ABSTRACT

Title of Document: THE ROLE OF READING COMPREHENSION IN LARGE-SCALE SUBJECT-MATTER ASSESSMENTS

Ting Zhang, Ph.D, 2013

Directed By: Professor, Judith Torney-Purta, Department of Human Development and Quantitative Methodology

Professor Emeritus, Robert J. Mislevy, Department of Human Development and Quantitative Methodology

This study was designed with the overall goal of understanding how difficulties in reading comprehension are associated with early adolescents' performance in large-scale assessments in subject domains including science and civic-related social studies. The current study extended previous research by taking a cognition-centered approach based on the Evidence-Centered Design (ECD) framework and by using U.S. data from four large-scale subject-matter assessments: the IEA TIMSS Science Study of 1999, IEA CIVED Civic Education Study of 1999, and the 1970s IEA Six Subject surveys in Science, and in Civic Education.

Using multiple-choice items from the TIMSS science and CIVED tests, the study identified a list of linguistic features that contribute to item difficulty of subject-matter

assessments through the Coh-Metrix software, human rating, and multiple regression analysis. These linguistic features include word length, word frequency, word abstractness, intentional verbs, negative expressions, and logical connectives. They pertain to different levels of Kintsch's reading comprehension model: surface level, textbase level, and situation model.

Integrating this item-level information into multiple regression analysis and Multidimensional IRT modeling, the study provided feasible methods (1) to estimate reading demand of test items in each subject-matter assessment, and (2) to partial out variance related to high level of reading demand of some test items and independent of the domain proficiencies that the subject-matter assessment was intended to measure. Overall, results suggested that reading demands of all test items in TIMSS Science and CIVED tests were within the reading capabilities of almost all of the students, and these two tests were not saturated with high reading demand.

In addition, multiple regression results from the earlier Six Subject Surveys showed that an independent measure of students' general vocabulary was highly correlated with their achievement in the domains of science and civic-related social studies. On average, boys outperformed girls in both subject domains, and students from home with ample literacy resources outperformed students from homes of few literacy resources. In the science assessment, interactions were found between gender and word knowledge, home literacy resources and word knowledge, meaning the correlation between vocabulary and science performances differed by gender and home background.

THE ROLE OF READING COMPREHENSION IN LARGE-SCALE SUBJECT-MATTER ASSESSMENTS

By

Ting Zhang

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 Judith Torney-Purta, Chair
Professor Robert J. Mislevy
Assistant Professor Meredith Rowe
Associate Professor Laura Stapleton
Associate Professor Jennifer Turner (Dean’s Representative)

© Copyright by
Ting Zhang
2013

Dedication

To my husband, Hua Song, this accomplishment would not have been possible without you standing behind me. Through my ups and downs you have been there to listen to me and encourage me to finish. I will forever be grateful for the sacrifices that you made in an effort to help me achieve this goal. I am also grateful to my parents, for being always supportive even when they disagreed with my decisions.

Acknowledgements

Note: This research was supported by a grant from the American Educational Research Association which receives funds for its "AERA Grants Program" from the National Science Foundation under Grant #DRL-0941014. Opinions reflect those of the author and do not necessarily reflect those of the granting agencies.

The pursuit of this degree would not have been possible without the support and encouragement of my mentors, committee members, faculty members, family, friends, and colleagues. Without all of them, this dream would not have come to fruition. I wish to express my appreciation to each of them, although these words seem quite incapable of expressing the depth of my gratitude.

First and foremost, I own my deepest gratitude to my advisor Judith Torney-Purta. Judith provided me invaluable guidance, insightful advice, consistent encouragement, exposure to diverse people, situations, and ideas. She transmitted her commitment, developed over more than forty years, to finding methods to establish the validity of international large scale assessments for secondary analysis. She always made herself available for consultations even on weekends and holidays during every stage of my graduate study. She spent too many hours to count supervising this study from conceiving, executing, revising, to editing the numerous drafts of this thesis. I cannot thank her enough for having the faith in my ability, being patient and flexible with me, and helping me to spread my wings through this process. I am truly blessed to have her as my mentor, and my deep appreciation to her is far beyond verbal expression.

I would like to thank my co-advisor Robert Mislevy, who sparked my interest, and guided me into this amazing interdisciplinary field of assessment and cognition. Bob provided me ongoing advice on theoretical and methodical issues, exciting research opportunities, and financial support since my first day at the University of Maryland. I feel honored to have been one of his advisees.

I also want to thank the members of my dissertation committee, who have been patient, flexible, and supportive as they guided me through this process. Particularly, I wish to thank for Dr. Meredith Rowe for sharing her extensive knowledge, and providing consistent encouragement, thoughtful and useful feedback. I appreciate the help of Dr. Laura Stapleton, who provided me statistical expertise support, and was willing to walk through some of my analytic plan. I want to thank Dr. Jennifer Turner, for agreeing to come on board for the journey, and providing invaluable input.

I would also like to express my genuine thanks to my former ETS internship mentors Dr. Xueli Xu and Dr. Matthias von Davier, who advised and assisted me with technical issues pertaining to the General Diagnostic Model and to the use of *mdltm* software.

I am indebted to the AERA Grant Program for funding this research. Especially, I would like to thank my AERA Grants Program mentor Dr. Jeremy Kilpatrick for his helpful advices and feedback. My thanks also extend to the AERA Grant Program officers, Dr. George Wimberly and Christina Farhart, for their help and support.

I would like to express my sincere thanks to Dr. Jack Schwille, who was very kind and generous to me when I visited the Michigan State University. He patiently answered my questions, and walked me through the documents of IEA Six Subject Surveys in Civic Education and Science.

I also own thanks to Dr. Patricia Alexander, and Dr. Ann Battle. Dr. Alexander introduced me to learning and cognition. I had fascinating experiences in her learning and educational psychology classes. She challenged and guided me to think deeply and critically about the conceptual issues behind statistics. Dr. Battle was my mentor when I was teaching Research Methods. She has consistently provided me moral and academic support at every step of my way, and shaped my ways of thinking about the relation between learning and teaching.

I also would like to thank Dr. Donald J Bolger, who shared his expertise in reading comprehension, and provided me valuable comments and suggestions.

Additional thanks to my former mentor and other professors in my master’s program at Texas Tech University: Dr. William Lan, Dr. Mary Tallent Runnels, Dr. Lee Duemer. They have believed in me, provided me consistent encouragement, and moral support at each step of my way.

I would also like to acknowledge the staff of the HDQM department, Eileen Kramer, Charm Mudd, Jo Peng, Cornelia Snowden, and Tony Ananeta for all their patient help. They made my graduate school experience very positive.

Finally I owe my heartfelt gratitude to my friends and colleagues: Meryl Barofsky, Alaina Brenick, Nicole Denmark, Melissa Duchene, Emily Grossnickle, Amy Ho, Alex List, Amanda Mason-Singh, Melissa Menzer, Shannon Michael, Danette Morrison, Lauren Musu-Gillette, Carlo Panlilio, Wendy Richardson, Shannon Russell, and Alexis Williams. They lent their expertise to me in a variety of areas, and their friendship and support have made it possible for me to achieve this difficult goal. Particularly, Emily and Wendy gave me suggestions about how to construct coding schemes for the dissertation, and how to train raters. They also coded test items for me. Lauren provided editorial assistance when I wrote my literature review. I am also deeply grateful for the loving support and encouragement from Meryl Barofsky, Amy Ho, Danette Morrison, and Shannon Russell.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Text-based large-scale educational assessments have been used to measure, document, and compare students' academic achievement, learning process, attitudes and beliefs. Results from the assessments provide a base from which policy makers, curriculum specialists, and researchers can better understand the performance of their educational systems. Educational assessments differ in the extent to which they are language dependent. Although a test may be designed to assess content proficiencies other than language or literacy, the measures of subject-matter achievement can be attenuated by complexity of the language usage in the assessment items. For instance, evidence from research on mathematical problems solving suggests that factors other than mathematical skill contribute to successful problem solving for students age 7 to 14 years (Abedi, Lord, Hofstetter, & Baker, 2000; Cummins, Kintsch, Reusser, &Weimer, 1988; Shaftel, Belton-Kocher, Glasnapp, & Poggio, 2006; Vilenius-Tuohimaa, Aunola, & Nurmi, 2008). One possible explanation is that much of the difficulty that students experience with verbal format problems can be attributed to difficulty in comprehending assessment tasks that contain abstract or ambiguous vocabulary and or complex sentence structure (Cummins, et. al., 1988). In fact, this problem is prevalent in a variety of subject areas, particularly school subjects such as science, mathematics, and social studies (Alexander & Kulikowich, 1991; RAND Reading Study Group, 2002; Wiley & Voss, 1999).

Existing but limited research indicates that misalignments of reading demands on assessment tasks (e.g., reading difficulty of test items) and the level of students' reading proficiency can adversely affect students' scores in a subject matter test.(e.g., Abedi et al., 2000; Alexander & Kulikowich, 1991; Cummins, Kintsch, Reusser, &Weimer, 1988; Garcia, 1991; Gorin & Embretson, 2006; O'Reilly & McNamara, 2007; Wiley & Voss, 1999). In addition, research on English language learners taking large-scale assessments suggests that abstract or complex language usage in a subject-matter assessment can lead to the underestimation of a student's content knowledge if the student is not proficient in the language of the assessment (e.g., Abedi & Lord, 2001; Abedi, Lord, Hofstetter, & Baker , 2000; Abedi et al., 2012; Haladyna, 2004). As a result, the interpretations of scores may not accurately reflect the psychological constructs of content knowledge and skills that the tests are intended to measure, hence construct validity of the assessment is undermined. Thus, it is critical when designing assessments to consider validity issues pertaining to reading demand and what affordances could be designed to enhance all students' comprehension.

Traditionally, readability formulas (e.g., Dale & Chall, 1948; Flesch, 1951; Fry, 1968; Gunning, 1968; Spache, 1953) have been used to assist in matching reading demands on assessment tasks with a reader's language and reading abilities. These formulas rely on a limited number of factors such as word length and sentence length. Validity and utility of these formulas have been questioned since the 1980s (Gordon, 1980; Rygiel, 1982; Templeton, Cain, & Miller, 1981; Wheeler & Sherman, 1983; Oakland & Lane, 2004). In addition, cognitive psychologists such as Kintsch (1998, 2005) argue that many analyses that employ these readability formulas do not reflect current understanding of comprehension processes in cognitive psychology. In addition, very little evidence from cognitive psychology supports the widespread practices for assessing readability.

Recent research on sources of reading complexity for students has incorporated a cognition-centered approach that studies features of texts and tasks that place varying demands on comprehension (e.g., Abedi et al. 2012; Embretson & Wetzel, 1987; Gorin & Embretson, 2006; Kirsch & Mosenthal 1990; Ozura, Rowe, O'Reilly, & McNamara, 2008). These studies utilized available theoretical frameworks drawn on advances in cognitive theories to model the nature and characteristics of comprehension processes when students read assessment tasks. This new cognitive approach provides a promising framework to model reading comprehension processes and gauge reading demands posed on assessment tasks. However, such an approach has mainly been used in reading assessments and math assessments. The link between this approach and subject-matter assessments such as science assessments and social studies assessments is still lacking.

This study is designed with the overall goal of understanding how students process assessment item questions and options in specific subject domains including science and civic-related social studies. Specifically, what features of assessment tasks are associated with students' comprehension and their test performance in subject matter domains? The aim is to suggest how educational assessments may be improved within the context of test validity as it is currently conceptualized (Kane, 1992; Messick, 1989; Mislevy, 2009). In the following sections, I will provide the background about current understanding of test validity for educational assessment and reading comprehension in subject-matter assessment. Then I will describe reading comprehension and factors that affect the difficulty of reading comprehension for individuals when they process assessment items in subject domain areas. Finally, I will discuss the relationship between task (test item) features that have been shown to be associated with reading comprehension and students' performance on subject-matter assessments (i.e., science assessments and civic-related assessments).

1.1 Test Validity for Educational Assessment

Educational assessment has long been used to document students' knowledge, skills, attitudes and beliefs. In an assessment, students' knowledge, skills, attitudes and beliefs are perceived as latent psychological attributes that influence what they say, do, or make in home, work, school, or social settings. Such psychological attributes are called constructs in measurement theory (Crocker & Algina, 1984). According to Crocker and Algina (1984), constructs are "products of the informed scientific imagination of social scientists who attempt to develop theories for explaining human behavior" (1984, p.4). By their nature constructs are hypothetical concepts, therefore, their existence can never be absolutely confirmed. Psychologists can make inferences about the degree to which a psychological construct characterizes an individual from observations of his or her behavior in a given context. Therefore, despite the fact that assessments are used in various domains and for different purposes, what they all have in common is the desire to reason from particular things students say, do, or make in a given context, to inferences about what they know or can do more broadly (Mislevy, Steinberg, & Almond, 2003). Assessment tasks (e.g., test items) are usually used as situational stimuli to evoke students' performance upon which a subsequent inference about what students know or are able to do can be drawn. Because of the importance of the assessment tasks for individuals and educational systems, test developers have an obligation to ensure that their tasks provide a valid measure of the intended construct with as little bias as possible.

In the third edition of *Educational Measurement*, Messick (1989) defined test validity as: “an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment” (p.13). In general, the perspective of test validity encourages researchers to collect and thoroughly evaluate all the evidence for and against the proposed interpretation of test scores (i.e. students’ performance in an assessment) in order to draw adequate and appropriate inferences with respect to the construct of interest (e.g., domain specific proficiencies).

Validation from this perspective can be strengthened by including testing of alternative explanations with respect to the validity of an interpretation (Cronbach, 1988; Kane, 1992). Kane (1992, 2006) suggests that by eliminating and reducing the plausibility of alternative explanations, we can increase our confidence that a desired interpretation regarding the specified construct is valid. For example, a math test that consists of word problems may make a large demand on reading comprehension. However if the reading demand on the math problems exceeds students’ capability for reading comprehension, their low comprehension impedes their performance on the test. Hence when students respond incorrectly to a math word problem, it is unknown whether their incorrect responses are due to lack of domain knowledge or inability to successfully comprehend test items and their choice-options (Homan, Hewitt, & Linder, 1994; Haladyna, 2004). If the latter explanation is true, then the reading comprehension becomes an alternative explanation for students’ poor performance in the math test. As a result, interpretations of test scores may not accurately reflect the construct (in this case, math knowledge) that the problem is intended to measure. In this case, eliminating the alternative explanation for students' poor test performance by reducing the excessive reading demand on the item-level can improve validity of the test.

In validity theory, a plausible alternative explanation for an interpretation of test scores is often referred as a threat to validity (Crooks, Kane, & Cohen, 1996). A number of threats can jeopardize the validity of educational assessment. One major threat is construct-irrelevant variance (Messick, 1989), a type of systematic error that attenuates the validity of interpretations and used of test scores. Reading comprehension is usually considered as a potential source of construct irrelevant variance in subject-matter assessments (Messick, 1984; Haladyna & Downing, 2004; Oakland & Lane 2004). Particularly, the National Research Council notes: "if a student is not proficient in the language of the test, her performance is likely to be affected by construct-irrelevant variance—that is, her test score is likely to underestimate her knowledge of the subject matter being tested" (Heubert & Hauser, 1999, p. 225). Therefore, Mislevy (1994, 2006, 2009) suggests that in order to ensure validity of the educational assessments, we need to develop a better understanding of students' cognitive capacities including reading comprehension in specific subject-matter areas by combining developments in cognitive psychology with advances in measurement theory. In the next section, I will review theories and research related to reading comprehension.

1.2 Reading Comprehension

Current cognitive theories from the perspective of constructivism conceptualize reading comprehension as iterative and reciprocal multicomponent cognitive processes that are constructed in the reader's mind. From this point of view, to successfully understand a text a reader must access the meaning of words, tie the meaning of words to a coherent sentence level representation, relate sentences to one another to build local coherence, and relate larger pieces of text to build a global coherent mental representation. In the end, the reader needs to integrate the representation with his or her prior knowledge in order to achieve deep level understanding. (Best, Ozuru, Floyd & McNamara, 2006; Duke & Carlisle, 2011; Kintch, 1998; RAND Reading Study Group, 2002).

Among reading comprehension theories, Kintsch's (1998) the Construction-Integration (CI) model has been considered to be a well-formulated one that has built a foundation for the development of other more sophisticated comprehension theories (McNamara & Magliano, 2009; Verhoeven & Perfetti, 2008). The CI model assumes comprehension is constructed and built on integrated mental models (i.e., schemas) derived from the text. To do this, the reader must activate concepts expressed in the text and form connections between activated concepts and relevant prior knowledge of words, concepts, ideas and personal experience. The comprehension processes are regulated by mental models (schemas) and constrained by contexts. In other words, the networks of concepts that are compatible with the context enhance the activation of one another, while concepts that are not compatible with the context lose activation (Kintsch & van Dijk, 1978; Stahl & Hiebert, 2006).

1.2.1 Factors that Influence Reading Comprehension

Many factors interactively contribute to successful comprehension of written assessment tasks. Some factors depend on the readers, such as reading skills and strategies, subject-matter knowledge, motivation, and interest. Some are inherent in the text, such as the difficulty of vocabulary, sentence length, text cohesion, and text genre. Others are specific to the assessment tasks and context, such as item format. These factors interact with each other during reading comprehension processes. As a result, individuals' reading comprehension varies as a function of their reading capacity and the nature of the source text and context.

Existing readability research suggests that at the item level, vocabulary difficulty and syntactic complexity are the most robust predictors of text readability (Klare, 1984; Haladyna, 2004; Abedi et al., 2012). These variables interact with reader-related factors such as working memory, domain knowledge, and reading skills during reading comprehension processes. If text features and item characteristics are not matched to a reader's knowledge and language ability level when the reader engages in a reading activity, the text may be too difficult for optimal comprehension to occur (RAND Reading Study Group, 2002). For instance, previous research (e.g., Abedi & Lord, 2001; O'Reilly & McNamara, 2007) found that students' performance on a subject-matter test can be attenuated by their deficiencies in reading skill even though they possess the required level of domain knowledge. Abedi and Lord (2001) found that simplifying linguistic complexity of a mathematics test without contaminating the construct improved the performance of students in low-level and average math classes, as well as English Language Learners and low SES students. This result implies that when we design a subject-matter assessment, it is necessary for us to ensure that the difficulty of language in which the test is written (language demand) is in alignment to students' reading abilities (associated with their grade level). In order to do so, we should pay attention to

(1) features of the text (e.g., the difficulty of vocabulary, sentence length, text cohesion, text genre) as well as item characteristics (e.g., items format), and (2) students' background. Features of text and items can either increase or decrease reading difficulty (reading demand) in interaction with the knowledge and abilities of the reader.

1.2.2 Group Differences in Reading

Previous studies show reading comprehension can vary with students' grade levels, gender, and language background (i.e., native speakers vs. English as second language learners).

Grade. Previous research suggests that lower level language skills such as word recognition, fluency, and oral language abilities reliably predict reading comprehension in the early elementary years. In the later elementary years, word recognition and fluency become less associated with reading comprehension. Higher level language skills such as semantic skill, and the use of comprehension strategies are more important determinants of reading comprehension by 5th or 6th grade (Duke & Carlisle, 2011).

Gender. Results of previous meta-analysis research revealed a clear pattern of gender differences in reading. That is, on average girls tend to have higher reading skill than boys across grades (e.g., Hyde & Linn, 1988; Lietz, 2006; Ryan & DeMark, 2002). A large number of national and international assessments in reading, including NAEP, Program for International Student Assessment (PISA), and the International Association for the Evaluation of Educational Achievement (PIRLS), confirm the reading advantage of girls over boys across grades.

Home Literacy Resources. Numerous studies found children' exposure to literacy and their literacy experiences at home are related to their cognitive development including word knowledge, reading competence, and conceptual knowledge. In general, studies suggest that there are reciprocal relations among children's home literacy resources, word knowledge, reading comprehension, and conceptual knowledge, and all contribute to their development of academic competence (e.g., Leseman & de Jong ,1998; Sénéchal & LeFevre, 2002; Stanovich,1986).

Language Background. Previous studies (e.g., Abedi, 2009; Abedi & Gándara, 2006; Abedi, Lord, Hofstetter, & Baker, 2000; Solano-Flores & Trumbull, 2003) have examined the influence of language complexity on English language learners' (ELL) performance in large-scale subject-matter assessments (e.g., NAEP math and science assessments). Overall, their studies suggested students' performance in subject-matter tests are confounded by their language background and English proficiencies. English language learners (ELLs) generally perform lower than non-ELL students on reading, science, and math. In addition, findings from these studies show that item-level text features influence students' performance in subject-matter assessments in different ways. High reading demand on subject-matter assessments (i.e. math and science assessments) has a higher impact on ELL students than on non-ELL students. The gap between the performance of ELL and non-ELL students grows as the level of reading demand of the test items increases in the areas of science and mathematics.

One way to minimize the impact of reading demand on students' test performance is reducing the level of unnecessary linguistic complexity of the assessment (e.g., Abedi, 2009; Abedi et al., 1997; Abedi, Lord, Hofstetter, & Baker, 2000). In other words, test designers may improve validity of subject-matter assessment by lessening the linguistic complexity unrelated to the construct being assessed in the content-based areas. Based on previous research and judgments of experts, Abedi and his colleagues identified several linguistic features associated with difficulty of text comprehension (reading demand): unfamiliar (or less commonly used) vocabulary, complex grammatical structures, and styles of discourse that include extra material, abstractions and passive voice (for more detailed descriptions of these features, see Abedi, 2009; Abedi et al., 1997). In subject-matter assessments, these features are likely to be construct-irrelevant because they increase the likelihood of misinterpretation and add cognitive load to readers; therefore such features are likely to interfere with the measure of a construct.

1.3 Reading Difficulty Modeling

Traditionally, researchers examine the impact of reading demand on test items through estimating the contribution of text and task specific features (which are associated with reading comprehension) on item difficulty (which usually refers to the proportion of students who provided a fully correct response to a test item). Reading assessment is a domain that has been studied extensively (e.g., Embretson & Wetzel, 1987; Gorin & Embretson, 2006; Kintsch & Kintsch, 2005; Ozura, Rowe, O'Reilly, & McNamara, 2008; Sheehan, 1997). Through statistical methods including multiple regression, factor analysis, and differential item functioning analysis, these studies overall have shown that some features of text and of items in assessments of reading (e.g. type of questions, sentence length, vocabulary difficulty, etc.) accounted for a significant amount of the variance in item difficulty. Results of these studies help to identify which types of text and item features are associated with students' comprehension processing of written test items. This information, in conjunction with cognitive theories of text processing and comprehension, affords researchers and educators insight into the nature of constructs that tend to be tapped by assessment items (Ozura, et al., 2008). However, limited studies have focused on large-scale assessments other than those intended to measure literacy or reading.

Recent research on reading assessment has incorporated a cognition-centered approach on text processing and comprehension (Embretson, 1998; Mislevy, 1994, 1995, 1999; Mislevy, Steinberg & Almond, 2003). This approach starts with defining what the test intends to measure (i.e., the construct) with a cognitive model that specifies students' representations of a domain in terms of requisite knowledge, skills, and abilities (KSA) (Mislevy, Steinberg & Almond, 2003). The approach then decomposes a task (i.e., taking a test) into a processing model (Embretson, 1998) and then examines the contribution of particular text processes and task features (including text specific features) to item responses. The results of this analysis potentially help to identify which types of task features (e.g., item format, proposition density, and sentence length) contribute most to the difficulty level of the tasks. Several studies based on this approach were conducted by Embretson and her colleagues on reading comprehension assessments (Embretson & Wetzel, 1987; Gorin & Embretson, 2006; Gorin, 2005; Ozura, et al., 2008). Overall, these studies have collectively shown that this type of theory-based analysis of test items provides useful information about the variability in test takers' reading comprehension as measured by these tests. Based on results from these studies a subset of specific reading comprehension item features have been identified as potential contributors to reading demand.

This new cognitive-psychometric approach provides a promising framework to modeling reading comprehension processes. However, such an approach has so far only been used in reading comprehension assessment. The link between this approach and subject-matter assessments in areas of science and social studies is still lacking. Testing theories suggest that reading comprehension can be a potential threat to the interpretation of test scores from subject-matter assessments. It is critical to look at this issue by making using of the advanced approach developed by cognitive psychologists and methodologists in the area of reading assessment.

