make an item

difficult?" that many interested in the process of learning would like to know (Lu, 2006).

In LLTM, the difficulty parameter of the Rasch model, β_j is postulated as a

linear model of the item characteristics: $\beta_j = \sum_{k=1}^{K} q_{jk} \eta_k = q'_j \eta,$ or in matrix notation as

shown in Sheehan and Mislevy (1990), $\beta = Q'\eta,$ where Q' is an n by k matrix of item

feature data and $\eta = (\eta_1,...,\eta_k),$ where η_k is the additive contributions of feature k to an

item's relative difficulty. In other words, the task difficulty parameter is estimated as

linear combinations of a smaller number of more elementary components. The

elementary components are defined to reflect differences in the cognitive processing

demands of the tasks in order to address the problem of why some items are more

difficult than others.

LLTM item parameters help describe the complexity in an item due to cognitive

factors. Ability combines additively with cognitive complexity to give the probability of

response to an item. Thus, the probability of a correct response under the LLTM is:

$$P_j(\theta_i) = P(x_{ij} = 1 \,|\, \theta_i, \beta_j, q_{jk}, \eta_k) = \frac{exp(\theta_i - \sum_k q_{jk}\eta_k)}{1 + exp(\theta_i - \sum_k q_{jk}\eta_k)}$$

Where

q_{jk} = represents the degree to which feature of k applies to item j,

$\quad q_{jk}$ is 1 if feature k is present in item j and 0 if not.

η_k = represents the contribution of feature k in item difficulty

θ_i = represents the ability of an examinee, in our case, the reader.

(For more on equation, see Embretson, 1983, and Lu, 2006)

Although multiple processing variables for items are postulated in LLTM, it is a

unidimensional model of individual differences. In other words, the ability parameter in

LLTM contains only one dimension. Application of LLTM requires two kinds of data:

responses to the intact item presented under the standard test instructions and coding for

each item on some variables that represent the theoretical complexity factors of the item.

LLTM presents advantages over other statistical procedures. For example, in

LLTM, the ordinal scaled numbers of solved items are transformed into interval scaled

parameters whereas ANOVA is based on the ordinal scaled numbers of solved items.

LLTM serves as a system of consecutively testing an alternative hypothesis compared to

the paired *t*-test, which would have to be applied many times, and thus increases type-I

risks. Further, LLTM postulates a non-linear relation between response probability and

the parameters, thus differing from other procedures that postulate a linear relation and can encounter floor and ceiling effects for test scores and item scores. In addition, examinee ability, θ_j and item difficulty, β_j are located on a common measurement scale, namely, positions on the latent trait. Also, the person ability and item difficulty have equal weight in determining the effective response potential, $P(x_{ij}=1)$, (see Embretson, 1983 for more).

LLTM has been applied in analyses of various sorts of items. For example, LLTM was applied with calculus items that are characterized in terms of the number and type of operations a student must carry out to solve a differentiation problem (Fischer, 1973). LLTM model was also applied with verbal ability items of the ASVAB test ([Armed Services Vocational Aptitude Battery], Mitchell, 1983), NAAL items ([NAEP Assessment of Adult Literacy], Sheehan & Mislevy, 1990), and items of matrix problems in the ART ([Abstract Reasoning Test], Embretson, 1998). Application of LLTM is also suggested for analyses of different testing conditions that may influence an item's difficulty (also see Kubinger, 2009).

Although it is possible to explain a meaningful portion of the variances in the difficulty of an item, standard LLTM parameter estimates obtained from the fitted regression model are not expected to account for all the variations in true task difficulties. Further, items with identical features are constrained to have identical difficulty estimates. In addition, when maximum likelihood is used in estimating, the item parameter is chosen at a point, the distribution of the parameter is not taken into consideration. Mislevy (1988) offered a less restrictive approach for incorporating cognitive processing information into a psychometric model. In his approach, key

aspects of the LLTM are combined with the exchangeability concept of Bayesian inference. In this modified LLTM approach, Empirical Bayesian (EB), the true task difficulties are expected to be distributed about the central values predicted by the fitted LLTM-like regression model.

In Bayesian terminology, the difficulty parameters of tasks with the same values of the item features constitute a set of "*exchangeable*" parameters. In this approach, task difficulty is posited to be multivariate normal with mean ($Q'\eta$) and variance ($\phi^2 I$) where the mean structure is defined as in the LLTM. The full Bayesian model of LLTM assuming examinee independence, local independence, and exchangeability is as follows:

$$P(X,\theta,\eta,\xi,\zeta) = \prod_i \prod_j p(X_{ij} \mid \theta_i, \eta_k) p(\theta_i \mid \zeta) p(\eta_k \mid \xi) p(\xi) p(\zeta)$$

where ζ = higher-level parameters for the prior of examinee parameter, θ and ξ = higher-level parameters for the prior of contribution of various features in item difficulty. The prior for each error term ε_j can be assumed to be normally distributed with mean of zero (Lu, 2006). Use of a full Bayesian approach allows one to expect the shrinkage to be toward the center of the subset to which the item belongs.

A measurement system needs to be congruent with the nature of a learning domain, at least at the level of models and approaches. The unique nature of each domain needs to be integrated when defining the model of measurement. As research has suggested, reading comprehension is influenced by various attributes of readers as well as comprehension tasks. Those attributes need to be incorporated when synthesizing the performance of students to provide more informative estimates of what students know and can do in deriving meaning from texts.

Conclusion

The purpose of the review presented in Chapter II was to present research on reader and task attributes, and measurement models relevant to assessments of reading comprehension. This review provides empirical support for choices of the operational definitions of the variables and the selection of techniques to measure those variables included in the current study. Although initial work has been carried out in this direction by applying the LLTM retroactively to understand item difficulty, studies have yet to be conducted that integrate the conceptualization of comprehension offered in the reading literature with a measurement model in order to synthesize the observations of comprehension performance. Such integration is expected to improve inferences derived from comprehension assessments. In addition, studies have yet to be conducted that represent the complexity of comprehension by combining reader and task attributes. Future studies need to take into consideration the constraints of testing in order to enhance our understanding of the construct as measured by tests.

In summary, the overall message from the literature is that, in order to provide more informative results, an assessment of comprehension should be designed that incorporates a model of comprehension offered in reading literature, as well as having the integral features of the construct integrated into the measurement model. To this end, a study was designed where measures were built following the conceptualization of comprehension presented in this literature review. And, the examinee responses obtained from the measure constructed for the study were analyzed by a measurement model that incorporated the conceptualization to see what we can learn about the difficulty of the comprehension question and the comprehension proficiency of the students. This

exploratory study illustrates an approach to comprehension testing by scaffolding it with

research and theory.

CHAPTER III

METHODOLOGY

In this chapter, the study's critical elements—participants, measures, procedures, data analyses, and the rationale applied in selecting key features of those elements—are described. In addition, a pilot study conducted in preparation for the current study is described. Accordingly, the chapter consists of two major sections: descriptions of the methodology of the pilot study and the current study.

Pilot Study

A pilot study was conducted to serve three purposes: to determine the amount of time the participants should be allowed to complete the study task, to evaluate the materials prepared for the measure, and to find out if any adjustment would be necessary in the measures or in the data-gathering procedures for the current study. In addition, the pilot data were necessary to conduct a dry run of the computer programs prepared for the statistical and psychometric analyses conducted for the current study.

It was important that the study participants had sufficient time to complete the task given to them. All the participants should have time to read directions to perform the task that entailed reading four text passages and reading eight questions about each text passage, selecting the best answer for those passage-related questions, and responding to the questions seeking information on reader attributes. The time allowed, however, could not be unlimited. The pilot study helped determine the amount of time sufficient for participants to complete the study task.

In addition, a cognitive lab interview conducted with the pilot study participants was valuable in finding out the participants' understanding of the directions given to

them, in understanding the participants' thinking process while performing the task, and in identifying whether any ambiguities were present in the study's tasks. In order to strengthen the items' validity, Haney and Scott (1987) strongly recommend talking individually with the pilot study participants to learn how they perceive the study task and how they reason about the cognitive items, an information that is difficult to obtain by item analysis that is based on psychometric properties (p. 363). Self-reported information from the pilot study participants about topic familiarity with and topic interest in the text passages was valuable as a support for the appropriateness of the study's task and the procedures for other students like themselves.

In addition, observation of participants while they performed the study task allowed the researcher to see whether they followed the directions given to them, whether they skipped any portion of the task, whether they were engaged in the task, and whether they reread the text passages.

Participants

The pilot study was conducted with 25 students; about 60% from grade 8, the target grade and the rest from grades 9 and 10. The pilot study participants included female as well as male students and students of different races as identified themselves by the students. None of the participants were receiving services for special education (e.g., students with disabilities or limited English language proficiency). The pilot study sample comprised students from both public ($n=16$) and non-public ($n=9$) schools.

In addition, the measures were tried out on three other students. Two students were asked to participate in order to check for passage dependency of the questions included in the comprehension measure. One student was asked to participate in order to

check whether the directions included in the various measures would be interpreted accurately.

Measures and Procedures

All measures that were applied in the current study had been tested in the pilot study. These measures were comprehension, topic familiarity, topic interest, and demographics. All measures were compiled in a booklet in which participants recorded their responses.

The pilot study participants were tested either individually or in a group. Those who participated individually (n=11) were interviewed upon completion of the study's task using a set of questions in order to understand their thoughts on the questions posed in the various study measures.

Analysis and Results

Researcher's observations. All 25 pilot study participants completed the task in an hour or less. The minimum time taken to complete was 40 minutes and the maximum time taken to complete was 55 minutes. The majority of the participants (n=16) completed the task within 50 minutes. All participants were seen to reread the text passages when answering the passage-related questions and no students showed obvious signs of boredom.

The students who participated in the cognitive lab interviews were able to explain their rationale for selecting response options and to describe the differences between the response options offered for the reader attribute questions. They also stated that they liked the questions about "yourself" referring to the questions asked in the reader

attribute measures, and mentioned that they liked the task and learned something new from reading the text passages. However, they found the questions asking about their expectations about content given the titles of the text passages difficult to answer and they were surprised when their expectations differed from what the texts actually described.

Participants' responses. The responses of the 25 pilot study participants were processed following the plan set for analyses of the current study. These processing helped to (a) establish data entry and data quality check procedure, (b) finalize data coding procedure, and (c) dry run all programs written for statistical analyses using SAS software and psychometric analyses using codes of WinBUGS, a software to conduct psychometric analyses in Bayesian framework.

The responses were checked to see whether any participants left any questions—either comprehension or reader attribute—unanswered before they were transferred from booklets to an electronic file. The data read by the SAS program were first crosschecked with hand-coded scorings for the purpose of establishing data consistency. Then, the responses for comprehension questions and reader attributes were coded following the coding scheme described in the measure section for the current study. For the data quality, the checks that were performed included: (a) classical item difficulty using p-values, (b) item discriminations using point-biserial correlations, and (c) internal consistency using Cronbach's alpha coefficients.

In sum, the results showed that no student answered all comprehension questions correctly and no student answered all questions incorrectly. Across the four text passages, the total number of items answered correctly ranged from 11 to 26 ($M = 19.92$,

SD = 3.85). The means of the individual passage total scores for four texts ranged from 4.28 (*SD* = 1.49) to 5.56 (*SD* = 1.29).

For the 32 comprehension questions, the *p*-values in the pilot study ranged from .16 to 1.00. Two items had perfect *p*-values and three had a *p*-value greater than .90. The Cronbach's alpha coefficients for questions asking about topic familiarity and topic interest differed among text passages. The Cronbach's alpha coefficient for a text passage titled *Kid Fights Cheater Meters and Wins* was relatively smaller than the other three text passages.

The specific values of these data quality indices were considered not important per se, especially in light of the small sample size. Rather, they served as checks for flaws in tasks or instructions—for example, finding many items with nearly all correct responses or nearly all incorrect responses would have called for adjusting the difficulty of the passages or the tasks.

Using the two reader-attribute variables, students were classified into four groups to see whether they were familiar with as well as interested in a text passage. More students reported having topic familiarity with one text passage titled *Going Green* compared to the other text passages. In general, more students reported to have an interest in the narrative type texts (*Cry of the Kalahari* and *Kid Fights Cheater Meters and Wins*) compared to the expository type texts (*Going Green* and *Shifting Sands*).

For the two participants who were asked to answer the comprehension questions without reading the relevant text passages, the 9th grader answered 10 of the 32 items correctly and the college student answered 14 correctly. They both took about 45 minutes to complete the exercise. In a passage-dependency study, Tuinman (1972-73)

found 50% of the items in five "widely used" reading comprehension tests were answered correctly when no reading of the passages was allowed. Tian (2006) reported that 34% of the items in the TOEFL comprehension test were answered correctly in a passage-out condition.

Revisions Recommended

Based on the pilot study, some revisions were made to the measures. Specifically, modifications were made to (a) the response options for five items that had *p*-values equal to or greater than .92, and (b) three reader-attribute measure questions in order to make them more meaningful. In addition, following a conversation with school personnel, it was decided that study participants would not be asked to sign the assent form because they are minors.

Current Study Method

Participants

One hundred and sixty ($n = 160$) students attending 8th grade in school year 2011-12 participated in the study. There were 66 male (about 41%) and 94 female (about 59%) participants. Among them, 85 were self-identified as White (i.e., 53% of the participants), 35 as Black (i.e., about 22%), and 40 (i.e., about 25%) as other races (e.g., Asian, Hawaiian/Pacific Islanders, Hispanic, or mixed). None of the participants was identified as a recipient of services for special education (i.e., students with disabilities or limited English proficiency). Among the participants, about 79% attended public and about 21% attended non-public schools located in a Mid-Atlantic state. All students participated with permission from their parent or guardian.

Students attending grade 8 are routinely assessed for their reading comprehension ability at the state and national levels in order to report the status and progress of student performance and to adopt educational policies. In the U.S., this practice has been mandated since 2000 by a law known as the Elementary and Secondary Education Act (ESEA, 2002). In addition, many studies that have been conducted on reading comprehension included students from this grade (e.g., Bray & Barron, 2003-2004; Johnston, 1984; Johnston & Pearson, 1982). This last grade of middle-school receives much attention perhaps because the students at this grade are required to read extensively and to read materials from different learning domains, a task difficult to accomplish without the comprehension of what is read.

Further, students in 8th grade, compared to those in 4th grade, another grade frequently included in comprehension assessments and studies, seemed to be more appropriate for exploring the effects of reader attributes, such as the effect of a reader's familiarity with and interest in what is being read on comprehension. The literature review conducted for the current study showed that a greater number of studies exploring the effects of interest in text comprehension of school level children included students from a higher rather than a lower grade level.

According to the Model of Domain Learning (Alexander, 1997), it is expected that students at grade 8 are most likely be acclimated learners. They may have very little domain knowledge and may not have deep-seated individual interests. However, they are expected to have had more exposure to a variety of texts and to have more developed interests than those students in earlier grades. It is possible that some 8th-grade students

are developing toward the stage when their interest is less transitory, and they have less difficulty in understanding and recalling what is read.

In the current study, no sampling weights were applied in any estimation. Further, no attempt was made to consider the sample a representative of the school district that cooperated in the study or the state where the students attended schools, in general. The students in the sample, however, were members of the population targeted for the assessment measure, that is, the 8th grade students. In addition, the sample size was considered sufficient to support the various statistical and psychometric analyses that were conducted to address the research questions. That is, the purpose of the study is to develop and implement an assessment design and psychometric method, rather than to attempt to estimate generalizable effects of the kinds that the method could be applied to study in future research.

Although a large sample is recommended to obtain stable parameter estimates using IRT or LLTM, this relatively small sample size was considered sufficient. Lord (1983) considered a sample size of between 100 to 200 students to be appropriate for the one-parameter logistic IRT model (1PL and AKA the Rasch model for the dichotomous items). Embretson (1983) said: "a size of 150 to 200 subjects is considered reasonable for exploratory study applying LLTM" (p. 190). Mislevy (1988) found a sample size of 150 sufficient when he explored inter-item similarities using LLTM.

The sample size used in the study supported the other analysis. A factor analysis was conducted to see whether the questions asking about reader attributes load on one factor. Gorsuch (1983) suggests using "five individuals to every variable, but not less than 100 individuals for any analysis" in conducting exploratory factor analysis (p. 332).

The current study had six variables related to reader familiarity and seven variables related to reader interest. The sample size of this study is comparable to the sizes used in other reading comprehension studies in which the sample size was as small as 16 and as large as a few hundred.

Measures

The current study used four measures: comprehension, topic familiarity, topic interest, and demographic profile. These multiple measures were constructed following the conceptual model of comprehension, presented in Figure 1, to address the research questions that were based on the premise that comprehension of a text passage is influenced by the attributes of a comprehension task as well as of the reader. This view is held by many reading researchers (e.g., Alexander, 1977a; 2005; Anderson, 1997; Duke, 2005; Kintsch, 1998; RRSG panel, 2002; Rumelhart, 1975; Wigfield, 1997; van den Broek et al., 1990). It should be noted that the study applied the most frequently used procedure, a paper-and-pencil mode, to assess comprehension. The choice of using this common technique did not allow for more extensive reflection on the processes of meaning-making.

Comprehension measure. The reading comprehension was measured using four text passages and eight questions about each of those passages (see Appendix A). Two of those passages were exposition and two were narrative. Each text dealt with a unique topic. Each passage-related question, referred to as a cognitive item, targeted one of two types of mental representation (i.e., textbase and situation model), and asked for information on different relations (i.e., causal, temporal, and spatial) that underlie the events in a situation described by a passage. Each cognitive item was followed by four

answer choices from which a reader selected the best answer. Each of the 32 cognitive items was scored as "1" if correct or "0" if incorrect.

This comprehension measure was constructed based on Kintsch's conceptualization where comprehension is said to vary by levels—the textbase, that is, a text-driven representation and the situation model, that is, a text and knowledge driven representation of a text. According to Kintsch's (1998) CI model of comprehension, the construction of textbase and situation model representations are posited to indicate a full comprehension of a text. An understanding of various dimensions or relations (i.e., temporal, spatial, and causal) is important for comprehension of what is read, because those relations or dimensions are components of a situation (Kintsch, 2004). The understanding of one relation may involve more complexity than the understanding of another relation. For instance, the understanding of a causal relation involves identifying at least two logically connected events.

All questions included in the comprehension measure were formatted as multiple-choice-question (MCQ), a format that is ubiquitous in contemporary comprehension tests. This format offers many advantages over others (e.g., constructed response). The format poses minimal burden on the examinees because they are not required to write when responding to the questions, the evaluation of examinee responses obtained through this format is more objective as no rater judgment is necessary, and the format is considered convenient operationally because scoring of responses is less time consuming (e.g., Campbell, 2005). Most importantly, this format was suitable for the measurement model, structured IRT that was applied in psychometric analyses. The aim where was to integrate a cognitive theory of comprehension with the measurement model.

Reader attributes measure. This measure had two components; one was designed to measure the readers' topic familiarity (see Appendix B) and the other was designed to measure the readers' topic interest (see Appendix C). The importance of these two attributes is emphasized in Alexander's MDL because they were found to contribute, individually or collectively to the comprehension of a text (Alexander, 2005).

These measures included questions that asked the readers to self-report about their familiarity with and interest in the topics of the text passages included in the comprehension measure. Some of the questions that asked about reader attributes were posed prior to the reading of any text passages included in the comprehension measure. These questions are referred to here as prereading attribute questions. Some of the questions were posed after reading of the last cognitive item of the last text passage and are referred to as postreading attribute questions. In addition, some questions about topic interest were posed immediately after reading the passage, but prior to reading any passage-related questions. They are referred to as immediate-reading questions.

Three prereading and three postreading questions were posed to determine readers' familiarity with the topic of a text passage. Two prereading topic familiarity questions asked readers how much they knew about two concepts identified as related to each text passage included in the comprehension measure. The other prereading familiarity question asked the readers what they expect a text to cover given its title. The postreading questions asked readers (a) whether they had previously read a passage similar to what they just read, (b) whether they knew a lot about the topic, and (c) how much their prior knowledge helped them understand what they read.

To measure readers' topic interest, three questions were posed as prereading, two questions were posed as immediate-reading, and two as postreading. Two prereading questions asked readers how interested they were in knowing about two concepts, also referred to in the familiarity question. One prereading question asked readers for their expected interest given the title of the text. One immediate-reading question asked how much they enjoyed the text and the other asked whether they would read the text again. One postreading question asked readers how much they would share with others what they read. The last postreading interest question asked readers how interesting they found the passage.

For the current study, multiple indicators for topic interest were suggested. In MDL, two forms of interest—personal and situational—are recognized (Alexander, 1997b). In addition, research indicates that interest measured prior to performing a comprehension task is different from that of the measure taken after the task. According to Ainley et al. (2002a), prior measurement represents an expectancy measure. It represents a reader's estimation of his/her likely psychological state. The interest measured after completion of the task requires the reader to reflect back. In addition, the measure of interest taken immediately after reading a passage allows a researcher to see whether interest has changed. Questions about interest posed immediately after reading a text, generally, are asked to counteract the recency effect and are included when comprehension is assessed by free recall (Cote, Goldman, & Saul, 1998).

In both cases, for topic familiarity and topic interest, two concepts per passage were chosen because some researchers suggest three concepts are sufficient for texts

where the length ranged from 3 to 5 pages (e.g., Langer, 1981; Zakaluk, Samuels, & Taylor, 1986). Each text passage selected for the study was approximately two pages.

In the current study, it was possible that topic interest influenced comprehension, and comprehension of a topic influenced the level of interest a reader expressed after reading a text. However, the study was not designed to disentangle those potential reciprocal effects. Rather, the purpose of this study was to illustrate through a psychometric model the passage-related interest as it influences comprehension. Having successfully implemented such a model, one could utilize it in studies in which substantive rather than methodological questions would be of primary interest.

The number of questions (six for topic familiarity and seven for topic interest) included for each reader attribute was considered sufficient. Gorsuch (1983) says that replication of factors is possible, "with at least four, preferably six variables per factor" (p. 332). The goal of the study was to summarize interest and familiarity variables to identify each student into simple dichotomous indicators for each passage, such as students who were more interested in a passage and those who were less interested in a passage. Similarly, students were identified as those who are more familiar and those who were less familiar with a passage.

Each reader attribute question had four response options. The response options differed in the extent to which a response was positive or negative. The response options used to indicate extent of topic familiarity or topic interest were: "Not at all," "A little," "Some," "A lot." Response options to indicate expectation of interest were: "No," "Maybe not," "Maybe," and "Yes." Data on self-reporting of attributes were collected

using a 4-point instead of a 5-point scale in order to avoid possibilities of misclassification of the responses or loss of information.

Demographic measure. In order to get a better description of the sample, a demographic measure was prepared (see Appendix D). The measure included questions on gender, race, and how much reading participants usually do. The questions on race and gender were posed after asking all other questions. The question on the extent of reading was posed prior to asking any other questions and offered response options: "Never or Hardly ever," "Sometimes," "Often," and "Every day or Almost every day." This question was included because disposition toward reading is considered an indicator of good readers (Valencia et al., 1989).

All questions on self-reporting were initially constructed by the study researcher following examples presented in the literature on reader attributes related to reading comprehension (e.g., Ainley et al., 2002a; Bray & Barron, 2003; Guthrie, McGough, & Wigfield, 1994; Schraw, 1997). The response options included in the measures had been used in prior studies and assessments that collected self-reported background information. These measures were reviewed by experts including an experienced teacher and tested in the pilot study before being applied in the current study.

Responses to the questions on topic interest, topic familiarity, and extent of reading were coded in three steps in order to dichotomize the responses.

Step 1: The four response options were coded so that a student who responded most positively were placed at the top 20% and a student who responded most negatively were placed at the bottom 20% on a scale. In this coding scheme, the most negative to most positive responses were coded as .1, .3, .7, and .9 respectively.

Step 2: The coded responses were summed to get a composite score representing an attribute, topic familiarity or topic interest. Each composite score was then divided by the number of variables related to the attribute in order to re-express the data in terms of scale score. For two passage-specific attributes, topic familiarity and topic interest, the responses were summed as related to a text passage. Then, the composite score for topic familiarity was divided by six, the composite score for topic interest was divided by seven, and the composite score for extent of reading was divided by five, respectively. The result was an average across all the questions contributing to the particular reader attribute, which was on a scale that ranged from 0 to 1.

Step 3: The obtained scaled score for each study participant was compared to the mean value of the distribution based on the study sample to determine whether the participant, for instance, was familiar or not familiar with a topic. If a participant's scaled score was higher than the mean value of the obtained distribution, the respective student was identified as familiar and was coded as "1," otherwise as "0." This coding scheme was also applied in determining whether a study participant was interested or not interested in a topic.

Materials

This section describes the materials included in the comprehension measure. Among the four passages, two texts were of narrative type and two were of expository type. Each passage described a distinct topic. These passages were obtained from 8th-grade reading test materials released to the public by the education agencies of various state and federal governments, which regularly conduct assessments on K-12 students.

This selection of passages from a pool of texts that have already been administered in formal assessments of 8th- grade student comprehension proficiency was expected to ensure sensitivity of the study materials, that is, appropriateness of the text to the population targeted for the current study. This expectation was based on the assumption that those passages were reviewed and validated for content and sensitivity by the agencies that administered them. In addition, selecting passages from materials released to the public guards against copyright violations by the researcher.

Those four passages were as follows. For the two narrative text passages, one text titled, *Kid Fights Cheater Meters and Wins*, downloaded from the web site of NCES[3], tells a story about a middle-school girl who sparked a revolt on a city's parking meters; the other narrative text titled, *Cry of the Kalahari*, downloaded from the web site of the Florida State Department of Education[4], recounts two researchers' adventure in the Kalahari desert that gave them the rare opportunity to learn about the natural behavior of wildlife. For the two expository texts, *Going Green*, downloaded from the web site of the Mississippi State Department of Education[5], describes the green movement in America by presenting a comparison between the early and late stages of the environmental movement. The other expository text, *Shifting Sands*, downloaded from the web site of the Ohio State Department of Education[6] , describes problems associated with the shifting sands of the Jockey's Ridge dunes located in North Carolina.

---

[3] http://nces.ed.gov/nationsreportcard/itmrlsx
[4] http://fcat.fldoe.org/pdf/releasepdf/grade8readingtestbook.pdf
[5] http://www.mde.k12.ms.us/acad/osa/LA_Grade_8_v2.pdf
[6] http://www.ode.state.oh.us/GD/DocumentManagement/DocumentDownload.aspx?DocumentID=15039

The selection of the text passages began with a passage bank, initially developed by Ought, Mugo, and Tirre (2010) and augmented by the researcher of the current study. All of the 90 text passages in the bank were reviewed again for possible selection in this study. The review focused on text type, text length, readability level and the language included in the text, to ensure that the passages included in the measure were suitable to address the research questions posed in the current study.

In the review of text passages, the text type was checked to ensure that only prose discourse was included. The language included in the text was checked to ensure that the study participants would not expend their text reading time in decoding words that required familiarity with a certain culture. The readability or grade level of each text passage was checked so the passages were not below the target grade level. The passage length, including the number of words and sentences, was checked to ensure that study participants could complete their reading of the texts within the time allowed for task completion. The length was also checked to ensure that a passage had enough content to construct the same number of comprehension questions from each passage. It is argued in the reading literature that short texts cannot provide much information on how people comprehend texts (e.g., Hannon & Daneman, 2001; Valencia et al., 1989). The length and grade level of the passages were checked using the Microsoft Office Word program. The program measures text length with a count of the number of words and sentences, and provides an estimation of readability using the Flesch-Kincaid formula for grade levels, which takes into, account the total words, total sentences, and total syllables.

This review process helped systematically eliminate some of the text passages from consideration. The review of text structure led to elimination of those passages that

were poems, dramas, speeches, letters, or excerpts from longer works. Texts that were biographies were also eliminated as some researchers consider them as of mixed type texts (Alexander & Jetton, 2000). The review of grade level (ranging from 1.3 to 11.8 with $M = 6.7$) led to elimination of those passages identified having a grade level lower than 8.0. The review of passage length (ranging from 91 words to 1820 words with $M = 690$) led to elimination of those passages that were shorter than 500 or longer than 1,200 words. This process of elimination narrowed the list to 19 text passages.

The short list of passages was then reviewed by two reading researchers and three teachers with experience teaching at the middle school level. The purpose of the review was to confirm the text type identified for the passages and reveal any issue that might require further elimination from the list. The teacher reviewers were also asked to approximate the amount of time a student from 8th grade might take to read each of those text passages in order to ensure that a selected text passage did not add any unnecessary burden on the study participants. Input from the teacher reviewers was also sought on topic familiarity to ensure that the prospective participants would not be familiar with all text passages selected for the study.

Based on comments from this group, the list was shortened to 11 text passages. In making the final selection of four passages, the goal was to make sure that each selected text described a distinct topic, the nature of the content was comparable across the passages, and the selected passages had information sufficient to construct the necessary number of items. The passage features of the four texts are described in Table 4.

Table 4

Features of the Four Text Passages Included in the Comprehension Measure

Text passages	Text type	Length in words	Number of sentence	Words per sentence	Grade level	Content type
Cry of the Kalahari	Narrative	923	48	19.1	8.4	Non-fiction
Going Green	Expository	686	38	16.6	10.8	Non-fiction
Kid Fights Cheater Meters and Wins!	Narrative	708	48	14.8	8.2	Non-fiction
Shifting Sands	Expository	575	28	20.5	9.0	Non-fiction

As noted, each text passage was followed by a set of cognitive items in order to assessing comprehension of the passage. An item targeting the textbase comprehension asked for information that could be derived from what is explicitly stated in the passage. Answering such an item requires some paraphrasing and integration of information given in sentences (i.e., connecting information or "bridging" the sentences). An item targeting the situation model of comprehension asked for information that required making inferences by integrating or reasoning through the information given in the passage. Of the total 32 cognitive items, half targeted the textbase and the other half targeted the situation model mental representations. Similarly, of the eight cognitive questions following each text passage, four targeted textbase and four targeted situation model comprehension.