1.4 Proposed Research

The field is generally lacking published research on international large-scale subject-matter assessments such as the International Association for the Evaluation of Educational Achievement (IEA) the Trends in International Mathematics and Science Study (TIMSS) and Civic Education (CIVED) Study examining the influence of reading comprehension on performance. This research contributes to the literature by employing modern cognitive models and psychometric methods to enhance the current state of knowledge regarding the role of reading comprehension in large-scale subject-matter assessments. My proposed study will integrate the cognitive literature on reading comprehension into a processing model to test validity in subject-matter assessments. The quality of items on the subject-matter test can be assessed based on the relation of the item difficulty to task features (Embertson & Wetzel, 1987; Gorin & Embretson, 2006).

A better understanding of the role of comprehension in subject-matter assessment as well as constructs measured by the assessments will lead to a more accurate and fine-grained interpretation of test scores, thus increase the test validity. It can also identify ways in which the abilities of groups disadvantaged in reading (e.g. boys, ELLs) can be better estimated.

The focus of this research is to examine large-scale subject-matter assessment items within the theoretical frameworks of reading comprehension theories. Utilizing a cognition-centered approach to text processing and comprehension (Embretson, 1998; Mislevy, 1994, 1995, 1999), I aim to measure, and partial out variance that is associated with reading comprehension in science, and civic-related social studies assessments. I pose the following questions:

1. To what extent do task features facilitate or hinder students' performance in subject-matter assessments including science and civic-related social studies?
    a. What task features pertaining to reading comprehension can be identified in each subject-matter assessment?
    b. At the item level, to what extent are these task features related to the difficulty level of test items in each subject-matter assessment?
2. To what degree do the average estimated scores of the domain-specific proficiency change after taking into account the reading demand of test items?
3. Does the relation between the reading demand and students' domain proficiency vary by gender and language status in each subject-matter assessment?

For the two additional subject-matter assessments that in addition measured students' general word knowledge, further research questions are:

4. Is there a relation between the measure of general word knowledge and students' achievement in the subject-matter assessment?
5. Does the relation between the students' general word knowledge and achievement vary by demographic factors in each subject-matter assessment?

1.5 The Research Approach

The study design follows a cognition-centered approach based on the Evidence-Centered Design (ECD) framework (Mislevy, Steinberg, & Almond, 2003). An introduction and more detailed descriptions about the ECD framework are presented in Chapter 3. Susan Embretson and her colleagues have employed a similar approach to investigate large-scale standardized reading tests (e.g., Embretson & Wetzel, 1987; Gorin & Embretson, 2006). Because of limitations to available data, this research is not designed to provide measures of person factors or even of a comprehension construct per se. Instead, I focus on features of texts and tasks that place varying demands on comprehension.

The first step of the cognitive-centered approach is to identify reading-related task features in the assessment domain based on the theoretical framework of reading comprehension theory (e.g., Kintsch, 1998). I had two reading experts and used a computational tool called Coh-Metrix (Graesser, et al., 2004) to identify task features that can account for comprehension demand of test items in four large-scale subject-matter assessments. For the purposes of this study, I focus on multiple-choice test items which were written in English and administrated to the U.S. students.

At a next step, I examine the degree to which selected task features contribute to the difficulty of test items through regression analyses. Analyses at this step inform me about types of features associated with the difficulty of items, and the amount of variance in test items explained by these task features. At the third step, a multidimensional IRT model (von Davier, 2005) is fit to the data to model subject-matter proficiencies and subreading demand. This approach allows me to better estimate domain proficiencies (in science and social studies) while taking the reading demand of test items into account.

Finally, students' background factors such as gender and language background are taken into account for subsequent analyses because previous studies indicate that these background factors were associated with reading proficiency (i.e., girls and native English speakers on average have higher reading proficiency compared with boys and English language learners).

Published research on international large-scale assessments such as TIMSS and CIVED examining the influence of reading comprehension on performance is generally lacking. Utilizing data from the IEA international large-scale assessments, the study contributes to the literature by employing modern cognitive models and psychometric methods to enhance the current state of knowledge regarding the role of reading comprehension in large-scale subject-matter assessments. Results of this research help to identify which types of comprehension-related item features contribute to the difficulty of items. This will be a practical contribution of the research. This information, in conjunction with cognitive theories of text processing and comprehension, can afford researchers and educators insight into the types of cognitive processing that are tapped by assessment items. This will be a theoretical contribution of the research.

Chapter 2: Review of Literature

This chapter will provide an overview of research related to the current study of subject-matter assessments in science and civic-related social studies. This review will include information with respect to text comprehension and factors that affect the difficulty of text comprehension for individuals when they process assessment items in these subject domain areas along with relevant material about gender, English language learners and the role of vocabulary in processing test items. In the literature search I used *reading comprehension*, *text comprehension*, *reading abilities*, *reading demand*, *science*, *mathematics*, *social studies*, *civic*, and *large-scale assessment* as key words for literature search. Another selection criterion that I used was that measurement tools used in the studies or reviews had to be written in English. By reviewing the relevant literature this review will address four important issues: (1) What is the nature of reading comprehension processing in subject-matter texts (i.e. science and social studies)? (2) What reading comprehension paradigm/theory can be used that will have theoretical validity for adolescent students' understanding of text passages and questions (including answer options) in subject-matter assessments (i.e., science and social studies)? (3) What text features can be understood as providing affordances to the reader constructing representations of text in subject-matter assessments? (4) What is the role of reader factors such as general vocabulary, status as an English language learner, gender, and home resources?

2.1 Educational Assessment and Validity

2.1.1 Educational Assessment

Educational assessment has long been used to document students' knowledge, skills, attitudes and beliefs. Although assessments are used in various domains and for different purposes, what they all have in common is the desire to reason from particular things students say, do, or make in a given context, to inferences about what they know or can do more broadly (Mislevy, Steinberg, & Almond, 2003). According to measurement theory (Crocker & Algina, 1984 in a review), assessment tasks, specifically test items, usually serve as situational stimuli to evoke students' performance upon which a subsequent inference about what students know or are able to do can be drawn.

2.1.2 Constructs in Subject-Matter Assessments

In an assessment, students' knowledge, skills, attitudes and beliefs are perceived as latent psychological attributes which characterize what they say, do, or make. Such psychological attributes are called constructs in measurement theory. According to Crocker and Algina (1984), constructs are "products of the informed scientific imagination of social scientists who attempt to develop theories for explaining human behavior" (1984, p.4). By nature constructs are hypothetical concepts, and their existence can never be absolutely confirmed. Psychologists can only make inference about the degree to which a psychological construct characterizes an individual from observations of his or her behavior in given context.

Subject-matter domains such as mathematics, science, and social studies differ in how instruction is provided, how students acquire and accumulate their knowledge, skills, and abilities, and how their knowledge, skills, and abilities develop over time (Webb, 2006). The differences have important implications for how subject-matter constructs should be identified and defined in assessments.

In his review paper regarding to assessment of content areas, Webb (2006) points out that in mathematics, students' conceptual understanding develops hierarchically. For example, students' understanding of numbers grows from whole numbers to integers to rational numbers and on to the real numbers. The implication to the test design is that to measure students' knowledge of a hierarchically structured content area requires attending to prerequisite knowledge as well as to more advanced knowledge that builds on the underlying concepts and skills.

Language arts, as Webb (2006) suggests, are less hierarchically structured than mathematics. The sophistication of language use gradually increases over grade levels through applying and practicing skills and procedures. Once students acquire necessary reading principles and skills, they can refine these skills. Complexity in language increases through broadening content-related and general vocabulary, using more sophisticated sentence structures, and requiring more complex analysis and inferences. In specifying content for tests in language arts, test developers are required to think about what makes the assessment more complex based on word usage, sentence structure, passage length, and the number of inferences required. They also need to take into account students' backgrounds and prior knowledge, which may strongly influence measures of competency in literacy.

On the other hand, sciences are distinct content areas that contain a variety of subfields such as biology, chemistry, physics, and earth science. Webb (2006) reviews and summarizes these subfields of science as following:

> As students develop understanding in each of these areas, they learn specific concepts and scientific principles that may or may not relate to concepts and principles in other areas. The scientific method or way of thinking is used throughout all areas of science, as are specific processes such as observing, reflecting, justifying, and generalizing. Early phases of learning in science begin with students experiencing different scientific phenomena in their environment. Students' understanding of science grows as a result of their involvement in performing increasingly complex experiments, and in making inquiries and observations. As they progress, they encounter scientific laws and principles of greater complexity. This knowledge builds on prerequisite content, but increased understanding enables students to branch out into the separate science areas as they advance through the curriculum. Developing tests of scientific knowledge requires that test designers attend to an increasing understanding of scientific inquiry while identifying the specific concepts and principles that comprise the different fields of science. (p. 158)

Similar to sciences, social studies contain distinct areas including history, civics, economics, and geography, etc. However, instruction in these fields is less hierarchically structured compared with mathematics and sciences. Specific content and topics are taught at particular grades. For example, students usually learn U.S. history in third or fourth grade, and world history in high school. Webb describes the nature of the content and the implication for test development in social studies as following:

> Across the disciplines that comprise social studies, a common expectation is that students acquire knowledge of civic responsibility and what is required to be a member of a democratic society. Inquiry in social sciences draws on applying skills from other content areas, including language arts, mathematics, and science. In developing tests in social studies, it is important to know what students have had the opportunity to learn in specific social studies fields, as well as skills that can be applied from other content areas. It is also important for test developers to be aware of the level of abstract thinking and the types of inferences students should be able to make in the different social studies disciplines. It is unreasonable to expect students to necessarily have the same competence in higher-order reasoning in one area of social studies (such as history) as in other areas of social studies (such as geography or economics). (p. 158)

Appropriate specification of what a test intends to measure (i.e., the construct) is critical at any level, from classroom assessments to large-scale assessments. Understanding the nature of the subject-matter construct based on cognitive or learning theories can help test designers to make important decisions with respect to what content to include on a test and what content to exclude. These decisions affect significantly the inferences that can be made based on students' responses, and hence have impacts on the validity of the test. In the next section, I will introduce the current view of test validity.

2.1.3 Current Views of Validity

According to Messick (1989), validation of a psychological construct is an investigative process "by which we (a) create a plausible argument regarding a desired interpretation or use of test scores; (b) collect and organize validity evidence bearing on this argument, and (c) evaluate the argument and the evidence concerning the validity of the interpretation" (p. 18). In general, current views of validity (Cronbach, 1988, Kane, 2006; Messick, 1989) suggest that researchers consider evaluation of the validity of the intended interpretations and uses of test scores as a process of evaluating an argument. The process requires an evaluation of all the available evidence for and against the proposed interpretation or use of test scores.

One way to strengthen the validation argument is to include the testing of alternative explanations with respect to the validity of an interpretation (Cronbach, 1988; Kane, 1992, Mislevy, 2009). Kane (1992, 2006) suggests that by eliminating and reducing the plausibility of alternative explanations, we can increase our confidence that a desired interpretation regarding test scores is valid. An alternative explanation to an interpretation of test scores is often referred as a threat to validity (Crooks, Kane, & Cohen, 1996). In a research review, Haladyna and Downing (2004) conclude that at least five major threats to validity deserve our attention: construct under-representation arising from poorly conceptualized or inadequately operationalized constructs, faulty logic of the causal inference regarding test scores, negative consequences of test score interpretations and uses, lack of reproducibility of test scores (over time), and construct-irrelevant variance. It is beyond the scope of this paper to discuss all threats to validity. Therefore, this review will particularly concentrate on one particular threat, namely construct-irrelevant variance that arises in language-based tests of knowledge and skill in domains such as science and social studies.

2.1.4 Construct-Irrelevant Variance

Construct-irrelevant variance (CIV) is a type of systematic error that biases the validity of interpretations and used of test scores (Messick, 1989). To better understand this concept, we must delve into the classic test theory which uses a linear model to describe the relationship among test scores, construct, and error variance. The model is:

$$X= T + E$$

Where X is the observed test scores for any student, T is the true score representing the construct that a test is intended to measure, and E is the error variance that consists of random error and systematic error, and it by definition is uncorrelated with true and observed scores (Crocker & Algina, 1986; Haladyna & Downing, 2004). Random error is associated with individual differences. Construct-irrelevant variance (CIV), on the other hand, is a type of systematic error that affects examinees differentially and leads to underestimation of individual examinee scores (Haladyna & Downing, 2004). Examples of construct irrelevant variance include inappropriate test administration, cheating , anxiety, fatigue, excessive reading demand on subject-matter assessments, and not considering the special problems of students with disabilities, second language learners, and students living in poverty when reporting group test results (Messick, 1984; Haladyna, 2002; Oakland & Lane 2004). I will elaborate the CIV due to reading comprehension in the following section.

2.1.5 Construct-Irrelevant Variance due to Reading Comprehension

Almost all educational assessments require language (Tourangeau, 2003). In subject-matter large-scale assessments, language commonly serves as a vehicle of communications between the assessment task and the test taker through text written or presented orally by an examiner (Oakland & Lane, 2004). When it comes to evaluate the subject-matter assessment, we often find that individuals' verbal abilities (including reading comprehension) are interwoven with what the assessment assesses (e.g., domain specific abilities). In addition, we may observe that some domain specific assessments make heavy demands on reading comprehension and others make less of a demand. The issue is: to what extent should reading comprehension influence test performance in subject-matter assessment?

In his book *Developing and Validating Multiple-choice Test Items*, Haladyna (2004) acknowledges that reading comprehension is necessary for subject-matter assessments. However, he points out that deficiencies in reading comprehension can interfere with students' performance in the test of subject matter and introduce bias into test interpretation. This is especially true when test takers with low reading proficiency (e.g., those learning the language in which the test is given). These students hence are likely to be subject to missing or incorrect responses not because of lack of required knowledge or skills but because of low reading comprehension.

This problem is prevalent among English language learners (ELLs). Abedi and his colleagues have conducted a set of studies to investigate the importance of reading comprehension in standardized reading and mathematics tests for K-12 students. In general, their results show that language demands in subject-matter assessments have larger impact on students with low reading proficiency, including ELLs, than students with high reading comprehension (e.g., Abedi, Hofstetter, Baker, & Lord, 2001; Abedi, Lord, Hofstetter, & Baker, 2000). Therefore, it is especially important to pay attention to special groups with low reading proficiencies when evaluating test scores of subject-matter. I will review and elaborate their studies in a later section on *Language Background*.

To summarize, how assessment tasks are written (e.g., the choice of vocabulary and sentence structure, and the amount of reading demands posed on a test) can influence the level of reading difficulty in the assessment. To the extent that reading difficulty exceeds a test taker's reading abilities, there may be interference with the test taker's demonstration of subject matter knowledge. As a result, the test score may not accurately reflect the test taker's achievement in the domain.

According to Messick (1989), if the construct to be measured does not include reading comprehension as an integral part of its definition, a test taker's reading comprehension level should not function to diminish test performance. Therefore, in order to ensure accurate and valid measures of student learning in content areas, it is necessary to understand the nature of the construct being measured as well as the students' comprehension processes when they read test items, so that unnecessary construct-irrelevant variance can be minimized. This study focuses on the role of reading comprehension in subject-matter assessments. Therefore, in the following section I will review reading comprehension theories which give us a way to understand the complex comprehension processes that are involved in processing subject-matter test items theoretically and systematically.

2.2 Reading Comprehension

2.2.1 Definition of Reading Comprehension

The RAND Research and Development Study Group that was led by Catherine Snow (2002) provides a definition of reading comprehension as "the process of simultaneously extracting and constructing meaning through interaction with written language" (p. 11). This definition reflects the current view from constructivists. That is, comprehension is not some type of static trait; rather comprehension of text consists of multidimensional and multilevel cognitive processes that are constructed in the reader's mind. To successfully understand a text, a reader must access the meaning of words, tie the meaning of words to a coherent sentence level representation, relate sentences to one another to build local coherence, and relate larger pieces of text to build a global coherent mental representation. In the end, the reader needs to integrate the representation with his or her prior knowledge in order to achieve deep level understanding. These processes are iterative and reciprocal (Best, Ozuru, Floyd & McNamara, 2006; Duke & Carlisle, 2011; Kintch, 1998). Although these processes are most complex for a long text or one with several topics, the same processes can be assumed to apply to short texts in which most test items are written.

2.2.2 Kintsch's Reading Comprehension Model

During the past several decades, numerous reading comprehension models have been developed to capture complex reading comprehension processes from the constructivist perspective. For example, the literature includes the construction-integration model (CI model, Kintsch, 1998), the structure building model (Gernsbacher, 1990, 1997), the resonance model (Myers, O’Brien, Albrecht, & Mason, 1994), the event-indexing model (Zwaan & Radvansky, 1998), the causal network model (Trabasso, van den Broek, & Suh, 1989), the constructionist theory (Graesser, Singer, & Trabasso, 1994), and the landscape model (van den Broek, Young, Tzeng, & Linderholm, 1999). Among these comprehension models, Kintsch’s (1988, 1998) CI model has been considered to be a complete and well-formulated one that has built a foundation for the development of other more sophisticated comprehension models (McNamara & Magliano, 2009; Verhoeven & Perfeiti, 2008 in reviews).

Originated from schema theory (Wilson & Anderson, 1986), the CI model assumes comprehension is constructed and built on integrated mental models (schemas) derived from the text. To do this, the reader must activate concepts expressed in the text and form connections between activated concepts and relevant prior knowledge of words, concepts, ideas and personal experience. The comprehension processes are regulated by mental models (schemas) and constrained by contexts. In other words, the networks of concepts that are compatible with the context enhance the activation of one another, while concepts that are not compatible with the context lose activation (Kintsch & van Dijk, 1978; Stahl & Hiebert, 2006).

In general, three levels of mental representations are involved during the comprehension processes: the surface structure (vocabulary and syntax), the propositional textbase (explicit meaning of the content), and the situation model (the coherent mental model). The surface code consists of vocabulary and syntax of the sentences. The propositional textbase contains explicit propositions in the text, such as statements, and idea units. The situation model (or what is sometimes called the mental model) is the referential microworld of what the text is about; it contains the people, setting, states, actions, and events that are either explicitly mentioned or inferentially suggested by the text.

2.2.3 Reading Comprehension Processes

In the CI framework, reading comprehension is viewed as iterative processes in which the reader is constructing mental models (which can be incoherent at the beginning) by activating meanings and concepts from text along with knowledge and personal experience that the reader brings to the situation. Generally speaking, reading comprehension consists of two phases: decoding and comprehension. In the decoding phrase, the individual words are perceptually and conceptually identified. The reader converts visual input into a linguistic mental representation, which contains a sequence of idea units, called propositions. The linguistic mental representation is the surface structure model. The next phase, comprehension, involves several interacting levels of processing: microstructure, macrostructure, and a situation model. Microstructure processes tie word meanings together. Macrostructure processes link and elucidate relations of individual sentences and groups of sentences to a global topic. A student who is asked to recall or focus on details from a text will rely both on the microstructure and macrostructure of the text. On the other hand, preparing a good summary would primarily reflect the macrostructure.

Microstructure and macrostructure together form the textbase model (i.e. the mental representation that the reader constructs of the text). A successful textbase model typically requires coherence building at both microstructure and macrostructure levels. Kintsch argues that the textbase model is only sufficient to support recall of text. Deep-level comprehension takes place when the reader integrates the textbase model with prior knowledge and personal experience. Kintch calls the final stage of comprehension a situation model, a mental representation of people, actions, events, and settings. Situation models emerge to the extent that the reader activates concepts, incorporates these concepts into the mental representation, and establishes connections between propositions (a network or hierarchy of concepts or idea units) in the mental representation (Graesser, Singer, & Trabasso, 1994).

The situation model can vary depending on the extent that the reader activates prior knowledge and integrates that knowledge into the textbase model (Kintsch, 1998; McNamara & Magliano, 2009, in a review). If the context is not compatible with the readers' mental models including the textbase model, it is less likely that the reader can activate prior knowledge and integrates that knowledge into existing models. For example, if the reader is not familiar with the characteristics of the text (e.g., genre, vocabulary difficulty), the text is less likely be able to call on prior knowledge and experience. Hence the situation level comprehension may not occur.

2.3 Factors that Affect Reading Comprehension

Many factors interactively contribute to successful comprehension of written assessment tasks, including students' cognitive ability, knowledge, motivation, interest, as well as features of the assessment tasks, including task description, question wording, item format, task goals and context (Kintsch & Kintsch, 2005; Schwarz, 1999; RAND Reading Study Group, 2002). In general, these factors can be summarized into three categories: factors that depend on the reader, factors inherent in the text, and factors specific to the context. Understanding factors that affect reading comprehension can afford researchers and educators greater insight into assessment designs in a variety of academic domains including reading, science, social studies, and mathematics. In the next section, I will review factors falling into these three categories, and discuss their implications for educational assessment.

2.3.1 Factors Dependent on the Reader

The reader brings his or her attributes, such as cognitive abilities, domain knowledge, motivation, and experience to tasks involving comprehension (Duke & Carlisle, 2011; Kintsch, 1998; RAND Reading Study Group, 2002; Verhoeven & Perfetti, 2008). Reader variables can be classified in a variety of ways. Snow, Corno, and Jackson (1996) provide a schema-like diagram that organizes reader factors into a hierarchy (Figure 1.1). Particularly, for cognition factors, many have been identified as developing over years and grade levels. In the early elementary years, word recognition, fluency, and oral language abilities have been found to reliably predict reading comprehension. In the later elementary years, word recognition and fluency become less associated with reading comprehension. Higher level language skills such as semantic skill, comprehension monitoring, and the use of comprehension strategies are more important determinants of reading comprehension by 5th or 6th grade (Duke & Carlisle, 2011).

Among these cognition factors in Figure 1.1, Kintsch’s approach (1998) identifies three as the most important for comprehension in the context of educational assessment: decoding skills, higher-level reading skills (knowledge of how including strategies and skills) and prior knowledge (knowledge of what). The review places an emphasis on how these factors are associated with reading comprehension, especially the comprehension of test items in large-scale subject-matter assessments.

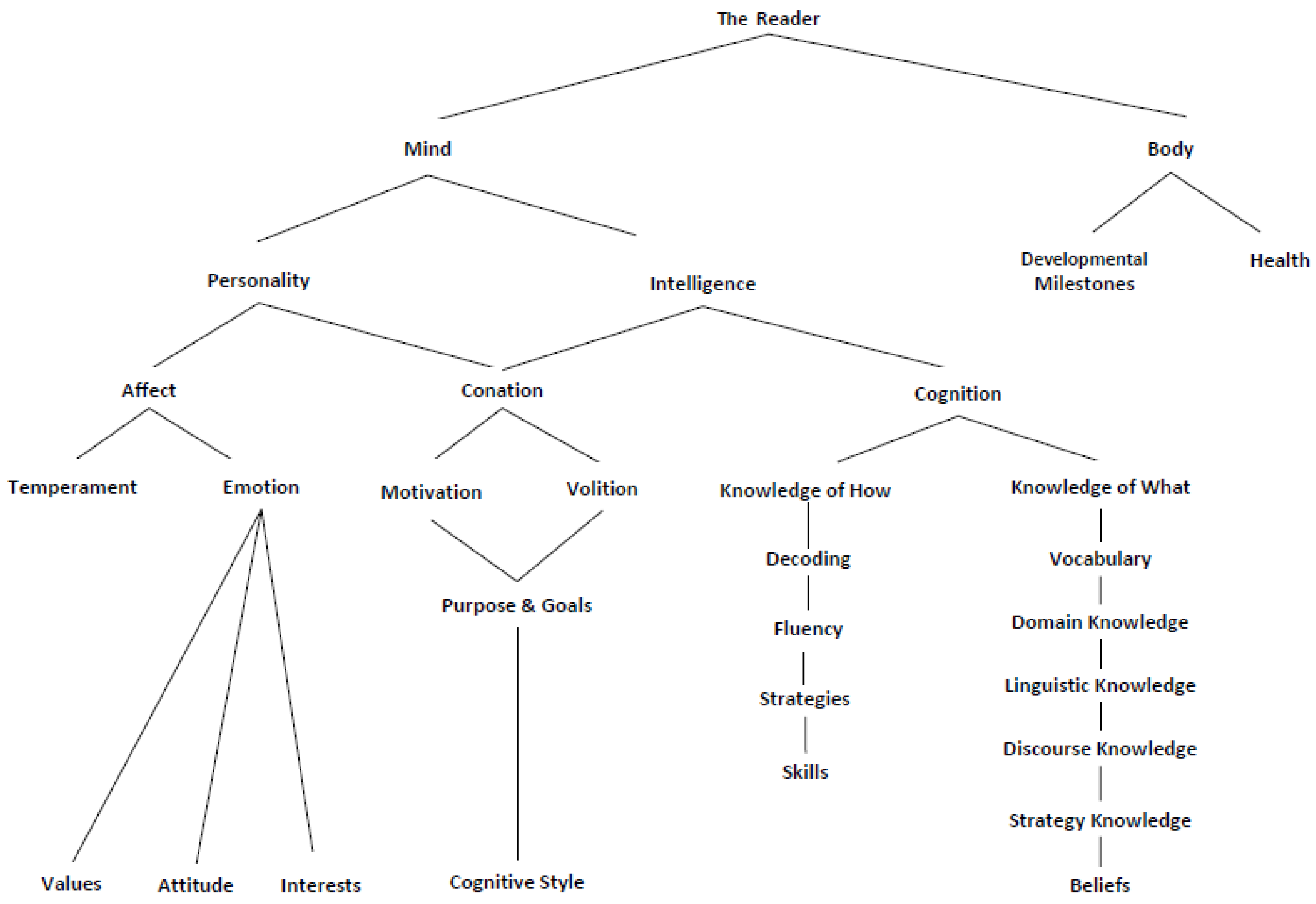


Figure 1.1. Reader variables. This diagram is adapted from Snow, Corno, and Jackson (1996) and Gaskins (2003)

Decoding Skills. Reading comprehension can be viewed as beginning with a bottom-up process with a variety of language skills involved. If we rank these comprehension-related skills from low to high in a comprehension processing chain, we find that decoding is the starting point of reading. Decoding is the perceptual and conceptual identification of individual words. Decoding skills are important to reading comprehension because rapid decoding and better word recognition free up working memory for higher-level cognitive processing, which can result in more accurate and complete representation of text (Kintsch, 1998). Decoding skills are usually associated with readers' capacity to read fluently (RAND Reading Study Group, 2002 in a review).

According to Duke and Carlisle (2011) in a review, in the early school years (especially the period between second and fifth grade), children's understanding of text is largely determined by their decoding skills and phonological awareness (e.g., awareness of sounds, and rhymes; understanding of the relation between written language and spoken language, Carroll, Snowling, Hulme, & Stevenson, 2003). However as their years of schooling increase, the amount of variance in reading comprehension explained by their decoding ability decreases. Wilson and Rupley (1997) conducted cross-sectional research investigating the association between phonemic knowledge and reading comprehension among students from grade 1 to 6. Research results suggest that children's ability to decode words appears to affect their comprehension in grade 2 and 3, but the effects diminished in the upper grades. It seems that higher-level reading skills drive comprehension when students become more fluent and automatic in reading. In addition, Storch and Whitehurst (2002) followed 626 children from preschool to 4th grade in a longitudinal study. They found that during early elementary school, a child's reading ability is mainly determined by his or her prior knowledge and phonological awareness the child brings from kindergarten. In the upper grades, reading accuracy and reading comprehension become separate abilities that are determined by different sets of skills.

In addition to decoding skill, other language skills associated with high-level meaning-based presentations are attributed to reading comprehension, such as knowledge of word meanings (Perfetti, 1985), inference making (Kintsch, 1998), comprehension monitoring (i.e., the metacognition of how well one understands, Baker & Brown, 2002), and knowledge about text structure (Kintsch, 1998).

Knowledge in Vocabulary. Among these reading skills, skill and knowledge of vocabulary (e.g. word identification, knowledge of word meanings) are the most essential (e.g., McKeown & Curtis, 1987; Perfetti, 1985; Stahl & Fairbank, 1986; Snow, 2010). Many psycholinguists and psychologists (e.g., Anderson & Freebody; 1981; Beck, McKeown, & Omanson, 1987; Perfetti,1985; 2010; Snow, 2010) especially those with specific subject-matter interests, view vocabulary as a core component that leads to successful comprehension.

Perfetti (1985, 2010) claims that vocabulary knowledge is the major source of reading ability. To address the importance of vocabulary to comprehension, he conceptualizes general reading skill as a triangle (i.e., the Golden Triangle; see Figure 2) that consists of three reading components: decoding, vocabulary, and comprehension. The three components reciprocally influence one another. In the Golden Triangle, vocabulary (specific knowledge of word meanings) plays the role of mediator mediating the relation between decoding and comprehension. Perfetti (2010) elaborates on the mediating relations: “The effects of decoding on comprehension are mediated by knowing the meaning of the decoded word. The effects of comprehension on decoding are mediated by achieving enough meaning from text to verify the identity of a decoded word" (p. 294). Limited empirical research in the area of science and social studies has addressed this issue. However, based on the literature cited above, one can anticipate that in large-scale science and social studies tests in which test items are usually composed of short texts, vocabulary is one of the most important issues because it is not only related to readers' comprehension, and also can be associated with their science and social studies knowledge when it is directly related to the content of testing.

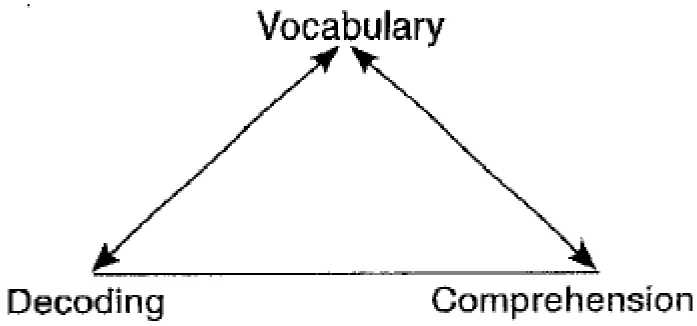


Figure 1.2. The Golden Triangle. Adapted from Perfetti (2010).

Reading Strategies. In addition to vocabulary knowledge, other language skills associated with high-level meaning-based presentations are attributed to reading comprehension. In a review of text comprehension, Kintsch and Kintsch (2005) provide a list of reading skills that have been shown to be effective in understanding of text (p. 84):

- Using words or imagery to elaborate the content.
- Rereading, paraphrasing, and summarizing in one's own words to clarify the content.

- Reorganizing the content into a hierarchical outline, diagram, or graph that shows the important relations between ideas.
- Consciously seeking relations between new content and existing knowledge (e.g., by self-explaining, forming analogies, hypothesizing, drawing conclusions and predictions, formulating questions, and evaluating the text for internal consistency and with respect to what one knows of the topic). The application of this and the previous three items in the list to reading test items has not been carefully investigated.
- Consciously monitoring one's ongoing knowledge, identifying the source for breakdown in comprehension, and attempting to resolve the problem rather than passively reading on through the text (for reviews of this literature, see Dansereau, 1985; Pressley, Wolshyn, & Associates, 1995). When reading test items, the corresponding process for passively reading on is probably guessing an answer (at least in a multiple-choice question).

These higher-level skills are important for comprehension because they aid the active construction of meaning from the text, and the deliberate linking of information derived from the text with prior knowledge and experience. Most of these higher-level reading skills are important for understanding when the text consists of relatively long passages. However, Kintsch provides limited discussion about what kind of higher-level reading skills are necessary for comprehension when the reader reads test items which are usually written in short sentences or even incomplete sentences.

Prior Knowledge. In the past three decades, research in the area of cognition and learning has reached a common agreement that the reader's prior knowledge plays a critical role in understanding of text (e.g. Alexander, 1997; Bransford, Brown, & Cocking, 1999; Kintsch, 1998; Perfetti, 1985; Shapiro, 2004; RAND Reading Study Group, 2002; Thompson & Zamboanga, 2004; Willoughby, Waller, Wood, & MacKinnon, 1993; Willoughby, Wood, & Khan, 1994). Experimental studies using college students (McNamara & Kintsch, 1996; McNamara, Kintsch, Songer, & Kintsch, 1996) show that prior knowledge, specific to the domain being assessed, is a key factor necessary for the reader to build a situation model in understanding science and social science texts.

Current educational psychologists view prior knowledge as a multi-dimensional construct that includes many types of knowledge, such as knowledge about content, content-specific vocabulary, knowledge in language syntactic, domain, the world, and cultures (Alexander, Kulikowich, & Schulze, 2004; Gaskins, 2003). Some types of knowledge can be informally acquired, and others are formally learned. Alexander, Kulikowich, and Schulze (2004) define the type of knowledge that is formally learned in school *as subject-matter knowledge*. The subject-matter knowledge has a variety of forms including domain knowledge and topic knowledge. Domain knowledge has been defined as knowledge broadly related to a particular field of study (Alexander, 1992; Alexander & Judy, 1988; Alexander, et al., 2004). Topic knowledge, on the other hand, concerns smaller units of knowledge than domain knowledge does. It is the knowledge related to a specific body of discourse (Alexander, Schallert, & Hare, 1991). For example, for a text on bacteria, the reader’s knowledge of biology or human immunology would be relevant to domain knowledge, and their knowledge of bacteria would be related to topic knowledge. Another example of topic knowledge is for students asked to read a passage about growth of democracy in Eastern Europe, their conceptual knowledge about democracy in regions such as this would be topic knowledge. In summary, topic knowledge tends to be more situationally specific to a text than domain knowledge. Compared with topic knowledge, domain knowledge is a broader form of subject-matter knowledge (Alexander, et. al., 2004).

A main purpose of subject-matter assessments such as International Association for the Evaluation of Educational Achievement (IEA) Third International Mathematics and Science Study (TIMSS) science and IEA Civic Education (CIVED) is to measure subject-matter knowledge including declarative knowledge and procedural knowledge (Li, Ruiz-Primo, & Shavelson, 2006; Torney-Purta, Lehmann, Oswald, & Schulz, 2001; Zhang, Torney-Purta, Barber, 2012). Under such circumstances, domain knowledge as well as topic knowledge, such as knowledge of content-related concepts or terminologies, are important parts of the construct. If a student does not provide a correct answer to a test item, we want to ensure that the incorrect response can be attributed to the student's deficiency in the subject-matter knowledge (the construct) that the item intends to measure, rather than something else such as their deficiencies in understanding the test item. Since reading comprehension plays an important role in such assessments, we want to find out at the item level, what kinds of task features, including text features, can afford students better chances to demonstrate their domain knowledge.

Research in assessment of reading comprehension has found that well-structured prior knowledge appears beneficial to reading comprehension because (a) the knowledge structures (knowledge representation) can help organize information in memory for later recall or use (Anderson & Pichert, 1978; Bower, Black, & Turner, 1979; Bransford &

Johnson, 1972, 1973); (b) such knowledge structures facilitate integration of new information from the text to what already exists. One implication for subject-matter assessment is that students with high subject-matter knowledge are more likely to comprehend texts better and remember them better than those with low domain knowledge given the level of reading demand on the text matches to readers' reading proficiency (Abedi, Lord, Hofstetter, & Baker , 2000; Kintsch & Kintsch, 2005).

Other Factors. In addition to decoding, higher-level reading skills, and prior knowledge, other reader-related factors also affect the reader's comprehension capacity, such as motivation, interest, and test taking strategies. Readers' motivation may impact their ability to read a difficult passage or complex item (Guthrie, & Wigfield, 1999). Struggling readers have been able to read text above their typical reading level when they have high interest in the subject matter (Allington & Cunningham, 2006). Students' motivation may also be related to their previous experience with text topic or genre (Alexander, Kulikowich, & Schulze, 1994). Certain features of a test may be motivating (e.g. use of cartoons in items). In spite of that, in a test we usually assume the students' motivation is to comprehend the items in order to answer correctly, whatever format used. In addition, the assessment context (high stakes vs. low stakes) can differentially influence how individuals or groups of students engage in the test-taking process (Heubert & Hauser, 1999; Ryan, Ryan, Arbuthnot, & Samuels, 2007).

Summary. In summary, reading comprehension consists of complex cognitive processes which involve several interactive levels of mental representations. Many attributes directly related to the reader have an impact on these processes (e.g., decoding skills, knowledge in vocabulary, reading strategies, subject-matter knowledge, motivation, and interest). According to Kintsch (2005), the most important attributes related to assessments appear to be decoding, higher-level reading skills including vocabulary knowledge and reading strategies, and subject-matter knowledge because they facilitate the reader in making sense of the text by constructing mental representations and integrating information from the text into the representations. In particular, Kintsch (1998) points out that these factors can compensate for one another to a considerable extent in domain specific areas. O'Reilly and McNamara (2007) conducted a cross-sectional experimental study with 1,651 high school students to investigate how science knowledge and reading skill (i.e., the ability to develop a coherent representation of the text that matches the intended message to the reader) relate to high school students' science achievement. Through multiple regression analyses, their results showed that the reading skill moderated the association between science knowledge and students' performance in a standardized science assessment. In other words, high reading skill compensated for some student's deficits in science knowledge in the science achievement test. Meanwhile, some students' performance on the science test was attenuated by their deficiencies in reading skill. In this case, poor reading skill interfered with some students' performance in the standardized science achievement test.

This example could characterize many subject-matter assessments. Therefore when we design a subject-matter assessment, it is necessary to ensure that the test is written using language that is in alignment to students' reading abilities (usually associated with their grade level). In a review of reading for learning science, Snow (2010) points out that "the major challenge to students learning science is the academic language in which the science is written" (p. 450). This statement can be generalized to other domains such as mathematics and social studies as well, because the central features of academic language such as grammatical sentence structure, sophisticated and abstract vocabulary, and precision of word choice are prominent features of the academic language in these domains. A key message that Snow (2010) delivers is that as educators we should ensure that students are able to read academic language in domain specific areas. One way to help students is teaching them skills to read the academic language in subject-matter domains, and another way is providing them enough practice so that they can get familiar with the central features of academic language.

In summary, reading comprehension can vary within an individual reader as a function of the particular text and context (intra-individual differences) (RAND Reading Study Group, 2002). In the next section, I will review and discuss how features inherent in text and context may affect the reading comprehension of assessment tasks.

2.3.2 Factors Inherent in the Text

Reading comprehension does not occur simply by constructing text meanings and integrating them into mental representations. In fact, the extent of comprehension varies within an individual reader as a function of the particular text and context. Features of the text (e.g., the difficulty of vocabulary, syntactic complexity, text cohesion, text genre) play an important role in reading comprehension. They can either increase or decrease reading difficulty in interaction with the knowledge and abilities of the reader. If many text features such as vocabulary and linguistic structure are not matched to a reader's knowledge and language ability level when the reader engages in a reading activity, the text may be too difficult for optimal comprehension to occur (RAND Reading Study Group, 2002).

Just as passages in textbooks can be written at different levels of difficulty, test items in subject-matter assessments likewise can be designed on a continuum from easy to difficult. At the text level, empirical studies in reading assessment found that features of text passages (i.e., vocabulary difficulty, sentence length, cohesion of sentences, etc.), and text genre (e.g., narrative or expository) account for a significant amount of the variance in item difficulty, but so far these studies have been limited to assessments of reading itself (e.g., Gorin & Embretson, 2005; Oakland & Lane, 2004; Ozura, et al., 2008). In the following sections I will review text-related factors that account for the understanding of texts in assessment tasks.

Vocabulary. A widely held view in reading research believes that readers with larger vocabularies understand texts better (Perfetti, 1985; 2011; RAND Reading Study Group, 2002, Snow, 2010). Empirical studies have consistently found reading comprehension has strong correlations with specific word recognition, and knowledge of word meaning in adolescents as well as adults (e.g., Holmes, 2009; Simmons, et. al., 2010). In addition, the latter two variables account for significant proportions of variance in reading comprehension (e.g. Carver, 2000). Among the vocabulary factors, experimental research also indicates that word recognition is related to familiarity of the word and semantic properties of words (e.g., their concreteness and abstractness). In terms of familiarity of a word, Adams (1990) found through experimental methods that readers recognize known words flashed on a screen more quickly and accurately than unknown words and nonsense words. In terms of the semantic properties of words, In order to investigate the influence of text variables including response options on item difficulty, Sheehan and Ginther (2001) examined test items in the Test of English as a

Foreign Language (TOEFL). They found that if the correct response consisted of frequent words, then the item was easier. Conversely, infrequent words in the distractors made the item easier because test takers were less likely to expend the time and effort to process the distractors. This finding suggests that the more familiar the word is to a test taker, the more likely such a word would activate relevant schema(s) in the test taker's mind. This is consistent with Kintch's CI model.

Sentence Structure. I use sentence structure to refer to two parts of sentence characteristics: (1) sentence length and vocabulary load: the number of words in a sentence, and (2) syntactic complexity, which is related to the grammatical connections of the words and the sentences. Both parts have to do with working memory. When a reader reads the text, he or she relies on working memory to process information in written sentences. In general, the accuracy of processing will be lower if sentences are longer because there are more ideas to process compared with reading shorter sentences (D'Arcy et al. 2005). Likewise, if sentences are syntactically complex, more effort has to be made to interpret meanings (Sigurd, Eeg-Olofsson, & Van de Weijer, 2004). Therefore, complex sentence structure can create working memory load, which increases the difficulty of text comprehension. Complexity of sentence structure has been used to predict problem difficulty in mathematics, although in a quite old study (e.g. Loftus & Suppes, 1972). In addition, traditional readability formulas (e.g., Dale & Chall, 1948; Flesch, 1951; Fry, 1968; Gunning, 1968; Spache, 1953) use sentence length (characterized by the number of words in a sentence) as an indicator to predict text difficulty.

Some researchers suggest that writing texts in short sentences decreases the cognitive demand in text processing. Similarly, simplification of some syntactic features may make problems easier for children to comprehend (De Corte, Verschaffel, & De Win, 1985; Marshall, 1995). Other researchers argue that some long sentences can be easily comprehensible if the reader is familiar with meanings of the words and the content. Therefore, rather than sentence length, current readability methods utilize propositional density (it is usually approximated by the number of verbs, adjectives, adverbs, prepositions, and conjunctions divided by the total number of words, Kintsch, 1974) of a sentence as an gauge of text difficulty (e.g., Embretson & Wetzel, 1987; Gorin & Embretson, 2006; Rowe, Ozuru & McNamara, 2006). In general, it seems that readers will benefit the most if the complexity of the sentence structure is aligned with their working memory ability (Mikk, 2008).

Text Coherence. Traditionally, difficulty of text passages has been gauged through the frequency or familiarity of the words, and the length or syntactic complexity of the sentences. Recent studies have shown that text coherence is also an important factor that relates to text difficulty (e.g. Kintsch & Kintsch, 2005; McNamara & Kintsch, 1996). According to McNamara and Magliano (2009), “a text is perceived to be coherent to the reader when the ideas connect to each other in a meaningful and organized manner. The text is less coherent when there are many conceptual and structural gaps in the text, and the reader does not possess the knowledge to fill them” (p. 312). Reading specialists believe that cohesive text is important to comprehension because it helps readers construct more coherent mental representations of text content. However, recent studies based on experimental methods showed that the cohesive text was only beneficial to students with low domain knowledge (McNamara & Kintsch, 1996; McNamara, Kintsch, Songer, & Kintsch , 1996: Voss & Silfies, 1996; O'Reilly & McNamara, 2007b). In other words, students with low domain knowledge understand better and learn more from cohesive texts. Contrariwise, students with high domain knowledge learn more from less cohesive texts. This counterintuitive finding has been called *reverse cohesion effect* (O'Reilly & McNamara, 2007b).

For example, McNamara, Kintsch, Songer, and Kintsch (1996) conducted an experimental study using junior high students to examine the effect of text coherence on science texts learning. In the study, participants first read four biology texts which had the same content but differed in coherence. Their comprehension of biology texts were then assessed through free recall, written questions, and a key-word sorting task. Results showed that the effects of text cohesion on comprehension interacted with the reader's prior knowledge (see also McNamara, 2001; McNamara & Kintsch, 1996; McNamara, Graessser, & Louwerse, 2011; O'Reilly & McNamara, 2007; Ozuru, Dempsey, & McNamara, 2009). Students who knew little about the domain of the text benefited from a coherent text, whereas high-knowledge students benefited from a less coherent text. One interpretation of this finding is that less coherent texts forced knowledgeable students to generate many inferences. Thus knowledgeable students were provoked by low coherent texts to work more actively to integrate their prior knowledge with the information from the text. This process resulted in a deep-level of understanding. This study has been replicated by O'Reilly and McNamara (2007b) using college students. Their experiment results further showed that low cohesive texts were especially beneficial to students with low reading skills but high domain knowledge.

Large-scale subject-matter assessments such as TIMSS and IEA Civic Education often consist of test items written with sentences shorter and less coherent than those that students normally read in school and everyday real life situations. Some sentences are even incomplete on purpose in order to elicit students' responses. Some scholars are critical that this type of text may not facilitate students' understanding because it is different from what they are familiar with, and students have few opportunities to develop skills to read it (e.g., Sternberg, 1991). However, evidence from research in the field of reading comprehension of longer passages supports the use of text that is relatively less coherent. This type of text does not prevent students (especially those who have high domain knowledge but low reading skill) from understanding the meaning of the text.

Text Genre. Text genre refers to a widely recognized class of text defined by function, sociocultural practices, and communicative purpose (Ravid & Tolchinsky, 2002). Two major types of text genres are expository text and narrative text. Narrative texts are constructed with some kind of story-line and usually contain topics that people are familiar with such as friendship, love, and family. Readers often have extensive experience and knowledge regarding what is described in a typical narrative text. Expository texts provide readers with information about concepts and events that may not be encountered in daily life or common place conversation. It usually presents specific scientific or historical facts, relations between facts, or both. A scientific research article usually belongs to the expository text category whereas a popular-science article can be considered a mixed text, found somewhere on a continuum between expository and narrative text, due to its periodic story-telling parts which include characters and events that appeal to a non-specialist (Alexander & Jetton, 2000).

Readers take different approaches and different strategies when reading narrative and expository texts. Readers generally read narrative text from beginning to end. In this way, they are gradually getting familiar with story elements including setting, characters, and problem development and resolution. When readers read expository text, they are likely to read it differently than a narrative text, and often need to apply reading strategies to assist their understanding (Duke, 2000; Guthrie & Mosenthal, 1986). Important characteristics of long passages of expository text include using features of the text such as a table of contents, index, heading, sub-heading, captions and glossaries to be able to locate information, explicit use of text structures such as problem/solution, compare/contrast, and cause/effect, and the inclusion of graphical elements such as maps and diagrams (Collin, 2007; Pappas & Pettegrew, 1998; Duke, 2000). Some texts have characteristics of both text genres. We call this type of text mixed. Texts related to social studies such as historical text are usually considered as mixed because it has characteristics of both narrative text and expository text (Eason, Goldberg, Young, Geist, & Cutting, 2012; McGraw, 1992; McNamara, et.al, in press).

Text genre has a profound impact on students' reading comprehension (e.g. Snow, 2010). In an experimental study, Best, Ozuru, Floyd, and McNamara (2006) asked 64 4th graders to read two expositive texts and two narrative texts taken from school textbooks. Topics of the two expositive texts were the *Heat* and the *Needs of Plant*; the narrative texts were *Moving* and *Orlando*. They found that 4th graders showed better comprehension of narrative than expository text (measured by multiple-choice questions).

In addition, students' domain knowledge moderated the effect of text genre on their reading comprehension. Students with higher domain knowledge showed better comprehension on expository texts, while the effect was not so substantial for narrative text. The implication for subject-matter assessment is that the effect of text genre is associated with reading comprehension and domain knowledge. Expository text appears to benefit students high in domain knowledge. However, students must develop reading ability and strategies in reading expository text so that they can construct an appropriate situation model and understand what they read. Most assessments outside of those of reading itself focus on expository text, but sometimes narratives are included for motivation (though with results that have not been carefully examined).

Summary. In general, previous studies reveal variables inherent in text can affect reading comprehension in general and specifically of assessment items. Such variables include vocabulary, syntactic complexity of sentence structure, text coherence, and text genre. In particular, existing readability research suggests that vocabulary load and syntactic complexity are the most robust predictors of text readability (Klare, 1984)

These variables interact with reader-related factors such as working memory, domain knowledge, and reading skills during reading comprehension processes. As a result, an individual's test performance varies as a function of nature of the source text and context. In the next section, I will briefly review characteristics that are specific to the context of educational assessment, and how these characteristics affect students' reading comprehension processes in a domain specific assessment.