Each cognitive question included in the study measure, whether targeting the textbase or the situation model of mental representation, asked about either a causal or a non-causal relation. A cognitive item that inquired about the causal relation asked for the

causal antecedent or the causal consequence relation. A cognitive item that inquired about the non-causal relation asked for information on a temporal relation or spatial relation. A pair of questions (one textbase and one situation model) targeted each of these four relations underlying the events of a situation presented in a passage. Thus, among the items, there were two sets of items if they were classified according to the type of mental representation, or four pairs if they were classified according to the type of relations. An item's identification following the classification scheme adopted in the study is indicated in Appendix A where the comprehension measure is presented.

The number of cognitive items, 32, was determined to be sufficient to represent the cognitive complexity identified for the study; there were replications of each text type, CI-model type, and inferential structure type. This number is within the range used in other tests and studies (as reviewed in Chapter II) that probed comprehension proficiency.

All cognitive questions were initially constructed by the study researcher. In constructing cognitive questions, the researcher consulted examples of textbase and situation model questions offered by others (e.g., Best, Floyd, et al., 2008; Magliano et al., 2007; Ozuru et al., 2009). The researcher also consulted suggestions offered by E. Kintsch (2005) for asking questions to probe aspects of comprehension as seen through the frames of CI theory, as well as research on relations embedded in events reviewed in the previous chapter.

The construction of response options used the following guidelines. For each cognitive item, four response options were constructed; one of those four options was considered correct, or the best answer for a target question. The three other options were

within the category of true but not pertinent, partially true, and false. In other words, the

choices were constructed so they would differ in the degree of plausibility. This

guideline was applied in constructing the response options by Johnston and Pearson

(1982) when they developed their comprehension measure, was recommended by E.

Kintsch (2005; personal communication, 2010), and was found applied in some test items

that were reviewed for this study. In constructing the items, attention was paid to the

guidelines offered by Haladyna (1997). Attention was also paid to the vocabulary

included in the stems and response options in order to ensure that the vocabulary would

be suited to the students as indicated in the list identified for the target population by

Taylor et al. (1989).

Six experts reviewed all cognitive items. They were three reading researchers

who are very familiar with Kintsch's model of comprehension and the relations

embedded in the events of a situation, two test specialists from a nationally known test

publisher who write multiple-choice items, and one individual who professionally

reviews comprehension items for test publishers. The purposes of the item review were

to make certain that the items fit the classification scheme (i.e., mental representation

crossed with dimensions or aspects of events), to detect any ambiguities in the stem or

answer choices, and to identify any feature, such as response position, length, or

language, that might distinguish the correct option from the distracters. The steps

followed in constructing the cognitive items as they fit into the two-dimensional

classification scheme are summarized in Table 5.

Table 5

Steps Followed in Developing Cognitive Items

- Identified the segments in a text relevant to the various dimensions or relations and identified the relational prepositions
- Constructed questions on each relation that targets the textbase level of mental representation
- Constructed questions on each relation that targets the situation model level of mental representation
- Constructed the response options, one as correct and three others as distracters
- Conducted a review of items constructed for one expository and one narrative text passages by two experts on CI-model of comprehension
- Constructed multiple items in some cases and asked the experts to indicate the item that fits the category best
- Constructed items for all the text passages selected for the study after receiving inputs from those two experts.
- Conducted item reviews by four experts. One expert was considered very knowledgeable in the CI model of comprehension and research on dimensions of situations. Two of them, as their profession write items for a testing corporation. The remaining one, as a profession reviews comprehension items for test publishers. The items were sent for review to ensure that each item fits the classification scheme, the response options meet the constraints of multiple-choice items, and to find any ambiguities
- Modified the items based on the inputs received from the reviewers. The modifications focused, in most cases, on structure of the response options, language of the stem and response options.
- Conducted review of the items by a teacher in order to see whether the language included is appropriate for the target students.

In order to alleviate possible order effects, the order of text passages presented to the students was counterbalanced. Therefore, four sets of booklets were prepared so every passage could be in every position. The order of the passages in the four sets were as follows: Set 1—*Shifting Sands*, *Kid Fights Cheater Meters and Wins*, *Going Green*, *Cry of the Kalahari*; Set 2—*Kid Fights Cheater Meters and Wins*, *Going Green*, *Cry of the Kalahari*, *Shifting Sands*; Set 3—*Going Green*, *Cry of the Kalahari*, *Shifting Sands*, *Kid Fights Cheater Meters and Wins*; Set 4—*Cry of the Kalahari*, *Shifting Sands*, *Kid Fights Cheater Meters and Wins*, *Going Green*.

Further, the order of presentation of the cognitive items targeting different mental representations and event dimensions was varied in the four text passages so a pattern was not apparent. In Appendix A, where the comprehension measure is presented, shows the order in which the questions were presented to participants.

Procedures

Grade 8 students attending public or non-public schools participated in the study. The study involved one session per participant. However, data were collected over several days as the participants attended different schools. The data were collected during March to June 2012. The majority of the students (approximately 97%) participated in a group setting (i.e., more than one student was in the room).

On the day the study was conducted, after the participants were given an introduction to the researcher and the study, they were then given instruction for how to proceed with the study. They were asked to carefully read the directions given in the booklet. At this point, booklets and pencils were distributed. Nearly all participants completed the task within an hour. Many of the participants completed the task in about

40 minutes. When the participants returned the booklets, they were thanked for their participation. The school and school district were thanked in writing. All instructions and materials administered in the data collection reflected revisions indicated necessary by the pilot study and the study's advisory committee. See Appendix E for verbatim instructions given to the participants.

After completing the data collection, the responses of the participants recorded in booklets were examined for stray marks as well as omit patterns, that is, missing answers for any question, and/or selecting a certain response option. In only a few cases, concerning three students, a total of four questions were left blank. After completing initial checks of data quality, responses to the cognitive items as well as the reader attribute questions were transferred from booklets to a data file. To maintain confidentiality, the data file did not include any information on name of the classroom teacher, school, or school district. Computer generated random numbers assigned to the participants' booklets were used as the identification numbers in order to protect the participants' identities.

Data Analyses

The data obtained for the study were analyzed applying statistical and psychometric procedures. The analyses for this cross-sectional and factorial study design were carried out in two phases. The Phase I analyses were conducted to describe the data and to check its quality in order to see whether critical assumptions were met for various analyses: Analysis of Variance (ANOVA), Factor analysis, t-test, and Item Response pattern analysis that were undertaken in Phase II. The analyses in Phase II, statistical and

psychometric in nature, were conducted to address two research questions identified in the purpose of the study.

Phase I. Step 1: A classical item analysis was conducted to obtain a *p*-value for each cognitive item (i.e., the proportion of students that answered an item correctly, or the mean of an item that was scored dichotomously). Then, a series of biserial correlations (i.e., item total correlation) were calculated to see the discriminating power of an item. In addition, the reliability of scores was checked using a measure for internal consistency as generally checked with Cronbach's coefficient of alpha.

Step 2: Descriptive analyses were performed to calculate mean, median, standard deviation, and skewness in order to describe the distribution of the number of correctly answered cognitive items. The score distribution was described for the total scores, that is, the total number of correctly answered cognitive items related to a passage as well as the number of items correctly answered across the four passages. The passage total scores ranged from 0 to 8 and the grand total scores ranged from 0 to 32, with 32 being the maximum.

In addition, there were sub-total scores, that is, the total number of correctly answered cognitive items for the task attributes. Those scores represented textbase mental representation (scores ranged from 0-16), situation model mental representation (scores ranged from 0-16), target relations—causal antecedent, causal consequence, spatial, and temporal (scores ranged from 0-8 for each relation), text type—expository and narrative (scores ranged from 0-16 for each type) and text topics—(scores ranged from 0-8 for each topic). The number of correctly answered items was also aggregated for those who were identified as familiar and not familiar; likewise for those who were

identified as interested and not interested in a topic. The information on score distributions was meant to check whether the analyses, *t*-test and ANOVA, intended for the study were tenable with the data.

Step 3: A set of inter-item correlations was calculated with questions that asked about topic familiarity and topic interest for each passage to see how those questions relate to each other and to see whether the correlation matrices would be factorable (i.e., whether a correlation matrix was not an identity matrix). These analyses were conducted prior to applying the factor analysis procedure, which was conducted to see whether the variables load on a factor. In addition, the reliability of scores was checked using Cronbach's coefficient of alpha.

Step 4: Another set of analyses was conducted to identify the distribution of the reader attributes of Familiarity and Interest. The distribution of scores was calculated to determine where to induce a cut-point for 0/1 scores on those two variables for each passage before conducting further analyses. In addition, a factor analysis was conducted to see whether classification based on differential weighting of the questions led to different results than the classification based on equal weighting of the questions. The number of students identified as familiar and not familiar or as interested and not interested, however, did not differ between the two ways of classification. Then, it was decided to carry on the analyses without using factor loadings in creating two variables for each passage: passage*familiarity and passage*interest.

It should be noted here that the indices of reliability that are generally seen in large-scale testing would not be tenable, because the sample size and the number of items included in the study were relatively smaller than what is generally included when

developing standardized individual tests. In this study, the comprehension tasks were developed to look at patterns over items to see whether the difficulty for an individual item appeared to be related to task attributes and to the reader attributes of familiarity with and interest in a passage topic.

Further, the current study was conducted to model the influence of reader-by-task factors, which are traditionally part of measurement error in true score and IRT models. For these reasons, low traditional reliabilities, in and of themselves, were not germane to the focus of the study, as they would have been if the study was meant simply to use total scores in some theoretical network. Instead, the purpose of these analyses was to check data quality, as certain errors in item writing or scoring manifest themselves in low reliability indices. Low indices would thus be employed here as triggers for further data quality checks.

Phase II. In this phase, a set of analyses was conducted to address the research question "In what ways do the attributes of task and reader influence comprehension of the 8th-grade students who participated in the study?" The goal of analyses was to detect the main effects of task and reader attributes.

To address the question, a set of *t*-test (paired and independent sample) and ANOVA were conducted with the task and reader attributes. The sources of variability for the task attribute were text topic, text type, mental representation of a text targeted in passage-related questions, and aspects of the situation described in a text that those questions asked about with focusing on causal and non-causal relations. The sources of variability for the reader attribute were topic familiarity and topic interest.

Step 1: For the main effect of text topic, a one-way repeated measure ANOVA was conducted with each text passage score as the dependent variable and four topics as the independent variable to see whether the total number of correctly answered questions varied by topic. Then, a set of paired *t*-test was conducted with the total number of correctly answered items as a dependent variable. In those *t*-tests, independent variables that had two levels were: text-type (expository and narrative types), text topics within a text type (i.e., expository text topics: *Going Green and Shifting Sands*, narrative text topics: *Cry of the Kalahari* and *Kid Fights Cheater Meters and Wins!*), the target mental representation (textbase and situation model), and the target relation (for which comparisons were made between: causal antecedent versus causal consequence, temporal versus spatial, and causal versus non-causal).

Step 2: For the main effects of reader attributes, a set of *t*-test was conducted. In this comparison, the dependent variable was the number of cognitive items correctly answered and independent variables were reader attributes, familiarity and interest, with each attribute coded dichotomously (i.e., familiar or not-familiar, interested or not-interested). The analysis was conducted for each attribute four times, once for each text passage with that passage's total score, (i.e., a sum over eight items). This was because a given student could be interested in some topics and not others, so the effect would not be seen if the analysis combined the scores across the passages. In addition, the study participants were classified into four groups following the final codings assigned to the reader attributes. Those groups were as follows: (a) Group 1 included those readers who were identified as having both familiarity and interest in the topic, (b) Group 2 included the readers who were identified to have familiarity but no interest in the topic, (c) Group

3 included the readers who were identified to have no familiarity but have interest in the topic, and (d) Group 4 included the readers who were identified to have neither familiarity nor interest in the topic.

These four groups were formed to examine the effects of two reader attributes. A between-group ANOVA was conducted four times, once for each text passage with that passage's total score, (i.e., a sum over eight items). As stated earlier, topic familiarity and topic interest of a given student might not be the same across the four passages, so the effect of those text passage-related attributes would not be seen if the analysis combined the scores across the passages. No interaction effect of those two attributes was examined either in ANOVA or in the psychometric analyses because students were not assigned randomly to the groups. The groups were determined from students' expressed topic familiarity and topic interest rather than formed by the researcher. As a result, the four groups were not balanced. A separate study with balanced groups would be necessary to rule out any confounding due to the correlation between topic familiarity and topic interest and for an in-depth exploration of the relation between those two reader attributes and comprehension.

Step 3: The next set of analyses was conducted on the joint distribution of individual cognitive items (i.e., patterns in student responses were analyzed to estimate item difficulty and student comprehension proficiency by applying an IRT model). In the psychometric analyses, where response to each of the 32 items was analyzed jointly, also examined the joint effect of the two sets of attributes, task and reader. This psychometric analysis was conducted in several steps to address the research question: "Do the task and reader attributes considered in the study contribute to the item difficulties and reader

comprehension proficiencies?" Based on the model of comprehension posited in the study that underlie the Belief Network shown in Figure 2, four sets of analyses were conducted to address the research question.

The analyses in step 3 were conducted applying the IRT Rasch model and the LLTM model (Fischer, 1973). Rasch model estimates were obtained as base estimates, which were compared with the estimates obtained by LLTM models. The analysis required comparing the estimates of item difficulty and examinee proficiency with and without incorporating the interaction of task and reader attributes as indicated in cognitive constructive views of comprehension. The LLTM analyses incorporated both item-specific attributes and reader*text specific attributes as shown in the Belief Network presented in Figure 2[7]. Item specific attributes included text-type, text-topic, and the nature of the items (i.e., if an item targeted the textbase or situation model, and if the item asked for a causal consequence or causal antecedent or spatial relation or temporal relation). Reader*text attributes included familiarity and interest. Thus, the network suggests that attributes of task and/or reader and comprehension proficiency causes the effect, that is, an item is answered correctly.

---

[7] In order to maintain clarity, the belief network, shown in the Figure 2, represents only 16 of the 32 cognitive items included in the comprehension measure.
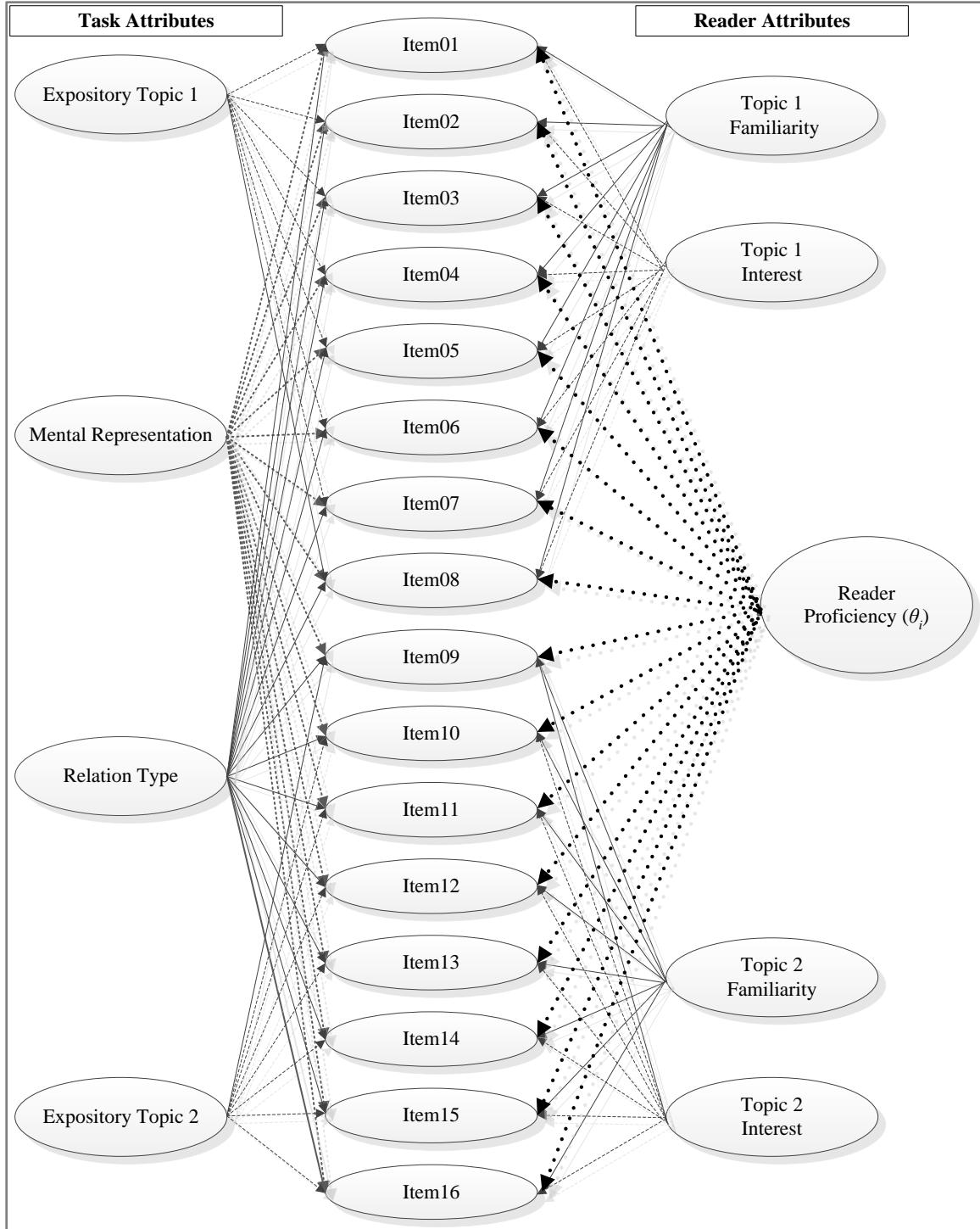
Figure 2. Belief Network Represented in the Pychometric Analyses

The focus of IRT and LLTM models in the estimation of proficiency is on the

probability that students at different points on the ability scale will correctly answer a

cognitive item. The probability of a given response, x_{ij} , is a function of the ability of an

examinee, θ_i , and parameters of an item, j. In the *Rasch* and LLTM models, there is one

parameter for each item, namely its difficulty, β_j. The probability that an examinee

makes a correct response to an item j according to the one-parameter model is:

$$P_j(\theta_i) = P(x_{ij} = 1 \mid \theta_i, \beta_j) = \frac{\exp(\theta_i - \beta_j)}{1 + \exp(\theta_i - \beta_j)}$$

   where:

   x_{ij} = denotes the response of an examinee i to item j; a response is 1 if correct and 0 if incorrect
   β_j = represents difficulty of the item
   θ_i = represents the ability of an examinee, in our case the reader.

In the above equation, it is not the individual item probabilities that are of primary

interest, but rather the item difficulties and the person proficiencies (Embretson, 1983).

   Under the assumption of local independence, the probability that a vector pattern,

$x_i = (x_{i1},...,x_{in})'$ represents responses to n items is:

$$P(X_i \mid \theta_i, \beta) = \prod_j P_j(\theta_i)^{x_{ij}} (1 - P_j(\theta_i))^{1 - x_{ij}}$$ where $\beta = (\beta_1,...,\beta_n)'$. Assuming independence of

responses over examinees, the probability of the data matrix: $X = (x_1...,x_N)'$ for N

examinees is the product of expressions like $P(X \mid \theta, \beta) = \prod_j P(x_i \mid \theta_i, \beta)$.

   In LLTM, the difficulty parameter of the Rasch model, β_j is postulated as a linear

model of the item characteristics: $\beta_j = \sum_{k=1}^{K} q_{jk} \eta_k = q_j'\eta$, or in matrix notation as shown in

Sheehan and Mislevy (1990), $\beta = Q'\eta$, where Q' is an n by k matrix of item feature

data and $\eta = (\eta_1,...,\eta_k)$, where η_k represents the additive contribution of feature, k to an

item's relative difficulty. In other words, the item difficulty parameter is estimated as linear combinations of a smaller number of more elementary components. The elementary components are defined to reflect differences in the cognitive processing demands of the tasks in order to address the problem of why some items are more difficult than others are. Thus, the probability of a correct response under the LLTM is:

$$P_j(\theta_i) = P(x_{ij} = 1 \mid \theta_i, \beta_j, q_{jk}, \eta_k) = \frac{exp(\theta_i - \sum_k q_{jk}\eta_k)}{1 + exp(\theta_i - \sum_k q_{jk}\eta_k)}$$

q_{jk} = represents the degree to which feature of k applies to item j,
  q_{jk} is 1 if feature k is present in item j and 0 if not.
η_k = represents the contribution of feature k in item difficulty
θ_i = represents the ability of an examinee, in our case, the reader.

Application of LLTM requires two kinds of data: responses to an item and coding for each item on some variables that represent the theoretical complexity factors of the item. In the LLTM analysis with item attributes only, items with identical features have identical estimates of item difficulty (Mislevy, 1988). With additional features from the readers, the LLTM analyses allow for the observation of an interaction between item-specific and reader*text-specific attributes. The patterns in student responses intended for analyses in step 3 were carried out in multiple stages.

Stage 3a: Applied the Rasch IRT model that incorporated no information about item or reader attributes. Therefore, in the model, there was no item matrix and the examinee matrix included the number of examinees and the number of items.

Stage 3b: Applied a modified IRT model that incorporated information on item attributes and item-by-reader attributes. LLTM analyses were carried out by applying the following three models:

1. LLTM model I where item parameters were modeled as depending on task attributes, but reader*topic familiarity and reader*topic interest were not included. Therefore, the item matrix included number of items by number of item attributes and the examinee matrix included number of readers by number of items;

2. LLTM model II where item parameters were modeled and included effects for the reader*topic familiarity and reader*topic interest. However, the dependence on task attributes was not included. Therefore, the item matrix included the number of technical items by the number of reader*topic familiarity and reader*topic interest variables plus an identity matrix of the items. The examinee matrix included the number of examinees by the number of technical items;

3. LLTM model III where item parameters were modeled as depending on task attributes and effects for reader*topic familiarity and reader*topic interest variables. Therefore, the item matrix included the number of technical items by item attributes plus reader*topic interest and reader* topic familiarity variables. The examinee matrix included the number of examinees by the number of technical items.

In the standard LLTM, parameter estimates obtained from the fitted regression model are not expected to account for all the variations in true task difficulties. Mislevy (1988) suggested an alternative. In this modified LLTM method, key aspects of the LLTM were combined with the exchangeability concept of Bayesian inference. In other words, the true task difficulties are expected to be distributed about the central values

predicted by the fitted LLTM-like regression model. In this approach, task difficulty is

posited to be multivariate normal with mean ($\boldsymbol{Q'\eta}$) and variance ($\phi^2\boldsymbol{I}$) where the mean

structure is defined as in the LLTM. Under this model, $\beta_j = \sum_{k=1}^{K} q_{kj}\eta_k + e_j = q_j'\eta + e_j$. The full

Bayesian model of LLTM assuming examinee independence, local independence, and

exchangeability is as follows:

$$P(X,\theta,\eta,\xi,\zeta) = \prod_i \prod_j p(X_{ij} \mid \theta_i,\eta_k) p(\theta_i \mid \zeta) p(\eta_k \mid \xi) p(\xi) p(\zeta)$$

where ζ = higher-level parameters for the prior distribution of the examinee

parameter, θ and ξ = higher-level parameters for the prior distribution of the

contribution of various features in item difficulty. The prior for each error term, ε_j was

assumed to be normally distributed with a mean of zero and a variance to be estimated

from the data.

For the LLTM analysis in this study, each of the 32 cognitive items had seven

surface features representing four item attributes. Each feature was represented by a

code; the positive feature was coded as "1," otherwise coded as "0." The codings for the

four representative items with different features are shown in Table 6.

Table 6

Coding of Items in LLTM Models

An item belongs to …	Code 1	Code 2	Code 3	Code 4	Code 5	Code 6	Code 7
an expository text, with topic A, targets textbase and causal consequence	1	1	0	0	1	0	0
an expository text, with topic B, targets textbase and causal antecedent	1	0	0	0	0	1	0
a narrative text, with topic A, targets situation model and spatial relation	0	0	1	1	0	0	1
a narrative text, with topic B, targets situation model and temporal relation	0	0	0	1	0	0	0

Note. Code 1 represents text-type, Code 2 represents expository topics, Code 3 represents narrative topics, Code 4 represents mental representation (textbase vs. situation model), Code 5 represents causal consequence, Code 6 represents causal antecedent, and Code 7 represents spatial relation.

In addition, each item had two other codes that indicated the reader attributes. An item coded as "1" to indicate the condition that a reader was familiar with the passage topic to which the item is associated with; otherwise the item was coded as "0." Similarly, an item was coded as "1" to indicate that a reader was interested in the passage topic, otherwise it was coded as "0." In the LLTM analyses, these additional variables were represented by creating four technical items for each real item. A technical item is a real item indicating the multiple conditions under which it is administered. Thus, the LLTM effect parameters would indicate change in difficulty induced by those conditions (Fischer & Formann, 1982). Therefore, there were 128 technical items to represent 32 cognitive items. For the LLTM analysis with reader attributes, both LLTM model II and

III, included those four "technical items" for each real item to distinguish the four groups formed using the data on the readers' interest and familiarity.

Note that an item in the familiarity*interest group associated with a given student depends on the passage in which it appears. A student could be interested in as well as familiar with some passage topics but not the others. In other words, for a given student, familiarity*interest group coding was the same for all eight items in a passage, but could differ from one passage to another.

Thus, in LLTM model I, the items were represented by a 32x7 matrix and the examinees were represented by a 160x32 matrix; in LLTM model II, the item matrix was 128x34 and the examinee matrix was 160x128; in LLTM model III, the item matrix was 128x9 and the examinee matrix was 160x128. It should be noted that 160 rows in the examinee matrix represented the student sample size included in the study.

All psychometric analyses—the IRT Rasch model and the three sets of LLTM— were conducted with WinBUGS[8], a computer program for Bayesian statistical analysis that uses Markov Chain Monte Carlo techniques. WinBUGS approximates the posterior distribution for each parameter with a chain of sampling from full conditional distribution through simulation. The coding for the Empirical Bayesian estimates had four parts: Part I specified the prior distribution of the difficulty contribution of the features, Part II modeled the item difficulty as a linear combination of features with or without the item-specific error, Part III specified the measurement model with the response probability of getting the item correct as a function of examinee parameters and item difficulties. The Bernoulli probability distribution was indicated for the item responses because the items

---

[8] http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml

were dichotomous (correct or incorrect). Part IV specified the examinee parameters as normally distributed and its precision (τ = inverse of variance) was modeled as a Gamma distribution (see Lu, 2006).

WinBUGS also provides Deviance Information Criterion (DIC), suggested by Spiegelhalter, Best, Carlin, and van der Linde (2002) to assess the model that best fits the obtained data. The DIC value was used to make comparisons among the four models—the Rasch simple model, the LLTM with only task attributes, the LLTM with reader but not task attributes, and the LLTM with both task and reader attributes.

It should be noted that the LLTM models that incorporated task or reader attributes are less complex and more constrained than the Rasch model. This is because LLTM models, when attributes are incorporated estimate fewer parameters than the Rasch model. In the case of the current study, the Rasch model estimated 32 item difficulties whereas the LLTM model I which incorporated only task attributes estimated 32-7=25 parameters (as there were seven variables representing the task attributes). This is because the 32 item difficulties were assumed to be a linear function of seven etas (i.e., η_k) representing those task attributes. The DIC comparison between the Rasch and LLTM I would show significant improvement only if the additional constraints (i.e., fewer parameters) did not degrade the model fit substantially given the fewer number of parameters to estimate.

In the current study, comprehension proficiency was considered as a continuous latent variable. The prior for comprehension proficiency was mean zero and a standard deviation to be estimated from the data. The prior for item parameters, a mild normal prior centered at zero was used as the LLTM model included effects for task attributes

and/or examinee by task attributes. All models were run at chain lengths of 25K MCMC

cycles. All WinBUGS reported estimates were additionally evaluated by calculating their

z-scores (posterior mean divided by standard deviation) in order to compare them across

the models. In addition, a visual examination was made with the item difficulties and

student proficiencies. For such examination, a scatter plot was produced. For items, the

examination was made to see whether the estimates from the Rasch and LLTM model I

scatter around the diagonal (45^0 angle).

Further, the relation between the two sets of estimates obtained by Rasch model

and LLTM model I was also examined for the proportion of variance in estimates of

Rasch that could be accounted for by the estimates of LLTM model I. A statistical

significance of the proportion was determined by a *t*-test. In addition, estimates of the

effect of each attribute of item and reader were also tested for their statistical

significance. The purpose of such a test was to indicate whether the proposed effects in

modeling student response patterns were statistically significant. For student ability or

proficiency, a visual examination was carried out for outliers. The observations of those

students' responses indicated that their performance was notably different when

considering their pattern of topic familiarity and topic interest.

All statistical significance was tested at $\alpha = .05$ as the current study was

considered as an exploratory study. Coding for the WinBUGS programs for the Rash

model and the LLTM model that included both reader and item attributes are presented in

Table 7 and Table 8.