2.3.3 Factors Specific to the Task

Reading does not occur in a vacuum. The dynamic interaction between reading comprehension and texts is embedded in a context. Broadly speaking, context refers to where readers read including classrooms, schools, home, neighborhoods, or the larger society. In a narrower sense, context involves reading activities in which readers are engaged. The RAND Research and Development Study Group (2002) define activity as including the purpose of reading, and the end to be achieved. For example, readers can engage in a variety of reading activities with different goals. Some may read a textbook in order to learn; some read a fiction book for entertainment; or others read a test item in order to provide or select an answer. Meanwhile, the purpose and consequence of a reading activity often intertwine with the reader's motivation (e.g. goal and interest) and abilities (e.g. reading fluency or metacognitive abilities). For example, when an activity is conducting a literature review for research, the reader may need to read multiple texts seeking certain information. In this scenario, the activity is also impacted by reader factors such as prior knowledge, reading strategies, and interest.

Large-scale assessments such as IEA TIMSS and Civic Education are standardized paper and pencil tests that were designed to measure students' achievement in specific domains and factors related to it across countries. Assessment results are compared across countries with the aim of gaining in-depth understanding of the effects of policies and practices within and across systems of education. Because of the purpose and consequence of these large-scale assessments, students are assumed to be motivated to a certain degree when they participated in these assessments, even though possible individual differences in motivation have not been acknowledged (e.g., Liu, Bridgeman, & Adler, 2012; O'Neil, Sugre, & Baker, 1995; Wise & DeMars, 2005).

In educational assessment, a common activity that students are involved in is the assessment task. The term *task* here refers to a goal directed human activity that is pursued in a specific manner and context (Haertel & Wiley, 1993). It describes "particular circumstances meant to provide the examinee an opportunity to take some specific actions that will produce information about what they know or can do more generally" (Mislevy, Steinberg, & Almond, 1999, p. 19). A task can thus include a long-term project such as a term paper, a think-aloud interview about an examinee's cognitive processes when solving a math problem, or a familiar multiple-choice or constructed-response item in a science assessment. Tasks are a central focus of educational assessment, because they evoke performance which is judged in relation to a standard and upon which subsequent feedback, decisions, prediction, or placement is based.

Generally speaking, tasks are used to elicit students' performance (e.g., students' item responses) upon which inferences about students' domain specific knowledge, skills, and abilities are drawn (Mislevy, Steinberg, & Almond, 2003). However, it is often the case that in a subject-matter assessment, a single task may tap into an additional set of skills that are not part of the domain, but rather construct-irrelevant skills that influence item responses. Thus, it is important to understand what task-related factors are related to construct-irrelevant skills so that we can come to a better interpretation of the construct that the assessment is intended to measure (i.e. domain specific knowledge, skills, and abilities). In the context of this review, reading comprehension is considered to be partially a construct-irrelevant skill in science and civic-related social studies assessments when it influence students differentially and leads to underestimation of individual examinee scores. In the following section, I will review how some task specific factors may be associated with students' comprehension of assessment items in a subject specific domain.

Item Format. A typical standardized subject-matter assessment can provide students with two types of response formats: multiple-choice and constructed-response. A constructed-response usually requires students to write their own answers. Multiple-choice, on the other hand, provides students a list of suggested response options which may include words, numbers, symbols, or phrases (Linn & Miller, 2005). Some researchers believe that constructed-response items can elicit students' higher-order cognitive abilities such as reasoning, analytical skills, and problem solving skills. But due to its subjectivity, constructed-response items are often harder to grade and can result in relatively low reliability. Standardized assessments usually favor multiple-choice items because multiple-choice tests are viewed as potentially more fair to individual test takers, since they are given a standard set of response options, and the correct answer is predetermined. Second, because multiple-choice questions can be answered quickly, more questions can be included in a single test, thus maximizing coverage of the domain being assessed. Third, the scoring procedure is relatively easy and more reliable (Bennett & Ward, 1993; Campbell, 1999; Haladyna, 2004). Limitations also exist for multiple-choice questions. An often cited criticism is that the diversity in prior knowledge and human experience across individuals can allow many possible answers to fit a question. This type of item forces students to choose among predetermined answers when other, more plausible options may exist or when a particular unspecified aspect of the situation may determine which answer is best. Consequently, multiple-choice test items can result in test performance that may reflect the extent to which students are able to construct meaning from text but may not fully reflect the students' subject-matter knowledge (Campbell, 1999; Ozura, Row, O'Reilly & McNamara, 2008; Pearson & Valencia, 1987).

Researchers have used both experimental and correlational approaches to examine the effects of question formats on reading comprehension. Experimental studies (e.g., Campbell, 1999; Cordon & Day, 1996; Karabenick, et al., 2007; Reich, 2009; Schoultz, Säljö, & Wyndhamn, 2001) have employed think-aloud procedures and asked participants (grade levels range from 7 to 11) to describe their thoughts while answering multiple-choice or constructed-response questions. Some studies have not detected differences in the cognitive processes underlying constructed-response and multiple-choice question responses (e.g., Campbell, 1999; Cordon & Day, 1996; Rodriguez, 2002).

Some correlational studies use factor analysis to examine the amount of common variance in reading comprehension shared by multiple-choice and constructed-response format questions. For example, Bridgeman and Rock (1993) performed a factor analysis on the Graduate Record Exam (GRE) analytical section and found no significant differences between the two formats with respect to their ability to measure factors (i.e., logical reasoning and analytical reasoning) underlying reading comprehension processes. In summary, results from these studies imply that the claim that only constructed-response items could elicit higher-order thinking skills may not be true. If written properly, multiple-choice items can evoke higher-level cognitive processes such as understanding, prediction, evaluation, and problem solving, at least when reading itself is the topic of interest (DeMars, 1998; Haladyna, 2004; Martinez, 1999).

In terms of reading comprehension in subject-matter assessments, Katz, Bennet, and Berger (2000) studied the influence of reading comprehension of stem-equivalent multiple-choice and constructed-response items on a set of 10 mathematics items from the SAT. In this study, 55 high school students were asked to think aloud about their problem-solving strategies after reading items. Results suggested that reading comprehension mediates format effects for problem-solving strategies as well as item difficulty. The researchers concluded that reading comprehension may have an overarching impact on students' item performance in the SAT math test.

After a review of assessment studies, Haladyna (2004) concludes that the results of item format depend on the nature of the construct. If a construct is knowledge based (e.g., students' basic conceptual knowledge about laws, political rights), the use of either multiple-choice or constructed-response format will yield in highly reliable scores. If a construct is skill based (e.g., reasoning and analyzing controversy in political opinions), responses on constructed-response items are often more trustworthy. However, multiple-choice items might serve better in a test because they have greater efficiency and can yield higher criterion validity when correlated with a measure with more fidelity. In addition, Haladyna (2004) suggests:

> The choice of an item format mainly depends on the kind of learning outcome you want to measure. If a domain knowledge or skill is conceptualized, the main validity concern is the adequacy of the sample of test items from this domain. Multiple-choice format provides the best sampling from the domain because the

> format allows more units of measurement/ wide range of coverage to the domain. (p. 62)

Item Alternatives. A standard multiple-choice test item is usually composed of two parts: a stem and a list of alternatives. A stem is an introductory statement that either asks a question or poses a problem, and it is often in the form of a question or an incomplete statement. Alternatives are solution options made of a single-correct or -best response to the question (answer) and several incorrect or inferior solutions (distractors). The purpose of the distractors is to appear plausible for those students who have not mastered the content being measured by the test item, but implausible for those who have achieved mastery of content. Furthermore, the correct answer should be the only plausible solutions to these students who have mastered the content (Burton, Sudweeks, Merrill, & Wood, 1991).

In a review of the effect of item alternatives on reading comprehension, Gorin (2002) concludes that some language variables embedded in item alternatives are correlated with item difficulty in standardized reading comprehension assessments (e.g., TOEFL, GRE-verbal section). These language variables include:

- Percent of content words (including verbs, nouns, adverbs, and adjectives) in the total text. The items contain more content words are assumed to be more difficult to process than the item with few content words. As the amount of content words associated with answering a question increases so does the demand on memory and cognitive process, which may lead to an increased item difficulty.

- Vocabulary level of the alternatives. According to a memory-type theory of processing (Sheehan & Ginther, 2001), an alternative is most likely to be selected as the correct answer when it is most highly activated in the individuals' mind. It is assumed that frequent words are more likely to be processed and activated in an examinee's long-term memory than infrequent words. Therefore, if the correct answer consists of frequent words, then the item is easier than those composed of infrequent words. Conversely, if the distractors are made up of infrequent words, then these distractors may not be processed by many examinees. Hence the item is easier than those that consist of alternatives with frequent words.
- The lexical and semantic similarity between the stem and the alternatives (in the multiple-choice format). Based on previous research using reading comprehension assessments (e.g., Embretson & Wetzel, 1987; Sheehan & Ginther, 2001), Gorin (2002) summarizes that when information highly elaborated in the stem appears in one of the alternatives then this information will be highly activated. High activation for the correct answer may decrease item difficulty, and high activation in distractors may increase item difficulty.

Item Question. The role of questioning in understanding and learning instructional texts has been explored using laboratory experimental research methods since 1960s, (see reviews by Allington & Weber, 1993; Anderson & Biddle, 1975; Memory, 1982; Pressley & Forrest-Pressley, 1985; Kintsch, 2005). Findings were often contradictory, and varied considerably depending on where the questions were located (before, interspersed, or after the instructional text), and what type of assessments were used to assessing learning outcomes. For example, learning outcomes were often measured in terms of how well learners remembered the text content. In addition, positive results were usually obtained for questions that targeted specific facts and required readers to recognize and recall facts or details from the text. Eileen Kintsch (2005) criticized this approach, and pointed out that these previous studies on questioning only measured shallow levels of comprehension processes (i.e. the surface level and textbase level in terms of Kintsch's model of comprehension), and the measures stemmed from a narrow view of learning which equated learning with memory. Eileen Kintsch (2005) provided guideline for formulating questions that map different levels of reading comprehension processes within the framework of Walter Kintsch' comprehension theory. A brief version of the guidelines is:

> Questions that require readers to recognize or recall facts or details from the text tap shallower levels of comprehension (e.g., specific facts or definitions of terms). Questions that ask learners to summarize or recall the gist of the content probe macro-level understanding of the text. Deeper level questions are those that probe a learner's ability to use the text content to solve problems, analyze relationships, or form connections among ideas in the text (Graesser & Person, 1994). In order to answer this type of question, learners have to form a mental model of the situation depicted in the text. That is, the situation model (Kintsch, 2005, p. 54).

Kintsch's (2005) guidelines were proposed for instructional purposes. The effectiveness of the questioning has not been consistently examined in the subject-matter assessments or with adolescent students. Large-scale subject-matter assessments such as

IEA TIMSS and IEA CIVED were designed to assess subject-matter knowledge including declarative knowledge and procedural knowledge. Students who took the tests were expect to use their own knowledge or skills to solve problems, analyze relationships, or make inferences based on what the test item asks. Therefore, item questions should focus on eliciting students' situation model in the domain being measured. Eileen Kintsch's guidelines can be used, however, to evaluate if item questions function the way they were supposed to in these large-scale subject-matter assessments.

Summary. In the previous sections, I reviewed factors inherent in the text, and dependent on the reader. In this section, I focus on task specific factors that affect how students interpret assessment items. Research results in general suggest multiple-choice items can be written in a way that is similar to constructed-response format items in terms of evoking cognitive processes such as understanding, prediction, evaluation, and problem solving. In a multiple-choice test, some language variables embedded in item alternatives are associated with item difficulty. For example, information most highly activated in an individual's mind is more likely to be processed. Hence the alternative that contains this information is more likely to be selected as a correct answer (even though it may be distractor). In addition, the amount of information that test takers have to process in an item contributes to item difficulty. The more demand on memory and cognitive process, the harder the item is. Finally, it is worth paying attention to the questions (which are usually located in item stems for multiple-choice items). Previous research suggests that how questions are asked in an item can influence students' understanding of the item as well as their test performance.

2.4 Group Differences in Reading

Many reader-related factors vary among readers, and between groups. Previous studies show that reading comprehension can vary with students' gender, SES status, and language background (i.e., Native speakers vs. English as Second Language learners). Due to the scope of this paper, I will focus on reviewing three factors: gender, language background, and home resource.

2.4.1 Gender

Results of previous meta-analysis research revealed a clear pattern of gender differences in reading. That is, on average girls tend to have higher reading skill than boys across grades (e.g., Hyde & Linn, 1988; Lietz, 2006; Ryan & DeMark, 2002). A large number of national and international assessments in reading, including NAEP, Programme for International Student Assessment (PISA), and the International Association for the Evaluation of Educational Achievement (PIRLS), confirm the reading advantage of girls over boys across grades. For example, Lynn and Mikk (2009) analyzed reading assessment data from the IEA PIRLS 2001 study, the PISA 2001, 2003, and 2006 study. Among these national represented samples (most of them were 15 years old), all four large-scale assessments found on average girls achieve higher reading achievement scores than boys. In terms of variation within gender, all four reading assessments showed that boys had greater variance in reading comprehension than the girls in all countries. One explanation of girls' higher achievement in reading is their deeper engagement in language related activities.

Other studies also showed that boys and girls have different preference for what to read. Generally, boys enjoy so called "masculinity" genres and topics including news, sport pages, science fiction, and special-interest books, whereas girls like narrative texts such as modern or classic fiction, romance stories, or song lyrics (e.g., Baker & Wigfield, 1999; Canadian Council on Learning, 2009; Guthrie, Wigfield, & Klauda, 2012; Logan & Johnston, 2009; Young & Brozo 2001).

Individuals can vary as a function of age and type of abilities within each gender group. In the meta-analysis of gender differences in verbal abilities in the United States, Hyde and Linn (1988) revealed that in reading comprehension, girls below the age of six outperformed boys, but among older children the gender differences were not very salient. In vocabulary, 6–10 year old girls outperformed boys, but among 11–18 year olds there was no gender difference. However, when it came to 19–25 year olds, men performed better than women. Analysis of gender differences by type of ability showed that women have higher average abilities than men in word fluency; men have higher average abilities in analogies. There were negligible gender differences detected in terms of reading comprehension, essay writing, and vocabulary. This included many types of studies and formats of measures, which are both a strength of meta-analysis and a weakness in applying it to a narrow area like that discussed in this section.

2.4.2 Language Background

According to the Test Standards, "…any test that employs language is, in part, a measure of the language skill" (p.91) of the examinee. Hence linguistic considerations are particular critical for test takers with diverse language backgrounds (Linn, 2002).

Previous research suggested that unnecessary linguistic complexity at the item level may hinder students (especially students with limited English proficiency) to demonstrate their knowledge of the construct being measured (e.g., Abedi, 2009; Abedi

& Gándara, 2006; Solano-Flores & Trumbull, 2003). A series of studies conducted by Abedi and his colleagues (e.g., Abedi, Hofstetter, Baker, & Lord, 2001; Abedi, Lord, Hofstetter, & Baker, 2000) have examined the influence of language complexity on students' performance in large-scale subject-matter assessments (e.g., NAEP mathematics and science assessments) with students' language background taken into account. Overall, results of their studies show that English language learners (ELLs) on average achieved lower scores than non-ELLs, particular on long and linguistically complex items. Compared to non-ELL students, test results from the ELLs showed more items omitted or not reached. In addition, the performance gap between ELLs and non-ELLs increases as the level of reading demand on the assessments increases. This implies that the language of the assessment can introduce construct-irrelevant variance that compromises the validity of interpretations and uses of the test scores when the focus of the test is other than literacy skills.

Abedi and his colleagues have identified several linguistic features that contribute to the difficulty of comprehending test items, including unfamiliar (or infrequent) vocabulary, complex grammatical structures, and styles of discourse that include extra material, abstractions and passive voice (for a more detailed descriptions of these features, see Abedi, 2009; Abedi et al., 1997 in reviews). According to Abedi, these features may increase the likelihood of misinterpretation, and add cognitive load to test takers, therefore they are likely to interfere with the measure of constructs in subject-matter. One way to minimize the impact of reading demand on students' test performance is reducing the level of unnecessary linguistic complexity of the assessment. In fact, evidence from experimental studies (e.g., Abedi, Courtney, & Leon, 2003; Abedi, Hofstetter, Baker, &

Lord, 2001; Abedi & Lord, 2001; Abedi, Lord, &Hofstetter,1998) suggests that students with limited English proficiencies benefited the most from the simplification of linguistic characteristics of test items in large-scale mathematics tests and science tests.

For example, Abedi and Lord (2001) studied the effect of linguistic complexity on eighth grade students' performance on NAEP mathematics items using an experimental method. In their study, two parallel forms of test items were randomly assigned to 1,031 eighth grade students in Southern California. The two parallel forms of test items consisted of the original version of NAEP items (with some items linguistically complex), and a modified version of items which were modified to reduce complexity of sentence structure, and to replace potential unfamiliar vocabulary (non-content words) with words that were likely to be more familiar to the students. The mathematical tasks and content words in the modified version were not changed. Test results show that on average ELL students had significant higher test scores on the modified items where the linguistic complexity of the item was reduced. The linguistic features that contribute to the difference are low frequency vocabulary and verbs in passive voice.

Another study (Abedi, Courtney, & Leon, 2003) tested 1,854 Grade 4 students and 1,594 Grade 8 students in 132 classes at 40 school (49.7% of students were ELLs) using NAEP science items and a few TIMSS multiple-choice items. Each student was provided with one of the four accommodations: a bilingual glossary, an English dictionary (words were customized and selected directly from test items), a modified test where the linguistic complexity of the items were reduced, or the standard test items. Results show that only the linguistically modified test items enhanced the ELL students' scores without impacting the non-ELLs' scores.

In conclusion, evidence from previous studies indicates that unnecessary linguistic complexity may affect validity of the subject-matter assessment outcomes and increase the achievement gap between ELL and non-ELL students in terms of their performance in subject-matter assessments. Abedi (2009) explains:

> It is extremely difficult for ELL students to understand test items that are complex in their linguistic structure. In such cases, ELL students with a fair level of knowledge of the content may not perform well not because of lack of content knowledge but because of difficulty understanding the assessment questions (p. 16).

Reading demands on subject-matter assessments have raised concerns about fairness for some groups of test takers. The Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1999) provide an entire chapter addressing the problems of students with diverse language backgrounds. The Standards urge cautions in test score interpretation and use when the reading demand of the test exceeds that linguistic abilities of the test takers. For example, the following standard:

> Standard 7.7: In testing applications where the level of linguistic reading ability is not part of the construct of interest, the linguistic or reading demands of the test should be kept to the minimum necessary for valid assessment of the intended construct. (AERA et al., 1999, p. 82)

To summarize, statistics in large-scale subject-matter assessments indicate achievement gaps in terms of gender, language and family background, etc. Identifying background factors that affect the performance gap may help gain insight into the nature of the construct in subject-matter assessments, and close the gaps.

2.4.3 Home Literacy Resources

Numerous studies found home literacy resources contribute to children's cognitive development including reading competence. Home literacy resources were often conceptualized as a construct that reflects the degree of exposure of children to literacy resources at home. The construct can entail the amount of time that children spend on shared reading, literacy activities, parents' literacy proficiencies, and family income. Some studies measure the construct by counting the number of books in children's homes.

Studies found that children's home literacy resources, word knowledge, reading comprehension, and conceptual knowledge were reciprocally related to each other (Leseman & de Jong, 1998; Lugo-Gil & Tamis-LeMonda, 2008; Sénéchal & LeFevre, 2002; Stanovich, 1986). For example, Sénéchal and LeFevre (2002) conducted a longitudinal study that followed 168 Canadian 4- and 5-year-old middle-class children for five years. They explored the relations among children's early home literacy experience, literacy skills, and reading achievement. They found that exposure to books was associated with children's vocabulary knowledge and listening comprehension. These language skills then predicted children's reading achievement in grade three. In addition, early literacy skills predicted children's word knowledge in grade one. Various relations they found in this study confirmed the reciprocal relation among home literacy resources (including home literacy activities that involve parents), lower level reading abilities (e.g., phonological awareness, word knowledge), and reading achievement.

2.5 Gauging the Impact of Reading Comprehension on Item Difficulty

According to literature reviewed here, the nature of the text and task specific features associated with reading comprehension can have an impact on item difficulty (which usually refers to the proportion of students who provided a fully correct response to a test item). Traditionally, psychometricians utilize statistical methods such as multiple regression, factor analysis, differential item functioning, and classification and regression tree analyses to estimate the contribution of various reading-related variables on item difficulty (Anastasi & Urbina, 1997; Oakland & Lane, 2004; Sheehan, 1997). Through these methods, variance associated with reading comprehension in specific subject areas is gauged in terms of the relation of item difficulties to reading demands of test items (e.g., vocabulary difficulty, complexity of sentences, text genre).

For example, Embretson and Wetzel (1987) developed a cognitive processing framework to analyze items contained in a large-scale reading comprehension assessment, which asked questions about text passages (i.e., the Army Services Vocational Aptitude Battery). Using a latent trait model called the linear logistic latent trait model, they examined item difficulty was analyzed in terms of various text features (e.g. number of words per sentence, Flesch's reading grade level, sentence length, and percent of content in the question stem and alternatives) and task specific features (e.g. the properties of the question stem, the response alternatives of items). Their results showed that several main variables influenced item difficulty in reading comprehension including percent of content words, the propositional density of the passage in the item stem, and the extensiveness of the inference required to map the question and answer onto the text passage. This study was replicated by Gorin and Embretson (2006) to analyze multiple-choice items in the Graduate Record Examination (GRE)-verbal session. Their findings suggest that item difficulty of GRE-verbal test was explained primarily by the extensiveness of the inference required to map the question and answer onto the text passage.

Similar analyses were performed by Ozura, Rowe, O'Reilly, and McNamara (2008) using hierarchical linear modeling (HLM) on reading comprehension items from the Gates-MacGinite Reading Test (GMRT) for the 7th-9th and 10th-12th grade levels. They applied the cognitive processing framework developed by Embretson and Wentzel (1987) to analyze 192 comprehension multiple-choice items from the GMRT. They estimated text features (e.g., number of propositions, number of words per sentence, word frequency) through the software Coh-Metrix, and coded item characteristics (e.g. the properties of the question stem, the response alternatives of items) using coding schemes developed by Embertson and Wetzel (1987) and Mosenthal (1996). The relation between item difficulty, text features, and item characteristic were examined through HLM. Their results were consistent with the previous studies. In addition, they found that the difficulty of items in the test for the 7th-9th grade level is primarily influenced by vocabulary difficulty—in particular, word frequency. On the other hand, the difficulty of items in the test for the 10th-12th grade level was not predicted by text features or item characteristics to a statistically significant extent.

Other studies that employed statistical item analyses on item difficulty showed that text characteristics such as sentence length, word frequency, type-token ratio (the number of unique words in a text [i.e., types] divided by the overall number of words [i.e., token]), and the degree of overlap between the text passages and test items had an impact on item difficulty (Oakland & Lane, 2004). Furthermore, task specific features such as number of plausible distractors, lexical overlap with distractors, and item type also affected the difficulty of reading comprehension items (Rupp, Garcia, & Jamieson, 2001).

2.6 Readability Methods

Readability methods also have been used to estimate the effect of reading demand on test difficulty.

2.6.1 Readability Formulas

Dale and Chall (1949) define readability as "the sum total (including the interactions) of all those elements within a given piece of printed material that affects the success a group of readers have with it. The success is the extent to which they understand it, read it at an optimum speed, and find it interesting" (p 23). Readability formulas (e.g. Dale & Chall, 1948; Flesch, 1951; Fry, 1968; Gunning, 1968; Spache, 1953) have been designed to identify the reading level of a text passage (with typically three or more sentences). These formulas tend to rely on two quantitatively measured qualities: vocabulary (e.g., word familiarity, number of letters or syllables within a word), sentence length, and paragraph length, (Oakland & Lane, 2004 in a review).

However, this readability approach has received criticism from educators and cognitive psychologists (e.g., Bertam & Newman, 1981; Helwig, et al., 1999; Kintsch & Kintsch, 2005; Oakland & Lane, 2004). Kintsch and Kintsch (2005) point out that this approach does not reflect cognitive psychologists' understanding of the comprehension process because the approach treats reading comprehension as an uni-dimensional and static construct, and only puts emphasis on surface-level difficulty of text such as sentence length and word difficulty but ignore factors such as cohesion, complexity of ideas, and required schemata.

2.6.2 Coh-Metrix

Theoretical advances in computational linguistics and discourse processing led to development of new tools for analyzing the difficulty of text. Coh-Metrix is one of the advanced computer tools. It was developed by Graesser, McNamara, and their colleagues based on Kintsch's reading comprehension theory (Graesser, McNamara, Louwerse & Cai, 2004; McNamara, Louwerse, McCarthy, & Graesser, 2010). Coh-Metrix measures text difficulty at various levels of language, discourse, and conceptual analysis and adjusts the output according to the targeted reader. It analyzes and measures text on the first five levels of discourse: *words, syntax, textbase, situation model,* and *genre* in addition to those measured by readability formulas. This computer tool can provide more than two hundred cohesion and readability measures considered to influence comprehension. The wealth of information provided by the Coh-Metrix about the textual features of passages also challenges researcher to decide which text feature are more relevant to their research topics (Elfenbein, 2011). One way to solve the problem is to conduct exploratory multiple regression analyses to find subsets and combinations of Coh-Metrix variables for a more parsimonious predictor of text or item difficulty. Using relevant reading comprehension theory or framework as guidance to carefully select Coh-Metrix variables is another way to deal with it. As a relatively new technique (in a field where many of the studies are decades old), this approach should be examined carefully for both its strengths and weaknesses.

2.7 Conclusion and Discussion

It is a common consensus that traditional assessments (especially paper and pencil tests) depend on language and literacy skills to a certain degree. In large-scale subject-matter educational assessments such as IEA's TIMSS and Civic Education, assessment items are written in a way to elicit test takers' content knowledge and skills. A prerequisite is that test takers must be able to recognize and understand the situations expressed in an item. Thus a reasonable level of reading comprehension has been acknowledged as part of the construct. However, if the reading demands of such assessments exceed the level of test takers' reading ability, reading comprehension may prevent students from demonstrating their true abilities (e.g., domain specific knowledge, abilities, and skills). In such circumstances, reading comprehension poses threats to interpretations of students' performance in the subject-matter assessment. In other words, it becomes a source of construct-irrelevant variance. The validity of the assessment is therefore in question.

The overall purpose of the current review is to understand how students comprehend test items in specific subject domains including science and civic-related social studies. Particularly, this review focuses on identifying text features and item characteristics that can be understood as providing affordances to facilitate the reader constructing accurate representations of text in subject-matter assessments.

Advances in cognitive theories such as Kintsch's the construction-integration (CI) theory have provided a feasible framework to model the nature and characteristics of comprehension processes when students read assessment tasks. According to the CI theory, reading comprehension involves complex and multilevel cognitive processes "that integrate information from the text that the students is reading with his or her background knowledge and experiences, subject to a multitude of contextual constrains" (Kintsch & Kintsch, 2005, p. 71). Kintsch (1998) suggests that a successful comprehension of the text depends on a variety of factors. Reader factors, text factors, and context factors all play a role.

Among the factors that depend on readers, it appears that decoding skills, knowledge of vocabulary (general and specific), and subject-matter knowledge can be identified as the most essential for a successful comprehension in subject-matter assessments, because these factors facilitate the reader in making sense of text by constructing mental representations and integrating information from the text into the representations. In addition, previous research found that prior knowledge and literacy skills (including reading comprehension and reading skills) can compensate for one another to a considerable extent. One implication for subject-matter assessments is students with high prior knowledge, especially high subject-matter knowledge, are more likely to comprehend texts and remember them better than those with low subject-matter knowledge when other variables such as reading skills, motivation, and text feature are controlled (Kintsch & Kintsch, 2005). This may apply especially to test questions that have extensive introductory materials or scenarios. However, low reading abilities, including poor decoding skill, have been shown by empirical studies to prevent some students from demonstrating their subject-matter knowledge over a variety of types of items. Hence it is important for assessment designers to evaluate and judge the extent to which the linguistic complexity of test items match with test takers' reading abilities. If reading comprehension cannot be justified as a vital part of the construct being measured, accommodations, such as providing the read-aloud accommodation, should be considered for students who are low in reading and decoding skills.

Providing a separate assessment section to test students' reading abilities and/or knowledge of vocabulary relevant to the subject-matter content, can be another way to evaluate the influence of reading ability on subject-matter knowledge. For example, the IEA Civic Education project conducted a large-scale international study in 1971 to assess young people's civic-related cognitive achievement and democratic attitudes (e.g., tolerance, support for civil liberties) (Torney, Oppenheim & Farnen, 1975). Participants were 30,000 adolescents (10-year-olds and 14-year-olds) from ten countries including the U.S.. In addition to the assessment of civic-related cognitive achievement, the researchers designed a separate scale measuring students' general vocabulary (synonyms and antonyms). Utilizing a dataset of nine countries, Schwille (1975) conducted a correlational study to investigate how factors related to students' home background, word knowledge, learning condition, and students' attitudes and interests predict their overall civic educational achievement. The multiple regression results showed that students' general word knowledge accounted for a substantial portion of variance in civic achievement across countries. This indicates that including a scale that measures an important component of reading ability may make possible a better prediction of performance on a cognitive subject-matter test. However, this analysis was conducted with different set of goals in mind and without the sophisticated analytic tools that are now available. This topic deserves further exploration.

Working with text features and context factors is another way to reduce the reading demands associated with construct-irrelevant variance in a subject-matter assessment. Abedi and his colleagues conducted a series of research on the effect of linguistic factors on English language learners’ performance on subject-matter assessments. Their results collectively suggest that reducing the linguistic complexity of the subject-matter assessment helps to provide a more valid assessment outcome for English language learners as well as native speakers of English at the lower tail of the academic achievement distribution (Abedi, 2009).

In general, recent research on reading comprehension assessment has incorporated a cognition-centered approach to text processing and comprehension (Embretson, 1998; Mislevy, 1994, 1995, 1999). One example is Mislevy’s evidence-centered design. This approach starts with defining what the test intends to measure (i.e., the construct) with a cognitive model that specifies students’ representations of a domain in terms of requisite knowledge, skills, and abilities (KSA) (Mislevy, Steinberg & Almond, 2003). The approach then decomposes a task (i.e., taking a science test) into a processing model (Embretson, 1998) and then examines the contribution of particular text processes and task features (including text specific features) to item responses. The results of this analysis potentially help to identify which types of task features (e.g., item format, proposition density, and sentence length) contribute most to the difficulty level of the tasks. Several studies based on this approach were conducted on reading comprehension or mathematics assessments (Abedi, et al., 2000; Embretson & Wetzel, 1987; Gorin & Embretson, 2006; Gorin, 2005; Ozura, et al., 2008). Overall, these studies have collectively shown that this type of theory-based analysis of test items provides useful information about the variability in test takers’ reading comprehension as measured by these tests. Based on results from these studies a subset of specific reading comprehension features at the item level have been identified as potential contributors to reading comprehension: unfamiliar (or less commonly used) vocabulary, sentence length, complex grammatical structures, and styles of discourse that include extra material, abstractions and passive voice.

This new cognitive-psychometric approach provides a promising framework to modeling reading comprehension processes. In spite of that, there has not been a link between this approach and international large-scale subject-matter assessments such as IEA TIMSS, IEA CIVED, IEA International Civic and Citizenship Education Study (ICCS), and PISA. In order to fill the gap, this study integrated the cognitive literature on reading comprehension into a processing model to be tested in subject-matter assessments. The quality of items on the subject-matter test was assessed based on the relationship of the item difficulty to the processing model (Embretson & Wetzel, 1987; Gorin & Embretson, 2006). Some psychometric methods, such as multiple regression, and item response theory models (Embretson, 1998; Tatsuoka, 2009; von Davier, 2008), allow for statistical modeling of associations predicted by cognitive theories.

Published research on international large-scale assessments such as TIMSS and CIVED examining the influence of reading comprehension on performance is generally lacking. Research is needed that employs modern cognitive models and psychometric methods to enhance the current state of knowledge regarding the role of reading comprehension in large-scale subject-matter assessments. Results of this research could help to identify which types of comprehension-related item features contribute most to the difficulty of items and to the relatively poor performance of some groups of individuals. This information, in conjunction with cognitive theories of text processing and comprehension, can afford researchers and educators greater insight into the types of cognitive processes that are tapped by assessment items. This research can also provide insight for test makers who want to maintain overall test discrimination and test sensitivity in domain proficiency assessments without compromising the theoretical validity of the assessment.

Chapter 3: Introduction to Method and Contemporary Test Design Framework

Designing and developing large-scale tests based on the science of human learning and cognition has been more and more appealing to educators, researchers, and practitioners (Leighton & Gierl, 2011). The 2001 National Research Council’s (NRC) report *Knowing What Students Know: The Science and Design of Educational Assessment* (KWSN; NRC, 2001) lays out a multidisciplinary assessment design approach that centralizes the role of cognition (including theories and methods of cognitive psychology) as the foundation and guiding ruler for test design and development. One of the test design frameworks that adopts this contemporary approach is Mislevy’s evidence-centered design (ECD, Mislevy, 2004; Mislevy, Steinberg, & Almond, 2003). I will review ECD in the following section, and discuss how to utilize this framework to guide my research design and interpretation in the next section.

3.1 Evidence-Centered Design

Evidence centered assessment design (ECD) was originally formulated at Educational Testing Service (ETS) by Mislevy, Steinberg, and Almond (2003) and may be seen as part of a long-standing tradition in educational assessment that revolves around validity arguments (Cronbach & Meehl, 1955; Kane, 1992; Messick, 1989, 1994; Spearman, 1904). ECD builds on developments in fields such as expert systems (Breese, Goldman, & Wellman, 1994), software design (Gamma, Helm, Johnson, & Vlissides, 1994), and legal argumentation (Tillers & Schum, 1991) to provide tools for building explicit assessment arguments that assist test designers in designing new assessments and understanding familiar ones (Mislevy & Riconscente, 2005). The ECD framework attempts to apply principles of evidentiary reasoning to handle the complexities of the validity argument associated with contextual features including item characteristics and text features in an assessment. The key idea is to lay out the assessment argument in evidentiary statements and structures. For example, Mislevy (1995, 1997, 2009) suggests that an assessment argument can be summarized as comprising: (a) a claim about a person possessing at a given level a certain targeted proficiency, (b) the data (e.g., test scores) that would likely result if the person possessed a certain level of the targeted proficiency, (c) the warrant (or rationale, based on theory and experience) that tells why the person's level in the targeted proficiency would lead to occurrence of the data, and (d) "alternative explanations" for a person's high or low test scores. Significant and credible alternative explanations might indicate that test validity is threatened (Messick, 1989). It is this fourth aspect of the theory that is most directly involved in this study.

Three pillars in the ECD serve as cornerstones of the framework: student model, task model, and evidence model. The Student Model contains cognitive and learning theories in regard to how students develop competence and represent knowledge in the subject domain. This model is usually based on empirical studies of students in a domain. The second model describes the tasks or situations that allow one to observe students' performance. The third model is an interpretation model that corresponds to the cognitive theory or learning theory in the student model. The model contains measurement (psychometric) models that represent a particular form of reasoning from evidence. For example, if a cognitive theory characterizes students' achievement as multiple dimensional rather than a single score, contemporary multidimensional statistical models such as Structural Equation Model or multidimensional IRT model may serve the purpose of the assessment guided by the theory. These measurement models provide explicit and formal rules for integrating the pieces of information from test tasks. In this model, assessment tasks, along with the criteria for evaluating students' responses, are carefully gauged as to what degree they elicit the knowledge and or cognitive process that the student model suggests are most important for competence in the domain.

Based on this ECD approach, the National Research Council (Pellegrino, Chudowsky, Glaser, & National Research Council (U.S.), 2001) urges researchers and test designers to conduct research to evaluate test tasks (items) that tap relevant knowledge (e.g., reading comprehension) and cognitive process (e.g., metacognitive strategies) through analysis of error (e.g., Gorin & Embretson, 2006; Mislevy, Steinberg, Breyer, Almond, & Johnson, 1999; Tourangeau, 2003) or cognitive interview (e.g., Karabenick, et al., 2007). The Council points out that "conducting such analysis early in the assessment development process can help ensure that assessments do, in fact, measure what they are intended to measure" (p. 7).

ECD has recently been extended by Hansen, Mislevy, Steinberg, Li, & Forer (2005) to integrate cognitive and learning theories and reason about how construct validity is affected by assessment misalignment (e.g., excessive amount of cognitive demand at the item level). This extension includes structures that provide scaffolds for test designers to use in reasoning and evaluating some core validity issues, specifically, the alignment between (a) what one intends to measure (i.e. construct) and (b) what one is actually measuring in an assessment. Generally speaking, an alignment is associated with positive evidence (e.g., construct-relevant variance) for test validity, and a misalignment is associated with negative evidence (e.g., construct-irrelevant variance and construct under-representation) that is likely to compromise the test validity. This approach has been applied to state-wide large-scale assessments in science (Haertel, et al., 2010; Zhang, et al., 2010). Using this framework, Mislevy and Yin (2009) also incorporate Kintsch's model in designing and evaluating assessment items in language and literacy. This approach is closely related to model-based approaches, such as those described generally in the next section and used specifically in the remainder of this chapter.

3.2 Cognitively-Based Statistical Models

3.2.1 Item Response Theory Models

A variety of statistical methods have been used in large-scale educational assessments to estimate students' cognitive proficiencies and item statistics such as item difficulty. One of the most commonly used methods is item response theory (IRT) modeling, a probability-based latent variable modeling approach that models individuals' item response patterns with item level characteristics (e.g., item difficulty, item discrimination) taken into account. A fundamental feature of IRT is that individuals' responses (incorrect and correct response) to test items are related to a unidimensional latent attribute θ, a statistical construct. In cognitive tests, the latent attribute is often called ability, skills, or proficiency measured by the test.

IRT models specify a relation between the latent attribute (latent variable) being measured and likelihood of specific observed responses in test performance. This relation usually can be described by up to three item parameters: item difficulty, item discrimination, and a pseudo guessing parameter (a parameter reflecting the probability that an examinee with a very low attribute level will correctly answer an item solely by guessing). IRT models have been widely used in the areas of measurement and testing to estimate individuals' possession of a latent attribute, such as cognitive skill or academic achievement In an IRT model, a latent attribute is inferred based on students' item responses and characteristics of the items in a test (Embretson & Reise, 2000). Compared with traditional methods such as Classical Test Theory, IRT models can provide more reliable estimates of test scores and more sophisticated information with respect to individuals' abilities and the quality of test items (Magno, 2009). In addition, as a latent variable modeling approach, IRT analyses provide model fit statistics that allow comparisons between models. Evidence regarding the fit of the model to data can be collected to support the interpretation and use of test scores, and to assist in the evaluation of test validity in the sense that the evidence-centered design requires. IRT models have been commonly used in large-scale educational assessments such as NAEP, IEA TIMSS and IEA Civic Education Study.

The early IRT applications were based on the assumption that the parameters describing examinees vary only on one dimension (Lord & Novick, 1968; Rasch, 1960). However, it quickly became evident that this assumption was often violated. For example, a standardized science assessment nowadays often tests multiple dimensions presenting students' multiple attributes such as science conceptual knowledge and procedural skills for solving problems. These dimensions by nature may be correlated with each other and/or hierarchically ordered. In this case, a simple unidimensional model may not be sufficient for describing the multivariate latent dimensions (attributes) being measured. In attempting to solve this problem, researchers have developed more advanced IRT-based statistical models such as multidimensional item response theory models. This advanced model permits multiple attributes to be estimated simultaneously and then compared within tests or within sets of items. I will review the multidimensional model in the following section.

3.2.2 Multidimensional Item Response Theory Model

Multidimensional IRT (MIRT) models are designed to overcome the dimensionality limitations of the traditional one-dimensional IRT models. As an extension of traditional IRT models, MIRT models can estimate multiple latent variables simultaneously based on students' item responses while taking into account item characteristics such as the difficulty and discriminating power of each test item. Conceptually speaking, multidimensional IRT can be viewed as a special form of confirmatory factor analysis (CFA) with multiple latent factors when the observed variables, the item response, are discrete. This type of model provides model fit indices as the CFA does. It allows researchers to conduct model comparisons and to investigate how well the model fits the data. It provides further evidence about the validity of the construct that the researcher examined.

More than a dozen MIRT models have been designed with this purpose (see Reckase, 2009 for a general review of the MIRT models). One of the MIRT models is General Diagnostic Model (von Davier, 2005). The GDM is not a single model. Instead, it is an overarching model framework that contains many logistic-type models including the one-parameter logistic model (Rasch, 1960), two-parameter logistic model, multidimensional IRT models, and cognitive diagnostic models.

Here is a statistical description of the GDM (von Davier, 2005, p. 6).

$$P(X = x \mid \beta_i, a, q_i, \gamma_i) = \frac{\exp\left[\beta_{xi} + \sum_{k=1}^{K} x\gamma_{ik}q_{ik}a_k\right]}{1 + \sum_{y=1}^{m_i} \exp\left[\beta_{yi} + \sum_{k=1}^{K} y\gamma_{ik}q_{ik}a_k\right]}$$

where

x = dichotomous or ordinal responses for each test item,

K = number of student proficiency variables,

k = the index variable for proficiencies,

i = the index variable for items,

q_{ik} = an entry in the Q-matrix (defined below),

a = discrete score determined before estimation and can be chosen by the user,

β_{yi}= item difficulty, for response category y of item i, and

γ_{ik} = slope parameter, for item i with respect to proficiency k.

The key component of the GDM is a_k, a latent variable with discrete user-defined skill levels with $a_k \in \{s_{k1}, ..., s_{sl}, ..., s_{kL_k}\}$, and s_{kl} is the discrete user-defined skill levels. The choice of a_k determines the function of the model. If K =1 and a_k is a continuous variable, the user obtains a traditional IRT model (Rasch or 2PL IRT model). When K ≥ 2, and the a_k is continuous, the user gets a multidimensional IRT model. When K ≥ 2, and the a_k is binary, the user obtains a diagnostic model. When users conduct multidimensional IRT model utilizing the GDM framework, the number of latent variables K should be greater than or equal to two, and the number of skill levels of each latent variable a_k is usually set to be greater than or equal to 5 to approximate a continuous (polytomous) distribution. In the GDM equation, the q_{ik} represents elements of a design matrix called the Q-matrix, which specifies the correspondence between latent variables and items. The entry q_{ik} = 1 means that the item i measures latent variable *k*, and q_{ik} = 0 otherwise. β_{yi} and γ_{ik} are item parameters to be estimated.

The GDMs were estimated through the *mdltm*, a software developed by Matthias von Davier, who also developed the GDM. An EM algorithm is implemented in the software. The *mdltm* can provide the following estimates: (1) latent attributes for each individual, (2) latent attributes for demographic groups specified in advance, and (3) item parameters of interest.

The *mdltm* software was designed for data from large-scale assessment programs like TIMSS, CIVED, PISA, or NAEP. Estimates can be obtained for a variety of models: unidimensional IRT (1PL and 2PL), multidimensional IRT, cognitive diagnostic models, latent class, and mixture IRT. The *mdltm* can be used for following data types: dichotomous / polytomous response data, matrix samples (data missing by design and at random), and weighted data.

The next section will describe the four datasets utilized in this dissertation. There have not been very many applications of the cognitive-centered approach on the four datasets because two were collected in 1999 and the other two in 1970 and 1971.

3.3. IEA Large-Scale Subject-Matter Assessments in Science and Civic Education

3.3.1 IEA TIMSS Science Study of 1999

Originating in the mid-1990s, the Third International Mathematics and Science Study has been one of the largest comparative international studies of educational outcomes. The purpose of the international assessment is to provide a base from which researchers, curriculum specialists and policy makers can better understand the quality and performance of their educational systems. The TIMSS 1999 compared the mathematics and science achievement of eighth grade students in 38 countries including the United States. The TIMSS science study for the eighth graders was designed to assess six content areas in accordance with the TIMSS science conceptual framework (Martin, et al., 2000). These areas are:

- Earth science
- Life science
- Physics
- Chemistry
- Environmental and resource issues
- Scientific inquiry and the nature of science

Across the six sub-disciplines, the performance expectations include understanding simple information, understanding complex information, theorizing, analyzing, and solving problems, using tools, routine procedures, and science processes, and investigating the natural world. Experts in subject-matter as well as pychometricians formulated these tests (Martin, et al., 2000).

Test Items. The TIMSS science test for the eighth grade students contains 143 test items representing a range of science topics and skills. Details about these test items are presented in Table 3.1, Table 3.2, and Appendix A. It is important to note that in particular, test items in TIMSS 1999 were assembled into eight different test booklets, and each student only took one booklet which contained both mathematics and science items.

Table 3.1 *Distribution of Science Items by Content*

	Item Type			
Reporting Category	Multiple-Choice	Short-Answer	Extended Reponses	Number of Items
Earth science	17	4	1	22
Life science	28	7	5	40
Physics	28	11		39
Chemistry	15	2	3	20
Environmental and resource issues	7	2	4	13
Scientific inquiry and the nature of science	9	2	1	12
Total	104	28	14	146

Source: Summarized into categories from Exhibit 3.6. Martin, Gregory, & Stemler, 2000.

Table 3.2 *Distribution of Science Items by Performance Category*

Performance Category	Percentage of Items	Total Number of Items	Number of Multiple-Choice Items	Number of Free-Response Items
Understanding simple information	39	57	56	1
Understanding complex information	31	45	30	15
Theorizing, analyzing, and solving problems	19	28	5	23
Using Tools, routine procedures and science processes	7	10	9	1
Investigating the natural world	4	6	4	2
Total	100	146	104	42

Source: Summarized into categories from Exhibit 3.7. Martin, Gregory, & Stemler, 2000.

Achievement Framework. Shavelson and his team at Stanford University have developed a conceptual framework to understand adolescent's science achievement in large-scale assessments (see details in Shavelson & Ruiz-Primo, 1999; Li *et al*., 2011). Drawing upon scientific research (e.g., Alexander & Judy, 1988; Bybee, 1997; Bennett & Ward, 1993; Pellegrino *et al*., 2001; Sadler, 1998), the framework addresses the connections among instruction, student learning, educational measurement, standards, and science curriculum. The framework conceptualizes science achievement as four types of knowledge: declarative knowledge or 'knowing that', procedural knowledge or 'knowing how', schematic knowledge or 'knowing why' and strategic knowledge or 'knowing when, where, and how knowledge applies'.

Shavelson and his team applied this framework to examine selected items and scores (the Booklet 8) in the 1999 TIMSS science test (Li *et al*., 2011). Through statistically modeling the underlying patterns of item scores using confirmation factor analysis, they compare their model based on this achievement framework with other competing models: one factor as general ability, two factors as format (multiple-choice and short-answer), or three factors using the performance expectation framework for TIMSS 1999 (Table 3.2). Results show that the model from Shavelson and his colleagues achieves the best model fit. This indicates that this science achievement framework that conceptualizes science achievement as four types of knowledge can be used to represent underlying structure of what the 1999 TIMSS science test assesses.

Gender Differences. In general, results from the 1999 TIMSS science study show that on average boys had significantly higher science scores than girls in 16 of the 38 countries that participated in the study including the U.S. Boys achieved higher scores in physics, earth science, chemistry, and environmental and resources issues. The gender gap in achievement is especially evident among high-performing students (Martin et al., 2000). Shavelson and his colleagues did not examine gender differences in the four types of knowledge, however.

Language Background. TIMSS science results show that how often students speak the language of test at home is correlated with their average science achievement. On average, students who always or almost always speak the language of test at home achieve higher science scores than those who speak it less frequently (Martin, et al., 2000).

Home Literacy Resources. In the TIMSS science 1999 studies, home literacy resources (the number of books per household, study aids, computer, study desk, dictionary, and parent's education levels) were generally used as an indicator of students' socioeconomic status, Internationally, students from homes with high level of education resources (more than 100 books; all three study aids: computer, study desk, and dictionary; and at least one parent finished university) on average had higher test performance than students from home with low level of resources (Martin, et al., 2000).

3.3.2 IEA Civic Education Study of 1999

The IEA Civic Education (CIVED) study of 1999 surveyed 14-year-old students, their schools, and civic-related teachers within the schools that they attended in 28 countries. The goal of the CIVED study was to identify and examine the ways in which young people are prepared to undertake their role as citizens in democracies. The CIVED cognitive test was designed to assess two types of civic-related knowledge: knowledge of content, and skills in interpretation (Torney-Purta et al., 2001).

Test Items. The CIVED cognitive test contains 38 multiple-choice items, 25 of which measure conceptual knowledge of content and 13 measure skills in interpretation (e.g., understanding the message of a political cartoon or the difference between a fact and an opinion). All 38 items are in the multiple choice form and were administered to all respondents in the survey. See Appendix A.