Table 7

WinBUGs Code for the Rasch Model

```
model 1PL;  #Rasch only no item attributes
{

   for (k in 1:I) {

                  b[k] ~ dnorm(0, 1);

                              }

                  for (j in 1:N) {
     for (k in 1:I) {
        pi[j,k] <- (exp(theta[j] - b[k])/(1 + exp(theta[j] - b[k])));        #Measurement
Model

                  r[j,k] ~ dbern(pi[j,k]);

      }
            theta[j] ~ dnorm(0,1);

                  }


            }

#init
#data
list(N=160, I=32,
r = structure(.Data = c(
1,1,0,0,1,1,1,0,1,1,1,0,1,1,0,0,1,1,0,0,1,1,1,1,0,0,1,1,1,1,1,1,
1,1,0,0,0,1,1,0,1,0,1,1,1,0,1,0,1,1,0,1,0,1,1,1,1,1,1,1,1,1,0,1,1,
1,1,1,0,0,0,1,1,1,1,1,0,1,1,1,0,1,1,1,0,0,1,1,1,1,1,1,1,1,1,1,0,
0,1,0,1,0,1,1,0,1,1,1,1,1,0,0,0,1,0,0,1,0,0,1,1,1,1,1,1,1,1,1,1,
……..
1,1,1,0,1,1,1,0,1,1,1,1,1,1,1,0,1,0,1,1,0,1,1,1,0,1,1,1,1,1,1,1,
1,1,0,1,0,0,1,0,1,1,1,1,1,0,0,1,0,0,1,0,0,0,1,0,1,0,1,1,0,1,1,0,
1,1,0,0,1,1,1,0,1,1,1,1,1,1,1,1,1,0,0,0,1,1,1,1,1,0,1,0,0,0,0,1,
1,1,0,0,1,1,1,0,1,1,1,1,1,0,1,0,1,0,0,1,1,1,1,1,1,0,1,0,0,0,1,1,
1,1,1,1,1,1,1,0,1,0,1,0,1,1,1,1,1,0,0,1,1,1,1,1,1,1,0,1,0,1,1,0,
0,1,0,0,0,1,1,1,1,1,1,1,1,1,1,1,1,1,0,1,0,1,1,1,1,1,1,1,1,1,1,0,1,0,
1,1,0,0,0,0,0,0,1,1,1,1,1,0,1,1,0,0,1,0,1,1,1,0,1,1,1,1,1,1,0
), .Dim=c(160,32)))
```

Table 8

WinBUGs Code for the LLTM Model III

```
model LLTM; # With item and examinee attribute LLTM III
{
  for (m in 1:F) {
                    eta[m] ~ dnorm(0,1);            # Feature parameters
                              }
    for (k in 1:I) {
                    b[k] <- inprod(q[k,],eta[]); #Item difficulty in linear combination of
features
                              }

                    for (j in 1:N) {
    for (k in 1:I) {
      pi[j,k] <- (exp(theta[j] - b[k])/(1 + exp(theta[j] - b[k])));         #Measurement
Model

                    r[j,k] ~ dbern(pi[j,k]);
    }
          theta[j] ~ dnorm(0,tautheta);  # Person parameters
                    }
                    tautheta ~ dgamma(1,1);
                    vartheta <- 1/tautheta;
                    }
#init
#data
list(N=160, I=128, F=9,
q=structure(.Data=c(
1,0,0,0,0,0,1,1,0,
….
0,0,0,1,0,0,0,0,0
), .Dim=c(128,9)),
r = structure(.Data = c(
NA,NA,NA,1,NA,NA,NA,1,NA,NA,NA,0,NA,NA,NA,0,NA,NA,NA,1,NA,NA,NA,1,NA,NA,N
A,1,NA,NA,NA,0,NA,NA,1,NA,NA,NA,1,NA,NA,NA,1,NA,NA,NA,0,NA,NA,NA,1,NA,NA,N
A,1,NA,NA,NA,0,NA,NA,NA,0,NA,1,NA,NA,NA,1,NA,NA,NA,0,NA,NA,NA,0,NA,NA,NA,1,
NA,NA,NA,1,NA,NA,NA,1,NA,NA,NA,1,NA,NA,NA,NA,NA,0,NA,NA,NA,0,NA,NA,NA,1,N
A,NA,NA,1,NA,NA,NA,1,NA,NA,NA,1,NA,NA,NA,1,NA,NA,NA,1,NA,
….
NA,NA,NA,1,NA,NA,NA,1,NA,NA,NA,0,NA,NA,NA,0,NA,NA,NA,0,NA,NA,NA,0,NA,NA,N
A,0,NA,NA,NA,0,NA,NA,NA,1,NA,NA,NA,1,NA,NA,NA,1,NA,NA,NA,1,NA,NA,NA,1,NA,N
A,NA,1,NA,NA,NA,0,NA,NA,NA,1,NA,NA,NA,1,NA,NA,NA,0,NA,NA,NA,0,NA,NA,NA,1,N
A,NA,NA,0,NA,NA,NA,1,NA,NA,NA,1,NA,NA,NA,1,NA,NA,NA,0,NA,NA,NA,1,NA,NA,NA
,1,NA,NA,NA,1,NA,NA,NA,1,NA,NA,NA,1,NA,NA,NA,1,NA,NA,NA,0
), .Dim=c(160,128)))
```

CHAPTER IV

RESULTS AND DISCUSSION

The current study investigated whether a conceptualization of comprehension offered in the reading research literature could be operationalized in an assessment of the construct. A comprehension measure was built for the study that integrated the premise that comprehension is influenced by both task and reader attributes. The data collected with this theoretically grounded comprehension measure were analyzed to address two research questions regarding the effects of task and reader attributes that have been identified as pertinent to the comprehension of a text.

Prior to analyzing the data in order to address the research questions, a set of analyses was conducted to check the quality of the data, hereafter referred to as Phase I analyses. In Phase II, a statistical approach was applied to address the first research question and a psychometric approach was applied to address the second research question. Results of analyses from both phases are presented here. Results of the Phase I analyses are presented first, followed by results of the Phase II analyses. In each case, results on task attributes are discussed, followed by results on reader attributes.

Phase I: Data Quality Check

Task Attributes

The set of analyses conducted to check the quality of the 32 cognitive items (i.e., 8 items per passage) included: (a) classical item difficulty test for which percent-correct or p-values were calculated; (b) item discrimination test for which biserial correlations were calculated; and (c) internal consistency test for which Cronbach's alpha coefficients

were calculated. The biserial correlation and Cronbach's alpha were calculated for the items across the four text passages and for the items within a respective passage. In addition, distributions of the scores, that is, the number of items correctly answered across the four text passages and in each passage, were examined to see whether they approximated normal distribution and whether there were any outliers.

Of the 32 cognitive items included in the study's comprehension measure, there was no item that was answered either correctly or incorrectly by all the students who participated in the study. As shown in Table 9, the classical item difficulty statistic, p_i (i.e., referred to as p-value that represents the proportion correct) fell within a range $0.09 < p_i < 0.89$ for the 32 cognitive items. The p-values of about 88% of the items remained within $0.20 \leq p_i \leq 0.80$, a range generally considered acceptable when constructing a standardized test. (The range may vary depending on a test's purpose, e.g., high-stakes, minimum competency.)

As shown in Table 9, the point-biserial correlations between item score and respective passage number correct scores ranged from $0.18 < r_i < 0.57$ (for $r_i = .16$ with $df = 158$, p-value $< .05$). A positive significant correlation for an item indicates that students who correctly answered the item also tended to score high overall and those who answered the item incorrectly tended to score low overall. Also shown in the table were correlations between the item score and across four passages total scores; those correlations ranged from $0.03 < r_i < 0.56$. The majority, about 85% of those correlations, were statistically significant at $\alpha = .05$.

Table 9

Classical Item p-values and Item-Total Correlations Within and Across Passages

Cognitive Items	*p*-values (*SE*)	Within Passage Total	Across Passage Total
Cry of Kalahari 01	0.49 (0.04)	0.51	0.34
Cry of Kalahari 02	0.64 (0.04)	0.53	0.42
Cry of Kalahari 03	0.89 (0.02)	0.37	0.34
Cry of Kalahari 04	0.69 (0.04)	0.54	0.46
Cry of Kalahari 05	0.68 (0.04)	0.52	0.51
Cry of Kalahari 06	0.65 (0.04)	0.53	0.28
Cry of Kalahari 07	0.60 (0.04)	0.52	0.49
Cry of Kalahari 08	0.49 (0.04)	0.57	0.54
Going Green 01	0.53 (0.04)	0.47	0.38
Going Green 02	0.47 (0.04)	0.18	0.03[†]
Going Green 03	0.29 (0.04)	0.45	0.31
Going Green 04	0.73 (0.04)	0.46	0.44
Going Green 05	0.49 (0.04)	0.52	0.52
Going Green 06	0.64 (0.04)	0.55	0.47
Going Green 07	0.71 (0.04)	0.56	0.51
Going Green 08	0.68 (0.04)	0.46	0.24
Kid Fights CM 01	0.71 (0.04)	0.56	0.41
Kid Fights CM 02	0.84 (0.03)	0.47	0.43
Kid Fights CM 03	0.51 (0.04)	0.53	0.56
Kid Fights CM 04	0.65 (0.04)	0.39	0.35
Kid Fights CM 05	0.56 (0.04)	0.32	0.13[†]
Kid Fights CM 06	0.78 (0.03)	0.39	0.28
Kid Fights CM 07	0.32 (0.04)	0.46	0.28
Kid Fights CM 08	0.26 (0.04)	0.34	0.16
Shifting Sands 01	0.66 (0.04)	0.52	0.31
Shifting Sands 02	0.23 (0.03)	0.47	0.37
Shifting Sands 03	0.09 (0.02)	0.27	0.12[†]
Shifting Sands 04	0.42 (0.04)	0.48	0.38
Shifting Sands 05	0.71 (0.04)	0.45	0.32
Shifting Sands 06	0.71 (0.04)	0.56	0.48
Shifting Sands 07	0.20 (0.03)	0.40	0.19
Shifting Sands 08	0.86 (0.03)	0.43	0.46

Note. All correlations presented in the above table are statistically significant at α=.05 except those that have †. A correlation, $r_i(df=158) > 0.154$ has $p < .05$.

For the overall measure, that is, taking into account all 32 cognitive items across the four text passages, Cronbach's alpha coefficient was 0.78 (see Table 10 for the standardized variables). This index with eight cognitive items given in a respective text passage varied among the four passages; the alpha coefficients ranged from 0.38 (for the narrative-type passage that described the topic *Kid Fights Cheater Meter and Wins*) to 0.60 (for the narrative-type passage that described the topic *Cry of the Kalahari*). The alpha coefficients for the two expository-type text passages, however, were similar, 0.44 and 0.48. According to Kehoe (1995, p.1), the Cronbach's alpha "values of as low as .5 are satisfactory for short tests (10-15 items)."

Table 10

Coefficients of Cronbach's Alpha for the Cognitive Items in Four Passages

Type	Overall	*Cry of the Kalahari*	*Going Green*	*Kid Fights Cheater Meters and Wins!*	*Shifting Sands*
Raw Variables	0.78	0.60	0.45	0.37	0.45
Standardized Variables	0.78	0.60	0.46	0.38	0.44

Note. The Cronbach's alpha was calculated using all 32 cognitive items for the *Overall* and eight items for each text passage. Among the eight items, each targeted different task attributes. For example, some items targeted different levels of mental representation, and of those items, each asked for a different relation among events.

Inter-item correlations among the 32 items that were scored as 0/1 was assessed with tetrachoric correlation, an indicator of unidimensionality, ranged from $-0.001 \leq r_i \leq + 0.592$. More than half of those correlations (about 65%) were statistically significant at $\alpha=.05$.

The distribution of the number of correctly answered items, taking into account all four text-passages, is shown in Figure 3. The minimum number of items correctly

answered was six and the maximum was 31. The distribution's mean was 18.16 and the standard deviation was 5.25. The distribution had negative skewness (-0.13) as well as negative kurtosis (-0.63). Both of these higher-order moments were within a range of ±1, an indication that the scores approximated a normal distribution.
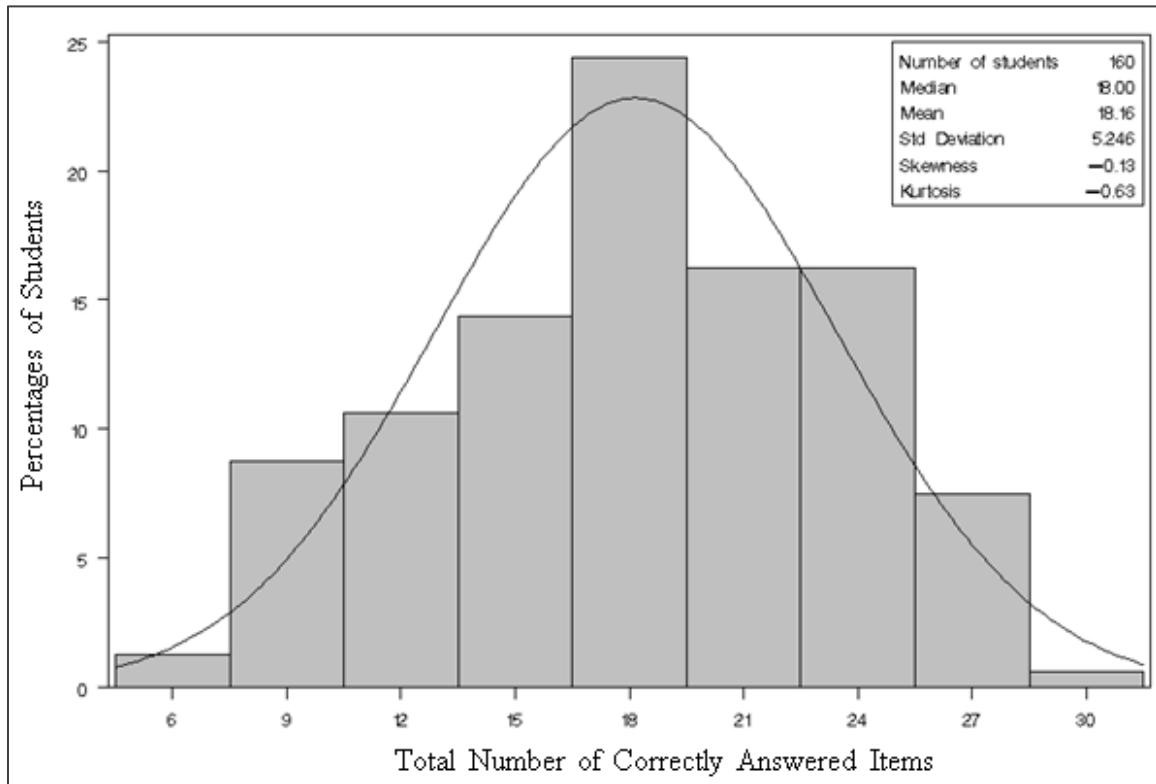
Figure 3. Frequency Distribution of the Items Correctly Answered across Four Passages

In addition, the distributions of the total number of correctly answered items in each text passage were checked. The bars, shown in Figure 4, indicate variations among the four passages, the average scores ranging from 3.87 to 5.14. Also seen in the figure, the mean (the dot) and median (the line in the box) did not coincide in any of the text passages; however, the skewness and kurtosis of each distribution remained within a range of ±1. In addition, there were no outliers.
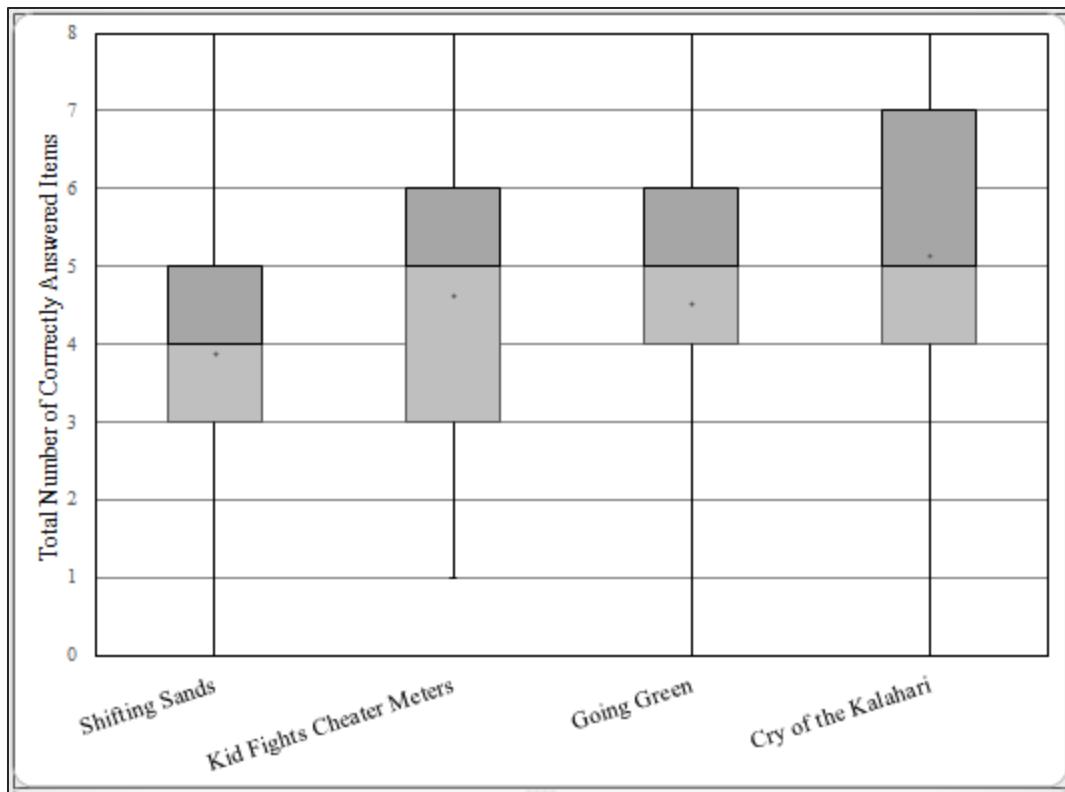
Figure 4. Distributions of Item Scores in the Four Text Passages

Reader Attributes

The next set of analyses was conducted to check quality of the data related to the

two text-passage specific reader attributes, topic familiarity and topic interest, both

considered important for the comprehension of a text passage. The data on these two

attributes were collected through a set of questions posed to the students; some were

posed prior to reading any text passages, some immediately after reading a text passage,

and some after the last cognitive item presented in the last text passage.

Analyses conducted with those data included: (a) an internal consistency test

using Cronbach's alpha coefficients and (b) the non-zero determinant test using Kaiser

Sampling Adequacy. Kaiser Sampling Adequacy is an indicator of whether variables are

factorable.

In addition, inter-item correlations among the questions included in reader attribute measures were calculated to see how the questions asked before reading a passage were related to the questions asked after reading the passage. These inter-item correlations were calculated for each of the four text passages. The correlations were also calculated to examine the relation between each attribute-related question and the total scores for a passage, that is, the number of correctly answered items given in a passage. Results of the data quality check are presented for topic familiarity followed by topic interest.

Topic familiarity. The familiarity measure assessed whether an 8th-grade study participant was familiar with the topic described in each text passage. To assess Familiarity, the measure consisted of six questions; three of which were posed prior to reading any text passages and three after the last cognitive item presented in the last text passage. Aside from the first two, all other questions posed to probe topic familiarity with the four text passages had the same wording. Results of the data quality checks for the familiarity attribute measure are presented in Tables 11-13.

As shown in Table 11, the Cronbach's alpha coefficient for the overall measure, that is, taking into account all 24 questions included in the topic familiarity measure was 0.82. The alpha coefficients based on six questions posed to each of four text passages ranged from 0.59 (for *Cry of the Kalahari)* to 0.65 (for *Shifting Sands*). Also shown in Table 11, the Kaiser Sampling Adequacy was 0.71 for the overall measure, and ranged from 0.64 to 0.68 for the four topics. A value > .5 of Kaiser Sampling Adequacy, in general, is considered to indicate that the questions are correlated enough to be factorable.

Table 11

Coefficients of Cronbach's Alpha and Kaiser Sampling Adequacy for Topic Familiarity Questions

Source	Across Four Passages	*Cry of the Kalahari*	*Going Green*	*Kid Fights Cheater Meters and Wins!*	*Shifting Sands*
Raw Variables	0.82	0.60	0.61	0.64	0.67
Standardized Variables	0.82	0.59	0.62	0.64	0.65
Kaiser Sampling Adequacy	0.71	0.64	0.68	0.64	0.66

As shown in Table 12 and Table 13, the magnitude of the inter-item correlations, in general, was higher among those questions that were asked of students after they had read each of the four text passages compared to those that were asked before they read any text passages. In addition, not all questions that asked about topic familiarity correlated statistically significantly with the total scores for a passage. This pattern varied by passages of different text-types. With both expository-type text passages, two questions correlated statistically significantly with the passage total scores. In contrast, in the case of the two narrative text passages, such correlations were not statistically significant.

Further, as shown in Figure 5, the scree plot based on taking into account all 24 questions included in the familiarity measure indicates a presence of one dominant factor, eigenvalue > 4.5. However, the second factor had eigenvalue ≈3.0 and several other factors had eigenvalue ≈1.

Table 12

Inter-item Correlations for Topic Familiarity Questions by Expository Texts

Text Passage	Questions		Shifting Sands						Total Score	Factor Loading
		QI	QII	QIII	QIV	QV	QVI		
Going Green	QI: "How much do you know about …?"	1	0.37	0.34	0.14	0.28	0.13	0.34	0.48
	QII: "How much do you know about …?"	0.26	1	0.24	0.05	-0.01	0.03	0.11	0.24
	QIII: "Do you know what … might cover?"	0.19	0.34	1	0.21	0.26	0.23	0.34	0.46
	QIV: "Did the following … discuss a topic you had read … before today?"	0.10	0.14	0.16	1	0.52	0.35	0.03	0.57
	QV: "Did … cover a topic you knew a lot about?"	0.24	0.22	0.15	0.39	1	0.47	0.08	0.79
	QVI: "How much did your prior knowledge…?"	0.06	0.10	0.11	0.34	0.43	1	-0.02	0.53
	Total score	-0.02	0.13	0.34	0.27	0.09	0.10		
	Factor Loading	0.34	0.47	0.38	0.52	0.68	0.53		

Note. Correlations presented above the diagonal represent the questions asked about the text passage *Shifting Sands* and those presented below the diagonal represent the text passage *Going Green*. A correlation, $r_i > .15$ with $df = 158$ has $p < .05$. All statistics presented in the table are based on $N=160$. Total Score represents total number of correctly answered cognitive items posed to the respective passage. All factor loadings shown in the table represent the first factor extracted without any rotation.

Table 13

Inter-item Correlations for Topic Familiarity Questions by Narrative Texts

Text Passage	Questions	QI	QII	QIII	QIV	QV	QVI	Total Score	Factor Loading
				Kid Fights Cheater Meters and Wins!					
Cry of the Kalahari	QI: "How much do you know about …?"	1	0.27	0.27	0.08	0.10	0.15	0.03	0.38
	QII: "How much do you know about …?"	0.19	1	0.10	0.04	0.16	0.06	0.09	0.20
	QIII: "Do you know what … might cover?"	0.18	0.11	1	0.28	0.24	0.26	0.07	0.42
	QIV: "Did the following … discuss a topic you had read … before today?"	0.04	0.03	0.15	1	0.61	0.38	-0.08	0.72
	QV: "Did … cover a topic you knew a lot about?"	0.17	0.07	0.14	0.51	1	0.41	-0.14	0.74
	QVI: "How much did your prior knowledge…?"	0.12	0.16	0.17	0.32	0.51	1	0.01	0.53
	Total score	-0.08	0.10	0.07	0.04	0.01	-0.01		
	Factor Loading	0.24	0.17	0.26	0.56	0.85	0.60		

Note. Correlations presented above the diagonal represent the questions asked about the text passage *Kid Fights Cheater Meters* and those presented below the diagonal represent the text passage *Cry of the Kalahari*. A correlation, $r_i > .15$ with $df = 158$ has $p < .05$. All statistics presented in the table are based on $N=160$. Total Score represents total number of correctly answered cognitive items posed to the respective passage. All factor loadings shown in the table represent the first factor extracted without any rotation.
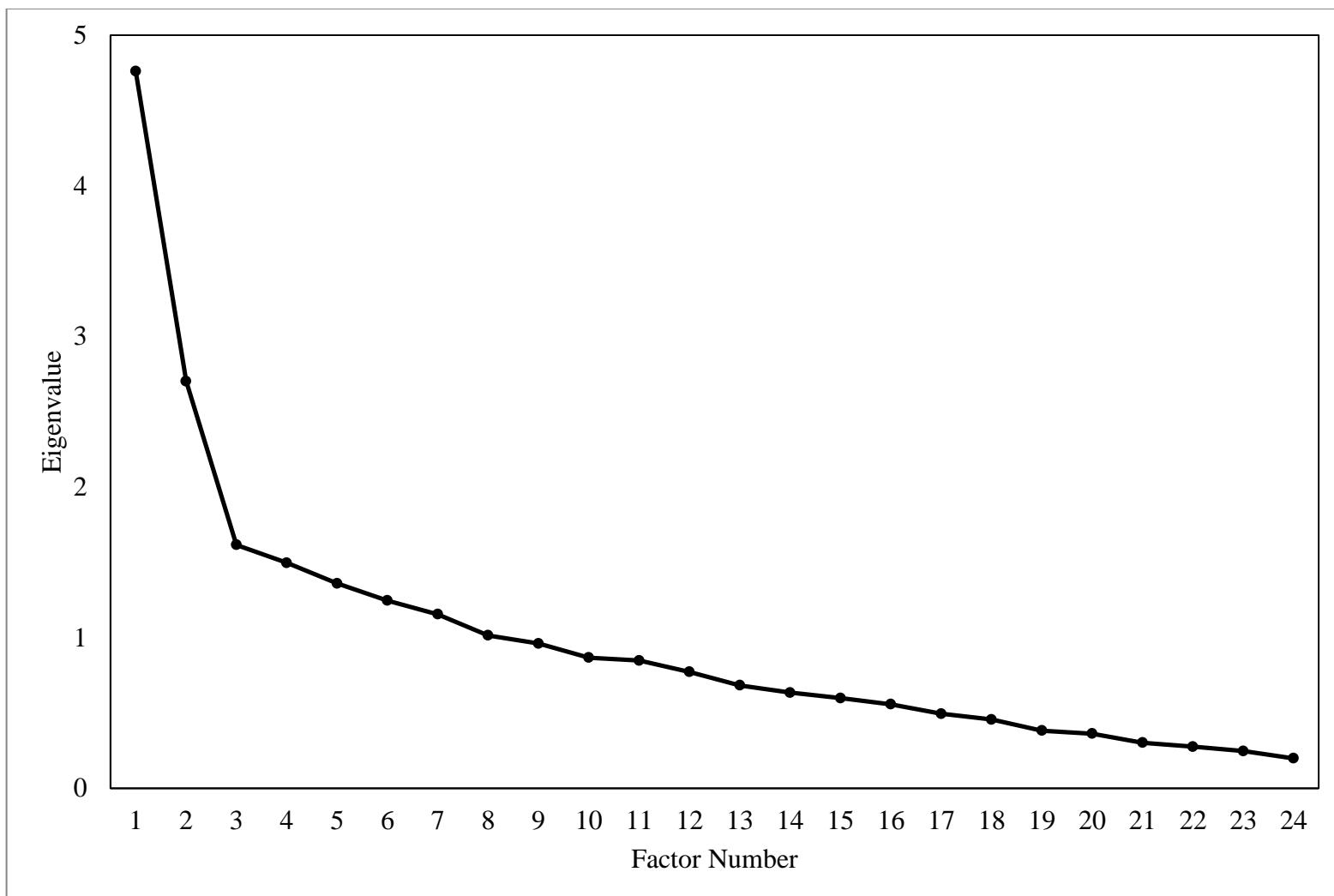
Figure 5. Eigenvalues with the 24 Questions included in the Topic Familiarity Measure.

Initial Factor Method: Iterated Principal Factor Analysis.

Further, as shown in Table 12 and Table 13, all the questions included in the familiarity measure had positive factor loadings on the first factor; factor loadings were relatively higher for the questions that were asked after the reading of a text passage compared to those that were asked before the reading of a text. Although all factor loadings were positive, no further analyses were conducted as the purpose here was not to develop a scale on topic familiarity.

Topic Interest. The interest measure assessed whether an 8th-grade study participant was interested in a topic described in the given text. The interest measure consisted of seven questions; three of them were posed before any passages were read, two immediately after a passage was read, and two after all passages were read. Similar to the familiarity measure, all but the first two questions posed to probe topic interest in the four text passages were worded the same. Results of the data quality check for the interest attribute measure are presented in Tables 14-16.

As shown in Table 14, the Cronbach's alpha coefficient for the overall measure, that is, taking into account all 28 questions included in the topic interest measure was 0.92. This index based on seven questions posed to each of four topics ranged from 0.78 (for *Kid Fights Cheater Meters and Wins!*) to 0.85 (for *Going Green*). The Kaiser Sampling Adequacy coefficients ranged from 0.79 to 0.81. These indices appear somewhat higher than those that were associated with the familiarity measure.

Table 14

Coefficients of Cronbach's Alpha and Kaiser Sampling *Adequacy for Topic Interest Questions*

Source	Across Four Passages	*Cry of the Kalahari*	*Going Green*	*Kid Fights Cheater Meters and Wins!*	*Shifting Sands*
Raw Variables	0.92	0.82	0.84	0.78	0.81
Standardized Variables	0.92	0.82	0.85	0.78	0.81
Kaiser Sampling Adequacy	0.85	0.81	0.79	0.81	0.79

As shown in Table 15 and Table 16, the magnitude of the inter-item correlations was relatively higher among those questions that were asked after reading text passages (questions IV, V, VI, and VII) than those that were asked before reading any text passages. Further, similar to the familiarity measure, not all questions asking about topic interest correlated with the respective passage total scores. However, the pattern varied by topic. As seen in Table 15 and 16, only one question asking about the expository-type text topic *Going Green* had a statistically significant correlation with the passage total scores. In the case of the other topics, more than one question had a statistically significant correlation with the passage total scores. The results of the factor analysis, the scree plot shown in *Figure 6*, indicated a presence of one dominant factor, eigenvalue > 8.0. However, one other factor had an eigenvalue ≈3.0 and several factors had an eigenvalue ≈1.