Achievement Framework. Schulz and Sibberns (2004) employed confirmatory factor analyses (CFA) and multidimensional IRT models to examine cognitive structures underlying the IEA CIVED test items. Their results suggested at least two latent dimensions (and perhaps more) underlying the test items. That is, the civic-related achievement items tap at least two types of knowledge: declarative knowledge (knowledge of content) and procedural knowledge (skills in interpretation). Research conducted by Zhang, Torney-Purta, & Barber (2012) also supported the dimensionality finding. Schulz and Sibberns' (2004) empirical analyses come out with essentially the classification of items shown by the columns of the Table 3.3. The distribution of items by conceptual category and items type is found in the same table.

Table 3.3 *Distribution of Civic Education by Content and Topic Category*

[All items were multiple-choice]

Topic Category	Knowledge	Procedural Skills
Democracy (concepts and institutions)	12	6
Citizenship	10	2
National identity and international relations	2	3
Social cohesion and diversity	1	2
Total	25	13

Source: Summarized into categories from Table A.1 Torney-Purta, et al., 2001.

Gender Differences. In general, results from the IEA CIVED suggest that gender differences are minimal in terms of civic knowledge of content and skills in interpretation in 27 of the 28 countries including the United States (Baldi et al., 2001; Torney-Purta et al., 2001).

Language Background. How often students speak the language of the test at home has been shown to be significantly correlated to their test performance in the CIVED assessment. On average, U.S. students who often or always speak English at home outperformed other students who speak less English at home (Torney-Purta et al., 2001; Wilkenfeld & Torney-Purta, 2012).

Home Literacy Resources. Results from CIVED studies show that home literacy resources (i.e., the number of books in the home) are positively correlated with students' achievement scores in 27 of the 28 countries (Hong Kong is the exception) (Torney-Purta et al., 2001).

3.3.3 IEA Six Subject Survey in Science

The six subject surveys included the IEA first international science study, which was concerned with students' achievement across the domain of science, instruction, students' attitudes, and the development of students' practical skills and understanding of the nature of science. The target populations were 10-year-old students, 14-year-old students, and students in the final year of secondary school in 18 countries including the United States. The achievement test focused on three content areas of science: biology, chemistry, and physics. In addition to the science achievement test, the researchers designed a separate test measuring students' word knowledge or vocabulary (Comber & Keeves, 1973).

Test Items. The science test for the 14-year-old students contains two booklets (Form A and Form B), and each student took one booklet that contains 40 multiple-choice test items representing a range of science topics and skills (see Appendix A). In addition, each student answered a separate general word knowledge test which contains 40 items asking respondents to label a pair of words as opposite or the same in meaning, for example "rare and habitual" or "create and originate" (Comber & Keeves, 1973; Thorndike, 1973). See Appendix B.

Gender Differences. Results from the science study show that on average boys had higher achievement scores than girls in all content areas covered by the science test across countries. However, the gender gap in achievement was considerably smaller in biology than in physics and the practical aspects of the subject (Comber & Keeves, 1973).

Home Literacy Resources. Number of books in home was shown to be positively correlated with achievement test scores in the Science test. Language status of students, however, is not available in the science dataset.

3.3.4. IEA Six Subject Survey in Civic Education

The first IEA study in civic education was conducted in 1971 to assess young people's civic-related cognitive achievement and democratic attitude (e.g., tolerance, support for civil liberties). The target populations were 10-year-old students, 14-year-old students, and students in the final year of secondary school from ten countries including the U.S. In addition to the assessment of civic-related cognitive achievement, the researchers designed a separate scale measuring students' word knowledge or vocabulary (Torney, Oppenheim & Farnen, 1975).

Test Items. The cognitive Civic Education test for the 14-year-old students contains 47 multiple-choice items that test students' conceptual knowledge in Civic Education (see Appendix A). In this study, students also took a general word knowledge test that contains 40 items (Torney, Oppenheim & Farnen, 1975). See Appendix B.

Gender Differences and Home Literacy Resources. Fourteen-year old boys scored higher than girls in about half the countries (including the U.S.). Resources at home were positively related to civic achievement in all the countries. Language status is also included in the questions but was not analyzed.

3.4 Summary and Research Questions

According to previous research, the extent to which a student comprehends test items influences his or her performance on the large-scale subject-matter assessments. The overall purpose of the current study is to understand the extent to which a student comprehends a given item (the question asked and the alternatives given) influence his or her performance on a large-scale subject-matter assessment. To achieve this goal, the present study utilized four low stakes large-scale subject-matter assessments in science and civic education for U.S. students. By examining test items in these assessments, this study focuses on identifying item-level factors that are associated with the student constructing accurate representations of test items as suggested by Kintsch's theory (1998). Eventually the aim is to suggest how the construct-irrelevant variance associated with reading demand can be efficiently minimized. Utilizing a cognition-centered approach to text processing and comprehension (Embretson, 1998; Mislevy, 1994, 1995, 1999) and the techniques described above, the present study aims to measure and understand variance in test scores that is associated with reading comprehension in science, and civic-related social studies assessments in the two data sets from 1999. The following research questions guide analyses of test items and students' data:

1. To what extent do task features facilitate or hinder students' performance in subject-matter assessments including science and civic-related social studies?
    a. What task features pertaining to reading comprehension can be identified in each subject-matter assessment?
    b. At the item level, to what extent are these task features related to the difficulty level of test items in each subject-matter assessment?

2. To what degree do the average estimated scores of the domain-specific proficiency change after taking into account the reading demand of test items?
3. Does the relation between the reading demand and students' domain proficiency vary by gender and language status in each subject-matter assessment?

For the two additional subject-matter assessments that measured students' general word knowledge, additional research questions are:

4. Is there a relation between the measure of general word knowledge and students' achievement in the subject-matter assessment?
5. Does the relation between the students' general word knowledge and achievement vary by demographic factors in each subject-matter assessment?

Chapter 4: Overall Research Design and Methodology

This chapter begins with an overview of research design. Next, I describe the relevant information pertaining to sampling, and conclude with a description of measures of each dataset used in the current study including item difficulty, text features, human ratings, and students' background information.

The overall research design followed a cognition-centered approach based on Evidence-Centered Design. A similar approach has been employed by Embretson and Wetzel (1987) and replicated by Gorin (2005) and Gorin and Embretson (2006) to analyze the GRE-verbal section, a large-scale standardized reading test. Ozuru, Row, O'Reilly, and McNamara (2008) also performed a similar analysis on the comprehension portion of the Gates-MacGinitie Reading Tests (GMRT) for the 7th-9th and 10th-12th grade students. All these studies drew on Kintsch's (1998) reading comprehension theory and focused on understanding the relations between reading comprehension and task features (including text features and item characteristics) using multiple-choice items in large-scale reading assessments. In order to understand the role of reading comprehension in domain specific assessments, the present study applied the cognition-centered approach to large-scale subject-matter assessments in science and social studies.

For each large-scale subject matter assessment proposed in this study, this approach started with defining what the test intends to measure (i.e., the construct) and reading comprehension based on the conceptual achievement frameworks from Kintsch (1998).

The second step was to identify individual-item characteristics and text features in each large-scale assessment based on previous research on reading comprehension and the theoretical framework of reading comprehension theory (e.g., Kintsch, 1998). In particular, text features that contribute to reading demand of test items were identified through a computational tool called Coh-Metrix (Graesser, et al., 2004). My current study used the second version of the Coh-Metrix issued prior to September, 2012. This version of Coh-Metrix (Coh-Metrix 2.0) was replaced late 2012 by a new version (Coh-Metrix 3.0) after my analysis had been completed.

The third step examined the degree to which identified text features and individual-item characteristics were related to the difficulty of test items through regression analyses. The difficulty level of each item (the item difficulty parameter) was estimated using a one-dimensional IRT model. Further analyses at this step informed me about the degree to which these reading-related task features explain the variance of item difficulty in domain specific tests. Information yielded from this step is helpful in identifying the level of reading demand in each item.

At the fourth step, the information about the level of reading demand of each item was incorporated into a multidimensional IRT model (von Davier, 2005). Items with high level of reading demand were modeled through the multidimensional IRT model. The modeling details (one-dimensional IRT and multidimensional IRT) and descriptions with respect to levels of reading demand are in the next chapter (more information about this model appears in von Davier, 2005).

This approach allows me to estimate domain proficiencies while taking reading comprehension components and items with high level of reading demand into account.

The hypothesis is that estimates of domain proficiencies are more accurate because the noise associated with high reading demand can be partialled out. A detailed description about this approach is presented in the next chapter.

In addition to the two datasets collected in 1999, I used the IEA Six Subject Surveys in Science (administrated in 1969), and in Civic Education (administrated in 1971). Both of these assessments included a separate test that measures students' knowledge of general vocabulary. I examined the relation between students' knowledge of general words (synonyms/antonyms) and achievement scores in science and civic-related social studies.

The combination of these steps should give a picture of the extent to which construct-irrelevant variance is associated with reading comprehension in large-scale subject-matter assessments in science and civic education. This evidence is potentially important to test developers, policy makers, and educators who are concerned with validity issues related to ethical evaluation and decision making based on students' test performance.

Finally, previous research suggests that students' language backgrounds and gender are related to the reading demand posed on the subject-matter test, and therefore associated with students' test performance in subject matter assessments. I examined whether the impact of reading comprehension varies by students' language background and gender in each subject-matter assessment.

4.1 Data Sources

The current study utilized the U. S. data from four large-scale subject-matter assessments the IEA CIVED civic education study of 1999 (ICPSR 3892), IEA TIMSS science study of 1999, and the 1970s IEA Six Subject surveys in Civic Education, and in Science. These four large-scale assessments are low stakes assessments for the students, meaning that there is no direct consequence for the test takers. Descriptions of these four datasets were presented in a previous section.

4.1.1 IEA CIVED 1999

The CIVED study involved a three-stage, stratified, clustered sample. At the first stage, communities were sampled with probability proportional to their representation in the population. In the second stage, schools were selected using a stratified random sample procedure, and an intact class of students within the school was randomly selected in the third stage for participation in the study. Additional details on the sampling design are described in Baldi et al. (2001). This sampling design produced a U.S. sample of 2811 14-year-old ninth graders from 124 public and private schools nationwide (Torney-Purta, Lehmann, Oswald & Schulz, 2001).

Given that the assessment did not involve a simple random sample (all students have an equal chance of selection), it is appropriate to apply sampling weights to account for different probability of selection due to using of the stratified sampling procedures. Applying the sample weigh, namely house weight, ensures the samples are representative of 14-year-old U.S. students, and therefore findings are generalizable to the national population. The U.S. data file that the current study employed is *bsusaf2*.

4.1.2 IEA TIMSS Science 1999

The TIMSS 1999 study involved a two-stage stratified sampling design. In the first stage of sampling, schools were selected through stratified random sample design. In the second stage, an intact classroom was randomly selected from the target grade in sampled schools. In the U.S., the sampling design resulted in 9072 14-year-old eighth grade students from 221 schools nationwide. The sample weight, *Student House Weight,* is usually applied to the student-level analysis to ensure the samples are representative of 14-year-old U.S. students, and results are generalizable to the national population (Gonzales & Miles, 2001). The student-level data file that the current study used is *BSAUSAm2*.

4.1.3 IEA Six Subject Survey in Civic Education

For the 1970s Six Subject Surveys in Civic Education, three-stage sampling was conducted in the U. S. Communities were randomly sampled at the first stage, and schools within the communities were selected randomly. At the third stage, students were sampled from the schools. The sample includes 3207 14-year-old students from 127 schools in the U.S.. Student weights per stratum were calculated and included in the student dataset to account for different probability of selection due to using of the stratified sampling procedures (Torney, Oppenheim, Farnen, 1975). The U.S. data file that the current study used is *DBMC3942_US_CV*, which contains variables from student-level, school-level, and community-level.

4.1.4 IEA Six Subject Survey in Science

The 1970s Science assessment utilized two-stage stratified probability sampling design. According to Comber and Keeves (1973), in the U.S., schools were randomly selected in the first stage of sampling "with a probability proportional to the size of school" (p. 43), and students were sampled "from within the school with a probability inversely proportional to the size of school, so that from each school approximately equal numbers of students would be drawn, although each student had the same nonzero chance of entering the sample" (p. 43). Student weights per stratum in the dataset account for different probability of selection due to using of the stratified sampling procedures Eventually, a total of 3398 14-year-old students from 137 schools were selected and involved in the survey. The U.S. data file that the current study employed is *dbm2942_US_SC_RL*, which contains variables from both student-level and school-level.

All four datasets are in the ICPSR's collections, and IRB requirements were checked before they were included there.

4.2 Measures

4.2.1 Materials

For the purposes of this study, I focused on multiple-choice test items which were written in English and on the datasets resulting from administration to nationally representative samples of U.S. students.

The targets of analyses were 104 multiple-choice items (for the 8th grade level) released from the science test of TIMSS 1999, 38 multiple-choice items (for the 8th grade level) from the CIVED 1999, 37 multiple-choice items (for the 14-year-olds) from the

1970s IEA Six Subject Survey in Science, and 47 multiple-choice items (for the 14-year-olds) from the 1970s IEA Six Subject Survey in Civic Education.

These multiple-choice test items have several features in common. First of all, all items were designed to present a high level of demand on subject-matter knowledge or skills (when compared to reading comprehension test items). Second, majority of items start with a short sentence followed by a question or an incomplete statement which calls for the answer. Finally, each item has four or five multiple-choice options (alternatives) with one presumed to be "the best" answer, and the others distractors. See Appendix A for a selection of these items.

4.2.2 Item Difficulty

The difficulty level of each item (item difficulty parameter) was estimated based on item responses of U.S. students using the one parameter (1PL) IRT model (Rasch Model, Rasch, 1960) through the *mdltm* software (von Davier, 2010). The item responses data from all four tests were coded as right and wrong with 1 representing a right response and 0 representing a wrong response. Sampling weights were applied when estimating the item difficulty parameter from each test. Descriptive statistics of the item difficulty of each data set are presented in the next chapter.

4.2.3 Text Features

Text-specific features that contribute to reading demand of test items were identified through Coh-Metrix version 2.0 (Graesser, et al., 2004). According to Graesser, et al., Coh-Metrix 2.0 was developed to analyze and measure text in categories related to the first five levels of the discourse-based Kintsch's reading comprehension theory:

words, syntax, textbase, situation model, and *genre* in addition to those measured by traditional readability formulas. The software can provide more than two hundred text cohesion and readability measures considered to influence comprehension. Examples of Coh-Metrix output are presented in Appendix C.

The present research mainly focuses on some key indices that were theoretically related to Kintsch's reading comprehension theory, and empirically known to affect comprehension difficulty. Information about the selected text features yielded from the Coh-Metrix 2.0 can be found in Appendix D. Extended theoretical information about the text features indices produced by the Coh-Metrix 2.0 can be found in Graesser, McNamara, Louwerse and Cai (2004), and McNamara, Louwerse and Graesser (2002).

One limitation of the Coh-Matrix 2.0 is that the readability formula including the Flesch Reading Ease and Flesch-Kincaid Grade Level may not yield reliable results when a text analyzed has less than 200 words. Therefore, when I analyzed test items which were written in short text less than 200 words, I used count text indices – average sentence length and the mean number of syllables per word – to substitute for the Flesch Reading Ease measure and Flesch-Kincaid Grade Level measure, because both formulas calculate text readability as a function of average sentence length and the mean number of syllables per word.

4.2.4 Item Characteristics and Human Rating

In addition to Coh-Metrix, I developed a coding scheme (presented in Appendix E) to code individual items including stems and response alternatives more holistically based on two classification coding systems developed by Mosenthal (1996) and Oruzu, et al. (2008).

Abstractness of the Item Question. The first coding system deals with the abstractness of the information requested by a question. I used Mosenthal's (1996) coding system which classifies abstractness of the item question into five levels (p. 1004). This was originally based on Kintsch's reading comprehension theory.

1. The first level, most concrete, asked for the "identification of persons, animals, or things."
2. The second level, the highly concrete class of questions, asked for the "identification of amounts, times, or attributes."
3. The third level, intermediate questions, asked for the "identification of manner, goal, purpose, alternative, attempt, or condition."
4. The fourth level, highly abstract, asked for the "identification of cause, effect, reason, or result."
5. The highest level, the most abstract questions, asked for the "identification of equivalence, difference, or theme."

Text Genre. The second coding system classifies item stem including the passage(s) and questions into three different text genres. The coding system is adapted from Ozuru, et al. (2008).

- *Narrative.* "Narrative passages tend to describe relatively mundane events with which most people have some familiarity from a personal perspective" (p.1006).
- *Expository.* "Expository passages tend to describe historical, social, and/or scientific facts from a nonpersonal, objective perspective" (p. 1006).
- *Mixed/Both.* The text contains characteristics of both genres, or some characteristics of narratives and some of expository.

Holistic Reading Difficulty. On top of these two coding items, an additional question in my coding scheme asked raters to rate holistically the level of reading difficulty for each item on a 5-point likert scale with “1” means very easy and “5” very difficult. When the rater rated a test item at a 3 or higher on reading difficulty, the rater was asked to identify where the difficulty was/were based on the following options:

1. The difficulty of vocabulary in the item stem.
2. The difficulty of vocabulary in the multiple-choice options. Please specify which option(s).
3. Complexity of grammar or syntax in the item stem.
4. Complexity of grammar or syntax in the multiple-choice options. Please specify which option(s).
5. Other. Please specify.

Other Ratings. When a rater had a rating 3 or higher on reading difficulty of an item, the rater was also asked to provide their opinions on two 3-point scales with respect to

1. Do you think the reason(s) you selected as causing the item difficulty is/are relevant to the content which the item assesses?

2. Do you think this item could be rewritten to reduce the reading difficulty, but still assess the relevant content?

Both 3-point likert scales range from 1 to 3 with “1” indicates yes, “3” no, and “2” somewhat.

Two reading experts were involved in this study to identify the item characteristics of multiple-choice items from IEA TIMSS science and CIVED assessments. One rater has been working in the area of reading research for many years, and also has expertise on civic education. The other rater is a Ph.D student with expertise in reading research. First, they were asked to code 20 items from IEA TIMSS science test and 20 items from CIVED test using the coding systems developed by the author. Then their inter-rater reliability was calculated. When their inter-rater reliability was .60 or greater, the senior rater continued to rate the rest of items (84 items from the TIMSS science and 18 item from the CIVED).

Inter-rater reliabilities of their coding were calculated using Kappa statistics (Cohen, 1960; Siegel & Castellan, 1988) and intra-class correlation (ICC, Shrout & Fleiss, 1979; McGraw & Wong, 1996) through SPSS 20. Hallgren (2012) provides an overview and tutorial with respect to how to compute and interpret these two types of inter-rater reliability statistics. According to Hallgren, Kappa statistics are often used to calculate the extent of agreement among raters beyond that expected by chance. The scale of items (subjects) coded by raters can be nominal or ordinal. The Kappa statistics range from -1 to 1, with 1 denoting perfect agreement, and -1 perfect disagreement. The Kappa statistic of 0 indicates completely random agreement. Landis and Koch (1977) provide commonly-cited guidelines for interpreting Kappa values. Kappa values from 0.00 to 0.20 indicate slight agreement; 0.21 to 0.40 denote fair agreement; 0.41 to 0.60 indicate moderate agreement; 0.61 to 0.80 suggest substantial agreement, and 0.81 to 1.00 denote almost perfect to perfect agreement.

The ICC is mostly used to calculate the magnitude of agreement among two or multiple raters on ordinal, interval, or ratio variables (i.e., items or subjects). The ICC used by the current study was derived from a two-way mixed ANOVA model. Based on guidelines provided by Hallgren (2012), the current study chose the mixed effect model because the raters were not randomly selected. ICC values range from 1 to less than -1, with 1 indicating perfect agreement, and 0 suggesting random agreement. ICC values can be less than -1 when there are more than two raters. When interpreting ICC values, a value of 0.75 or greater suggests an excellent agreement. ICCs of 0.60 to .74 indicate good agreement, and 0.40 to 0.59 denote fair agreement. An ICC below .40 suggests a poor agreement (Cicchetti, 1994).

The coding results and inter-rater reliabilities are reported in the next chapter.

4.2.5 Personal Factors

To examine the association between comprehension processes in subject-matter assessments and students' personal level factors, I incorporated gender, language background, home literacy resources, and general word knowledge (this measure can only be found in the 70s Six Subject Surveys) into my analyses. These personal level factors have been shown related to reading comprehension by previous research.

All four assessments provide information about students' gender. In my analyses, I recoded the gender variables in all four datasets with “0” representing boys and “1”girls.

Language background variable measures how often a student speaks the language of test at home. The responses include “1”—always or almost always, “2”—sometimes, and “3”—never. I created from this variable a dummy-coded variable in which “0” indicates English language learners (ELLs) who sometimes and never spoke the language of test at home, “1” represents non ELLs who always or almost always spoke the language of test at home. Among the four assessments of interest, the Six Subject Survey in Science did not measure students' language background. In the Six Subject Survey in Civic Education, fewer than 80 ELLs participated in the assessment. Therefore, I replaced the language background variable with the home literacy resources variable in my analyses of the Six Subject Surveys in Civic Education and Science datasets.

IEA Six Subject Surveys in Science and in Civic Education both provided a separate scale of general word knowledge. The word knowledge scale contains 40 items that measure students' knowledge of general vocabulary. See Appendix B for examples. My review of reading comprehension literature suggests that vocabulary knowledge is a core reading component that affects reading comprehension. Therefore, when I analyzed each Six Subject Surveys dataset, I utilized the general word knowledge scale as one commonly considered aspect of reading comprehension.

Chapter 5: Results

In the previous chapter, I described the research design, measures, and statistical procedures used in the present study. In the current chapter I describe the analysis results with respect to my five research questions step by step. Each research question concludes with a brief summary of the findings. An extensive summary and discussion is presented in Chapter 6.

5.1 Research Question 1

My research question one asks to what extent reading-related task features contribute to the item difficulty of large-scale subject-matter assessments. I used two large-scale subject-matter assessments to answer this question: the CIVED test of 1999 and the TIMSS science test of 1999. Two series of multiple regression analyses were designed to provide evidence with respect to the degree of association between task features and item difficulty.

5.1.1 IEA Civic Education Test of 1999

My first set of analyses was conducted using the IEA CIVED test items. In this section, I begin with describing variables of interest: item difficulty, and task features pertaining to reading comprehension in civic education items. Then I report the statistical results from my data analyses.

Item Difficulty. The criterion variable, item difficulty of the 38 test items from the IEA CIVED, was estimated based on U.S. students' item responses (dichotomous, coded as right/wrong) using the Rasch model (1PL IRT model; Rasch, 1960) through the *mdltm* software (von Davier, 2010). The sample size is 2786. Twenty five cases (students) were excluded from the analysis because these cases have missing scores on all 38 test items.

For the Rasch model, the estimated item difficulty values indicate the location on the ability scale where a student has a 50 percent chance of choosing the correct answer. The Rasch model assumes that guessing is a part of the ability and that all items have equivalent discriminations, so that the probability of a person getting an item correct is only described by a single parameter (b_i), item difficulty, and the person's hypothetical ability, θ. I chose the Rasch model is because utilizing the Rasch model makes the item difficulty parameter easier to interpret than those from two-parameter (2PL) and three-parameter (3 PL) IRT models. House weights in the IEA civic student-level data set were applied when I conducted parameter estimation using Rasch model.

To compare the fitness of the Rasch model to data, I also applied a 2PL IRT model through the *mdltm* software. Table 5.1.1.1 shows model fits of these two IRT models. In general, the model fit results indicate that there is no noticeable difference between the Rasch model and 2PL IRT model in terms of the log likelihood, Akaike information criterion (AIC), and Bayesian information criterion (BIC).

Table 5.1.1.1 *Model Fit of IRT Models for the CIVED Data*

Model	# of parameters	Log-Likelihood	AIC	BIC
Rasch (1PL)	50	-53465.6883	107031.37669	107327.99480
2PL	88	-52826.4221	105828.84417	106350.89204

Note. AIC = Akaike information criterion
BIC = Bayesian information criterion

Descriptive statistics for the item difficulty from the Rasch model are shown in Table 5.1.1.2. Item difficulty values yielded from the *mdltm* software are reverse from the typical way the Rasch model is scaled. That is, the higher the “difficulty value” from the *mdltm*, the easier the item is. The average difficulty of the 38 CIVED test items is .001 with a standard deviation of .732. The range is from -1.82 to 1.26. Among the 38 test items, item 33, 24, 23, 2, and 5 are the easiest, and item 22, 27, 29, 34, and 21 are the most difficult.