Further, as shown in Table 15 and Table 16, all the questions included in the interest measure had positive factor loadings on the first factor; factor loadings were relatively higher for the postreading questions compared to the prereading questions.

Again, similar to the familiarity measure, no further analyses were conducted based on the factor loadings of topic interest questions as the purpose here was not scale development on topic interest.

These data quality checks on cognitive items and questions included in the reader attribute measure appeared to indicate no presence of anomalies. With respect to the cognitive items, there was no perfect or zero scores either for an item or for a student. Most of the items had positive point-biserial correlations, and the overall homogeneity appeared to be sufficient. With respect to the reader attribute questions, homogeneity for the overall and for each topic were sufficient, most of the inter-item correlations were positive and statistically significant, and the questions had positive loadings on the first factor. These results seemed reasonable to proceed with the Phase II analyses outlined in Chapter III in order to address the study's two research questions.

Table 15

Inter-item Correlations for Topic Interest Questions by Expository Texts

| Text Passage | Questions | | | | | *Shifting Sands* | | | | |
		QI	QII	QIII	QIV	QV	QVI	QVII	Total Score	Factor Loading
Going Green	QI: "How much would you like to know about …?"	1	0.35	0.35	0.35	0.31	0.39	0.31	0.08	0.53
	QII: "How much would you like to know about …?"	0.61	1	0.26	0.23	0.26	0.24	0.20	0.04	0.37
	QIII: "How interesting does each of the following titles seem to you?"	0.46	0.49	1	0.34	0.18	0.24	0.24	0.21	0.39
	QIV: "How much did you enjoy reading the passage ..?"	0.40	0.29	0.47	1	0.68	0.54	0.66	0.30	0.80
	QV: "How much would you like to read the passage again?"	0.31	0.28	0.39	0.69	1	0.59	0.57	0.12	0.75
	QVI: "How much would you like to discuss this text passage with others (e.g., family or friends)?"	0.42	0.28	0.37	0.50	0.56	1	0.69	0.12	0.76
	QVII "How interesting was the passage you read today?"	0.39	0.22	0.30	0.56	0.57	0.76	1	0.23	0.79
	Total Score	0.10	0.14	0.05	0.16	0.11	-0.00	-0.03		
	Factor Loading	0.63	0.57	0.60	0.73	0.72	0.76	0.75		

Note. Correlations presented above the diagonal represent the questions asked about the text passage *Shifting Sands* and those below the diagonal represent the questions asked about the text *Going Green*. A correlation, $r_i > .154$ with $df = 158$ has $p < .05$. All statistics presented in the table are based on $N=160$. Total Score represents total number of correctly answered cognitive items posed to the respective passage. All factor loadings shown in the table represent the first factor extracted before any rotation.

Table 16

Inter-item Correlations for Topic Interest Questions by Narrative Texts

Text Passage	Questions									*Kid Fights Cheater Meters and Wins!*		
		QI	QII	QIII	QIV	QV	QVI	QVII	Total Score	Factor Loading
Cry of the Kalahari	QI: "How much would you like to know about …?"	1	0.14	0.11	0.20	0.28	0.19	0.23	0.22	0.29
	QII: "How much would you like to know about …?"	0.45	1	0.10	0.20	0.26	0.26	0.23	0.08	0.30
	QIII: "How interesting does each of the following titles seem to you?"	0.18	0.29	1	0.28	0.24	0.24	0.25	0.04	0.32
	QIV: "How much did you enjoy reading the passage ..?"	0.28	0.27	0.45	1	0.65	0.59	0.66	0.35	0.79
	QV: "How much would you like to read the passage again?"	0.29	0.28	0.38	0.69	1	0.54	0.57	0.14	0.74
	QVI: "How much would you like to discuss this text passage with others (e.g., family or friends)?"	0.26	0.20	0.34	0.57	0.61	1	0.77	0.14	0.80
	QVII: "How interesting was the passage you read today?"	0.21	0.26	0.39	0.77	0.58	0.58	1	0.17	0.86
	Total Score	0.03	0.19	0.26	0.30	0.10	0.12	0.32		
	Factor Loading	0.40	0.47	0.50	0.87	0.77	0.70	0.79		

Note. Correlations presented above the diagonal represent the questions asked about the text *Kid Fights Cheater Meters* and those presented below the diagonal represent the questions asked about the text *Cry of the Kalahari*. A correlation, $r_i > .154$ with $df = 158$ has $p < .05$. All statistics presented in the table are based on $N=160$. Total Score represents total number of correctly answered cognitive items posed to the respective passage. All factor loadings shown in the table represent the first factor extracted before any rotation.
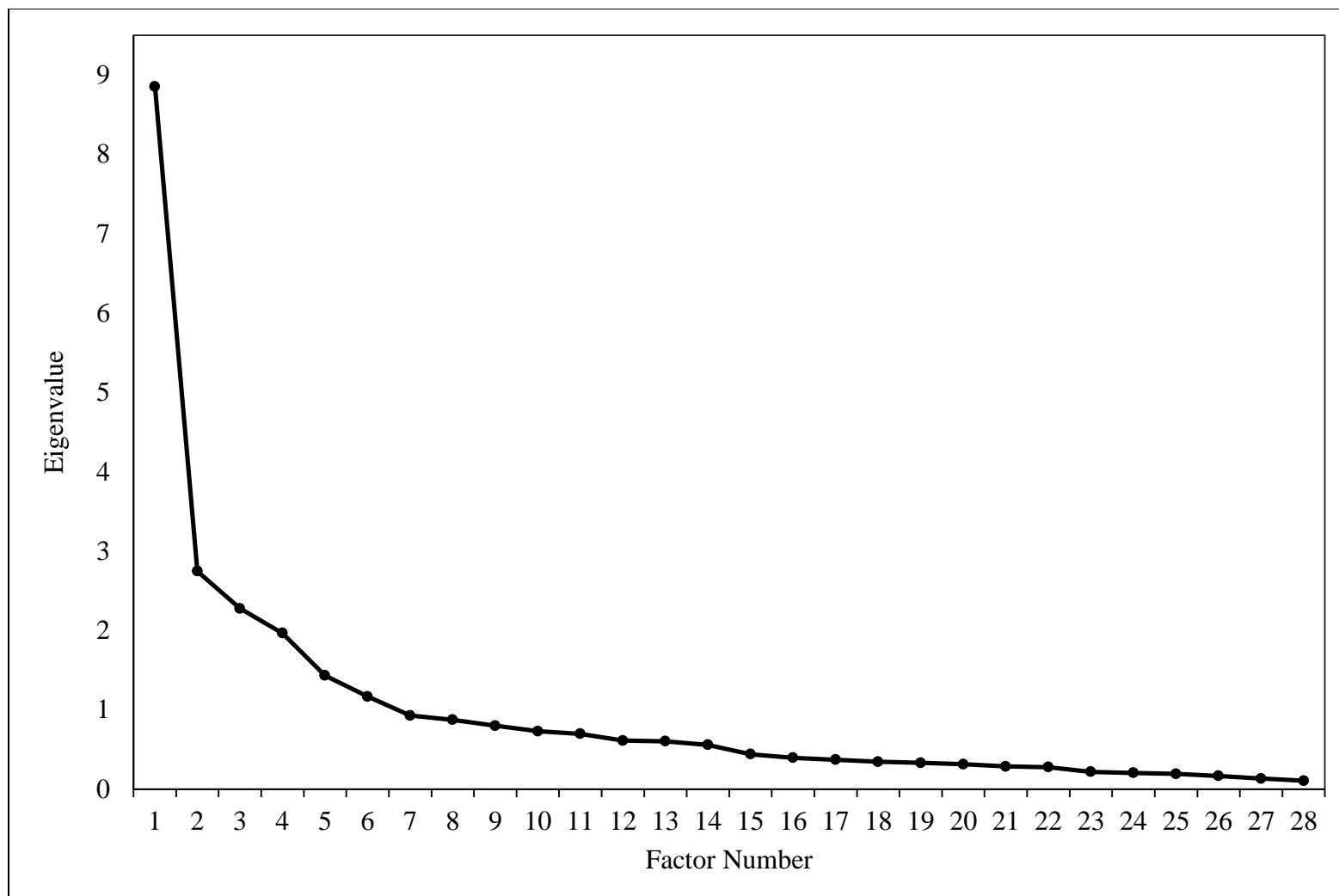
Figure 6. Eigenvalues with the 28 Questions included in the Topic Interest Measure.

Initial Factor Method: Iterated Principal Factor Analysis

Phase II: Research Question #1

Research question 1 was, "In what ways do the attributes of task and reader influence the comprehension of the 8th-grade students who participated in the study?" This question was addressed by means of a set of statistical analyses. In this approach, study participants' performances on the text passage-related questions, referred to as cognitive items were analyzed. The focus of the analysis was on the number of items correctly answered for a text passage, in other words, the focus was on correctly answered items at the aggregate level.

Task Attributes

The set of analyses conducted with respect to the task attributes examined whether the total number of correctly answered items varied by:

1a. the topic and type of text passage included in the comprehension measure;

1b. the level of mental representation of a text and the type of relation targeted in the items included in the comprehension measure.

To examine the effect of topic as a text attribute, a one-way repeated measure ANOVA was conducted. The results indicated that the number of items answered correctly, on average, varied among the text passages, $F(3,477) = 28.40$, $p < .01$, $d = .73$. With respect to the topics included in each text type, a paired t-test was conducted. As shown in Table 17, the results indicated a statistically significant difference in the number of items correctly answered between the two topics. With respect to the two expository topics, more items, on average, were answered correctly from the text passage *Going Green* ($M = 4.53$, $SD = 1.73$) than from the text passage *Shifting Sands* ($M = 3.87$, $SD =$

1.53), $t(159) = 5.00$, $p < .01$. With respect to the narrative topics, more items, on average, were answered correctly from the text passage *Cry of the Kalahari* ($M = 5.14$, $SD = 1.91$) than from the text passage *Kid Fights Cheater Meters and Wins* ($M = 4.63$, $SD = 1.57$), $t(159) = 3.75$, $p < .01$.

Table 17

Means and Standard Deviations of Item Scores by Text Passage Attributes

Item Attributes	M (Max)	SD	t- (p-value)	Effect size
Topic: Expository				
Going Green	4.53 (max=8)	1.73	5.00 (<.001)	0.39
Shifting Sands	3.87 (max=8)	1.53		
Topic: Narrative				
Cry of the Kalahari	5.14 (max=8)	1.91	3.75 (<.001)	0.29
Kid Fights Cheater Meters and Wins	4.63 (max=8)	1.57		
Text Type				
Expository	8.40 (max=16)	2.81	6.64 (<.001)	0.46
Narrative	9.76 (max=16)	3.04		

*Note. N=160 for all paired *t*-test presented in the table. Effect size represents Cohen's *d*.*

With respect to the attribute of text type which included expository and narrative texts, the results of the paired *t*-test, as shown in Table 17, indicated that the number of items answered correctly, on average, was higher for the narrative texts, ($M = 9.76$, $SD = 3.04$) than for the expository texts, ($M = 8.40$, $SD = 2.81$), $t(159) = 6.64$, $p<.01$.

It should be noted that the observed differences in the number of items correctly answered were not in line with other properties of the texts described in Table 4, the chapter on Methodology. The text passage *Cry of the Kalahari* which is somewhat longer (length=923 words) than the other text passages, had the highest average score and the text passage *Shifting Sands* which is somewhat shorter (length=575) than the other text passages, had the lowest average score. In addition, the average scores associated

with the text passage *Kid Fights Cheater Meters and Wins* which is somewhat less difficult (Grade Level=8.2) and the text passage *Going Green* which is somewhat more difficult (Grade Level=10.8) were more or less the same [(*M* = 4.63, *SD* = 1.57) and (*M* = 4.53, *SD* = 1.73)].

The results of the analyses conducted to examine the effect of the other two item attributes, mental representations of a text targeted in the items and types of relation asked in those items are presented in Table 18. With respect to the attribute of mental representation, the results of a paired *t*-test suggested that the study participants, on average, correctly answered more items that were identified as targeting the textbase (*M* = 10.00, *SD* = 2.96) than those items that were identified as targeting the situation model mental representations, (*M* = 8.16, *SD* = 2.79), *t*(159) = 9.84, *p* <.001.

Table 18

Means and Standard Deviations of Item Scores by Cognitive Item Attributes

Item Attributes	*M* (*Max*)	*SD*	*t*-(*p-value*)	Effect size
Mental Representations				
Textbase	10.00 (max=16)	2.96	9.84 (<.001)	0.64
Situation Model	8.16 (max=16)	2.79		
Relation: Causal			6.70 (<.001)	0.49
Causal Antecedent	5.06 (max=8)	1.87		
Causal Consequence	4.21 (max=8)	1.58		
Relation: Non-causal			8.11 (<.001)	0.68
Spatial	5.03 (max=8)	1.48		
Temporal	3.86 (max=8)	1.94		
Relations			1.63 (0.106)	0.13
Causal	9.27 (max=16)	3.06		
Non-Causal	8.89 (max=16)	2.94		

Note. *N*=160 for all paired *t*-test presented in the table. Effect size represents Cohen's *d.*

Two analyses were conducted to examine the effect of the item attribute of types of relations among the events of a situation described in a text passage. First, a one-way

repeated measure ANOVA was conducted with the number of correctly answered items that asked about one of the four relations: causal antecedent, causal consequence, temporal, and spatial relations. The results indicated that the number of items correctly answered, on average, varied by the relations, $F(3,477) = 34.09$, $p < .01$, $d = .80$. Additional analysis showed that the participants, on average, correctly answered more items that asked for a causal antecedent relation ($M = 5.06$, $SD = 1.87$) than those that asked for a causal consequence relation ($M = 4.21$, $SD = 1.58$), $t(159) = 6.70$, $p < .01$. The results also suggested that the participants, on average, correctly answered more items that asked for a spatial relation ($M = 5.03$, $SD = 1.48$) than those that asked for a temporal relation ($M = 3.86$, $SD = 1.94$), $t(159) = 8.11$, $p < .01$. However, when the items were grouped into questions asking about causal relation ($M = 9.27$, $SD = 3.06$) and non-causal relation ($M = 8.89$, $SD = 2.94$), the apparent difference was not statistically significant, $t(159) = 1.63$, $p > .05$. (It is important to note here that the study participants were allowed to reread text passages when answering the passage-related questions, an assessment practice meant to minimize the burden on memory.)

The effect size (Cohen's d) associated with each respective comparison is shown in Table 17 and Table 18. Effect sizes associated with those comparisons ranged from 0.13 to 0.68, that is, within a range of somewhat small to medium. In sum, the results suggested that the total number of correctly answered items varied by the text topic and text type as well as the level of mental representation and type of relation targeted in the items. In other words, the comprehension of 8th-grade study participants seemed to be impacted by the cognitively-relevant nature of the task performed.

Reader Attributes

The set of analyses conducted with respect to the two reader attributes, topic familiarity and topic interest, treated as dichotomous variables, examined whether the total number of items correctly answered for a passage varied between:

1d. the study participants who were identified as familiar with and those identified as not familiar with the topic of a passage given to read;

1e. the study participants who were identified as interested in and those who were not interested in the topic of a passage they were given to read.

With respect to Topic Familiarity, study participants were classified as familiar or not familiar with a topic based on their responses to the questions asked in the familiarity measure following the procedure described in the chapter on Methodology. The number of students identified as familiar or not familiar with a topic, the number of items the students answered correctly for the text passage describing the topic, and the statistical comparisons made between those two groups are presented in Table 19.

As can be seen in Table 19, more students reported not being familiar with three of the four topics. In the case of one topic, *Going Green*, more students reported being familiar with the topic. This finding is consistent with what was gathered in the cognitive lab interview conducted during the pilot study. Further, as shown in Table 19, in the case of both expository-type text topics, the students who were identified as familiar with the topic, on average, answered more items correctly than those who were identified as not familiar with the topic. In the case of *Shifting Sands*, readers familiar with the topic had an average score, 4.24 ($n = 70$, $SD = 1.65$) higher than those readers who were identified as not familiar with the topic, 3.58 ($n = 90$, $SD = 1.37$), $t(158) = 2.78$, $p < .01$. In the case

of *Going Green*, readers familiar with the topic had an average score, 4.86 ($n = 100$, $SD =$ 1.63), higher than those who were not familiar with the topic, 3.98 ($n = 60$, $SD = 1.77$), $t(158) = 3.18$, $p < .01$. Results for both topics of the narrative-type texts showed no statistically significant differences between the groups, $p > .05$.

Table 19

Means and Standard Deviations for Familiar and not Familiar Readers

Text Topics	Familiar			not-Familiar			t-value	Effect Size
	n	*Mean*	*SD*	*n*	*Mean*	*SD*		
Cry of the Kalahari	75	5.17	2.06	85	5.11	1.79	0.22	0.03
Going Green	100	4.86	1.63	60	3.98	1.77	3.18*	0.52
Kid Fights Cheater Meters and Wins!	71	4.59	1.61	89	4.65	1.54	0.24	0.03
Shifting Sands	70	4.24	1.65	90	3.58	1.37	2.78*	0.44

Note. $N = 160$. *$p < .05$. Effect size represents Cohen's *d*.

With respect to Topic Interest, similar to the topic familiarity attribute, study participants were classified as interested and not interested in a topic. The number of students reported as interested or not interested in a topic, the average number of items the participants answered correctly for the respective topic, and the statistical comparisons made between those two groups for each of the four topics are presented in Table 20.

Table 20

Means and Standard Deviations for Interested and not Interested Readers

Text Topics	Interested			not-Interested			t-value	Effect Size
	n	*M*	*SD*	*n*	*M*	*SD*		
Cry of the Kalahari	90	5.47	1.70	70	4.71	2.09	2.51*	0.40
Going Green	80	4.60	1.78	80	4.46	1.70	0.50	0.08
Kid Fights Cheater Meters and Wins!	89	4.83	1.60	71	4.37	1.49	1.90	0.30
Shifting Sands	81	4.10	1.54	79	3.63	1.50	1.94	0.31

Note. $N = 160$. *$p < .05$. Effect size represents Cohen's *d*.

As can be seen in Table 20, in general, more students reported as being interested in all topics except one. In the case of *Going Green*, an equal number of students, $n = 80$, reported as being interested and not interested in the topic. Those who were identified as being interested in the narrative-type text topic *Cry of the Kalahari*, on average, answered more items correctly, ($M = 5.47$, $SD = 1.6$, $n = 89$) than those who were identified as not interested in the topic, ($M = 4.71$, $SD = 2.09$, $n = 70$); $t(158) = 2.51$, $p < .05$. As seen in Table 20, the apparent differences between those two groups were not statistically significant in the case of other topics when evaluated against the critical value, 1.96 for a two tailed test at $\alpha = .05$.

Another set of analyses was conducted to see the joint effect of the two reader attributes, topic familiarity and topic interest. In those analyses, the study participants were classified into four reader groups: (a) "YY," which included those readers who were identified as having both familiarity with and interest in the topic; (b) "YN," which included the readers who were identified as having familiarity but no interest in the topic;

(c) "NY," which included the readers who were identified as having no familiarity but interest in the topic; and (d) "NN," included the readers who were identified as having neither familiarity nor interest in the topic. The number of readers in each of the four groups and the average number of items answered correctly are presented in Table 21. It should be noted that a given student could be and often was in different groups on four topics.

Table 21

Means and Standard Deviations for Four Reader Groups by Text Topics

Groups	Cry of the Kalahari	Going Green	Kid Fights Cheater Meters and Wins	Shifting Sands
	n *M (SD)*	*n* *M (SD)*	*n* *M (SD)*	*n* *M (SD)*
NN	52 4.87 (1.94)	42 3.90 (1.64)	50 4.32 (1.49)	60 3.42 (1.39)
NY	33 5.48 (1.46)	18 4.17 (2.09)	39 5.08 (1.51)	30 3.90 (1.30)
YN	18 4.28 (2.47)	38 5.08 (1.57)	21 4.48 (1.50)	19 4.32 (1.66)
YY	57 5.46 (1.84)	62 4.73 (1.67)	50 4.64 (1.66)	51 4.22 (1.67)
F (*p*- value)	2.53 (0.06)	3.81 (0.01)	1.81 (0.15)	3.28 (0.02)

Note. Group labeled, YY= the student were identified as familiar with and interested in the topic, YN= the student identified as familiar with but not interested in the topic, NY= the student identified as not familiar with but interested in the topic, NY= the student identified as neither familiar with nor interested in the topic. All *F*-statistics shown in the table have $df_1=3$, $df_2=156$.

As shown in Table 21, the number of students identified for the four reader groups varied by topic. The number identified for a group varied within a topic and across the four topics. A one-way ANOVA was conducted for each text passage to examine whether the groups varied in the average number of items answered correctly.

Presented in Table 21, ANOVA results indicated that the groups within a text topic varied in the number of items answered correctly. However, such differences were statistically significant for only two topics of the expository-type text passages.

No additional analyses were conducted to determine which group had the highest average scores among all groups. It should be noted that these student groups were not predetermined and balanced, but rather arose from the students' self-perceived interests and familiarities. In other words, this was not a between-subject experimental design with an aim to compare main effects across equivalent groups. Rather, this classification was elicited to see how the psychometric tool, LLTM would model the within-subject effect of the two reader attributes in estimating ability of the students who participated in the study.

In summary, the results of the analyses on the reader attributes showed that the number of items correctly answered varied between those who were identified as familiar with and those identified as not familiar with a topic of a passage. However, the variations were limited in the case of expository-type text passages included in the study. Further, the number of items correctly answered varied between those who were identified as interested in and those who were not interested in the topic of a passage. However, such variations were also limited to some of the topics included in the study. Although this study was not designed to test between-subject effects, the patterns found here are suggestive that the influence of topic familiarity and topic interest depends on the text type and text topic.

Phase II: Research Question #2

The second research question investigated in this study was, "Do the task and reader attributes considered in the study contribute to the item difficulties and reader comprehension proficiencies?" The analyses to address this question used cognitive items. However, the analyses focused on items and students at an individual level. In a latent-variable based psychometric analysis as generally conducted in an assessment, the focus of core analyses was on the estimation of an examinee's proficiency. Such an analysis performs joint scaling of examinees and items.

Four IRT models were applied to examine the effects of the task and reader attributes on estimates of item difficulty (β_i) and student proficiency (θ_j). The IRT Rasch model was applied to obtain the initial estimates of β_i and θ_j. These estimates were compared with those obtained by LLTM models. Three LLTM models were applied; those models vary in what attributes were incorporated. In model I, only task attributes were incorporated; in model II, only reader attributes were incorporated; and in model III, both task and reader attributes were incorporated. In addition, the estimates of effects (η_k) of task and reader attributes obtained by LLTM models were evaluated for their statistical significance.

These analyses were conducted using Markov Chain Monte Carlo methods within the framework of Bayesian Inference using WinBUGS, software for conducting Bayesian analyses. The program produced posterior means, standard deviations, and sampling errors. Relatively loose priors were used so the data could speak. All models were run in one chain with 25K iterations. In order to ensure that the model converged, the first 5K iterations were discarded as "burn-in cycles" (i.e., initial cycles to get the sequence of

draws that approximate a variable's posterior distribution into the right range, or "reach stationarity").

The estimated posterior means of item difficulty (β_i) from the two models, Rasch and LLTM model I are shown in Figure 7. The difficulties of the 32 cognitive items estimated by the Rasch model were within $-2.5 < b_i < +3.0$, whereas the difficulties estimated by the LLTM model I were within $-1.0 < b_i < +1.0$. The correlation between the two sets of estimates was, $r(30) = 0.50$, associated $t(31) = 3.08$, $p < .0.01$. In other words, the four item attributes, text type, text topic, mental representation, and relation accounted for about 25% of the variances in the difficulties of the 32 items included in the study. This finding is consistent with previous LLTM studies of reading comprehension, where task features were found to account about 20% to 30% variance in item parameters. This amount of variance, as seen in the present study, indicates significant systematic effects of theoretically predicted determinants of difficulty, but passage and task specific determinants of difficulty remained. Studies in which LLTM estimates of difficulty predicted most of the variance of item difficulty were seen in more restricted and controlled tests, such as arithmetic operations.
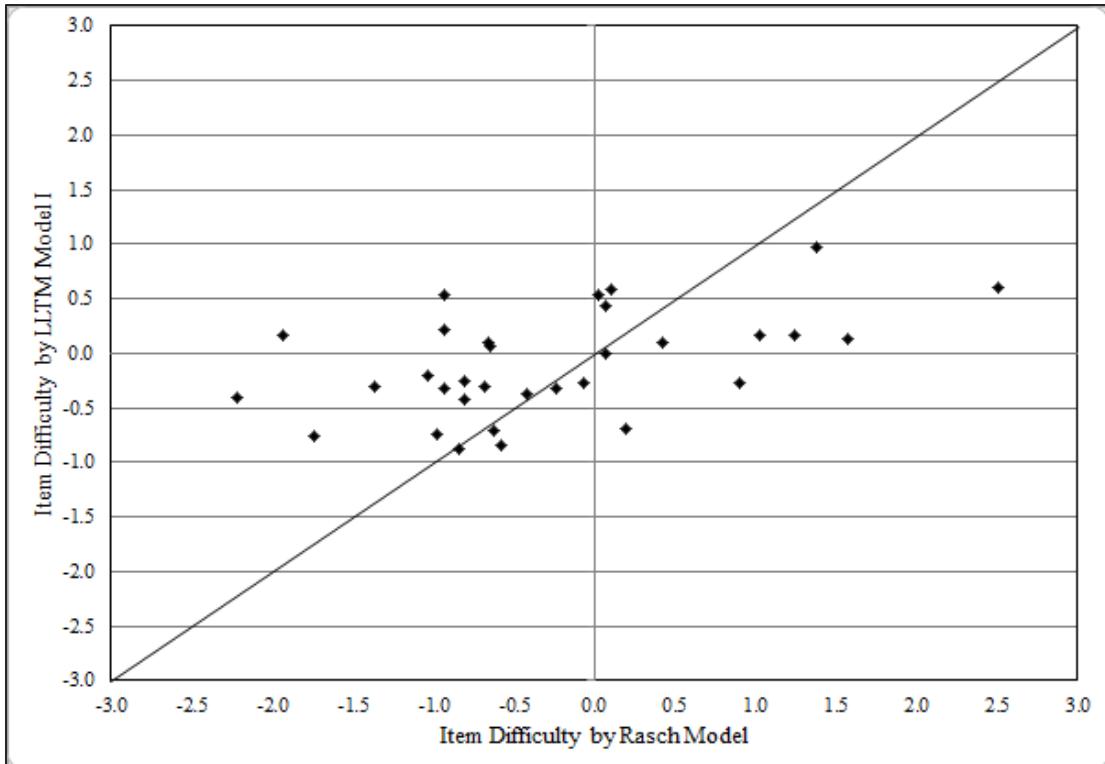
Figure 7. Item difficulties (b_j) Estimated by Rasch Model and LLTM Model I

Further examination of the items that were off of the diagonal, (see Figure 7), showed that some of the items with larger deviation from the Rasch model were among those that had the proportion correct relatively lower (e.g., $p_i < 0.20$) or higher (e.g., $p_i >$ .80). Item difficulty estimates from the LLTM item-effect-only model were shrunk back toward the average difficulty from the unconstrained difficulties of the Rasch model.