The normality of the distribution of the item difficulty values yielded from the Rasch model was examined through histograms, Kolmogorov-Smirnov test, and Shapiro-Wilk test. Overall, normality tests suggested that the distribution of 38 item difficulty values from the Rasch model did not adversely violate the normality assumption and *mdltm* estimates of Rasch item difficulty parameter were hence used as the criterion variable in multiple regression.

.

Table 5.1.1.2 *Descriptive Statistics for Item Difficulty and Task Features of the CIVED Test*

		Stem			
Variable	M	SD	Min	Max	N
Item difficulty	.001	.732	-1.827	1.264	38
DENSNP	247.021	71.823	.0000	388.889	38
HYNOUNaw	4.800	.826	2.889	6.583	38
HYVERBaw	1.197	.548	.500	3.500	38
Question-abstractness	3.030	.716	2	4	38
READASL	11.584	3.800	5.000	21.000	38
READASW	1.600	.240	1.200	2.111	38
WORDCacw	349.364	41.723	271.833	446.333	38
		Key			
Variable	M	SD	Min	Max	N
Item difficulty	.001	.732	-1.827	1.264	38
CONLGni	9.042	33.121	.000	166.667	38
FRQCRmcs	68.82	78.338	2	333	38
HYNOUNaw	4.685	.785	2.875	6.000	38
HYVERBaw	1.789	1.3120	.000	5.650	38
READASW	1.842	.523	1.000	4.000	38
WORDCacw	386.997	63.237	288.000	584.000	38
		Distractors			
	M	SD	Min	Max	N
Item difficulty	.001	.732	-1.827	1.264	38
DENLOGi	19.051	39.311	0.000	138.889	38
DENNEGi	8.532	25.167	0.000	111.111	38
DENSNP	305.529	54.110	208.333	425.926	38
HYVERBaw	1.341	0.702	0.000	3.193	38
FRQCRacw	1483.354	1878.510	125.111	7254.211	38
INTEi	28.897	37.977	0.000	120.370	38

Note. Item difficult: higher numbers = less difficult;

CONLGni = Incidence of negative logical connectives;
DENLOGi= Logical operator incidence score (and + if + or + cond + neg);
DENNEGi = Number of negations;
DENSNP = Noun phrase incidence;
FRQCRacw = Raw frequency of content words;
HYNOUNaw = Mean concreteness values of nouns;
HYVERBaw = Mean concreteness values of verbs
INTEi = Incidence of intentional actions, events, and particles;
Question-abstractness = the abstractness of an item question. This is an item characteristic rated through human coding;
READASL = Average Words per Sentence;
READASW =Average Syllables per Word;
WORDCacw = Concreteness, mean for content words.

Text Features. Text features were obtained from the Coh-Metrix version 2.0. Examples of Coh-Metrix output and code values are presented in Appendix C.

I conducted text analyses with respect to the stem, correct response, and distractors of each test item. On average, the 38 items contain 2.03 (complete and/or incomplete) sentences (SD= 2.15) on stems, and each alternative followed an item stem contain only 1.00 (complete/incomplete) sentence (SD = .00).

The Coh-Metrix 2.0 package provides measures of text features that pertinent to different levels of Kintsch’s reading comprehension theory including the surface level, textbase level, and situation level (Graesser, et al., 2004). Appendix D presents descriptions of some key text features yielded from Coh-Metrix, and the level of Kintsch’s reading comprehension model that each feature is likely to tie to.

Initially, about fifty indices yielded from Coh-Metrix were considered. However, such a large number of variables are unlikely to produce a satisfactory multiple regression results given the relative small number of items (*n* = 38). Preliminary analyses of bivariate correlations were therefore used to screen out variables that would be unlikely to contribute to the prediction. Only variables that show statistically significant correlations with the criterion variable—item difficulty—at a = .05 (two-tailed), were retained for the subsequent analysis. Among the remaining text features, I limited cohesive indices in my analysis because CIVED items were designed to be short with the purpose of eliciting students' domain knowledge or skills (and not necessarily cohesive).

The bivariate correlation approach is a heuristic, given that Coh-Metrix provides vastly more potential predictors than observations (*n* =38 for the CIVED test). It is possible that some combinations of predictors with low bivariate correlations could provide better prediction than just restricting attention to ones with significant bivariate correlations. However, given that I looked for features of tasks that have both theoretical and practical meaning, unusual combinations of predictors with small correlations would likely be hard to interpret and would offer no guidance to test developers. By limiting my attention to feature variables with significant bivariate correlations, I made sure that the functions that I ended up with would be more interpretable and actionable.

Three bivariate correlations between item difficulty and text features were computed with respect to item stems, correct responses, and distractors respectively.

Item Characteristics. Two reading experts coded item characteristics based on the coding scheme described in chapter 4. Appendix E presents the coding scheme. These two raters rated the first twenty items of each assessment (TIMSS science and CIVED), and the senior rated rest of the items. Among the quantitative rating items (i.e., *item 1, 3, 4, 5* and *6*), *item 1, 5,* and *6* are on a 5-point scale, and *item 3* and *4* are on a 3-point scale. Their inter-rater reliabilities were calculated using Kappa Statistics (Siegel & Castellan, 1988) and intra-class correlation (ICC, Shrout & Fleiss, 1979) through SPSS 20. Results are presented in Table 5.1.1.3.

Table 5.1.1.3 *Inter-Rater Reliabilities of CIVED Item Ratings (n=20)*

	Kappa	ICC
Item 1	0.061	-0.375
Item 3	-0.161	-0.053
Item 4	-0.063	-0.107
Item 5	0.185	0.593

Note. Kappa = Siegel & Castellan's Kappa
ICC = Intra-class correlation

The inter-rater reliability (ICC) of *Item 1, 3,* and *4* were lower than .50. Therefore, none of them were used for subsequent analyses. *Item 1* asks raters to evaluate holistically the level of reading difficulty of each CIVED test item. *Item 2* and *3* are follow-up questions when a rater rates a CIVED test item at a 3 or higher on item difficulty. *Item 4* asks raters to evaluate whether a test item could be rewritten to reduce the reading difficulty, but still assess the relevant content.

Item 5, abstractness of the item question, was used for subsequent analyses as an item characteristic. The scheme of Item 5 was developed based on Mosenthal's (1996) study. There were five levels in this scheme ranging from the "most concrete" to the "most abstract". First, the two raters independently coded the first 20 CIVED items. The average measures of the ICCs between the two raters were initially less than .40, which suggests a poor agreement. After rating, the raters reflected that they were not very sure about their ratings because the rating description itself was not very tangible. To increase the concreteness of the item, I provided the two raters with examples from Mosenthal (1996), which led to the increase of the ICC to .593, meaning a fair agreement. Next, the senior rater rated rest of the 16 items from the CIVED using the same coding scheme, and the full set of the senior rater's rating was used for statistical analysis. The descriptive statistics of the item characteristic “abstractness of the question” were presented in Table 5.1.1.2.

In addition to the item characteristics, one type of text feature, namely text genre, was coded by two reading experts based on a coding system adapted from Ozuru, et al. (2008). The raters were asked to classify the text of each item (including the stem and alternatives) into three different text genres: narrative, expository, mixed/both. For the first 20 items of each assessment, two raters coded all of them as “expository.” The senior rater rated rest of items and classified all of them as expository. This step of ratings provides important evidence with respect to the genre of CIVED test, as well as other large-scale subject-matter assessments similar as CIVED. Because ratings are constant and indicate that the text genre is expository across items and tests, I did not use text genre for any statistical analysis.

Regression Analysis. In next step, multiple regression analyses were conducted through SPSS 20 with respect to item stems, correct responses, and distractors separately. The purpose of these analyses was to identify a reasonable number of predictors across all three types of predictors. For each analysis, text feature and item characteristic predictors that have statistically significant correlations with the item difficulty were entered into the regression model to further examine their associations with item difficulty values. Some of those variables did not survive with a significant regression coefficient when entered simultaneously. I conducted model comparisons in regard to item stems, correct responses, and distractors. The best models were decided based on the model fit indices including R squares, and regression coefficients.

Before multiple regression analyses, I examined bivariate correlations among pairs of predictors in order to detect possible multicollinearity in multiple regression analysis. When a bivariate correlation between predictors was high (r > .80) and reading comprehension theories suggest that the two predictors were similar in terms of function (e.g., both measure readability of the text), I removed one predictor from the multiple regression model to avoid the impact of multicollinearity. In addition, I used variance inflation factors (*VIF*) to detect possible multicollinearity in multiple regression analysis. The VIF is a common method in multiple regression analysis that helps detect multicollinearity. The method measures how much the variances of the estimated regression coefficients are inflated as compared with when the predictors are not linearly correlated.

Stems. Table 5.1.1.4 presents the bivariate correlations among item difficulty, text features from the Coh-Metrix 2.0, and item characteristic (*Item 5*) coded by human raters. Initially, text features that have significant bivariate correlations with the item difficulty were entered into a multiple regression model. They are readability indices: average words per sentence (READASL), and average syllables per word (READASW); vocabulary index: mean concreteness of words in a text (WORDCacw); and syntactic index: mean number of modifiers per noun-phrase (DENSNP). I also added the item characteristic identified by human raters: the abstractness of the item question (*Item 5*). Moreover, because my raters emphasized in their ratings that vocabulary played an important role in the reading difficulty of CIVED test items, I entered two more Coh-Metrix indices—HYNONaw and HYVERBaw—into the regression model. HYNONaw measures average abstractness of nouns, and HYVERBaw measures average abstractness of verbs. According to McNamara, et al., (2005), an abstract word is one with few distinctive features and few attributes that can be pictured in the mind. The detailed descriptions of these text features are presented in Appendix D, and descriptive statistics of the variables are showed in Table 5.1.1.4.

Table 5.1.1.4 *Correlations of Selected Task Features from the CIVED Test Item Stems*

	Item difficulty	DENSNP	HYNOUNaw	HYVERBaw	Question-abstractness	READASL	READASW	WORDCacw
Item difficulty	--							
DENSNP	.415**	--						
HYNOUNaw	.231	.246	--					
HYVERBaw	.189	.145	.158	--				
Question-abstractness	.040	-.015	-.010	.060	--			
READASL	-.279	-.112	-.119	-.014	.583**	--		
READASW	-.473**	-.281	-.262	-.089	-.276	.142	--	
WORDCacw	.319	-.029	.189	.582**	-.128	-.060	-.006	--

Note. Item difficult: higher numbers = less difficult;
DENSNP = Noun Phrase Incidence Score (per thousand words);
HYNOUNaw = Mean concreteness values of nouns;
HYVERBaw = Mean concreteness values of verbs
Question-abstractness = the abstractness of an item question. This is an item characteristic rated through human coding;
READASL = Average Words per Sentence;
READASW =Average Syllables per Word;
WORDCacw = Average concreteness of content words in a text.
* *P* < 0.05. **p < .001 (2-tailed).

After model comparisons, the best multiple regression model for stems is presented in Table 5.1.1.5. Three text features significantly predict the item difficulty: the frequency of noun phrases—DENSNP (**β** = .317, *p* < .05), word length—READASW (*β* = -.382, *p* < .05), and the mean abstractness of content words—WORDCacw (*β* = .326,

$p < .05$). The item characteristic, the abstractness of the item question, is not significantly related to the item difficulty ($r = .040$, $p = .813$, two-tailed), therefore is not retained in the final best model.

Table 5.1.1.5 *Multiple Regression Results for Item Difficulty of CIVED Item Stems*

	B	SE	Beta	t	sig
(Constant)	-.928	1.176		-.789	.435
DENSNP	.003	.001	.317	2.320	.026
READASW	-1.167	.417	-.382	-2.797	.008
WORDCacw	.006	.002	.326	2.485	.018
R^2	.416				
Adjusted R^2	.365				

Dependent Variable: Item difficult: higher numbers = less difficult;
Note. DENSNP = Noun Phrase Incidence Score (per thousand words);
READASW =Average Syllables per Word;
WORDCacw = Average concreteness of content words in a text.

Among the statistically significant predictors, DENSNP is the frequency of noun-phrase constituents per 1000 words. The higher the frequency score, the more noun phrases are contained in the analyzed text. READASW, the average number of syllables per word, is a readability index that reflects word length. WORDCacw measures the average concreteness value of all content words in a text that match a word in the MRC Psycholinguistics Database (Coltheart, 1981). Concreteness measures in terms of ratings of whether content words are more or less abstract or concrete. Content words are nouns, adverbs, adjectives, main verbs, and other categories with rich conceptual content (McNamara, et al., 2005). The more concrete a word is, the higher the score.

The linear combination of these text features are significantly related to item difficulty, F (4, 32) = 8.078, $p < .01$. Variance of item difficulty explained (R^2) is .416, and adjusted R^2 is .365, which indicates that more than 36 percent of the variance of item difficulty is accounted for by the reading-related text features after taking into account the number of predictor variables in the model.

In general, results suggest that text features that measure the surface level of reading comprehension significantly predict item difficulty of the CIVED test items. First, lengthy words are associated with difficult items. Previous studies suggest that lengthy words usually take more space in the reader's work memory; therefore, lengthy words increase reading difficulty. Second, the average concreteness of all content words in a stem predicts the item difficulty. The more concrete content words (including nouns, adverbs, adjectives, main verbs) in an item stem, the less difficult the item to read. Third, results also reveal that the frequency of noun phrases is related to item difficulty. In other words, when an item stem contains one or more noun phrases, the item appears to be easier than items that contain no noun phrases. The possible explanation is that noun phrases may aid readers in chunking the information into fewer units so as to increase their short-term memory capacities. Therefore, when the text lengths are similar (about two or three sentences per item), items that contain noun phrases are easier to process than other items. Results also show that the average concreteness of all content words in a stem predicts the item difficulty. The more concrete content words (including nouns, adverbs, adjectives, main verbs) in an item stem, the easier the item to read.

Correct Responses. Table 5.1.1.6 presents the bivariate correlations among item difficulty and text features from the CIVED test correct-responses. Initially, readability index—average word length (READASW), and syntactic complexity index—the number of negative logical connectives (CONLGni) were entered into a multiple regression model as predictors because these two text features showed significant bivariate correlations with the item difficulty. What's more, because the raters suggested that difficult vocabulary contributed to the reading difficulty of CIVED test items, I added three more Coh-Metrix features that measure the concreteness of vocabulary into the regression model: mean concreteness of words in a text (WORDCacw), average abstractness of nouns (HYNONaw), and average abstractness of verbs (HYVERBaw).

Table 5.1.1.6 *Correlations of Selected Task Features from the CIVED Item Correct Responses*

	Item difficulty	CONLGni	FRQCR mcs	HYNOU Naw	HYVER Baw	READ ASW	WORD Cacw
Item difficulty	--						
CONLGni	.276	--					
FRQCRmcs	.141	-.161	--				
HYNOUNaw	.016	-.084	-.123	--			
HYVERBaw	.080	.068	-.288	.208	--		
READASW	-.373*	-.095	-.343*	.090	-.120	--	
WORDCacw	-.133	.049	-.031	.094	-.200	.222	--

Note. Item difficult: higher numbers = less difficult;
CONLGni = Incidence of negative logical connectives;
FRQCRacw = Raw frequency of content words;
HYNOUNaw = Mean concreteness values of nouns;
HYVERBaw = Mean concreteness values of verbs
READASW = Average syllables per word;
WORDCacw = Average concreteness of content words in a text .
* p < 0.05. **p < .001 (2-tailed).

After model comparisons, the best multiple regression model for correct responses is presented in Table 5.1.1.7. The results reveal that only word length (READASW) significantly predicts the difficulty levels of items (β = -.373, p < .05). Variance of item difficulty explained (R^2) is .139, and adjusted R^2 is .115, which means that about 12 percent of the variance of item difficulty is accounted for by the average word length of correct responses after taking into account the number of predictor variables in the model.

Table 5.1.1.7 *Multiple Regression Results for Item Difficulty of CIVED Item Correct Responses*

	B	SE	Beta	t	sig
(Constant)	.964	.414		2.326	.026
READASW	-.522	.217	-.373	-2.412	.021
R^2	.139				
Adjusted R^2	.115				

Dependent Variable: Item difficult: higher numbers = less difficult;
Note. READASW = Average Syllables per Word

Distractors. Because each CIVED test item contains three distractors, I analyzed the distractors individually using the Coh-Metrix 2.0, and merged outcome values of three distractors to a single distractor by summing values of the three distractors across rows.

Table 5.1.1.8 presents the bivariate correlations among item difficulty and text features from the CIVED test correct-responses. Initially, bivariate correlations show that the item difficulty is significantly correlated with the following variables: syntactic complexity indices: DENSNP, DENNEGi, and DENLOGi, indices of word information: HYVERBaw, FRQCRacw, and FRQCLacw, and a situation model index: INTEi.

Table 5.1.1.8 *Correlations of Selected Task Features from the CIVED Item Distractors*

	Item difficulty	DENLOGi	DENNEGi	DENSNP	FRQCRacw	HYVERBaw	INTEi
Item difficulty	--						
DENLOGi	.353*	--					
DENNEGi	.335*	.753**	--				
DENSNP	-.296	-.279	-.273	--			
FRQCRacw	.381*	.527**	.576**	-.240	--		
HYVERBaw	-.335*	-.179	.033	-.027	-.203	--	
INTEi	-.405*	-.111	.099	.124	-.079	.641**	--

Note: Item difficult: higher numbers = less difficult;
DENLOGi= Logical operator incidence score (and + if + or + cond + neg);
DENNEGi = Number of negations;
DENSNP = Noun phrase incidence;
FRQCRacw = Raw frequency of content words;
HYVERBaw = Mean concreteness values of verbs;
INTEi = Incidence of intentional actions, events, and particles;
* *P* < 0.05. ***p* < .001 (2-tailed).

Regression results show that the item difficulty is significantly associated with the frequency of negative expressions in the distractors (DENNEGi, $\beta = .379$, $p < .05$), and the frequency of intentional actions, events, and particles (INTEi, $\beta = -.442$, $p < .05$). Table 5.1.1.9 presents results of analyses for distractors. The linear combination of reading-related task features is significantly related to item difficulty, $F(2, 35) = 7.717$, $p < .01$. Variance of item difficulty explained (R^2) is .306, and adjusted R^2 is .266. In summary, my regression results indicate that item distractors containing (a) negative expression(s) tend to be easy. An example of a distractor containing a negative expression is “The United Nation has its own flag even though it is not a country.” It should be noted that this is a true statement but was not the correct answer. Another example is “People with very low incomes should not pay any tax.” After having securitized the CIVED items, I also found that item alternatives that contain negative expressions are most likely to be distractors. That explains why previous analyses on correct responses did not show statistically significant correlation between negative expressions and item difficulty.

Likewise, the predictor—INTEi—is an index associated with the situation model in reading comprehension processes. The intentional content reflects the extent to which sentences are related by intentional particles (e.g., in order to, so that, for the purpose of, by means of, by, wanted to), actions, and events. Coh-Metrix estimates intentional actions and events by counting the number of main verbs that are intentional (actions which are performed in pursuit of goals) based on WordNet (Fellbaum, 1998). The higher the counts in a text, the more the text is assumed to carry goal-driven content (McNamara, et al., 2005). My regression results show that distractors containing goal-driven content were associated with difficult items. One possible explanation is that these main verbs are likely to activate students' existing schema and to draw out prior knowledge that may or may not reflect the intention of the test item. Examinees are likely to select an incorrect answer if these activated schemas do not match with the purpose of the item.

Table 5.1.1.9 *Multiple Regression Results for Item Difficulty of CIVED Item Distractors*

Model	B	SE	*β*	t	Sig
(Constant)	.154	.131		1.169	.250
DENNEGi	.004	.001	.379	2.680	.011
INTEi	-.003	.001	-.442	-3.124	.004
R^2	.306				
Adjusted R^2	.266				

Dependent Variable: Item difficult: higher numbers = less difficult;
Note. DENNEGi=Number of negative expressions;
INTEi = Incidence of intentional actions, events, and particles.

Summary. Overall, my results suggest that when students took this standardized test in civic education, the difficult levels of items are predicted by linguistic features (i.e., word length, word concreteness, and syntactical complexity of text) pertaining to lower level comprehension processes. On top of that, my results also reveal the difficulty levels of 38 civics items are associated with a text feature pertaining to the higher level of reading comprehension: the situation model. This has to do with main verbs used in item alternatives. It seems that some intentional verbs are likely to activate students' prior knowledge. When distractors contain these verbs, the item tends to be difficult. The Coh-Metrix version 2.0 only provides a numerical index about the frequency of the intentional verbs. More specific information such as what exactly the intentional verbs are that made a difference is not provided however.

I was also interested in how text features in a science test may influence students' comprehension of science items. Next, I applied the same procedures to the TIMSS science items to investigate how comprehension-related features at the item level contribute to item difficulty levels.

5.1.2 TIMSS 1999 Science Test

My second study analysis was conducted using 104 multiple-choice items from the IEA TIMSS 1999 science test.

Item Difficulty. Item responses (dichotomous, coded as right/wrong) from 9072 U.S. students were used for IRT analysis. Item difficulty values of the 104 test items from the TIMSS science were estimated using the Rasch model (Rasch, 1960) through the *mdltm* software (von Davier, 2009). In addition, a 2PL IRT model was applied to the same data through the *mdltm* software.

The TIMSS test involved a booklet design (i.e., matrix sampling design), meaning not all students answered all 104 questions, and each student was only administrated a small proportion of the 104 items (for detailed descriptions about the booklet design of TIMSS assessment of 1999, see Gonzales & Miller, 2001). In the TIMSS student-level dataset, an item that was not assigned to a student was marked as "not administrated" and coded as "8", and the *mdltm* software treats this item as if it was not administrated to the student when the software estimates item difficulty based on students' item responses. House weights in the IEA TIMSS 1999 science dataset were applied when I conducted parameter estimations using the Rasch model and 2PL model.

Table 5.1.2.1 demonstrates model fit statistics of these two IRT models. The results indicate the 2PL IRT model has slightly better model fit than the Rasch model in terms of the log likelihood, AIC, and BIC. On the other hand, the Rasch model is more parsimonious than the 2PL model. The current study chose the Rasch model for subsequent analyses because the item difficulty from it is relatively easy to interpret.

Table 5.1.2.1 *Model Fit of IRT Models for the TIMSS Science Data*

Model	# of parameters	Log-Likelihood	AIC	BIC
Rasch (1PL)	116	-143662.9249	287557.84972	288382.95169
2PL	220	-142366.4905	285172.98095	286737.82952

Note. AIC = Akaike information criterion
BIC = Bayesian information criterion

Normality distributions of the item difficulty values were examined through histogram, Kolmogorov-Smirnov test, and Shapiro-Wilk test. Results suggested that the item difficulty values from the 1PL model were normally distributed and therefore were used as the criterion variable in multiple regression.

The average difficulty of the 104 science test items is 0.00 with a standard deviation of .979, and the difficulty values range from -2.170 to 2.228. Among the 104 test items, item s012007, s012010, s012035, s012026, and s012037 are five top easy items, and item s022094, s022275, s022106, s012009, and s012047 are the five top difficult ones.

Text Features. Text feature variables were obtained from Coh-Metrix version 2.0. Appendix C presents examples of Coh-Metrix output and index values.

I conducted text analyses through Coh-Metrix with respect to the stem, correct responses, and distractors of each test item separately. Initially, for each part of items, about fifty variables yielded from Coh-Metrix 2.0 were considered. However, such a large number of variables is unlikely to produce a satisfactory multiple regression results given the relative small size of the items (number of items = 104). Preliminary analyses of bivariate correlations were therefore used to screen out variables that would be unlikely to contribute to the prediction. Only variables that showed statistically significant correlations with the criterion variable—item difficulty—at α = .05 (two-tailed) were retained for the subsequent analyses. Descriptive statistics for selected text features are presented in Table 5.1.2.2.

On top of text features identified by Coh-Metrix 2.0, one kind of text feature, namely text genre, was rated by two reading experts based on a coding system adapted from Ozuru, et al. (2008), and all 104 science items were rated as expository (two raters classified texts of the first 20 science items as expository. The senior rater rated the rest of the 84 science items, all as expository).

Three bivariate correlations between text features and item difficulty were computed through SPSS 20 in regard to item stem, correct response, and distractors respectively.