The posterior means of estimated reader or student proficiencies, θ_i are shown in three scatter plots presented in Figures 8, 9, and 10. These plots with estimated proficiencies after standardization indicate how the abilities estimated by the Rasch and LLTM models compare. The proficiencies of the study participants estimated by the Rasch model and LLTM model I (i.e., the model that included only task attributes), as shown in Figure 8, have lined up more closely to the diagonal.
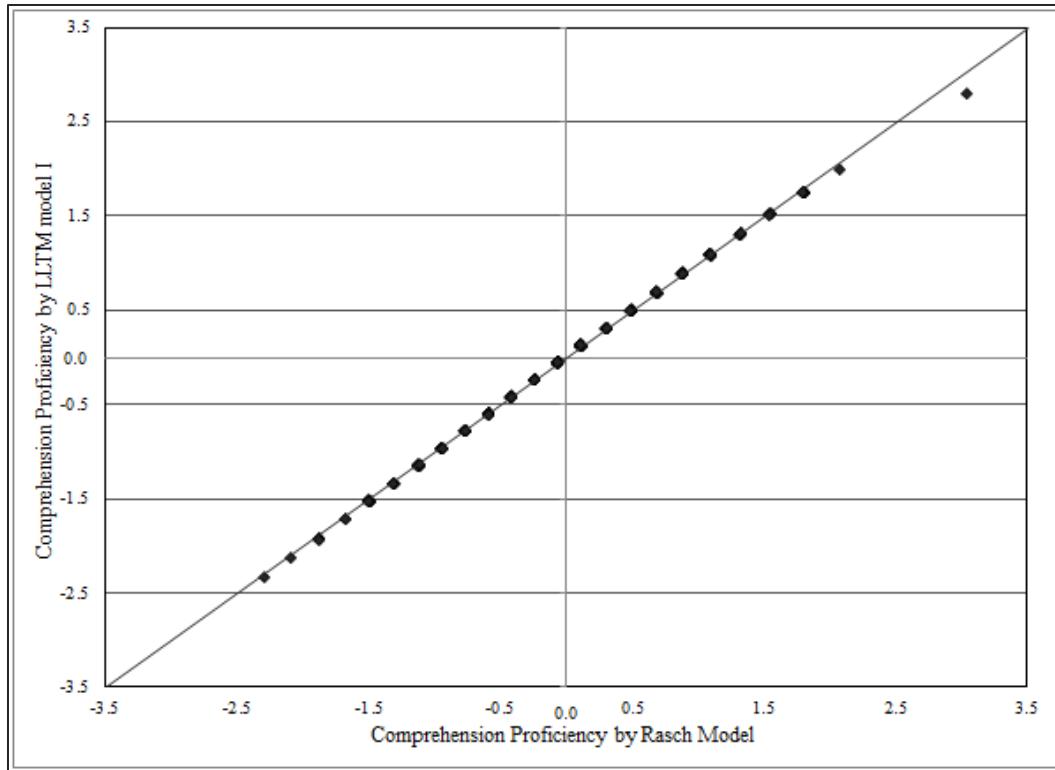
Figure 8. Comprehension Proficiencies (θi) Estimated by Rasch and LLTM Model I

In contrast, proficiencies estimated by the other two LLTM models, LLTM model II (i.e., the model that included only reader attribute) as shown in Figure 9, and the proficiencies estimated by LLTM model III (i.e., the model that included both task and reader attributes) as shown in Figure 10, did not line up on the diagonal. More specifically, the plots in Figure 9 and Figure 10 indicate that some student proficiencies estimated by the Rasch model changed under the LLTM models that included reader attributes.
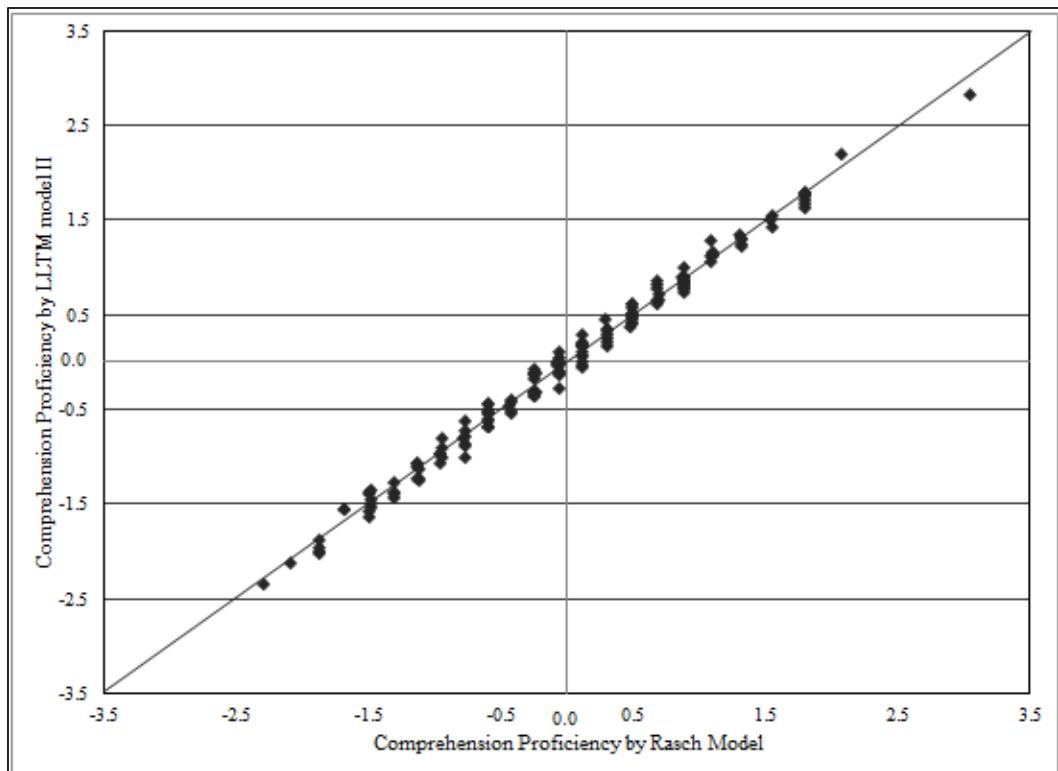
Figure 9. Comprehension Proficiencies (*θi*) Estimated by Rasch and LLTM Model II
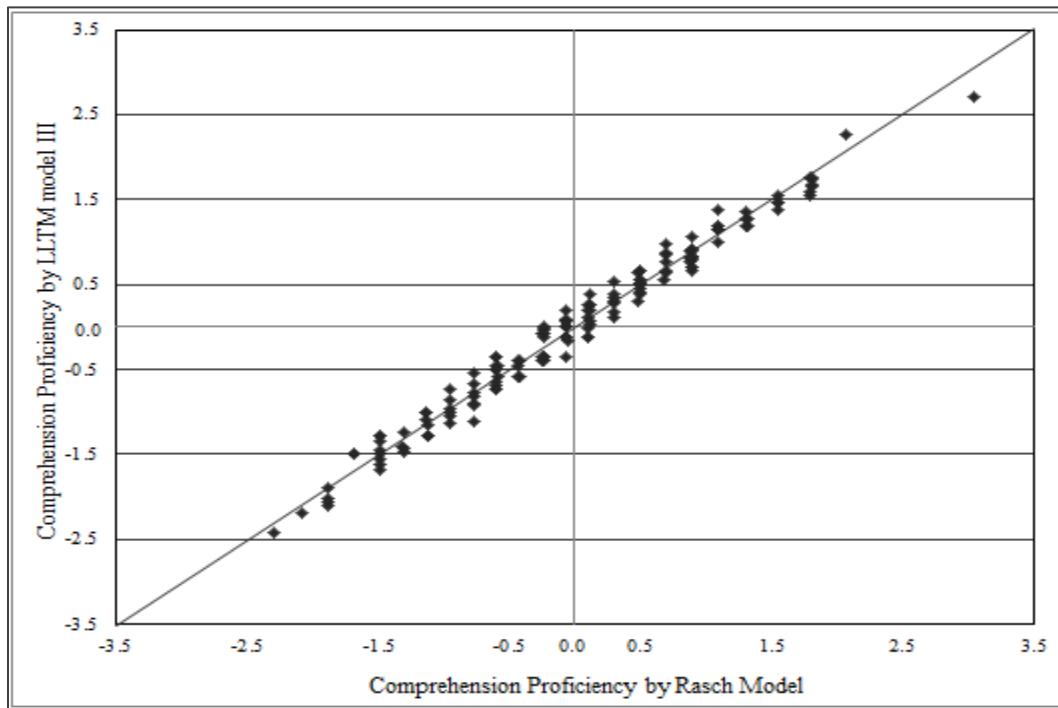

Figure 10. Comprehension Proficiencies (*θi*) Estimated by Rasch and LLTM Model III

To illustrate the magnitude of differences in comprehension proficiencies estimated by the Rasch and LLTM models, a summary of the differences is presented in Figure 11. As the figure shows, the differences in the proficiencies estimated by the Rasch and LLTM models II and III were more apparent than differences between the Rasch and the LLTM model I, which included only task attributes. The difference between the two z-scores represents the area under the normal curve.

As can be seen in Figure 11, the differences in the estimated proficiencies between the Rasch and LLTM model I, which did not include reader attributes, remained within one percentile. The differences between the Rasch and LLTM model II, which included only reader attributes, remained within five percentile for nearly all readers. The differences between Rasch and LLTM model III, which included both item and reader attributes, exceeded five percentile for about a quarter of cases of the readers included in the study.
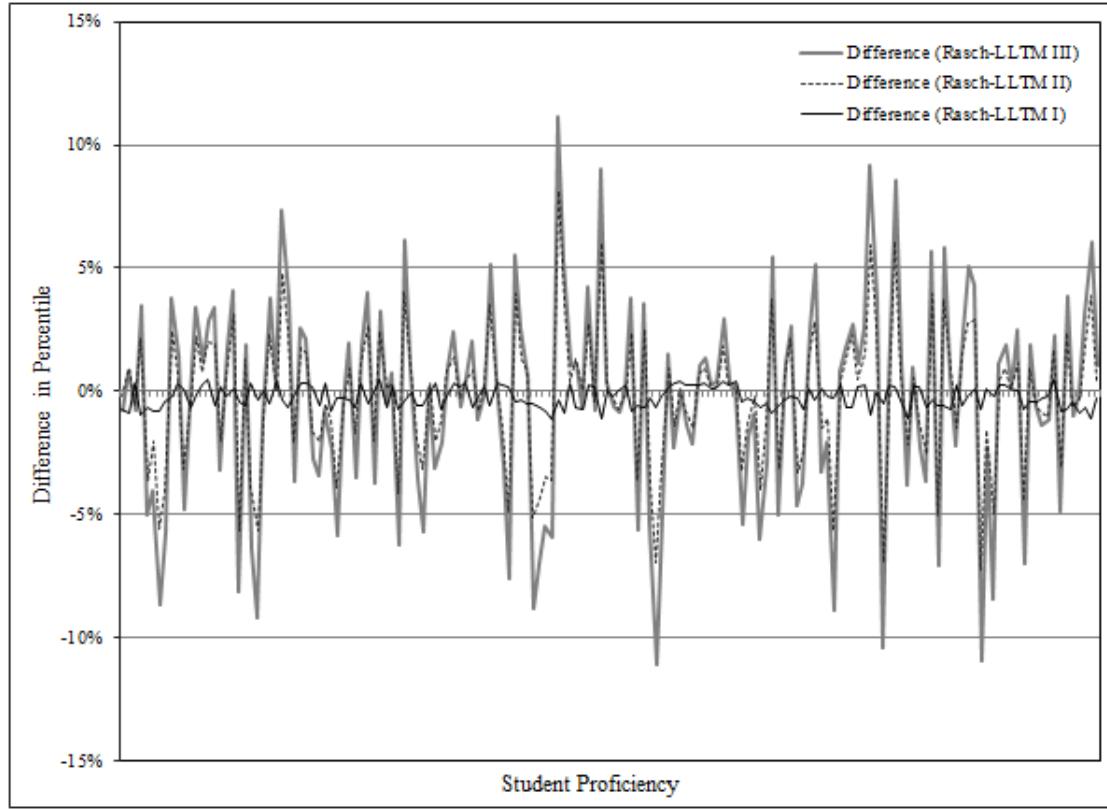
Figure 11. Differences in Proficiencies (θi) Estimated by Rasch and LLTM Models

To instantiate the differences, the abilities estimated by the three models—Rasch, LLTM II, and LLTM III—were examined further for readers who had the same number of items correct. As can be seen in Table 22, estimates of proficiency obtained in the Rasch model remained the same for the students who correctly answered the same number of items, shown in the column with label "Total Items Correct." Variations in estimates of proficiencies found using LLTM II and LLTM III models depended on whether the students had familiarity with and/or interest in topics they read.

Table 22

Proficiencies Estimated by Three Psychometric Models

Student ID	Total Items Correct	Text Topics				Psychometric Models		
		Cry of the Kalahari	*Going Green*	*Kid Fights Cheater Meters*	*Shifting Sands*	Rasch θi	LLTM II θi	LLTM III θi
XXX052	10	NN	NN	NN	NN	-1.49	-1.35	-1.29
XXX057	10	YY	NY	NN	NN	-1.49	-1.52	-1.54
XXX126	10	YY	YY	YY	YN	-1.50	-1.57	-1.61
XXX107	15	NN	NN	NN	NN	-0.60	-0.44	-0.35
XXX002	15	NY	YN	NN	YY	-0.60	-0.62	-0.64
XXX114	15	YY	YY	YY	YY	-0.60	-0.68	-0.72
XXX015	19	NN	NN	NN	NN	0.11	0.30	0.40
XXX008	19	YN	NN	NN	YY	0.11	0.21	0.26
XXX016	19	YY	YY	YY	YY	0.11	0.07	0.03
XXX087	24	NN	NN	NN	NN	1.09	1.28	1.37
XXX024	24	YY	YN	YN	YY	1.09	1.13	1.15
XXX018	24	YY	YY	YY	YY	1.09	1.06	1.01

Note. YY= the student identified as having familiarity with and interested in the topic, YN= the student identified as having familiarity with but not interested in the topic, NY= the student identified as not familiar with but interested in the topic, NN= the student identified as neither familiar with nor interested in the topic. Proficiency estimates (θi) presented in the above table represent their standardized scores.

The effects, η_k of the four item attributes, represented by seven variables, were modeled by LLTM I, two reader attributes were modeled by LLTM II, and all six attributes, represented by nine variables, were modeled by LLTM III. Estimates of those effects are presented in Tables 23-25. In addition to the posterior means and standard deviations, the table shows a z-score associated with each η_k in order to evaluate the

contribution of an attribute. The z-score was a ratio between the posterior mean and the standard deviation reported by WinBUGS. These z-scores were evaluated against the critical value, 1.96 for a two-tailed test at $\alpha = .05$ (see Embretson & Gorin, 2001 for more on such a comparison).

As can be seen in Table 23, in LLTM model I, a higher difficulty was estimated for items of the expository-type texts compared to items of the narrative-type texts. Further, within expository-type texts, a higher difficulty was estimated for items related to the topic *Shifting Sands* compared to items related to the topic *Going Green*. A higher difficulty was estimated for items that targeted the situation model mental representation compared to items that targeted the textbase mental representation. In addition, with LLTM model I, a higher difficulty was estimated for the items that asked about temporal relation compared to the items that asked about causal-consequence, causal-antecedent, or spatial relation. However, the effect, η_k due to the narrative topics was not statistically significant, $p > .05$.

Table 23

Effects (η_k) of Task Attributes Estimated by the LLTM Model I

Attributes	Posterior Mean	Standard Deviation	Z-score
eta[1]: Text Type	0.536	0.078	6.89*
eta[2]: Expository Topics	-0.376	0.084	-4.47*
eta[3]: Narrative Topics	0.103	0.078	1.32
eta[4]: Mental representation	0.434	0.059	7.39*
eta[5]: Causal-Consequence compared to Temporal	-0.371	0.077	-4.83*
eta[6]: Causal-Antecedent compared to Temporal	-0.863	0.078	-11.07*
eta[7]: Spatial compared to Temporal	-0.845	0.078	-10.86*

Note. *$p < .05$

As shown in Table 24, with LLTM model II, a higher item difficulty was estimated for a reader who was not familiar with the topic described in the text. However, the amount by which this model estimated a higher difficulty of items when a reader was not interested in the topic was not statistically significant. These two reader attribute-related variables accounted for about 12% of the variances in the item difficulties (represents a proportion of between condition variance over the between plus within condition variances).

Table 24

Effects (η_k) of Reader Attributes Estimated by the LLTM Model II

Attributes	Posterior Mean	Standard Deviation	Z-score
eta[1]: Familiarity with topic	-0.204	0.082	-2.49*
eta[2]: Interest in topic	-0.122	0.079	-1.54

Note. *p < .05

Similarly, significant effects of the attributes were estimated with LLTM model III. In this model, the effects of all task and reader attributes considered in the current study were statistically significant, see Table 25. With LLTM model III, the effect of topics included in the narrative-type text was also statistically significant. In addition to item attributes, in the LLTM model III, a higher difficulty was estimated if readers were not familiar with the topic described in a text. Further, as also shown in Table 25, the model estimated a higher difficulty of the items for a reader who was not interested in the text topic. These attributes accounted for about 38% of the variances in the item difficulties. It should be noted that, within an examinee, effects are averaged over all the examinees, items, and text passages. This average amount of contribution by which

items become easier or harder for an examinee depends on whether or not the examinee

is familiar with or interested in a text topic.

Table 25

Effects (η_k) of Task and Reader Attributes Estimated by the LLTM Model III

Attributes	Posterior Mean	Standard Deviation	Z-score
eta[1]: Text Type	0.584	0.078	7.48*
eta[2]: Expository Topic	-0.366	0.085	-4.32*
eta[3]: Narrative Topic	0.177	0.079	2.22*
eta[4]: Mental representation	0.469	0.058	8.05*
eta[5]: Causal-Consequence compared to Temporal	-0.307	0.078	-3.94*
eta[6]: Causal-Antecedent compared to Temporal	-0.801	0.079	-10.16*
eta[7]: Spatial compared to Temporal	-0.782	0.079	-9.86*
eta[8]: Familiarity with topic	-0.270	0.071	-3.80*
eta[9]: Interest in topic	-0.147	0.072	-2.03*

Note. * $p < .05$

The relation of the estimates by LLTM, a model that estimated item difficulty

without incorporating the error terms (i.e., a fixed-effects rather than a random effects

model), to the estimation by Rasch was examined using the Deviance Information

Criterion (DIC), a fit index adapted for MCMC estimation. For the Rasch model,

WinBUGS reported DIC = 5775.90, whereas WinBUGS reported DIC = 6479.91 for

LLTM model I, DIC = 5775.29 for LLTM model II, and DIC = 6472.25 for LLTM

model III. Ideally, the smaller the DIC, the better is the fit of the data. As explained in

the chapter on methodology, these higher DIC values for the LLTM models I and III

compared to the Rasch model, however, did not indicate a substantial degrade in the model fit given those models estimated fewer parameters.

The fit of the Rasch and LLTM models, was also examined to see how the distributions of estimates of difficulty parameter, b_i was spread out in each case. The spread, between the 2.5 percentile (i.e., lower point) and the 97.5 percentile (i.e., higher point) of the distribution of each respective item's estimated difficulty is shown in Figure 12. There are 32 pairs of bars. In the case of each pair, the first bar represents the estimate obtained by the Rasch model and the second bar represents the estimate obtained by the LLTM model I. As it is apparent in *Figure 12*, the distributions of the LLTM model estimates, in general, have a narrower spread.
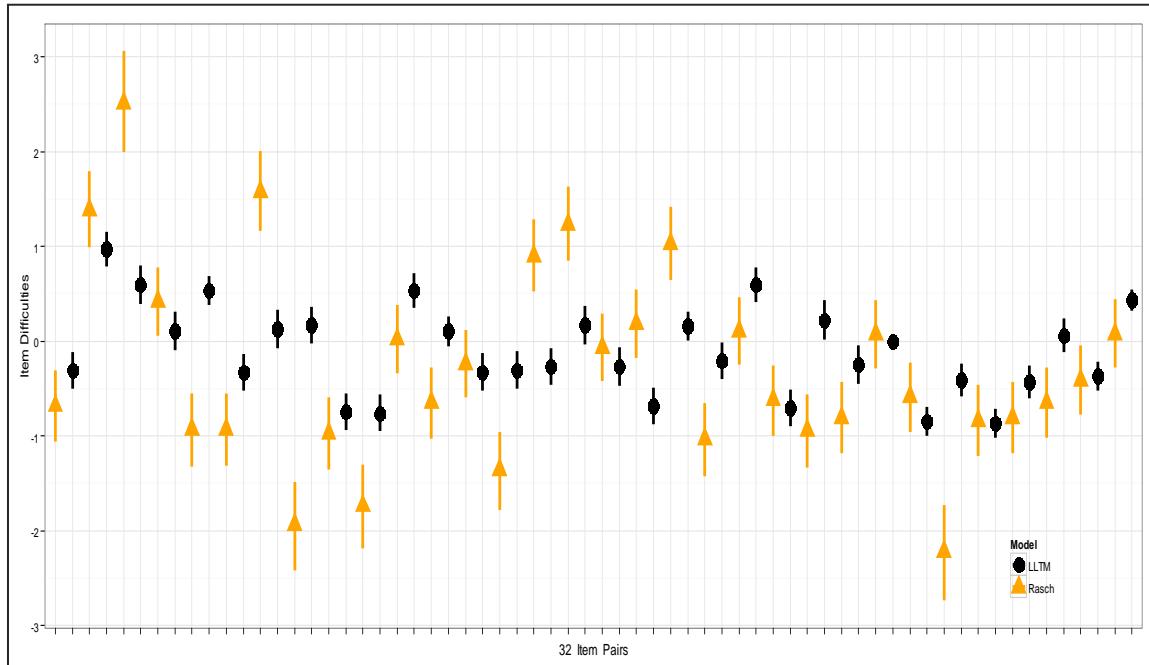

Figure 12. Distributions of the 32 Item Difficulties Estimated by Rasch and LLTM
Models

To sum up, it was possible to model the effects of attributes of task and reader, the two major components of reading comprehension using a psychometric tool. In other

words, results presented here suggest that the task and reader attributes considered in the current study contributed to the items' difficulties and the examinees' comprehension proficiencies.

CHAPTER V

SUMMARY, IMPLICATIONS, AND FUTURE RESEARCH

In this chapter, the results presented in Chapter IV are summarized in light of the purpose of the study presented in Chapter I and the conceptual framework described in Chapter II. In addition, implications of the study are discussed and some suggestions are offered for future research in order to continue the pursuit of the current study.

Summary

This study was an attempt to address a concern raised regarding comprehension assessments that are conducted on students attending K-12 schools in the U.S. The assessment of reading comprehension, a precondition to learning from text, has a long history as testing in this domain began in the early 20th century (Pearson & Hamm, 2005; Segel, 1935; Smith, 1965). Reading comprehension assessment, nonetheless, has long been a source of dissatisfaction among researchers and educators. Issues have been raised from early on to recent years (e.g., Gates, 1921; RRSG, 2002). A concern that many have raised is that comprehension assessments are not grounded in a clearly articulated conceptualization of the construct offered in the reading research literature (e.g., Johnston, 1984; Keenan et al., 2008; Magliano et al., 2007; Sarroub & Pearson, 1998; Valencia et al., 1989). Further, little effort has been made to integrate the cognitive view of comprehension into the measurement model generally applied in synthesizing the comprehension performance that is observed in an assessment.

In order to demonstrate that it is viable to develop an assessment of comprehension following a conceptualization of the construct suggested in the reading literature, in this study, the construction of the comprehension measure and the ensuing

analyses were framed following Kintsch's (1998) CI model and Alexander's (1997) MDL. The design of the measure and the analytic approach were grounded on a premise that reading comprehension of a text passage is influenced by attributes of the task performed and attributes of the reader as related to a text passage.

The resulting data, the comprehension performance of 160 grade 8 students, were analyzed to address two research questions regarding the effects of attributes of key comprehension components. Separate analytical approaches were applied in addressing those questions. These approaches differed in the focus of analysis and served different purposes. The key findings for each research question are summarized next[9].

Research Question #1

The first research question investigated was: "In what ways do the attributes of task and reader influence the comprehension of the 8th-grade students who participated in the study?" For this question, the analyses focused on four task attributes followed by two reader attributes and used a statistical approach.

The comprehension task constructed for the current study was comprised of two elements: a set of text passages and a set of comprehension questions about those text passages. In the study, two passage-level attributes—type and topic—were considered. There were two topics in each text type, expository and narrative. The patterns observed are as follows. Comprehension of 8th grade study participants varied by text topic, that is, the number of items correctly answered varied among the four text topics. Further, comprehension varied by text type, that is, the study participants, on average, correctly

---

[9] When stating the results, a reference to more, higher, lower, or varied indicates a statistical comparison that is significant at a minimum of $\alpha = .05$.

answered more items from the narrative-type texts than from the expository-type texts. In addition, the number of correctly answered items varied between the two topics included within a text type.

In the CI model of comprehension, Kintsch (1998) suggests that what a reader reads affects comprehension. More specifically, researchers suggest that text type affects comprehension (e.g., Best, Floyd, et al., 2008; Best, Ozuru, et al., 2006; Tun, 1989; Wolfe & Woodwyk, 2010; Zwaan, 1994). It has been recommended that texts of different types be considered when exploring comprehension, as it is a complex process (e.g., Pressley & Block, 2002). It has also been suggested that the content of a text affects comprehension (e.g., Cervetti et al., 2009; Magliano et al., 2007). It is recommended that both the type and topic of a text be taken into consideration when analyzing comprehension because we cannot assess "comprehension, but comprehension of …" (Duke, 2005; p. 98).

In the current study, two other attributes of a comprehension task were considered when constructing the questions that a study participant was asked in order to demonstrate his or her comprehension. Those attributes were the levels of mental representation of a text passage and the types of relation entailed in the events of a situation described in the passage. The two levels of mental representation considered were textbase and situation model. With respect to the item attribute of relation among the events, the passage-related questions presented to probe comprehension asked about four relations, namely, temporal, spatial, and causal, with the last of these classified into relations involving causal antecedent and causal consequence.

In this study, it was observed that comprehension of the 8th-grade study participants varied by level of text representation. The participants correctly answered more items among those that targeted textbase mental representation than those items that targeted the situation model representation of a text.

In the CI model, Kintsch (1998) also suggests that what a reader is asked about what has been read affects comprehension. According to the CI model, the textbase, a simpler level of mental representation of a text generally is dominated by the propositions in the text. In contrast, the situation model, a higher-level representation of a text, is a blend of a text-driven and a knowledge-driven representation of the text; such a representation generally results when a reader integrates his or her rich and relevant knowledge with the information from the text being read. Magliano et al. (2007) observed higher scores for textbase items than for situation model items.

In this study, it was also observed that study participants' comprehension of the different relations varied. The general pattern observed is as follows. Participants in the study, on average, correctly answered more of those items that asked for a spatial relation than those that asked for a temporal relation. Further, more items were correctly answered among those that asked for a causal antecedent relation than among those that asked for a causal consequence relation. However, when those passage-related questions were grouped into causal and non-causal, the difference in the number of items correctly answered between those two broad types of relation was not statistically significant.

Reading researchers suggest that a reader needs to identify various relations embedded among the events of a situation described by a text in order to comprehend it (e.g., Graesser et al., 2003; Kintsch, 1998; Rinck et al., 1997; Therriault & Raney, 2007;

Zwaan & Radvansky, 1998). An understanding of the relations among the events helps a reader develop coherence among the various aspects of a situation described in a text. Nevertheless, according to Kintsch (1998), the activation processes of these relations vary. Magliano et al. (1993) observed differential elevated times when a text presented discontinuities in these relations; time was more elevated for temporality and causality, and less so for spatiality.

With respect to a reader, the primary component of comprehension, two text passage-specific attributes were considered in the current study. Those attributes were a reader's familiarity with a text topic and interest in a text topic. In the MDL (Alexander, 2005), these two attributes are considered critical for the meaning a reader derives from reading a text. In this study, those reader attributes were assessed by means of a set of questions that asked readers about their familiarity with or interest in a given topic.

In this exploratory study, with respect to the reader attribute of Topic Familiarity, it was observed that the students who were identified as familiar with a topic tended to answer correctly more items than those who were not familiar with the topic. Such a difference was statistically significant in the case of both expository-type text topics. In the case of both narrative-type text topics, the apparent difference between those who were familiar and those who were not familiar with a topic was not significant.

Topic familiarity is an indicator of prior knowledge, which is considered pivotal in the comprehension of what is read (e.g., Alexander, 2003; RRSG, 2002; Rumelhart, 1980; van den Broek et al., 1999). In the CI model of Kintsch (1998), prior knowledge is considered to be an integral factor of comprehension. Many researchers concerned with the assessment of comprehension suggest that explaining student performance in light of

this attribute is vital (e.g., Durkin, 1993; Johnston, 1984; Pearson & Hamm, 2005; Pearson & Johnson, 1978).

With respect to the reader attribute of Topic Interest, fewer of the study participants reported being interested in topics of the expository-type texts compared to the narrative type texts included in the study. Further, results showed that study participants who reported being interested in the narrative topic *Cry of the Kalahari* tended to answer more items correctly than those participants who were identified as not being interested in the topic. However, this pattern of difference between the groups, interested and not-interested was not observed in the case of other topics included in the study.

Topic interest is also credited with contributing to the comprehension of a text (e.g., Hidi, 2001; Magliano, Durik, & Holt, 2011). In explaining predictors of comprehension, prior research reported that topic interest accounted for a significant portion of variance in comprehension performance (e.g., Ainley et al., 2002a; Alexander et al., 1994; Baldwin et al., 1985; Bray & Barron, 2004).

Analyses were also conducted by taking into consideration of both reader attributes, topic familiarity and topic interest. In those analyses, each study participant was classified into four groups, once for each topic, based on whether the participant was familiar with and/or interested in a topic. In the case of both expository-type text topics, the four groups were found to vary in the number of items correctly answered. In the case of narrative-type text topics, the apparent difference among the four groups was not statistically significant.

Thus, the current study showed that the attributes of task and reader influence the comprehension of the 8th-grade students who participated in the study. However, the influence seemed to depend on what was read, that is, type of text and topic read. The influence also depends on what was asked, that is, level of representation and type of relation targeted. The influence of reader attributes, topic familiarity and topic interest, appeared to depend also on what was read, that is, the type of text and topic read. The patterns of influence observed in the study are in concert with what is suggested in the reading literature about comprehension, considered here as the foundation of a theoretically grounded comprehension assessment.

Research Question #2

The second research question investigated in this study was: "Do the task and reader attributes considered in the study contribute to the item difficulties and reader comprehension proficiencies?" This question was addressed by means of a psychometric approach where analysis performs joint scaling of examinees and items.

The goal of an assessment is to estimate proficiencies of individual examinees (i.e., an examinee's ability to respond correctly to an item given that item's difficulty). Therefore, it is necessary to use an analytic model to ascertain the effects of task and reader attributes, individually and jointly, on every item posed to the examinees. For this psychometric analysis, a cognitively based IRT model known as the Linear Logistic Test Model (LLTM; Fischer, 1973) was applied to see how those attributes of task and reader affect the estimates of students' comprehension proficiencies.

Three LLTM models were applied that differed in which comprehension components were being considered. In addition, the Rasch model was applied to get the

base estimates. In the Rasch model, item difficulties are not explained by task attributes and comprehension proficiencies are not associated with reader attributes.

In LLTM model I, four attributes of the study's comprehension task were represented by seven variables. Of those seven, one variable compared the effect of two text types, expository versus narrative. Similarly, one variable represented the effect of two topics included within the expository-type texts and one variable represented the effect of two topics included in the narrative-type texts. Of the remaining four variables, one represented the effect of mental representations of a text targeted in the items and three represented the four relations asked about by the items where a contrast was made with respect to the temporal relation.

Results showed that item difficulty was moderated due to all variables except one. In this model, items' relative difficulties were not moderated due to the topics included within the narrative-type text. Further, the effects of six variables differed in magnitude. The effects in the order of magnitude, from high to low, were as follows. Item difficulty was relatively higher if it asked for a temporal relation compared to a spatial relation or a causal antecedent relation. Item difficulty was relatively higher if the item targeted a situation model mental representation compared to the textbase. Further, difficulty was higher if the item was from an expository-type text, if the item asked for a temporal relation compared to a causal consequence relation, and if the item was about the text that described the topic, *Shifting Sands* compared to the topic *Going Green*. Those four task attributes (represented by seven variables) collectively contributed to about one quarter of the variance in difficulties of the items included in the study.

Both of the reader attributes, topic familiarity and topic interest were incorporated into LLTM models II and III in order to modeling item difficulty and student proficiency. In model II, where only the reader attributes were included, results showed a significant effect for familiarity but not for the interest attribute. In other words, the difficulty of an item relatively decreased if a reader was identified as familiar with a topic, by an amount reflected as an average over students in an effect parameter in the LLTM model. These two reader attributes collectively accounted for about twelve percent of the variance in the difficulties of the items.

In LLTM model III, these two reader-attributes were incorporated in addition to those of the task attributes. That is, nine variables were included in modeling item difficulty and student proficiency. All attributes were found to contribute to the difficulties of items. In other words, even effects due to narrative-type text topics and to a reader's interest in a topic were statistically significant. These nine variables (i.e., six attributes that were considered in the study) collectively accounted for about thirty-eight percent of the variance in effective item difficulties across all the items included in the study's comprehension measure.

In addition, it was observed that student proficiencies estimated by the Rasch model changed when the LLTM model was applied. As expected, the changes were negligible in the case of LLTM model I, which incorporated only item attributes. This is because all task attributes were the same for all readers. The changes were more conspicuous in the case of LLTM models II and model III which incorporated reader attributes. In those models, the circumstances of having familiarity and interest, separately and additively, rendered a task easier for a student. In other words, students

answering the same number of items showed higher effective proficiency when having topic familiarity and topic interest than when not familiar and/or not interested.

To sum up, the current study showed that task and reader attributes contribute to the item difficulties and comprehension proficiencies. Effects of the attributes as revealed in the psychometric approach were in accord with what was observed through the analyses conducted for the first research question. Since the psychometric approach was able to model both task and reader attribute effects that were in line with what was observed in the statistical approach, it is perhaps appropriate to say that it is viable to design a measure and define a compatible psychometric model for comprehension assessment following a conceptualization of the construct grounded on reading research.

Implications

This study investigated the viability of developing an assessment of reading comprehension following a premise of the construct held by many reading researchers. For this study, the assessment measure and the psychometric approach were grounded on Kintsch's (1998) Construction-Integration model and Alexander's (1997) Model of Domain Learning. The results of the study hold implications for comprehension assessment and research.

Comprehension Assessment

This study sets an example for designing an assessment of comprehension that is more aligned with a theory of the construct. It shows how to interlace the essential elements of an assessment with a cognitive view of the construct that is empirically and theoretically grounded. In this study, the major decisions in constructing the measure

(e.g., selecting text passages of various types and topics, developing cognitive questions) and in defining the analytic approach were backed by research in order to make the assessment process more transparent. As noted in ECD assessment principles (Mislevy et al, 2003), if the argument and critical decisions made in an assessment are not backed by research, it is difficult to re-examine whether the measure provides evidence and to determine whether the inferences drawn from the assessment results are warranted. Thus, this exploratory study illustrates how methodologies involved in a comprehension assessment can grow out of theories of the construct.

Further, by successful application of the LLTM, a cognitive-based IRT model, the study shows how we can integrate views of the construct into psychometric models. In this study, by expanding the LLTM model, attributes of the reader in addition to attributes of the task were incorporated in the psychometric analysis. This innovative approach shows how we can build a network among the attributes of various comprehension components. Such a network gives us a means to explain variance in the difficulty of comprehension items and variance in the comprehension proficiency of readers. This enhanced analytic capability enables us to have more insights into the measurement of comprehension.

This new way of designing comprehension assessment supports what many have yearned to see. For example, Alexander and Fox (2004) wanted to see comprehension tests developed by incorporating the advances in our knowledge about reading comprehension. An assessment, when conceptualized within the framework of a theory of comprehension is valuable in many ways including enhancing our understanding of that theory, says Kintsch and Kintsch (2005). Further, as stated in the NRC Report

(2001), a more coherent picture of the assessed ability, in our case, comprehension ability seems likely if tests are built by incorporating a model of the construct.

In addition, the analysis of comprehension tests vis-à-vis a theory of the construct, as done in this study when reviewing tests, puts forward a new way to evaluate tests in order to improve test-construct alignment. Tests are evaluated, in most cases, for their psychometric properties (e.g., reliability and validity indices) or for the alignment between test framework and test content. Sheehan and Mislevy (1990) contend that such evaluations are not sufficient to guarantee neither a "'good' test, nor a 'valid' one" (p. 270). As Carpenter and Paris (2005) explained, those psychometric properties could be influenced by factors unrelated to the construct measured. In addition, content alignment is not enough as validity encompasses more than content relevance (Messick, 1995). As Gorin and Embretson (2006) said:

> Traditional examinations of the construct measured by reading comprehension assessment have focused on evidence provided by correlations with external variables. However, patterns of relations among text scores primarily support the significance of a construct, rather than its meaning. The cognitive analysis provides important information regarding the substantive meaning of the construct underlying reading comprehension questions. The results provide a cognitive model that gives meaning to the construct of reading comprehension, including a list of construct relevant item features useful for the development of future items. (p. 405)

Finally, the findings of this study, task and reader attributes tended to influence comprehension whether analyzed at an aggregate or individual item level, supports for a

reform in developing comprehension tests and reporting results from those tests that others have advocated (e.g., Bray & Barron, 2003; Johnston, 1983; Snow & Lohman, 1989; RAND, 2002). Additional information from comprehension assessments can benefit test-result users in various ways. Information on the relation between task-related attributes and comprehension would help researchers and educators more fully understand what students know and can do given text passages to read. Information on the relation between reader-related attributes and comprehension can be found helpful in the diagnosis of students' failure and success in text comprehension. Such an additional understanding can help determine what kind of support could be offered, if necessary, to foster children's continued engagement in reading in order to promote learning from texts and to enhance the students' ability to comprehend.

Comprehension Research

This study also enriches the reading research literature. The results of the study strengthen the premise that attributes of task as well as reader influence comprehension. Further, the evidence that shows the effects of various attributes on estimates of proficiencies helps us more fully understand reading comprehension and broadens the horizon of research on this complex phenomenon.

In the current study, multiple key components of comprehension and the joint effects of attributes of those components were considered. Comprehension was explored using multiple text passages and multiple questions where the questions shared the same attributes across the passages. In addition, reader attributes, topic familiarity and topic interest, were also taken into consideration when designing the assessment measure and defining the analytic approach. This comprehensive design to explore comprehension

overcomes the shortcomings of traditional ways to explore comprehension. According to Kintsch and Rawson (2005), reading research studies to date are considered limited in enhancing the understanding of how the various components work together as a system to give rise to comprehension, because those studies generally explore the various comprehension components and their attributes in isolation.

This study can be used as an example to set up tasks in order to evoke evidence about comprehension and as a prototype to model the effects of various attributes of tasks and readers on comprehension. In addition, this study can serve as an example for constructing a measure to probe familiarity with and interest in texts that vary in types and topics for 8th-grade students.

In spite of these implications, one needs to take into consideration the limitations of this exploratory study when interpreting its findings. In this study, no claim has been made about the best way or one singular way to design and develop a comprehension assessment aligned with the reading literature. In addition, the form of meaning upon which this study has focused, that is, textbase and situation model representations of relations among events of a situation described in a text, is one of many possible forms of meaning a reader needs for complete comprehension.

Issues for Future Research

To complete the endeavor of developing a comprehension assessment grounded on a theory of the construct, several avenues for future research are suggested. These suggestions are offered given this study's limitations, which were identified in Chapter I.

To continue beyond this study, studies with expanded samples representing the target population would allow obtaining more stable estimates of the parameters from the

LLTM models. Replicating this study with a larger sample size that gives more

information on the four topic-familiarity*topic-interest reader groups would allow more

insight into the changes in estimates of comprehension proficiencies due to reader

attributes. More information on reader groups would allow for additional differentiations

in reader effect modeling. Further, studies conducted with a systematically drawn

representative sample and a different set of text passages that fits the framework of the

attributes considered in the present study would strengthen the assertion of the viability

of developing a theory-based comprehension assessment. Such studies will also allow

generalizations about what students at the 8th grade know and can do when asked to read

texts.

To continue exploring the effects of reader attributes on comprehension of texts,

studies are necessary to establish a scale to measure topic familiarity and topic interest of

readers. This study used self-reported information for those reader attributes and was

limited in verification of the information. As this study has shown, the quality indices of

the two measures were not parallel. Further, not all questions included in the measures

had statistically significant correlations with the total scores. A scale that is established

using a variety of topics and a large representative sample of the target population might

help minimize the uncertainty associated with self-reported information. With more

consistent information, it might be possible to discern a pattern in the relation between

reader attributes and comprehension proficiency estimates that was observed in the study.

In establishing a reader attribute measure, nevertheless, the constraints of an assessment

context, such as limited time, burden on students, should be taken into consideration.

Further, we need studies to improve upon the LLTM approach. In the current study, all task and reader attributes were treated as dichotomous variables. We need to identify ways so task and reader attributes in LLTM could be treated as a continuous variable without adding complexities to interpretations of the obtained results. We also need to explore the model fit metric, DIC that WinBUGS uses so that the comparisons between less and more constrained models are transparent. Finally, we need to explore sample sizes that would be necessary for the application of LLTM when the model incorporates multiple reader attributes and when the primary interest of the application is to examine interactions among those attributes.

More importantly, the need for a comprehensive theory of comprehension cannot be ignored. We need a model that portrays the interplay among the various comprehension components, including their attributes, to develop a theory-based comprehension assessment. We need to identify the underlying mechanisms of the effects of various task and reader attributes on comprehension in order to understand proficiency estimates. A comprehensive model can enrich our understanding of what students know and can do when reading texts.

Conclusion

While we are still making efforts toward a full understanding of comprehension, the lack of complete understanding cannot be a pretext for not building reading comprehension tests grounded on research. As this study has demonstrated, a theory of the construct of comprehension can be integrated jointly into the test design and a compatible psychometric tool. Comprehension assessments based on conceptualizations shared by reading researchers and test developers, as suggested by ECD principles, have

implications for what we measure and how we measure. An assessment of

comprehension built on a cognitive understanding within a principled assessment design

framework can guide task design and psychometric modeling, and more efficiently serve

the purpose of an assessment, namely, to provide educators with feedback on student

learning. Studies like this need to continue so before long we can have the next

generation of comprehension assessments that many have yearned to see for so long.

Appendix A: Comprehension Measure

Instruction:

In this section, you will read text passages and answer questions about the passages. For each question, you will have four answer choices. Fill in the circle beside the choice that you think is correct. [The booklet distributed to the participants provides instructions in detail.]

<u>Shifting Sands</u>

The carpeted holes of a miniature golf course lie buried beneath many feet of sand.

Now and again, depending on how the winds blow, the turrets of a small castle that once guarded the eighth hole reappear above one edge of the massive sand dunes named Jockey's Ridge.

Jockey's Ridge is the largest natural sand dune system in the eastern United States. Nearly 100 feet tall and lacking vegetation, the dunes cover more than 400 acres on the thin strip of land in North Carolina known as the Outer Banks.

The sand is blown back and forth by the prevailing winds, so the dunes are always moving. In the summer, lighter southwest winds move the sands one way, then stronger winter winds from the northeast move them back again. In fact, these winds cause the dunes to move three to six feet southwest each year.

Over time, man-made structures have been no match for the shifting sands. In the late 1880s, one hotel was covered by sand before it could be completely dismantled and moved to a nearby site. More recently, the brick-lined holes of a miniature golf course were buried under the sand.

The Jockey's Ridge dunes were originally formed when strong hurricane storm waves picked up sand from the shoals off the North Carolina coastline and deposited it on the shore. Each year the winds took it from there, and, grain by grain, a series of dunes developed.

Jockey's Ridge is a popular tourist attraction. During most of the year, the dunes are filled with visitors running up, rolling down, and hang gliding from the immense

sand hills. Climbers enjoy sunrises over the ocean and sunsets over the seaway at the beginning and end of each day.

The sand is hot in the summer. It can be thirty degrees hotter than the air temperature. The dunes attract many lightning strikes with their high profile along the flat shoreline. When lightning hits the sand, the temperature can reach 15,000°C, or twice the temperature of the surface of the sun, melting the quartz sand into silica glass tubes called fulgurites, which can be as long as six feet.

Over the last quarter of a century, Jockey's Ridge has moved several hundred feet southwest and has begun to flatten out, threatening to bury nearby houses and roads. Scientists and local residents are studying the process carefully, trying to understand why the dunes now seem to be steadily moving and shrinking, and what should be done about it.

Many geologists believe that the dune migration is a natural process, perhaps the result of changes in wind patterns over recent years. Others believe that commercial and residential development of the barrier islands has interfered with the cycle of shifting sand on Jockey's Ridge.

Debate continues about what to do with the moving dunes. Should at-risk areas of the dunes be fenced in, or should dump trucks be used to keep moving the sand back where people want it to be? Should plants and grass be grown on the dunes to keep them from moving, or should the dunes be allowed to go whichever way the wind blows?

While each proposal for handling the moving sands of Jockey's Ridge has its own set of problems and issues, one thing is certain. The turrets of the castle that used to decorate the now-buried miniature golf course are an eerie reminder of the vastness and power of the moving dunes.

1. Where did the sands of Jockey's Ridge originate?

    Textbase
    Spatial

    a.) in the shoals of South Carolina brought by ocean waves

    b.) in the coastline of Outer Banks migrated by a natural process

    c.) in the shoals of North Carolina deposited by strong storm waves

    d.) in the sandbar on Barrier Islands deposited by strong winter winds

2. How long does it take for dunes like Jockey's Ridge to develop?

    a.) a few days

    Situation Model
    Temporal

    b.) a few years

    c.) a quarter of a century

    d.) a century or more

3. What would most likely happen if the southwest winds stop blowing?

    a.) Dunes will shift further southwest.

    Situation Model
    Causal Consequence

    b.) Dunes will shrink and flatten out.

    c.) Dunes will stay in their current location.

    d.) Dunes will move toward the northwest.

4. Why do Jockey's Ridge dunes attract lightning?

    Situation Model
    Causal Antecedent

    a.) because their sky is filled with hang gliders

    b.) because they are taller than other objects in the area

    c.) because they cover a vast flat area along the shoreline

    d.) because their temperature is hotter than the surface of the sun

5. How long ago did Jockey's Ridge dunes start to shrink and flatten out?

    (a.) about 15 years ago

    (b.) about 25 years ago

    (c.) about 45 years ago

    (d.) about 75 years ago

> Textbase
> Temporal

6. Why are Jockey's Ridge dunes a problem?

    (a.) They puzzle those studying dunes.

    (b.) They put the nearby residences in danger.

    (c.) They take away the lands from the Outer Banks.

    (d.) They attract too many visitors spoiling the dunes.

> Textbase
> Causal Antecedent

7. Where would a visitor find Jockey's Ridge dunes in 25 years if the dunes keep shifting?

    (a.) farther to the northeast

    (b.) more toward the ocean

    (c.) farther to the northwest

    (d.) more toward the seaway

> Situation Model
> Spatial

8. What happened to the man-made structures built on Jockey's Ridge?

    (a.) They were moved to a nearby island.

    (b.) They were taken apart by their owners.

    (c.) They were blown away by high winds.

    (d.) They were buried underneath the dunes.

> Textbase
> Causal Consequence

Kid Fights Cheater Meters and Wins!

Ellie Lammer wasn't trying to spark a revolt, she just wanted a haircut. That was in the fall of 1997. Ellie was 11 years old at the time, and she was getting her tresses trimmed in her hometown of Berkeley, California. When Ellie and her mom returned to their car, they found a parking ticket stuck to the windshield. It didn't seem possible: Less than an hour earlier, Ellie had pumped an hour's worth of coins into the meter. But now the needle was at zero, and Ellie's mom owed $20.

Feeling cheated, Ellie dropped another nickel in the meter and twisted the knob. The needle clicked over to the four-minute mark. Ellie stared at her watch while her mom watched the meter. Less than three minutes later, all of the time had expired. There it was: proof that they'd been cheated. The city tore up the ticket when Ellie's mom complained about the meter.

But the experience left Ellie wondering how many other meters were inaccurate. Six months later, she decided to find out. She'd been looking around for a good science-fair project—and that meter in Berkeley still bothered her. So armed with a bag of nickels and a stopwatch, she hit the streets.

Ellie didn't have the time or money to test every meter, so she focused on a sample of 50 meters located in different parts of the city. To avoid inconveniencing motorists, she did her research after 6 P.M. and on Sundays, when the meters were not in use. She put in eight minutes' worth of nickels in each meter, then measured how much time it really gave.

The results were not pretty. Ellie's findings suggested that more than nine out of every ten meters in the city were inaccurate—and that every fourth parking meter was running out of time too quickly. With 3,600 parking meters in the city, that meant a lot of undeserved tickets. As Ellie wrote in her science-project report, "I learned which meters cheat you and which meters cheat the City of Berkeley. But I learned that almost all meters cheat someone, so beware."

When the science fair rolled around, Ellie presented her findings with computer-generated charts and graphs. Her classmates weren't very interested in her project. "It's not like they have to drive a car or put money in a parking meter," she explains. But her project was a huge hit with parents. More than 50 of them lined up that night to share their own parking-meter horror stories with Ellie.

After that, word about Ellie's meter project spread fast. Within a few weeks, Ellie got a call from local politician Diane Woolley. At the time, Berkeley was considering replacing its meters with more accurate digital ones. Ellie shared her findings at city hall, and the politicians were impressed. "We don't get reports this thorough when we pay

consultants hundreds of thousands of dollars," one remarked. Based on Ellie's study, they decided to purchase 2,000 new meters.

The California state legislature also decided to crack down on cheater meters. After Ellie presented her findings, they enacted "Lammer's Law," which requires California's 26 counties to test the accuracy of parking meters. Any meter found to be inaccurate must be fixed or dismantled.

California Governor Pete Wilson signed the law on November 1, 1998. At the time, he commented, "Ellie's ingenuity and dedication has earned her the gratitude of those Californians who've dug through their purses and pockets in search of exact change to feed the meters, only to return to find their cars bearing the dreaded green envelope of a parking ticket."

Ellie became a celebrity. She was in newspapers all over the country and featured on local television news during the summer and fall of 1998. CNN did a story about her. She was even a guest on the Late Show with David Letterman. "It was kind of a weird moment of being a celebrity," she says.

Ellie is proud of the work she's done. But she doesn't see meter monitoring as her life's work: "Right now I don't mind being known as the parking-meter girl, but I'm sure that later in life I'll want something different."

1. Where did Ellie live when she became a celebrity for her science project?

Textbase
Spatial

   a. She lived in a city that has about four thousand parking meters.

   b. She lived in a city seldom visited by motorists on Saturdays.

   c. She lived in a city that has thousands of consultants working on meters.

   d. She lived in a city located in California that had no television station.

2. Why did Ellie's mom find a parking ticket stuck on her car windshield?

Textbase
Causal Antecedent

   a. She parked next to a hair salon.

   b. She parked next to a faulty meter.

   c. She parked blocking other motorists.

   d. She parked without paying the meter.

3. What was most likely the last step Ellie took prior to writing her science project report?

> Situation Model
> Temporal

- a She collected data from the inaccurate meters in the city of Berkeley.

- b She determined the number of meters she needed to test for her study.

- c She shared her study results at the science fair using charts and graphs.

- d She calculated the possible number of malfunctioning meters in the city.

4. According to the passage, when does the city of Berkeley get most motorists in need of parking?

> Textbase
> Temporal

- a. every day of the week during the daytime except Saturday

- b. every day of the week during the daytime except Sunday

- c. every day of the week during the nighttime except Saturday

- d. every day of the week during the nighttime except Sunday

5. Why did media feature a story on a young girl named Ellie?

> Situation Model
> Causal Antecedent

- a. She conducted a well-designed science fair project.

- b. She met famous politicians including Governor Wilson.

- c. She uncovered a problem with the city's parking meters.

- d. She saved many Californians from undeserved parking tickets.

6. Where did Ellie go to present her findings to Diane Woolley?

> Situation Model
> Spatial

- a. Ellie went on The Late Show to present her findings.

- b. Ellie went to Berkeley City Hall to present her findings.

- c. Ellie went to the Longfellow School to present her findings.

- d. Ellie went to the California State Capitol to present her findings.

7. What did the California state legislature do to resolve the issue reported by Ellie?

<div style="float:right; border:1px solid #ccc; padding:4px;">
Textbase

Causal consequence
</div>

   a.) They asked their counties to replace every meter with a digital one.

   b.) They dismantled all inaccurate meters based on Ellie's findings.

   c.) They invited Ellie to the signing ceremony of Lammer's Law.

   d.) They passed a law named after Ellie to monitor the local meters.

8. What is most likely to happen if a parking meter is inaccurate?

<div style="float:right; border:1px solid #ccc; padding:4px;">
Situation Model

Causal Consequence
</div>

   a.) The city that owns the meter will earn money.

   b.) The motorist who uses the meter will lose money.

   c.) Either the owner or user of the meter will lose money.

   d.) Both the owner and user of the meter will lose money.

<u>Going Green</u>

One of the most frequently discussed topics in the United States today is environmental awareness. Many Americans are becoming increasingly concerned about the environment and are searching for more ways of "going green." The phrase "going green" refers to the act of becoming involved in environmental protection and adopting the beliefs of environmentalism. Environmentalist ideals have existed for a very long time. Recently environmentalism has been known as the Green Movement. This movement encourages recycling, conservation, and protection of the environment.

Early Environmentalism: The Sierra Club and the United States Forest Service

In the early history of the United States, the country appeared to have plentiful natural resources. The New World seemed full of unmapped wilderness and endless forests, rivers, and animals. As time passed, however, some people began to realize that waste and pollution threatened the abundance of resources. Environmental efforts became more public and more organized in the late nineteenth century. Concerned citizens worried about the diminishing American wilderness and the harmful effects of the Industrial Revolution.

The year 1892 saw the founding of the Sierra Club, the oldest formal environmental organization in the United States. The Sierra Club first focused on preserving prominent natural areas such as Yellowstone National Park and the Grand Canyon. The organization has taken on many environmental battles during its long existence—fighting against the reduction of Yosemite National Park, arguing for the establishment of national forest parks, and urging the creation of a national park service. Today the Sierra Club has more than a million members.

Another important environmental organization, the United States Forest Service (USFS), was established in 1905. The USFS worked to increase the number and the size of national forests. In the five years following the service's founding, the country more than doubled the number of national forests. The number of forests grew from 60 to 150; and the total area of national forests more than tripled, moving from 56 million acres to 172 million acres. Gifford Pinchot, the first forest service director, spread the idea of conservation and encouraged Americans to use natural resources wisely. He is now widely considered the father of American forestry.

The Rise of Environmentalism in the 1960s and 1970s

Environmentalism began to assume its current shape during the 1960s. In 1962 Rachel Carson published her book *Silent Spring*, which posed questions regarding the safety of certain pesticides. Despite the controversy that rose up around the book, Silent

Spring became a driving force behind many people's interest in environmental protection. This rise in interest led to a rapid increase in environmentalist activity in the early 1970s.

An Earth Day celebration was held on April 22, 1970. More than 20 million Americans participated in events around the country to demonstrate their commitment to a healthy environment. Earth Day is an annual observance honoring the modern environmental movement and is still celebrated almost forty years later. The founding in 1971 of the environmental groups, Friends of Earth and Greenpeace came soon after the first Earth Day. The first international Earth Summit, in 1972, was attended by representatives from 113 countries.

The Environmental Protection Agency (EPA) was also established in the 1970s. The agency's initial employees were assembled from several departments within the federal government. The EPA's first administrator, William D. Ruckelshaus, took immediate and dramatic steps to begin remedying the country's pollution problems. The agency continues to monitor and protect America's natural resources in five key areas:

- clean air and global climate change,
- clean and safe water,
- land preservation and restoration,
- healthy communities and ecosystems, and
- compliance and environmental stewardship.

Environmentalism Today

The intensity of the environmental movement slowed down slightly during the following decade but grew again during the 1990s as concerns regarding climate change increased. In the first decade of the twenty-first century, the issue of global warming has become a popular subject for news outlets, politicians, and the American public. The Green Movement continues to gain followers who believe that America's citizens share responsibility for the environment. Because of the continued work of individuals, schools, public and private groups, and corporations, environmental issues are receiving the attention they deserve.

1. Why does America have so many organizations devoted to protecting the environment?

   Situation Model

   Causal Antecedent

   a. It is a country that needs to protect various industries.

   b. It is a country that has many types of natural resources.

   c. It is a country that leads others in preserving resources.

   d. It is a country that has many people looking for jobs.

2. Where do supporters of the Green Movement come from?

   Textbase

   Spatial

   a. about fifty countries

   b. the United States of America

   c. nations around the world

   d. agencies within the government

3. Which of the following environmental organizations was the first to be founded in the 20th century?

   Textbase

   Temporal

   a. Sierra Club

   b. Silent Spring

   c. United States Forest Service

   d. Environmental Protection Agency

4. How did Rachel Carson's book impact the Green Movement?

   Textbase

   Causal Consequence

   a. It inspired people to expend more natural resources.

   b. It offered ideas for celebrating the first Earth summit.

   c. It led the EPA to put a ban on the use of certain pesticides.

   d. It raised awareness among people in saving the environment.

5. What was the most recent decade when interest in the Green Movement lessened in the United States?

> Situation Model
>
> Temporal

   a. 1891 to 1900

   b. 1960 to 1969

   c. 1980 to 1989

   d. 2001 to 2010

6. Why did millions of Americans take part in the first Earth Day celebration?

> Textbase
>
> Causal Antecedent

   a. They wanted to honor the environmental movement.

   b. They wanted to establish a new environmental group.

   c. They wanted to raise funds to protect the environment.

   d. They wanted to show support for a cleaner environment.

7. How would wise use of natural resources benefit America?

> Situation Model
>
> Causal Consequence

   a. America's local climate would get warmer.

   b. America's drinking water would become purer.

   c. America's abundant resources would last longer.

   d. America's natural areas would be protected for ever.

8. Where would you most likely find members of the EPA working to reduce pollution?

> Situation Model
>
> Spatial

   a. Chesapeake Bay

   b. Yosemite National Park

   c. Audubon Wildlife Refuge

   d. Black Hills National Forest

Cry of the Kalahari by Mark J. and Delia D. Owens

My left shoulder and hip ached from the hard ground. I rolled to my right side, squirming around on grass clumps and pebbles, but could not get comfortable. Huddled deep inside my sleeping bag against the chill of dawn, I tried to catch a few more minutes of sleep.

We had driven north along the valley the evening before, trying to home in on the roars of a lion pride. But by three o'clock in the morning they had stopped calling and presumably had made a kill. Without their voices to guide us, we hadn't been able to find them and had gone to sleep on the ground next to a hedge of bush in a small grassy clearing. Now, like two large armyworms, our nylon sleeping bags glistened with dew in the morning sun.

Aaoouu—a soft groan startled me. I slowly lifted my head and peered over my feet. My breath caught. It was a very big lioness— more than 300 pounds—but from ground level she looked even larger. She was moving toward us from about five yards away, her head swinging from side to side and the black tuft on her tail twitching deliberately. I clenched a tuft of grass, held on tight, and froze. The lioness came closer, her broad paws lifting and falling in perfect rhythm, jewels of moisture clinging to her coarse whiskers, her deep-amber eyes looking straight at me. I wanted to wake up Delia, but I was afraid to move.

When she reached the foot of our sleeping bags, the lioness turned slightly. "Delia! S-s-s-h-h-h—wake up! The lions are here!"

Delia's head came up slowly and her eyes grew wide. The long body of the cat, more than nine feet of her from nose to tuft, padded past our feet to a bush ten feet away. Then Delia gripped my arm and quietly pointed to our right. Turning my head just slightly, I saw another lioness four yards away, on the other side of the bush next to us . . . then another . . . and another. The entire Blue Pride, nine in all, surrounded us, nearly all of them asleep. We were quite literally in bed with a pride of wild Kalahari lions.

Like an overgrown house cat, Blue was on her back, her eyes closed, hind legs sticking out from her furry white belly, her forepaws folded over her downy chest. Beyond her lay Bones, the big male with the shaggy black mane and the puckered scar

over his knee— the token of a hurried surgery on a dark night months before. Together with Chary, Sassy, Gypsy, and the others, he must have joined us sometime before dawn.

We would have many more close encounters with Kalahari lions, some not quite so amicable. But the Blue Pride's having accepted us so completely that they slept next to us was one of our most rewarding moments since beginning our research in Botswana's vast Central Kalahari Desert, in the heart of southern Africa. It had not come easily.

As young, idealistic students, we had gone to Africa entirely on our own to set up a wildlife research project. After months of searching for a pristine area, we finally found our way into the "Great Thirst," an immense tract of wilderness so remote that we were the only people, other than a few bands of Stone Age Bushmen, in an area larger than Ireland. Because of the heat and the lack of water and materials for shelter, much of the Central Kalahari has remained unexplored and unsettled. From our camp there was no village around the corner or down the road. There was no road. We had to haul our water a hundred miles through the bush veld, and without a cabin, electricity, a radio, a television, a hospital, a grocery store, or any sign of other humans and their artifacts for months at a time, we were totally cut off from the outside world.

Most of the animals we found there had never seen humans before. They had never been shot at, chased by trucks, trapped, or snared. Because of this, we had the rare opportunity to know many of them in a way few people have ever known wild animals. On a rainy-season morning we would often wake up with 3000 antelope grazing around our tent. Lions, leopards, and brown hyenas visited our camp at night, woke us up by tugging the tent guy ropes, occasionally surprised us in the bath boma, and drank our dishwater if we forgot to pour it out. Sometimes they sat in the moonlight with us, and they even smelled our faces.

There were risks—we took them daily—and there were near disasters that we were fortunate to survive. We were stranded without water, battered by storms, and burned by droughts. We fought veld fires miles across that swept through our camp—and we met an old man of the desert who helped us survive.

We had no way of knowing, from our beginnings with a third-hand Land Rover, a campfire, and a valley called "Deception," that we would learn new and exciting details about the natural history of Kalahari lions and brown hyenas: How they survive droughts

with no drinking water and very little to eat, whether they migrate to avoid these hardships, and how members of these respective species cooperate to raise their young. We would document one of the largest antelope migrations on earth and discover that fences are choking the life from the Kalahari.

1. What did Mark do first after he heard the soft groan of a lion?

    Textbase
    Temporal

    a. He whispered to Delia to wake up.

    b. He clenched a tuft of grass and froze.

    c. He huddled deep inside his sleeping bag.

    d. He raised his head to look toward his feet.

2. What does the passage suggest about a pristine area?

    Textbase
    Spatial

    a. It is an area untouched by civilization.

    b. It is an area accessible only by Land Rover.

    c. It is an area inhabited by a variety of animals.

    d. It is an area densely populated by human beings.

3. What was the main advantage of the site that Mark and Delia selected for their research?

    Situation Model
    Spatial

    a. The site gave them chances to experience various natural hazards.

    b. The site allowed them to study the natural behavior of wild animals.

    c. The site allowed them to see many kinds of animals not seen in zoos.

    d. The site gave them easy access to their daily needs of food and water.

4. Why did the Blue Pride approach Mark and Delia without fear and did not harm them?

Textbase

Causal Antecedent

    a. The researchers had fed those starving wild animals.

    b. Those wild animals were friendly and harmless to others.

    c. Those wild animals had never before encountered humans.

    d. The researchers allowed those animals to sniff and lick them.

5. Why were there no roads where Mark and Delia set up their camp?

    a. Bands of Stone Age Bushmen used only trails.

Situation Model

Causal Antecedent

    b. Few people came to live in this inhospitable area.

    c. Veld fires and extreme heat destroyed the roads.

    d. Botswana designated the Kalahari as a wildlife refuge.

6. How have fences most likely affected the Kalahari wildlife?

Situation Model

Causal Consequence

    a. by keeping the animals from moving freely

    b. by limiting the illegal hunting and capture of animals

    c. by protecting the animals from droughts and storms

    d. by helping the animals avoid conflicts with each other

7. What is the effect on the animals of the hardship of living in the Kalahari?

Textbase

Causal Consequence

    a. They learn to survive with little to eat and drink.

    b. They become aggressive toward human beings.

    c. They learn to live in harmony with other species.

    d. They become curious about strangers to their habitat.

8. How long do you think Mark and Delia stayed in Africa to learn about wildlife in the Kalahari?

> Situation Model
>
> Temporal

   a. a few weeks

   b. two months

   c. one rainy-season

   d. more than a year

Appendix B: Topic Familiarity Measure

Read each question and the response choices carefully. Fill in the circle that best describes you.

I. How much do you know about the following topics?

	Fill in one circle for each row			
	Not at all	**A little**	**Some**	**A lot**
1. Dunes: what they are, where they are	○	○	○	○
2. Deserts: what they are, where they are	○	○	○	○
3. Earth Day: when and why we observe it	○	○	○	○
4. Parking a car: where and how it is done	○	○	○	○
5. Politicians: who they are, what they do	○	○	○	○
6. Recycling: what it is and how we do it	○	○	○	○
7. Seas and bays: what they are, where they are	○	○	○	○
8. Wild animals: where and how they live	○	○	○	○

II. Do you know what a text with the following title might cover?

	Fill in one circle for each row			
	Not at all	**Maybe not**	**Maybe**	**A lot**
1. Going Green	○	○	○	○
2. Shifting Sands	○	○	○	○
3. Cry of the Kalahari	○	○	○	○
4. Kid Fights Cheater Meters and Wins	○	○	○	○

III.   Did the following passage discuss a topic you had read about it before today?

	Fill in one circle for each row			
	No	Maybe not	Maybe	Yes
1. Going Green	○	○	○	○
2. Shifting Sands	○	○	○	○
3. Cry of the Kalahari	○	○	○	○
4. Kid Fights Cheater Meters and Wins	○	○	○	○

IV.   Did the following passage cover a topic you knew a lot about it?

	Fill in one circle for each row			
	No	Maybe not	Maybe	Yes
1. Shifting Sands	○	○	○	○
2. Cry of the Kalahari	○	○	○	○
3. Kid Fights Cheater Meters and Wins	○	○	○	○
4. Going Green	○	○	○	○

V.   How much did your prior knowledge help you to understand the information described in the passage?

	Fill in one circle for each row			
	Not at all	A little	Some	A lot
1. Kid Fights Cheater Meters and Wins	○	○	○	○
2. Going Green	○	○	○	○
3. Shifting Sands	○	○	○	○
4. Cry of the Kalahari	○	○	○	○

Appendix C: Topic Interest Measure

Read each question and the response choices carefully. Fill in the circle that best describes you.

I. How much would you like to know about the following topics?

Fill in one circle for each row

	Not at all	A little	Some	A lot
1. Dunes: what they are, where they are	○	○	○	○
2. Deserts: what they are, where they are	○	○	○	○
3. Earth Day: when and why we observe it	○	○	○	○
4. Parking a car: where and how it is done	○	○	○	○
5. Politicians: who they are, what they do	○	○	○	○
6. Recycling: what it is and how we do it	○	○	○	○
7. Seas and bays: what they are, where they are	○	○	○	○
8. Wild animals: where and how they live	○	○	○	○

II. How interesting does each of the following titles seem to you?

Fill in one circle for each row

	Not interesting	A little interesting	Some interesting	Very interesting
1. Going Green	○	○	○	○
2. Shifting Sands	○	○	○	○
3. Cry of the Kalahari	○	○	○	○
4. Kid Fights Cheater Meters and Wins	○	○	○	○

IIIs. How much did you enjoy reading the passage?

	Not at all	A little	Some	A lot
	Fill in one circle for each row			
Shifting Sands	○	○	○	○

IVs. How much would you like to read the passage again?

	Not at all	A little	Some	A lot
	Fill in one circle for each row			
Shifting Sands	○	○	○	○

IIIk. How much did you enjoy reading the passage?

	Not at all	A little	Some	A lot
	Fill in one circle for each row			
Kid Fights Cheater Meters and Wins	○	○	○	○

IVk. How much would you like to read the passage again?

	Not at all	A little	Some	A lot
	Fill in one circle for each row			
Kid Fights Cheater Meters and Wins	○	○	○	○

IIIg. How much did you enjoy reading the passage?

	Not at all	A little	Some	A lot
	Fill in one circle for each row			
Going Green	○	○	○	○

IVg. How much would you like to read the passage again?

	Fill in one circle for each row			
	Not at all	**A little**	**Some**	**A lot**
Going Green	○	○	○	○

IIIc. How much did you enjoy reading the passage?

	Fill in one circle for each row			
	Not at all	**A little**	**Some**	**A lot**
Cry of the Kalahari	○	○	○	○

IVc. How much would you like to read the passage again?

	Fill in one circle for each row			
	Not at all	**A little**	**Some**	**A lot**
Cry of the Kalahari	○	○	○	○

V. How much would you like to discuss this text passage with others (e.g., family or friends)?

	Fill in one circle for each row			
	Not at all	**A little**	**Some**	**A lot**
1. Kid Fights Cheater Meters and Wins	○	○	○	○
2. Going Green	○	○	○	○
3. Cry of the Kalahari	○	○	○	○
4. Shifting Sands	○	○	○	○

VI. How interesting was the passage you read today?

	Fill in one circle for each row			
	Not interesting	**A little interesting**	**Somewhat interesting**	**Very interesting**
1. Cry of the Kalahari	◯	◯	◯	◯
2. Shifting Sands	◯	◯	◯	◯
3. Kid Fights Cheater Meters and Wins	◯	◯	◯	◯
4. Going Green	◯	◯	◯	◯

Appendix D: Demographic Profile Measure

Read each question and the response choices carefully. Fill in the circle that best

describes you.

I. How often do you do the following?

	Never or Hardly ever	**Sometimes**	**Often**	**Everyday or Almost every day**
1. I do homework that requires reading	○	○	○	○
2. I take tests/quizzes about what I read in my class	○	○	○	○
3. I read on my own for enjoyment	○	○	○	○
4. I read on my own to learn new things	○	○	○	○
5. I read on my own to make or fix something	○	○	○	○

II. Which of the following categories best describes you? (*You may fill in more than one circle*)

○ American Indian or Alaska Native

○ Asian

○ Black or African American

○ Native Hawaiian or Other Pacific Islander

○ White

III. Are you female or male?

○ Female

○ Male

Appendix E: Instructions Given to Participants

Good morning / Good afternoon!

I am …, a researcher from the University of Maryland, College Park.

Today you will participate in a study about reading.

You are invited to this study because you are among many students who are currently

attending grade 8.

You will complete a booklet, *like this as holding one*.

The booklet has directions to follow, text passages to read, questions to answer. Some

questions are about the passages and some are about you.

Please answer all questions included in the booklet the best you can.

If you have any problems following the directions, or you are not sure of the task you

are asked to do, please raise your hand.

If you are done before others, please stay at your seat until I collect the booklet from

you.

Your responses will be used for research purposes. Your participation in the study is

very important.

Do you have any question?

Please open your booklet and proceed.

Thank you for your support, *as collecting the booklet*.

Appendix F: Institutional Review Board Approval

UNIVERSITY OF
MARYLAND
INSTITUTIONAL REVIEW BOARD

To: Principal Investigator, Dr. Patricia A. Alexander, Human Development
Student, Taslima Rahman, Human Development

From: James M. Hagberg
IRB Co-Chair
University of Maryland College Park

Re: IRB Protocol: 11-0393 - Reading comprehension and its assessment: Aligning operationalization with conceptualization of the construct

Approval Date: June 23, 2011

Expiration Date: June 23, 2012

Application: Initial

Review Path: Expedited

> Per an e-mail received on 5/24/2012 from Andrea Dragan, UMCP IRB, this application, the IRB Net #320281 is approved through June 22, 2013.

The University of Maryland, College Park Institutional Review Board (IRB) Office approved your Initial IRB Application. This transaction was approved in accordance with the University's IRB policies and procedures and 45 CFR 46, the Federal Policy for the Protection of Human Subjects. Please reference the above-cited IRB Protocol number in any future communications with our office regarding this research.

Recruitment/Consent: For research requiring written informed consent, the IRB-approved and stamped informed consent document will be sent via mail. The IRB approval expiration date has been stamped on the informed consent document. Please note that research participants must sign a stamped version of the informed consent form and receive a copy.

Continuing Review: If you intend to continue to collect data from human subjects or to analyze private, identifiable data collected from human subjects, beyond the expiration date of this protocol, you must submit a Renewal Application to the IRB Office 45 days prior to the expiration date. If IRB Approval of your protocol expires, all human subject research activities including enrollment of new subjects, data collection and analysis of identifiable, private information must cease until the

Renewal Application is approved. If work on the human subject portion of your project is complete and you wish to close the protocol, please submit a Closure Report to irb@umd.edu.

Modifications: Any changes to the approved protocol must be approved by the IRB before the change is implemented, except when a change is necessary to eliminate an apparent immediate hazard to the subjects. If you would like to modify an approved protocol, please submit an Addendum request to the IRB Office.

Unanticipated Problems Involving Risks: You must promptly report any unanticipated problems involving risks to subjects or others to the IRB Manager at 301-405-0678 or jsmith@umresearch.umd.edu

Additional Information: Please contact the IRB Office at 301-405-4212 if you have any IRB-related questions or concerns. Email: irb@umd.edu

The UMCP IRB is organized and operated according to guidelines of the United States Office for Human Research Protections and the United States Code of Federal Regulations and operates under Federal Wide Assurance No. FWA00005856.

1204 Marie Mount Hall
College Park, MD 20742-5125
TEL 301.405.4212
FAX 301.314.1475
irb@umd.edu
http://www.umresearch.umd.edu/IRB

Appendix A: Parental Permission Form

<u>University of Maryland College Park</u>

Page 1 of 2

Initials _____ Date _____

Project Title	Reading comprehension and its assessment: Aligning operationalization with conceptualization of the construct
Purpose of the Study	This research is being conducted by **Dr. Patricia A. Alexander** at the University of Maryland, College Park. We are inviting you to participate in this research project because you are the parent (or guardian) of a child who is attending 8th grade, the target population for the study. The study investigates ways to improve current reading comprehension assessments that appear inadequate to explain performance differences between students within- and across text passages.
Procedures	The study will be conducted in a setting (e.g., in library, or in a classroom) at your child's school. Your child will be asked to read a set of text passages (e.g., narrative and expository texts) and answer questions related to those passages. In addition, the child will be asked a set of questions to indicate his/her familiarity with- and interest in those passages. You child will also be asked a couple of questions on demographics (i.e., gender and race/ethnicity). All questions have response choices. No additional writing is required except selection of the responses. This procedure is expected to take about 50 minutes for a child to complete.
Potential Risks and Discomforts	There are no known risks associated with participating in this study. The study will be conducted in a setting at your child's school. Your child will participate along with other 8th graders selected from the school.
Potential Benefits	This research is not designed to help the child who is participating in this study personally; but this research will assist the reading researchers and test specialists improve the kind of tests that are given in schools to measure reading comprehension.
Confidentiality	All responses are confidential. The data obtained from your child will be combined with the data obtained from others when reporting the results. Your child's name will not be used in the storage or reporting of the results. The information obtained from your child will be coded and transferred to a data file that will be stored in a password protected computer file in the office of the investigator. The booklet distributed to collect information will have computer generated identification number and will be destroyed when the data are transferred to a file.
Right to Withdraw and Questions	Your child's participation in this research is completely voluntary. Your child may choose not to take part at all. If your child decides to participate in this research, he or she may stop participating at any time. If your child decides not to participate in this study or if your child stops participating at any time, your child will not be penalized or lose any benefits to which your child otherwise qualifies. If you have questions, concerns, or complaints, or if you need to report an injury related to the research, please contact the investigator, Patricia A. Alexander at: Department of Human Development, 3304 Benjamin Building, College Park, MD 20742, or call 301-405-2821.
Participant Rights	*If you have questions about your rights as a research participant or wish to*

University of Maryland College Park

	report a research-related injury, please contact: **University of Maryland College Park** **Institutional Review Board Office** **1204 Marie Mount** **College Park, Maryland, 20742** **E-mail: irb@umd.edu** **Telephone: 301-405-0678** *This research has been reviewed according to the University of Maryland, College Park IRB procedures for research involving human subjects.*
Statement of Consent	*Your signature indicates that you are at least 18 years of age; you have read this consent form or have had it read to you; your questions have been answered to your satisfaction and you voluntarily allow your child to participate in this research study. You will receive a copy of this signed consent form.* *If you agree to participate, please sign your name below.*
Name of the Child Participant	
Signature and Date [Parent or Guardian]	**PARENT OR GUARDIAN NAME** [Please Print]	
	PARENT OR GUARDIAN SIGNATURE	
	DATE	

IRB APPROVED
EXPIRES ON

JUN 2 3 2012

UNIVERSITY OF MARYLAND
COLLEGE PARK

References

Adams, M. J., & Collins, A. (1985). A schema-theoretic view of reading. In H. Singer & R. B. Ruddell (Eds.), *Theoretical models and processes of reading* (3rd ed., pp. 406-425). Newark, DE: International Reading Association.

Afflerbach, P. (2007). *Understanding and using reading assessment, K-12*. Newark, DE: International Reading Association.

Ainley, M., Hidi, S., & Berndorff, D. (2002). Interest, learning, and the psychological processes that mediate their relationship. *Journal of Educational Psychology*, *94*(3), 545-561.

Ainley, M., Hillman, K., & Hidi, S. (2002). Gender and interest processes in response to literary texts: Situational and individual interest. *Learning and Instruction*, *12*, 411-428.

Alexander, P. A. (1997a). Mapping the multidimensional nature of domain learning: The interplay of cognitive, motivational, and strategic forces. In M. L. Maehr & P. R. Pintrich (Eds.), *Advances in motivation and achievement* (Vol. 10, pp. 213-250). Greenwich, CT: JAI Press.

Alexander, P. A. (1997b). Knowledge-seeking and self-schema: A case for the motivational dimensions of exposition. *Educational Psychologist*, *32*(2), 83-94.

Alexander, P. A. (2003). The development of expertise: The journey from acclimation to proficiency. *Educational Researcher*, *32*(8), 10-14.

Alexander, P. A. (2005). The path to competence: A lifespan developmental perspective on reading. *Journal of Literacy Research*, *37*(4), 413-436.

Alexander, P. A., & Fox, E. (2004). A historical perspective on reading research and practice. In R. B. Ruddell & N. J. Unrau (Eds.), *Theoretical models and practices of reading* (5th ed., pp. 33-68). Newark, DE: International Reading Association.

Alexander, P. A., & Jetton, T. L. (1996). The role of importance and interest in the processing of text. *Educational Psychology Review*, *8*(1), 89-121.

Alexander, P. A., & Jetton, T. L. (2000). Learning from text: A multidimensional and developmental perspective. In M. L. Kamil, P. B. Mosenthal, P. D. Pearson & R. Barr (Eds.), *Handbook of reading research* (Vol. III, pp. 285-310). Mahwah, NJ: Lawrence Erlbaum Associates.

Alexander, P. A., Kulikowich, J. M., & Schulze, S. K. (1994). The influence of topic knowledge, domain knowledge, and interest on the comprehension of scientific exposition. *Learning and Individual Differences*, *5*(4), 379-397.

Alexander, P. A., Schallert, D. L., & Hare, V. C. (1991). Coming to terms: How researchers in learning and literacy talk about knowledge. *Review of Educational Research*, *61*(3), 315-343.

Alvermann, D. E., Smith, L. C., & Readence, J. E. (1985). Prior knowledge activation and the comprehension of compatible and incompatible text. *Reading Research Quarterly*, *20*(4), 420-436.

Anderson, R. C. (1977). The Notion of schemata and the educational enterprise: General discussion of the conference. In R. C. Anderson, R. J. Spiro & W.E. Montague (Eds.), *Schooling and the acquisition of knowledge* (pp. 415 -431). Mahwah, NJ: LEA.

Anderson, R. C. (1982). How to construct achievement tests to assess comprehension. *Review of Educational Research*, *42*(2), 145-170.

Anderson, R. C., & Pearson, P. D. (1984). A schema-theoretic view of basic process in reading comprehension. In P. D. Pearson (Ed.), *Handbook of reading research* (pp. 255-291). Mahwah, NJ: Lawrence Erlbaum Associates.

Anderson, R. C., Reynolds, R. E., Schallert, D. L., & Goetz, E. T. (1977). Frameworks for comprehending discourse. *American Educational Research Journal*, *14*(4), 367-381.

Artelt, C., Schiefele, U., & Schneider, W. (2001). Predictor of reading literacy. *European Journal of Psychology and Education*, *16*(3), 363-383.

Ausubel, D. P. (1963). *The psychology of meaningful verbal learning*. NY: Grune & Stratton.

Baer, J., Baldi, S., Ayotte, K., & Green, P. J. (2007). *The reading literacy of U.S. fourth-grade students in an international context: From the 2001 and 2006 Progress in International Reading Literacy Study* (PIRLS) (NCES 2008-017). National Center for Education Statistics, Institute of Education Science, U.S. Department of Education. Washington, DC.

Baldwin, R. S., Peleg-Bruckner, Z., & McClintock, A. H. (1985). Effects of topic interest and prior knowledge on reading. *Reading Research Quarterly*, *20*(4), 497-504.

Baldi, S., Jin, Y., Skerner, M., Green, P. J., & Herget, D. (2007). *Highlights from PISA 2006, Performance of U.S. 15-year old Students of Science and Mathematics Literacy in an International Context* (NCES - 2008-D16). National Center for

Education Statistics, Institute of Education Sciences, U. S. Department of Education, Washington, D.C.

Bandeira de Mello, V., Blankenship, C., & McLaughlin, D. H. (2009). *Mapping State Proficiency Standards Onto NAEP Scales: 2005-2007* (NCES 201-456). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education, Washington DC.

Beckner, W., & Cornett, J. D. (1972). *The secondary school Curriculum: Content and structure*. Scranton, PA: Intext Educational Publishers.

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

Best, R., Ozuru, Y., Floyd, R. G., & McNamara, D. S. (2006). Children's text comprehension: Effects of genre, knowledge, and text cohesion. *Proceedings of the Seventh International Conference on Learning Science*, ISBN:0-8058-6174-2, 37-42.

Bloom, B. (1956). *Taxonomy of Educational Objectives: The classification of educational goals*. NY: Longman Inc.

Bormuth, J. R. (1970). *On the Theory of Achievement Test Items*. Chicago, IL: The University of Chicago.

Bray, G. B., & Barron, S. (2003-2004). Assessing reading comprehension: the effects of text-based interest, gender, and ability. *Educational Assessment*, *9*(3&4), 107-128.

Brewer, W. F. (1980). Literary theory, rhetoric, and stylistics: Implications for psychology. In R. J. Spiro, B.C. Bruce & W. F. Brewer (Eds.), *Theoretical issues*

in reading comprehension (pp. 221-239). Hillsdale, NJ: Lawrence Erlbaum Associates.

Brown, H. A. (1914). The measurement of the efficiency of instruction in reading. *Elementary School Teacher*, *14*(10), 477-490.

Brown, H. A. (1916). *The measurement of ability to read.* Department of Public Instruction, Bureau of Research, Concord, NH.

Caldwell, J. S. (2002). *Reading Assessment*. NY: The Guilford Press.

Campbell, J. R. (2005). Single instrument, multiple measures: Considering the use of multiple item formats to assess comprehension. In S. G. Paris & S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp.347-368). Mahwah, NJ: Lawrence Erlbaum Associates.

Carpenter, R. D., & Paris, S. G. (2005). Issues of validity and reliability in early reading assessments. In S. G. Paris & S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp.279-304). Mahwah, NJ: Lawrence Erlbaum Associates.

Carroll, J. B. (1972). Defining language comprehension: Some speculations. In R. O. Freedle & J. B. Carroll (Eds.), *Language comprehension and the acquisition of knowledge* (pp. 1-29). NY: John Wiley & Sons.

Carver, R. P. (1974, July). Two dimensions of tests: Psychometric and edumetric. *American Psychologist*, 512-518.

Cervetti, G. N., Bravo, M. A., Heibert, E. H., Pearson, P. D., & Carolyn, A. (2009). Text genre and science content: Ease of reading comprehension and reader preferences. *Reading Psychology*, *30*(6), 487-511.

Cote, N., Goldman, S. R., & Saul, E. U. (1998). Students making sense of informational text: Relations between processing and representation, *Discourse Processes*, *25*(1), 1-53.

Crocker, L., & Algina, J. (2006). *Introduction to classical and modern test theory*. Belmont, CA: Wadsworth Thomas Learning.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 282-302.

Davis, F. B. (1944). Fundamental factors of comprehension of reading. *Psychometrika*, *9*(3), 185-197.

Davis, F. B. (1964). *Educational measurements and their Interpretation*. Belmont, CA: Wadsworth Publishing Company.

Davis, F. B. (1968, Summer). Research in comprehension in reading. *Reading Research Quarterly*, *3*(4), 499-545.

Davis, F. B. (1972). Psychometric research on comprehension in reading. *Reading Research Quarterly*, *7*(4), 628-678.

Dewey, J. C. (1913). *Interest and effort in education*. NY: Houghton Mifflin Company.

Dewey, J. C. (1915). *The school and society & the children and the curriculum*. Miami, FL: BN Publishing Company.

Dewey, J. C. (1935). The acquisition of facts as a measure of reading comprehension. *Elementary School Journal*, *35*(5), 346-348.

Dochy, F. J. R. C. (1992). *Assessment of Prior Knowledge as a determinant for future learning: The use of prior knowledge state tests and knowledge profiles*. Uitgeverij Lemma B.V. - Utrecht.

Dochy, F. J. R. C., & Alexander, P. A. (1995). Mapping prior knowledge: A framework for discussion among researchers. *European Journal of Psychology of Education*, *10*, 225-242.

Dochy, F. J. R. C., Serger, M., & Buehl, M. M. (1999). The relation between assessment practices and outcomes of studies: The case of research on prior knowledge. *Review of Educational Research*, *69*(2), 145-186.

Drum, P. A., Calfee, R. C., & Cook, L. K. (1981). The effects of surface structure variables on performance in reading comprehension tests, *Reading Research Quarterly*, *16*(4), 486-514.

Duke, N. K. (2005). Comprehension of what for what: Comprehension as a Nonunitary Construct. In S. G. Paris & S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp.93-104). Mahwah, NJ: Lawrence Erlbaum Associates.

Durkin, D. (1993). *Teaching them to read*, 6th ed. Boston, MA: Allyn & Bacon.

Ebel, R. L. (1972). *Essentials of educational measurement*. NJ: Prentice-Hall, Inc.

Embretson, S. E (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, *93*, 179-197.

Embretson, S. E. (1994). Applications of cognitive design systems to test development. In C. R. Reynolds (Ed.), *Cognitive assessment: A multidisciplinary perspective* (pp. 107-135). NY: Academic Press.

Embretson, S. E. (1998). A cognitive design systems approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, *3*(3), 380-396.

Embretson, S. E., & Gorin, J. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement*, *38*(4), 343-368.

Embretson, S. E., & Wetzel, C. D. (1987). Component latent trait models for paragraph comprehension tests. *Applied Psychological Measurement*, *11*(2), 175-193.

Entin, E. B., & Klare, G. R. (1985). Relationships of measures of interest, prior knowledge, and readability to comprehension of expository passages. *Advances in Reading/Language Research*, *3*, 9-38.

Farr, R., & Carey, R. F. (1986). *Reading: What can be measured?* Newark. DE: International Reading Association.

Fischer, G. H. (1973). The Linear Logistic Test model as an instrument in educational research. *Acta Psychologica*, *37*, 359-374.

Fischer, G. H., & Formann, A. K. (1982). Some applications of Logistic Latent Trait Models with Linear Constraints on the Parameters. *Applied Psychological Measurement*, *6*(4), 397-416.

Fletcher, C. R., & Chrysler, S. T. (1990). Surface forms, textbase, and situation models: Recognition memory for three types of textual information. *Discourse Processes*, 13, 175-190.

Fox, E., & Alexander, P. A. (2009). Text Comprehension: A retrospective, perspective, and prospective. In S. E. Israel & G. G. Duffy (Eds.), *Handbook of research on reading comprehension* (pp. 227-239). NY: Routledge Taylor & Francis Group.

Francis, D. J., Fletcher, J. M., Catts, H. W., & Tomblin, J. B. (2005). Dimensions affecting the assessment of reading comprehension. In S. G. Paris & S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp.369-394). Mahwah, NJ: Lawrence Erlbaum Associates.

Freeman, F. N. (1939). *Mental tests: Their history, principles, & applications*. Chicago, IL: Houghton Mifflin Company.

Gates, A. I. (1921). An experimental and statistical study of reading and reading tests. *Journal of Educational Psychology*, *XII*(6), 303-314, 378-391, 445-464.

Gates-MacGinitie Test, Level 3 through 10/12: Retrieved June 2009 from http://www.riverpub.com/productss/gmrt/details.htm

Gibson, E. J., & Levine, H. (1975). *The Psychology of reading*. Cambridge, MA: The MIT Press.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes. *American Psychologists*, *18*, 519-521.

Goodykoontz, B. (1937). The place of reading in the curriculum. In G. M. Whipple (Ed.). *The 36th Yearbook of the National Society for the Study of Education* (Part I, pp. 41-62). Bloomington, IL: Public School Publishing Company.

Gorin, J. S. (2005). Manipulating processing difficulty of reading comprehension questions: the feasibility of verbal item generation. *Journal of Educational Measurement*, *42*(4), 351-373.

Gorin, J. S., & Embretson, S. E. (2006). Item difficulty modeling of paragraph comprehension items. *Applied Psychological Measurement*, *30*(5), 394-411.

Gorsuch, R. L. (1983). *Factor Analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Gough, P. B. (1972). One second of reading. In J. F. Kavanagh & I. G. Mattingly (Eds.), *Language by ear and by the eye* (pp. 331-58). Cambridge, MA: MIT Press.

Grabe, W. (1999). Developments in reading research and their implications for computer-adaptive reading assessment. In M. Chalhoub-Deville (Ed.), *Studies in language testing* (pp.11-46). Cambridge, England: Cambridge University Press.

Graesser, A. C. (1981). *Prose comprehension beyond the word*. NY: Springer-Verlag.

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

Graesser, A. C., & Bertus, E. L (1998). The construction of causal inferences while reading expository texts on science and technology. *Scientific Studies of Reading*, *2*(3), 247-269.

Grasser, A. C, Golding, J. M., & Long, D. L. (1996). Narrative representation and comprehension. In R. Barr, M. L. Kamil, P. Mosenthal & P. D. Pearson (Eds.), *Handbook of reading research* (II, pp. 171-205). Mahwah, NJ: Lawrence Erlbaum Associates.

Graesser, A. C., McNamara, D. S., & Louwerse, M. M. (2003). What do readers need to learn in order to process coherence relations in narrative and expository text? In A. P. Sweet & C. E. Snow (Eds.), *Rethinking Reading Comprehension* (pp. 82-98). NY: The Guilford Press.

Gray, W. H. (1917). *Studies of elementary-school reading through standardized tests*. Chicago, IL: The University of Chicago Press.

Gray, W. H. (1937a). The nature and types of reading. In G. M. Whipple (Ed.). *The Thirty-sixth yearbook of the National Society for the Study of Education* (Part I, pp.23-38). Bloomington, IL: Public School Publishing Company.

Gray, W. H. (1937b). The nature and organization of basic instruction in reading. In G. M. Whipple (Ed.). *The thirty-sixth yearbook of the National Society for the Study of Education* (Part I, pp. 65-131). Bloomington, IL: Public School Publishing Company.

Guthrie, J. T., & Wigfield, A. (2005). Roles of motivation and engagement in reading comprehension assessment. In S. G. Paris & S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp.187-213). Mahwah, NJ: Lawrence Erlbaum Associates.

Guthrie, J. T., McGough, K., & Wigfield, A. (1994). Measuring reading activity: An inventory (Instructional Resource No. 4). Athens, GA: National Reading Research Center.

Haladyna, T. M. (1997). *Writing test items to evaluate higher order thinking*. Boston, MA: Allyn and Bacon.

Hambleton, R. K., & Swaminathan, H. (1985). *Item Response Theory: Principles and applications*. Boston, MA: Kluwer Nijoff Publishing.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of *Item Response Theory*. London: Sage Publications.

Hampton, S., & Resnick, L. B. (2009). *Reading and writing with understanding: Comprehension in fourth and fifth grades*. Newark, DE: International Reading Association.

Haney, W., & Scott, L. (1987). Talking with children about tests: An exploratory study of test item ambiguity. In R. O. Freedle & R. C. Duran (Eds.), *Cognitive and*

Linguistic Analyses of Performance (pp. 298-368). NY: Ablex Publishing Corporation.

Hannon, B., & Daneman, M. (2001). A new tool for measuring and understanding individual differences in the component processes of reading comprehension. *Journal of Educational Psychology*, *93*(1), 103-128.

Hidi, S. (1990). Interest and its contribution as a mental resource for learning. *Review of Educational Research*, *60*, 549-571.

Hidi, S. (2000). An interest researcher's perspective: The effects of extrinsic and intrinsic factors on motivation. In C. Sansone & J. M. Harackiewicz (Eds.) *Intrinsic and extrinsic motivation* (pp. 309-339). NY: Academic Press.

Hidi, S. (2001). Interest, reading, and learning: Theoretical and practical considerations. *Educational Psychology Review*, *13*(3), 191-209.

Hidi, S., & Anderson, V. (1992). Situational interest and its impact on reading and expository writing. In K.A. Renninger, S. Hidi & A. Krapp (Eds.), *The role of interest in learning and development* (pp. 215–238). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hidi, S., & Renninger, K. A. (2006). The four-phase model of interest development. *Educational Psychologist*, *41*(2), 111-127.

Huey, E. B. (1908/1968). *The psychology and pedagogy of reading*. Cambridge, MA: The M.I.T. Press.

Illinois Reading Assessment Framework Grades 3-8 (State Assessment Beginning Spring 2006). Retrieved February 2010, from http://www.isbe.state.il.us/assessment/pdfs/iaf_reading.pdf

Iowa Test of Basic Skills Tests (ITBS). Retrieved June 2009, from

    http:/www.education.uiowa.edu/itp/itbs/itbs_about_5-8.htm

Jetton, T. L., & Alexander, P. A. (2001). Interest assessment and the content area literacy

    environment: Challenges for research and practice. *Educational Psychology*

    Review, *13*(3), 303-318.

Johnston, P. H. (1983). *Reading comprehension assessment: A cognitive basis*. Newark,

    DE: International Reading Association.

Johnston, P. H. (1984). Prior knowledge and reading comprehension test bias. *Reading*

    Research Quarterly, *19*(2), 219-239.

Johnston, P., & Pearson, D. P. (1982). *Prior knowledge, connectivity, and the assessment*

    of reading comprehension. Technical Report No. 245, Center for the Study of

    Reading, University of Illinois at Urbana-Champaign.

Johnstone, C. J., Thurlow, M. L., Moen, R. E., Matchett, D., Hausmann, K. E., & Scullin,

    S. (2007). *What do state reading test specifications specify?* Minneapolis, MN:

    University of Minnesota, Partnership for Accessible Reading Assessment.

    Retrieved January 2009 from http://www.readingassessment.info

Judd, C. H. (1914). Reading Tests. *Elementary School Teacher*, *14*(8), 365-373.

Kane, M. T. (1992). An argument-based approach to validation. *Psychological Bulletin*,

    112, 527-535.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (4th

    ed., pp. 17-64). Westport, CT: American Council for Education/Praeger.

Keenan, J. M., Betjemann, R. S., & Olson, R. K. (2008). *Reading comprehension tests*

    vary in the skills they access: Differential dependence on decoding and oral

comprehension, ISSN: 1088-8438 print/1532-799X online. Copyright @2008 Taylor & Francis Group, LLC.

Kehoe, J. (1995). Basic item analysis for multiple-choice tests. *Practical Assessment, Research & Evaluation*, *4*(10). Retrieved November 2011, from http://PAREonline.net/getvn.asp?v=4&n=10

Kelly, F. J. (1916). The Kansas silent reading tests. *The Journal of Educational Psychology*, *VII*(2), 63-80.

Kendeou, P., & van Den Broek, P. (2007). The effects of prior knowledge and text structure on comprehension process during reading of scientific texts. *Memory & Cognition*, *35*(7), 1567-1577.

King, A. (2007). Beyond literal comprehension: A strategy to promote deep understanding of text. In D. S. McNamara (Ed.), *Reading comprehension strategies*. NY: LEA, Taylor Francis Group.

Kintsch, E. (2005). Comprehension theory as a guide for the design of thoughtful questions, *Topics in Language Disorders*, *25*(1), 51-64.

Kintsch, W. (1980). Learning from text, levels of comprehension, or: Why anyone would read a story anyway. *Poetics*, *9*, 87-98.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, *95*(2), 163-182.

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

Kintsch, W. (2004). The Construction-Integration model of text comprehension and its implications for instruction. In R. B. Ruddell and N. J. Unrau (Eds.), *Theoretical*

models and processes of reading (5th Ed., pp.1270-1328). Newark, DE: International Reading Association.

Kintsch, W., & Kintsch, E. (2005). Comprehension. In S. G. Paris & S. A Stahl (Eds.), *Children's reading comprehension and assessment* (pp. 71-92). Mahwah, NJ: Lawrence Erlbaum Associates.

Kintsch, W., & Rawson, K. A. (2005). Comprehension. In M. J. Snowling & C. Hulme (Eds.), *The science of reading: A handbook* (pp. 209-226). Malden, MA: Blackwell Publishing LTD.

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

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

Kintsch, W., & Young, S. (1980). Selective recall of decision relevant information from texts. *Memory and Cognition*, *12*(2), 112-117.

Krapp, A. (2002). Structural and dynamic aspect of interest development: Theoretical consideration from an ontogenetic perspective. *Learning and Instruction*, *12*, 383-409.

Krapp, A., Hidi, S., & Renninger, K. A. (1992). Interest, learning, and development. In K. A Renninger, S. Hidi & A. Krapp (Eds.), *The role of interest in learning development* (pp.3-25). Hillsdale, NJ: Lawrence Erlbaum Associates.

Kubinger, K. D. (2009). Applications of the Linear Logistic Test Model in psychometric research. *Educational and Psychological Measurement*, *69*, 232-244.

LaBerge, D., & Samuels, S. (1974). Toward a theory of automatic information processing in reading. *Cognitive psychology*, *6*, 293-323.

Langer, J. A. (1981). From theory to practice: A prereading plan. *Journal of Reading*, *25*, 152-56.

Langer, J. A. (1984, Summer). Examining background knowledge and text comprehension. *Reading Research Quarterly*, *XIX* (4), 468-480.

Lee, G. (2002). The influence of several factors on reliability for complex reading comprehension tests. *Journal of Educational Measurement*, *39*(2), 149-164.

Lee, J., Grigg, W., & Donahue, P. (2007). *The Nations Report Card: Reading 2007* (NCES - 2007-496). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education, Washington, D.C.

Lorch, R. F., & van den Broek, P. (1997). Understanding reading comprehension: Current and future contributions of cognitive science. *Contemporary Educational Psychology*, *22*, 213-246.

Lord, F. M. (1983). Small N justifies Rasch model. In D. J. Weiss (Ed.), *New horizons in testing: Latent Trait Test Theory and Computerized Adaptive Testing* (pp. 51-61). NY: Academic Press, Inc.

Lu, R. (2006). The Linear Logistic Test Model and its extension. Unpublished Manuscript.

Lucas, P. A. & McConkie, G. W. (1980). The definition of test items: a descriptive approach. American *Educational Research Journal*, *17*(2) 133-140.

Magliano, J. P. , Baggett, W. B. , Johnson, B. K., & Graesser, A. C. (1993). The time course of generating causal antecedent and causal consequence inferences. *Discourse Processes*, *16*, 35-53.

Magliano, J. P., Durik, A. M., & Holt, J. K. (2011, March). *Assessing the impact of topic interest on comprehension processes*. Law School Admission Council Research Report 11-02.

Magliano, J. P., Millis, K., Ozuru, Y. & McNamara, D. S. (2007). A multidimensional framework to evaluate reading assessment tools. In D. S. McNamara (Ed.), *Reading comprehension strategies: Theories, interventions, and technologies* (pp. 107-136). NY: Lawrence Erlbaum Associates.

Mandler, J. M. (1986). On the comprehension of temporal order. *Language and Cognitive Processes*, *1*(4), 309-320.

Mandler, J. M, & Johnson, N. S. (1977). Remembrance of things parsed: Story structure and recall. *Cognitive Psychology*, *9*, 111-151.

Maria, K. (1990). *Reading comprehension instruction: Issues & strategies*. Parkton, MD: York Press.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp.13-103). NY: American Council on Education/MacMillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*, 741-749.

Meyer, B. J. F., & Rice, G. E. (2002). The structure of text. In P. D. Pearson (Ed.), *Handbook of reading research* (pp. 319-350). Mahwah, NJ: Laurence Erlbaum Associates.

Millis, K. K., & Cohen, R. (1994). Spatial representations and updating situation models. *Reading Research Quarterly*, *29*(4), 369-380.

Millis, K. K., Morgan, D., & Graesser, A. C. (1990). The influence of knowledge-based inferences on the reading time of expository text. In A. C. Graesser & G. H. Bower (Eds.), *Inferences and text comprehension* (pp.197-212), NY: Academic Press.

Minsky, M. (1975). A framework for representing knowledge. In P. H. Winston (Ed.), *The psychology of computer vision* (pp. 211-277). New York: McGraw-Hill.

Mislevy, R. J. (1988). Exploiting auxiliary information about items in the estimation of Rasch item difficulty parameters. *Applied Psychological Measurement*, *12*(3), 281-296.

Mislevy, R. J. (2006). Cognitive psychology and educational assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 257-306). Westport, CT: American Council on Education/Praeger.

Mislevy, R. J., & Huang, C. (2006). Measurement models as narrative structures. Center for the Study of Evaluation Report 680. Retrieved November 2012, from http://www.cse.ucla.edu/products/reports/r680.pdf,

Mislevy, R. J., & Rahman, T. (2009). Design Pattern for assessing cause and effect reasoning in reading comprehension (PADI Technical Report 20). Menlo Park, CA: SRI International.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessment. *Measurement: Interdisciplinary Research and Perspectives*, *1*, 3-62.

Mitchell, K. (1983). *Cognitive processing determinants of item difficulty on the verbal subtests of Armed Forces Vocational Aptitude Batteries and their relationships to success in army training*. Technical Report 598. U.S. Army Research Institute for the Behavioral and Social Sciences.

Monroe, W. S. (1918). A report on the use of the Kansas Silent Reading Tests with over one hundred thousand children. *The Journal of Educational Psychology*, *IX*, 599-608.

Murphy, P. K., & Alexander, P. A. (2002). What counts? The predictive powers of subject-matter knowledge, strategic processing, and interest in domain-specific performance. *The Journal of Experimental Education*, *70*(3) 197-214.

National Assessment of Educational Progress (NAEP). Reading Framework, Retrieved February 2010 from http://www.nagb.org/publications/frameworks/reading03.pdf

National Reading Panel Report (2000). *Teaching children to read*. NIH Pub No. 00-4754, National Institute of literacy.

National Research Council Report (2001). *Knowing what students know: The science and design of educational testing*. National Research Council, Washington, DC: National Academy Press.

Norman, D. A., & Rumelhart, D. E. (1975). Memory and knowledge. In D. A. Norman & D. E. Rumelhart (Eds.), *Explorations in cognition* (pp. 3-32). San Francisco, CA: W. H. Freeman and Company.

North Carolina End-of-Grade Test of Reading Comprehension-6 (NCDPI, 2004).

Retrieved February 2010 from

http://www.ncpublicschools.org/docs/accountability/testing/eog/TISG6.pdf

Ohtsuka, K., & Brewer, W. F. (1992). Discourse organization in the comprehension of

temporal order in narrative texts. *Discourse Processes*, *15*, 317-336.

Oklahoma School Testing Program (OSTP-Reading). Test Specifications. Retrieved

February 2010, from

http://sde.state.ok.us/acctassess/pdf/OMAAP/Itemspecs/Gr4Reading.pdf

http://bixbyps.org/education/sctemp/75e30fd86089d5cd430730a14cc1e7e4/12661

02968/Grade_6_Reading_Item_Specifications.pdf

Ought, B., Mugo, M., & Tirre, W. (2010). *Predicting Difficulties of National Assessment*

of Educational Progress (NAEP) Reading items with Coh-Metrix and Readability

Indices. AIR paper presented at the American Psychological Science conference,

Boston, MA.

Ozuru, Y., Dempsey, K., & McNamara, D. S. (2009). Prior knowledge, reading skill, and

text cohesion in the comprehension of science texts. *Learning and Instruction*, *19*,

228-242.

Palincsar, A. S., Magnusson, S. J., Pesko, E. & Hamlin, M. (2005). Attending to the

nature of subject matter in text comprehension assessments. In S. G. Paris & S. A.

Stahl (Eds.), *Children's reading comprehension and assessment* (pp.257-278).

Mahwah, NJ: Lawrence Erlbaum Associates.

Pearson, P. D. (1985, April). Changing the face of reading comprehension instruction.

The Reading Teacher, 724-738.

Pearson, P. D. (2000). Reading in the twentieth century. In T. L. Good (Ed.), *American Education: Yesterday, today, tomorrow, Ninety-ninth Yearbook of the National Society for the Study of Education* (pp. 152-208). Chicago, IL: University of Chicago Press.

Pearson, P. D., & Hamm, D. N. (2005). The Assessment of Reading Comprehension: A review of practices-past, present, and future. In S. G. Paris & S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp.13-70). Mahwah, NJ: Lawrence Erlbaum Associates.

Pearson, P. D., & Johnson, D. D. (1978). *Teaching Reading Comprehension*. NY: Holt, Rinehart and Winston.

Pilonieta, P. (2011). The expository text primer: A teacher's resource guide for using expository text. *The NERA Journal*, *46*(2), 45-51.

Pintner, R. (1913, March). Oral and silent reading of 4th grade pupils. *Psychological Review*, 333-337

Pressley, M., & Block, C. C. (2002). What comprehension instruction could be? In C. C. Block & M. Pressley (Eds.), *Comprehension instruction research-based best practices*. NY: Guilford.

Program for International Reading Literacy Study (PIRLs). Retrieved February 2010, from http://timss.bc.edu/pirls2006/tech_rpt.html

RAND Reading Study Group (2002). *Reading for understanding: Toward an R&D program in reading comprehension*. Santa Monica, CA: RAND.

Readence, J. E., & Moore, D. W. (1983, January). Why question? A historical perspective on standardized reading comprehension tests. *Journal of Reading*, *26*(4), 306-313.

Recht, D. R. & Leslie, L. (1988). Effect of prior knowledge on good and poor readers' memory of text. *Journal of Educational Psychology*, *80*(1), 16-20.

Renninger, K. A. (2000). Individual interest and its implications for understanding intrinsic motivation. In C. Sansone & J. M. Harackiewicz (Eds.), *Intrinsic and extrinsic motivation: The search for optimal motivation and performance* (pp. 375-407). NY: Academic Press.

Renninger, K. A., & Hidi, S. (2002). Student interest and achievement: Developmental issues raised by a case study. In A. Wigfield & J. S. Eccles (Eds.), *Developmental achievement motivation* (pp.173-195). NY: Academic Press.

Rinck, M., Hahnel, A., Bower, G. H., & Glowalla, U. (1997). The matrices of spatial models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(3), 622-637.

Rinck, M., Williams, P., Bower, G. H., & Becker, E. S. (1996). Spatial situation models and narrative understanding: Some generalizations and extensions, *Discourse Processes*, *21*, 23-55.

Rinsland, H. D. (1935). *Manual for constructing objective tests and improving grading in elementary and high school subjects.* NY: John S. Swift Co., Inc.

Robinson, H. A., Faraone, V., Hittleman, D. R., & Unruh, E. (1990). *Reading comprehension instruction 1783-1987*. Newark, DE: International Reading Association.

Roller, C. M., & Schreiner, R. (1985). The effect of narrative and expository organizational instruction on sixth-grade children's comprehension of expository and narrative prose. *Reading Psychology: An International Quarterly*, *6*, 27-42.

Romanes, G. (1884). *Mental evolution in animals*. NY: D. Appleton and Company

Ruddell, R. B., & Unrau, N. J. (2004). Reading as a meaning-construction process: The reader, the text, and the teacher. In Ruddell, R. B. & Unrau, N. J. (Eds.), *Theoretical models and process of reading* (5th ed., pp. 1462-1521). Newark, DE: International Reading Association.

Rumelhart, D. E. (1975). Notes on a schema for stories. In D. G. Bobrow & M. A. Collins (Eds.), *Language, thought, and culture* (pp. 211-237). NY: Academic Press.

Rumelhart, D. E. (1977). Introduction to human information processing. NY: John Wiley & Sons.

Rumelhart, D. E. (1980). Schemata: Building Blocks of Cognition. In R.J. Spiro, B.C. Bruce & W.F. Brewer (Eds.), *Theoretical issues in reading comprehension* (pp.33-58). Hillsdale, NJ: Lawrence Erlbaum Associates.

Rupp, A. A., Ferne, T., & Choi, H. (2008). How assessing reading comprehension with multiple-choice questions shapes the construct: A cognitive processing perspective. *Language Testing*, *23*(4), 441-474.

Samuels, J., & Kamil, M. L. (1984). Models of the reading process. In P. D. Pearson (Ed.), *Handbook of reading research* (pp. 185-224). Mahwah, NJ: Lawrence Erlbaum Associates.

Sarroub, L. K., & Pearson, P. D. (1998). Two steps forward, three steps back: The stormy history of reading comprehension assessment. *Clearing House*, *72*, 97-105. Faculty publications: Department of Teaching, Learning and Teacher Education, University of Nebraska, pp. 87-105. Retrieved June 2008 from http//digitalcommons.unl.edu/teachlearnfacepub/39.

Schank, R. C. (1982). *Reading and Understanding: Teaching from the perspective of artificial intelligence*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals, and understanding*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Schiefele, U. (1996). Topic interest, text representation, and quality of experience. *Contemporary Educational Psychology*, *21*, 3-18.

Schiefele, U. (1999). Interest and learning from text. *Scientific Studies of Reading*, *3*(3), 257-279.

Schiefele, U. (2009). Situational and individual interest. In K. R. Wentzel & A. Wigfield (Eds.), *Handbook of motivation at school* (pp.197-221). NY: Routledge Taylor & Francis Group.

Schraw, G. (1997). Situational interest in literary text. *Contemporary Educational Psychology*, *22*, 436-456.

Schraw, G., Bruning, R., & Svoboda, C. (1995). Sources of situational interest. *Journal of Reading Behavior*, *27*(1), 1-17.

Segel, D. (1935). *Elementary school graduating examinations*. Bulletin No. 16, Washington, DC: United States Printing Office.

Sheehan, K., & Mislevy, R. J. (1990). Integrating cognitive and psychometric models to document literacy. *Journal of Educational Measurement*, *27*(3), 255-272.

Simons, H. D. (1971). Reading comprehension: The need for a new perspective. *Reading Research Quarterly*, *6*(3), 338-363.

Singer, M. (1994). Discourse inference processes. In M. A. Gernsbacher (Ed.), *Handbook of Psycholinguistics* (pp. 479-515). NY: Academic Press.

Smith, N. B. (1965). *American reading instruction*. Newark, DE: International Reading Association.

Smith, W. A. (1922). *The reading process*. NY: McMillan Company.

Snow, C. E. (2003). Assessment of reading comprehension. In A. P. Sweet & C. E. Snow (Eds.), *Rethinking reading comprehension* (pp.192-206). NY: Guilford Press.

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology in educational measurement. In R. L. Lin (Ed.), Educational Measurement, 3rd edition (pp. 263-331). NY: McMillan.

Spache, G. D. (1963). *Toward better reading*. Champaign, IL: Garrard Publishing Company.

Spearman, C. (1904a). The proof and measurement of association between two things. *American Journal of Psychology*, *15*, 72-101.

Spearman, C. (1904b). "General Intelligence" objectively determined and measured. *American Journal of Psychology*, *15*, 201-292.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of Royal Statistical Society*, *64*(4), 583-639.

Spyridakis, J. H., & Wenger, M. (1991). An empirical method of assessing topic familiarity in reading comprehension research. *British Educational Research Journal*, *17*(4), 353-359.

Starch, D. (1915). The measurement of efficiency in reading. *The Journal of Educational Psychology*, *VI*(1), 1-24.

Stevens, K. (1980). The effect of topic interest on the reading comprehension of higher ability students. *Journal of Educational Research*, *73*, 365-368.

Sweet, A. P. (2005). Assessment of Reading Comprehension: The RAND reading Study Group Vision. In S. G. Paris & S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp.3-12). Mahwah, NJ: Lawrence Erlbaum Associates.

Tarchi, C. (2010). Reading comprehension of informative texts in secondary school: A focus on direct and indirect effects of reader's prior knowledge. *Learning and Individual Differences*, *20*, 415-420.

Taylor, S. E., Frackenpohl, H., White, C. E., Nieroroda, B.W., & Browning, C. L. (1989). *EDL core vocabularies in Reading, Mathematics, Science, and Social Studies*. Orlando, FL: Steck-Vaughn Company.

Therriault, D. J., & Raney, G. E. (2007). Processing and representing temporal information in narrative text. *Discourse Processes*, *43*(2), 173-200.

Thorndike, E. L. (1917a). *Thorndike's scale Alpha for measuring the understanding of sentences.* Teachers College, Columbia University.

Thorndike, E. L. (1917b). Reading as reasoning: A study of mistakes in paragraph reading. *Journal of Educational Psychology*, *8*(6), 323-332.

Thorndyke, P. W. (1977). Cognitive structures in comprehension and memory of

   narrative discourse. *Cognitive Psychology*, *9*(1), 77-110.

Tian, S. (2006). Passage dependency of reading comprehension items in the GEPT and

   TOEFL. *The Reading Matrix*, *6*(1), 66-84.

Tobias, S. (1994). Interest, prior knowledge, and learning. *Review of Educational*

   Research, *64*(1), 37-53.

Traxler, A. E. (1970). Values and limitations of standardized reading tests. In R. Farr

   (Ed.), *Measurement and evaluation of reading* (pp. 220-229). NY: Harcourt,

   Brace, & World.

Tri-State New England Test Specifications for Reading (TSNE, 2003). Retrieved

   February 2010, from:

   http://www.ed.state.nh.us/Education/doe/organization/curriculum/NECAP/ReadingTest.p

   df

Tuinman, J. J. (1972-1973, Summer). Inspection of passages as a function of passage-

   dependency of the test items. *Journal of Reading Behavior*, *5*(3), 186-191.

Tun, P. A. (1989). Age differences in processing expository and narrative text. *Journal of*

   Gerontology: Psychological Sciences, *44*, 9–15.

Tyler, R. W. (1983, Summer). Educational Assessment, can we have one. *Educational*

   Measurement: Issues and Practice, 15-23.

Valencia, S. W., & Pearson, P. D. (1987, April). Reading assessment: Time for a change.

   The Reading Teacher, 726-733.

Valencia, S. W., Pearson, P. D., Peters, C. W., & Wixson, K. K. (1989). Theory and

   practice in statewide reading assessment: Closing the gap. *Educational*

   Leadership, *47*, 57-63.

Valencia, S. W., Stallman, A. C., Commeyras, M., Pearson, P. D., & Hartman, D. K. (1991). Four measures of topical knowledge: A study of construct validity. *Reading Research Quarterly*, *26*(3), 204-233.

van den Broek, P. (1990). Causal inferences and comprehension of narrative texts. *The Psychology of Learning and Motivation*, *25*, 174-194.

van den Broek, P., Kendeou, P., Kremer, K., Lynch, J., Butler, J., Lorch, E. P. (2005). Assessment of comprehension abilities in young children. In S. G. Paris & S. A. Stahl (Eds.), *Children's Reading Comprehension and Assessment* (pp.107-130). Mahwah, NJ: Lawrence Erlbaum Associates.

van den Broek, P., Young, M., & Tzeng, Y. (1999). The Landscape model of reading: Inferences and the online construction of a memory representation. In H. van Oostendrop & S. R. Goldman (Eds.), *The Construction of mental representation during reading* (pp. 71-98). Mahwah, NJ: Lawrence Erlbaum Associates.

van Dijk, T. A., & Kintsch, W. (1983). Observation on the status of experimental research on discourse comprehension. *Strategies of Discourse Comprehension* (pp.21-60). NY: Academic Press.

Venezky, R. L. (1977, May). Research on reading processes: A historical perspective. *American Psychologist*, 339-345.

Warren, W. H., Nicholas, D. W., & Trabasso, T. (1979). Event chains and inferences in understanding narratives. In R. O. Freedle (Ed.), *New directions in discourse processing* (pp. 23-37). Norwood, NJ: Ablex Publication Corporation.

Washington Assessment of Student Learning (WASL, 2006). Test specification for grade 3-high school reading WASL. Retrieved February 2010, from http://www.k12.wa.us/assessment/wasl/Readingtestspecs/TestandItemSpecsv2006

Weaver, C. A., & Kintsch, W. (1996). Expository text. In R. Barr, M. L. Kamil, P. Mosenthal & P. D. Pearson (Eds.), *Handbook of reading research* (Vol. II, pp. 230-245), Mahwah, NJ: Lawrence Erlbaum Associates.

Webb, J. R. (1856). *Webb's Normal Reader, No. 4: To teach correct reading, to improve and expand the mind, and to purify and elevate the character*. Sheldon, Blakeman & Co.

Webb, J. R. (1858). *Webb's Normal Reader, No. 3: A guide to correct articulation with examples for sight practices, and incentives to mental cultivation*. NY: Sheldon, Blackman & Co.

Wharton-McDonald, R., & Swiger, S. (2009). Developing higher order comprehension in the middle grades: A retrospective, perspective, and prospective. In S. E. Israel & G. G. Duffy (Eds.), *Handbook of Research on Reading Comprehension* (pp. 510-530). NY: Routledge Taylor & Francis Group.

Whipple, G. M. (1925). *The Twenty-fourth Yearbook of the National Society for the Study of Education*, Part I. IL: Public school publishing company.

Wigfield, A. (1997). Reading Motivation: A Domain-Specific Approach to Motivation. *Educational Psychologist*, *32*(2), 59-68.

Willis, A. I. (2008). *Reading comprehension research and testing in the U.S.: Undercurrents of race, class, and power in the struggle for meaning*. NY: Lawrence Erlbaum Associates.

Wilson P. T., & Anderson, R. C. (1986). What they don't know hurt them: The role of prior knowledge in comprehension. In J. Orasanu (Ed.), *Reading comprehension: From research to practice* (pp. 31-48). Hillsdale, NJ: Lawrence Erlbaum Associates.

Wolfe, M. B. W. (2005). Memory for narrative and expository text: Independent influences of semantic associations and text organization. *Journal of Experimental Psychology: Learning, Memory, and Cognitio*n, *31*, 359–364.

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

Wolfe, M. B. W., & Woodwyk, J. M. (2010). Processing and memory of information presented in narrative or expository texts. *British Journal of Educational Psychology*, *80*, 341–362.

Yen, W., & Fitzpatrick, A. R. (2006). Item Response Theory. In R. Brenan (ed.), *Educational Measurement* (4th Ed., pp. 111-153). Westport, CT: American Council on Education/Praeger.

Zakaluk, B. L., Samuels, S. J., & Taylor, B. M. (1986). A simple technique for estimating prior knowledge: Word association. *Journal of Reading*, *30*(1), 56-60.

Zwaan, R. A., Magliano, J. P., & Graesser, A. C. (1995). Dimensions of situation model construction in narrative comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(2), 386-397.

Zwaan, R. A. (1994). Effect of genre expectations of text comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 920–933.

Zwaan, R. A., Langston, M. C., & Graesser, A. C. (1995). The construction of situation models in narrative comprehension: An event-indexing model. *Psychological Science*, *6*(5), 292-297.

Zwaan, R. A., & Oostendorp, H. V. (1993). Do readers construct spatial representations in naturalistic story comprehension. *Discourse Processes*, *16*, 125-143.

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, *123*(2), 162-185.