Table 5.1.2.2 *Descriptive Statistics for Item Difficulty and Task Features of the TIMSS Science Test*

		Stem			
Variable	M	SD	Min	Max	N
Item difficulty	.000	.979	-2.170	2.228	104
CONADpi	12.333	25.843	.000	136.364	103
DENSPR2	.0668	.128	.000	.667	104
DENLOGi	24.452	38.3854	.000	166.667	103
READASW	1.434	.205	1.000	2.000	104
INTEi	17.932	32.031	.000	200.000	103
Graphic features	.29	.455	0	1	104
Question-abstractness	2.70	.974	1	4	104
		Correct Responses			
Variable	M	SD	Min	Max	N
Item difficulty	.003	.977	-2.170	2.228	96
FRQCLacw	1.755	.654	.000	3.200	96
READASW	1.697	.663	1.000	5.000	96
		Distractors			
Variable	M	SD	Min	Max	N
Item difficulty	.003	.977	-2.170	2.228	96
CONLGni	14.785	47.932	.000	200.000	96
DENNEGi	28.118	84.133	.000	450.000	96
FRQCLmcs	4.065	1.7647	.000	9.040	96
FRQCRmcs	169.550	219.216	0	130	96
READASL	13.572	9.622	2.89	47.00	96
READASW	5.424	1.958	1.000	11.500	96
WORDCmcs	1050.89	415.252	218	2037	89

Note. Item difficult: higher numbers = less difficult;
DENNEGi = Number of negations;
DENSNP = Noun phrase incidence;
DENSPR2 = Ratio of pronouns to noun phrases;
CONLGPi = Incidence of positive logical connectives;
CONLGni = Incidence of negative logical connectives;
DENNEGi = Number of negations;

FRQCLacw = Log frequency of all content words in the text;
FRQCLmcs = A mean of minimum LOG frequency scores among all of the content words in each sentence;
FRQCRacw = Raw frequency of content words;
FRQCRmcs = A mean of minimum frequency scores among all of the content words in each sentence;
HYNOUNaw = Mean hypernym values of nouns;
INTEi = Incidence of intentional actions, events, and particles;
Question-abstractness = the abstractness of an item question. This is an item characteristic rated through human coding;
READASL = Average Words per Sentence;
READASW = Average syllables per word;
SYNNP = Mean number of modifiers per noun-phrase;
SYNHw =Mean number of higher level constituents per word;
WORDCacw = Average concreteness of content words in a text;
WORDCmcs = Average low-concreteness words across sentences

Item Characteristics. Two reading experts coded item characteristics based on the coding scheme showed in the Appendix E. These two raters rated the first twenty items of the TIMSS science test, and the senior rated rest of the items.

The inter-rater reliabilities of rating items were calculated using Kappa statistics and intra-class correlation through SPSS 20. Table 5.1.2.3 presents results.

Overall, the ICC results indicate that the two raters achieved good agreement on the *Item 1*, *3*, and *5*. Particularly, the two raters reached excellent agreement on *Item 5* (the Abstractness of the Item Question). The coding scheme is presented in Appendix E.

Table 5.1.2.3 *Inter-Rater Reliabilities of TIMSS Science Item Ratings (n =20)*

	Kappa	ICC
Item 1	.184	.731
Item 3	.244	.784
Item 4	.225	.630
Item 5	.460	.841

Note. Kappa = Siegel & Castellan’s Kappa
ICC = Intra-class correlation

Graphic Features. Graphic features are pervasive in scientific texts. One distinguishing characteristic of the TIMSS Science test is the use of graphic features. Among the 104 science multiple-choice items, 30 items contain graphs, diagrams, or tables.

Contrasting to the belief that graphic features benefit comprehension, previous research (e.g., Mayer, 1993; Shah, Mayer, & Hegarty, 1999) suggest that graphics features can either facilitate or hinder text comprehension. For example, students can have difficulty interpreting quantitative information depicted in a bar graph or a scatterplot. In addition, Harp and Mayer (1998) found through an experimental study that when interesting but irrelevant graphic features were added to texts, students actually demonstrated worse memory and learning of the content than when the interesting information was not presented. They explained that irrelevant information is likely to drawing readers' attention away from the content that they are supposed to focus on.

To investigate the effect of graphics and table on students' test performance in the TIMSS Science test, I added graphic features as an additional variable in my analyses to predict the difficulty of items. To create this graphic feature variable, I coded items that contain graphs, diagrams, and/or tables as "1", and items that merely contain text as "0". The descriptive statistics of the graphic features variable are presented in Table 5.1.2.2.

Regression Analysis. In the next step, regression analyses were conducted through SPSS 20 for the item stems, correct responses, and distractors separately. The purpose of these analyses is to explore the effect of text features and item characteristics on item difficulty across all three parts of three items. For each analysis, task feature predictors which showed significant correlations with the item difficulty were entered into the regression model to further examine their associations with item difficulty values of science items. I also incorporated item characteristics—abstractness of item questions (from two raters) and graphic representation—into multiple regression analyses.

Before multiple regression analyses, I examined bivariate correlations among pairs of predictors in order to detect possible multicollinearity in multiple regression analysis. When a bivariate correlation between predictors was high (r > .80) and reading comprehension theories suggest that the two predictors were similar in terms of function (e.g., both measure word frequency), I removed one predictor from the multiple regression model. In addition, I used variance inflation factors (*VIF*) to detect possible multicollinearity in multiple regression analysis.

Stems. Table 5.1.2.4 shows the bivariate correlations among item difficulty, text features from the Coh-Metrix 2.0, and two item characteristic coded by human raters. Initially, text features with significant bivariate correlations with the item difficulty were entered into a multiple regression. They are readability features: READASW (average word length); features measuring syntactic complexity of the text: CONADpi, DENSPR2, DENLOGi; and a measure of situation model: INTEi. In addition, I added two item characteristics—the abstractness of questions and graphic features—into the initial multiple regression model.

After model comparisons, the best multiple regression model of stems is presented in Table 5.1.2.5. Two item characteristics—the abstractness of item question and graphic features—did not significantly predict the item difficulty values of TIMSS Science items. Therefore they were not retained in the final best regression model.

Table 5.1.2.4 *Correlations of Selected Task Features from TIMSS Test Item Stems*

	Item difficulty	CONADpi	DENSPR2	DENLOGi	INTEi	READ ASW	Graphic features	Question-abstractness
Item difficulty	--							
CONADpi	.204*	--						
DENSPR2	.172	.085	--					
DENLOGi	.209*	.589**	.082	--				
INTEi	.328**	.092	-.007	-.073	--			
READASW	-.191	.063	-.190	-.091	.025	--		
Graphic features	.032	.059	-.093	-.119	.186	-.007	--	
Question-abstractness	-.098	.021	.025	-.019	.115	-.004	.240*	--

Note. Item difficult: higher numbers = less difficult;

CONADpi = Incidence of positive additive connectives;

DENSPR2 = Ratio of pronouns to noun phrases;

DENLOGi = Logical operator incidence score (and + if + or + cond + neg);

INTEi = Incidence of intentional actions, events, and particles;

Question-abstractness = the abstractness of an item question. This is an item characteristic rated through human coding;

READASW = Average Syllables per Word.

Table 5.1.2.5 *Multiple Regression Results for Item Difficulty of TIMSS Science Test Item Stems*

Model	B	SE	*β*	t	Sig
(Constant)	.934	.640		1.459	.148
DENLOGi	.006	.002	.218	2.403	.018
INTEi	.011	.003	.348	3.853	.000
READASW	-.883	.437	-.183	-2.020	.046
R^2	.195				
Adjusted R^2	.171				

Dependent Variable: Item difficult: higher numbers = less difficult;
Note. DENLOGi = Logical operator incidence score (and + if + or + cond + neg);
INTEi = Incidence of intentional actions, events, and particles;
READASW = Average syllables per word.

Overall, the linear combination of reading-related task features is significantly correlated with item difficulty, *F (*3, 99) = 8.014, $p < .01$. The R^2 is .195, and the adjusted R^2 is .171, indicating that about 17% of the variance of item difficulty is accounted for by the reading-related text features after taking into account the number of predictors in the model. The results show that word length (average syllables per word) significantly predicted item difficulty ($\beta = -.183$, $p < .05$), meaning the more lengthy words in a stem, the more difficult the item was. Furthermore, a syntactic complexity feature, DENLOGi ($\beta = .218$, $p < .05$), is significantly associated with item difficulty when taking the other text features into account. Syntactic complexity index DENLOGi measures the frequency of logical operators that express logical reasoning in a text. The logical operators include *and*, *or*, *not*, *if*, *then* (McNamara, et al., 2005). The results reveal that item stems containing logical connectives such as *and, or, not if, then* are relatively easy, and students tend to answer them correctly.

Finally, at the item stem level, intentional verbs are significantly related to the item difficulty when other predictors are controlled (β = .348, p < .05). The INTEi index, which counts the frequency of intentional verbs, is a text cohesion feature pertaining to the situation model in Kintsch's reading comprehension theory. The index measures the intentional content and reflects the extent to which sentences are related by intentional particles (e.g., in order to, so that, for the purpose of, by means of, by, wanted to), actions, and events (McNamara, et al., 2005). Coh-Metrix 2.0 version measures the frequency of intentional actions and events by counting the number of intentional verbs based on a lexical database—WordNet (Fellbaum, 1998; Miller, Beckwith, Fellbaum, Gross, &Miller, 1990). To interpret the regression coefficient associated with the INTEi, I went back and scrutinized the original data and test items. I found that the TIMSS science items that contain verbs consistent with descriptions of intentional verbs in the Coh-Metrix Manual (McNamara, et al., 2005) were relatively easy, which means students tended to answer them correctly. This piece of evidence supports my previous hypotheses that the intentional verbs were likely to activate students' prior knowledge. That is, when the intentional verbs are in the item stem, they tend to activate students' existing knowledge that matches what the item intended to draw out. Therefore, the item is easier.

Below are some examples of item stems from the TIMSS science test. Coh-Metrix 2.0 identified these stem as containing high value of intentional verbs.

- Humans interpret seeing, hearing, tasting and smelling in the
- Fanning can make a wood fire burn hotter because the fanning
- The picture shows the three main layers of the Earth. Where is it the hottest?

- Why do mountain climbers use oxygen equipment at the top of the world's highest mountains?
- Which best describes the movement of the plates that make up Earth's surface over millions of years?

Correct Responses. Table 5.1.2.6 shows the bivariate correlations among item difficulty and text features from the TIMSS science test correct-responses. Among the 104 science multiple-choice items, 18 of them have alternatives that consist of graphic features and no text. I only analyzed alternatives that contain texts. Therefore the sample size for correct responses was reduced to 96. Bivariate correlations show that only two text features have significant relations with item difficulty: word length (READASW) and word frequency (FRQLacw). I entered them into a multiple regression as predictors. Results (Table 5.1.2.7) show that only 3% of variance in item difficulty is explained by these two features. In addition, F test (*F (*2, 93) = 2.510, *p* = .087) indicates that the linear combination of these two features cannot explain the item difficulty of test items to a statistically significant degree.

Table 5.1.2.6 *Correlations of Selected Task Features from TIMSS Test Correct Responses*

	Item difficulty	FRQCLacw	READASW
Item difficulty	--		
FRQCLacw	.194	--	
READASW	-.196	-.491**	--

Note. Item difficult: higher numbers = less difficult;
FRQCLacw = Celex, logarithm, mean for content words (0-6);
READASW = Average syllables per word
***p* < .01 level (2-tailed); * *p* <.05 level (2-tailed)

Table 5.1.2.7 *Multiple Regression Results for Item Difficulty of TIMSS Science Test Correct responses*

Model	B	SE	*β*	t	Sig
(Constant)	-.002	.522		-.004	.996
FRQCLacw	.193	.173	.129	1.113	.268
READASW	-.196	.171	-.133	-1.147	.254
R^2	.051				
Adjusted R^2	.031				

Dependent Variable: Item difficult: higher numbers = less difficult;
Note. FRQCLacw = Log frequency of all content words in the text;
READASW = Average syllables per word.

Distractors. Distractors from 96 multiple-choice items were used for the analysis. Each item contains three distractors. I analyzed distractors separately using Coh-Metrix 2.0. For each item, I then collapsed the outcome values for distractors to a single number by summing outcome values of the three distractors across rows.

Text features that have significant bivariate correlations with item difficulty were employed as predictors in multiple regression. Table 5.1.2.8 presents the bivariate correlations. These features are readability indices including average words per sentence (READASL), and average syllables per words (READASW), syntactic indices—CONLGni and DENNEGi, and word frequency indices—FRQCRmcs and FRQCLmcs. Particularly, FRQCRmcs is the frequency of content words in a text, and FRQCLmcs is the log frequency of content words. Bivariate correlations among the predictors show that these two indices are highly correlated with each other (*r* = .78), and they both measure the word frequency in a text. I kept FRQCRmcs **(**raw frequency of content words), and removed the FRQCLmcs (log raw frequency of content words) from the multiple regression model because the former is easier to interpret.

Table 5.1.2.8 *Correlations of Selected Task Features from TIMSS Test Distractors*

	Item	CONLGni	DENNEGi	FRQCLmcs	FRQC	READ	READ
	difficulty				Rmcs	ASL	ASW
Item difficulty	--						
CONLGni	.199	--					
DENNEGi	.187	.509**	--				
FRQCLmcs	.248*	-.026	-.005	--			
FRQCRmcs	.246*	.061	-.028	.780**	--		
READASL	.192	.440**	.301**	-.036	-.123	--	
READASW	-.183	-.124	-.077	-.062	-.065	-.262**	--

Note. Item difficult: higher numbers = less difficult;
CONLGni = Incidence of negative logical connectives;
DENNEGi = Number of negations;
FRQCLmcs = A mean of minimum LOG frequency scores among all of the content words in each sentence;
FRQCRmcs = A mean of minimum frequency scores among all of the content words in each sentence;
READASL = Average Words per Sentence;
READASW = Average syllables per word;

After model comparisons, results (Table 5.1.2.9) suggest that the reading-related task features for distractors explained about 10 percent of variance in item difficulty (R^2=. 094, *F (*2, 93) = 4.829, p < .01). Word frequency (FRQCRmcs, β = .234, p < .05) and the frequency of the negative logical connectives (CONLGni, β = .184, p < .05) significantly predicted item difficulty. FRQCRmcs initially computes the lowest frequency score among all of the content words in each sentence (McNamara, et al., 2005). The results suggest that TIMSS Science distractors that contain frequent words are associated with items that are relatively easy. Additionally, distractors with negative logical connectives (CONLGni) are associated with relatively easy items. One example of such distractor is “no change in pulse but a decrease in breathing rate”. Another example is “from either his father or his mother, but not from both”. One possible explanation is that these negative connectives were likely to increase cognitive load when students processed the option. Students, especially those who were not highly motivated to take the test, were likely to ignore such alternatives in an item. Instead, they focused on other alternatives (including the correct responses) that were easier to process. As a result, these items were relatively easy. After having securitized the science items, I also found that item alternatives that contain negative logical connectives are most likely to be distractors. That explains why the current study did not find that negative logical connectives predicted item difficulty when analyzing the correct responses of the TIMSS Science items.

Table 5.1.2.9 *Multiple Regression Results for Item Difficulty of TIMSS Science Test Distractors*

	B	SE	β	t	Sig
(Constant)	-.229	.124		-1.846	.068
CONLGni	.004	.002	.184	1.862	.066
FRQCRmcs	.001	.000	.234	2.369	.020
R^2	.094				
Adjusted R^2	.075				

Dependent Variable: Item difficult: higher numbers = less difficult;
Note. CONLGni = Incidence of negative logical connectives;
FRQCRmcs = A mean of these minimum frequency among all of the content words in each sentence.

Summary. In summary, results from the TIMSS science multiple-choice items show that at the stem level, the difficulty levels of science items are associated with task features, especially linguistic features including word length, logical connectives, and intentional verbs. On average, lengthy words are related to difficult items. Logical connectives, on the other hand, are associated with easy items. Students are more likely to answer items correctly if descriptions in item stem are connected by a logical connective such as *and, or, not if, then*. In addition, items are relatively easy if they include intentional verbs at the stem level.

In correct item responses, word frequency significantly predicts item difficulty of TIMSS test items. In average, the more frequent the words contain in correct responses, the easier item appears to be.

With respect to distractors, I had similar findings. That is, in distractors frequent words are associated with items that are relatively easy. Moreover, negative logical connectives predict item easiness.

In my analyses, I did not find statistically significant correlations between item difficulty and item characteristics (human rating) including graphic features and the abstractness of item questions when other text features were controlled in multiple regressions.

5.2 Research Questions 2 and 3

My analyses in the previous section contributes important evidence to understand the role of reading comprehension in subject-matter tests including CIVED and TIMSS science by providing a list of linguistic features and estimates of their impact on item difficulty. Furthermore, the substantial amount of variance explained by these linguistic features suggests that reading is a vital part of the subject-matter assessments when answering the test item questions. In this section, I move from the item level to the student level to find out:

(1) How to quantify levels of reading demand?

(2) To what extent will an advanced statistical model—a multidimensional IRT model--partial out the noise associated with high level of reading demand?

(3) To what degree do the average estimated scores of the domain-specific proficiency change after taking into account the reading demand of test items?

(4) Does the relation between the reading and students' domain proficiency vary by gender and language status in each subject-matter assessment?

5.2.1 Reading Demand

Levels of reading demand were quantified through the regression method. That is, the level of reading demand corresponding to each test item was predicted by text features which were previously identified as significantly related the item difficulty.

The CIVED Test of 1999. In the previous section, I identified task features that significantly predicted item difficulty of the CIVED test using multiple regressions analyses. At the stem level, these text features were (1) the number of noun phrases, (2) the average syllables per word, and (3) the average concreteness of content words (identified by Coh-Metrix 2.0). At the correct-response level, the predictor was the average syllables per word. I combined these significant features from stems and correct-responses in a final regression model and used these features to predict the item difficulty. By doing so, each item had a predicted value yielded from the multiple regression as a function of these salient linguistic features that correspond to this item. The predicted value of items from the multiple regression model were used as the predicted reading difficulty (predicted reading demand) of these items.

I did not include features from distractors in to the final model for two reasons. First of all, my sample size is small (n = 38). Second, in well-constructed tests such as the CIVED, the features of all multiple-choice options are fairly similar – for example, there are few multiple-choice items where there is one choice with a strikingly different sentence complexity or length (item-writing guides advise against this). Therefore, in order to reduce the number of predictors, I only picked the keyed alternatives as a representative of all item alternatives.

Overall, significant test features explained 42.2% of variance in the item difficulty (adjusted *R* square = .352). Through the multiple regression analysis, I obtained a standardized predicted difficulty as a function of salient text features. The predicted variable was used as the predicted reading difficulty of the 38 items. The mean of the standardized predicted difficulty is 0.00 and standard deviation is 1.00. The predicted difficulty values range from -.2.626 to 1.715 with lower values leaning towards difficulty and higher values easiness. The predicted difficulty values are normally distributed based on Histogram, Kolmogorov-Smirnov test, and Shapiro-Wilk test.

The TIMSS Science Test of 1999. The reading difficulty of 104 TIMSS Science items was predicted through the same method as I used for the CIVED test items. In previous multiple regression analyses, I identified that only stem level linguistic features were significantly related to the item difficulty of the 104 items. These features are (1) the frequency of logical operator, (2) number of intentional actions, events, and particles, and (3) average syllables per word. By regressing these features on item difficulty, I obtained a standardized predicted variable of difficulty as a function of these three salient linguistic features. Then this predicted variable was used as the predicted reading difficulty of the 104 items.

Overall, the three linguistic features only explain 20 % of variance of item difficulty (whereas linguistic features on the CIVED test items explain 42.2% of the variance of item difficulty). The standardized predicted difficulty range is from -1.464 to 4.255 with a mean of 0.00 and standard deviation of 1.00.

In summary, these pieces of evidence suggest that the reading demand of the science items explains one fifth of variance in overall item difficulty. Hence it is likely that reading demands of test items influence students' test performance. It is the true for the civics test as well for the science test. Grounded on these findings, a further inference is that students' likelihood of correct responses on test items depends on their possession of two or more correlated proficiencies: reading and domain proficiencies.

5.2.2 Multidimensional IRT Modeling

Next, I applied an advanced statistical model—multidimensional IRT (MIRT) models (von Davier, 2005)—to students' item responses to model the potential multiple latent proficiencies underlying their responses. The assumption is that items with high level of reading difficulty require at least two latent proficiencies: subject-matter proficiency and high level of linguistic ability that is independent of the subject proficiency. If variance associated with the high level of reading demand can be separated by a multidimensional IRT model, it is hypothesized that the standard error of estimated scores of the subject-matter proficiency would change.

I tested this assumption by fitting two types of IRT models to each subject-matter assessment (the TIMSS science test of 1999 and the CIVED test of 1999). One is a two-dimensional MIRT model that assumes that there are at least two latent proficiencies underlying the items: high level of reading ability and subject-matter proficiency. Items with low reading demand depend on just one attribute (the subject-matter proficiency), and items with high reading demand depend on that subject-matter attribute and also the reading attribute. I also fitted a one-dimensional Rasch IRT model to each assessment assuming that reading ability is a part of the subject-matter proficiency. Rasch model is often used in large-scale educational assessments to estimate the subject-matter achievement scores. I examined whether there are differences between the domain achievement scores from these two runs in terms of standard error of estimate. To ensure that these two types of IRT models (one-dimensional and multidimensional) are comparable, I set them on the same scale, and restricted the item parameters so that both models are 1PL models. I will describe the model specifications including restrictions for each assessment in the following sections.

The CIVED Test of 1999. My first set of analyses focused on the 38 multiple-choice items from the CIVED assessment of 1999. The data source is item responses from 2786 U.S. students on 38 CIVED items. Twenty five cases (students) were excluded from the analysis because these cases have missing scores on all 38 test items. House weights in the IEA CIVED 1999 dataset were applied when I conducted parameter estimations and student level statistical analyses.

I applied a multidimensional IRT (MIRT) model from the General Diagnostic Model (GDM) framework to the data through the *mdltm* softwar (von Davier, 2009). The MRT assumes one general CIVED achievement attribute, and a reading attribute. For many large-scale subject-matter assessments such as CIVED and TIMSS, the users are often interested in students' achievement scores yield from the test. Therefore, I focus on the CIVED achievement attribute and examine to what degree the reading attribute is correlated with it.

The relation between the latent attributes and items was specified through the design matrix, or Q matrix (see Table 5.2.1). Items with low levels of estimated reading difficulty (reading demand) are modeled as depending on just the CIVED achievement attribute, and items with high reading difficulty depend on the CIVED attribute and also a reading attribute.

To determine the high level of reading difficulty, I first sorted the items based on their predicted difficulty values with higher values leaning to easiness and lower values difficulty. This sorting was therefore not an ordering on actual item difficulty, but rather an ordering based on prevalence of linguistic features that tend to make items difficult. For example, an item that contains more lengthy and abstract words may have higher predicted difficulty than another item that has less lengthy and more concrete words when other linguistic features of these items are equivalent.

I then used the 25th percentile as a cutoff criterion to decide whether an item has high predicted reading difficulty or not (I also tried 1/3 but the 25th percentile provided the best model fit when I fit the 2-dimensional Rasch model to the data). After sorting, 28 items with predicted reading difficulty of -.440 and greater were set as depending only on the CIVED achievement dimension, and other 10 items depend on both CIVED achievement and reading dimensions in the Q matrix (see Table 5.2.1). In Table 5.2.1, "1" indicates that the item depends on the attribute and "0" otherwise. I labeled these items below the 25th percentile as linguistically complex, and items above as linguistically simple.

Table 5.2.1 *The Q-Matrix of 2-dimensional IRT Model of CIVED Test*

Items	CIVED Achievement	High Level Reading
1	1	0
2	1	0
3	1	0
4	1	0
5	1	0
6	1	0
7	1	0
8	1	0
9	1	1
10	1	0
11	1	1
12	1	1
13	1	0
14	1	0
15	1	1
16	1	0
17	1	1
18	1	0
19	1	1
20	1	0
21	1	1
22	1	1
23	1	0
24	1	0
25	1	0
26	1	0
27	1	1
28	1	0
29	1	1
30	1	0
31	1	0
32	1	0
33	1	0
34	1	0
35	1	0
36	1	0
37	1	0
38	1	0

Finally, a 2-dimensional 1PL IRT model from the General Diagnostic Model (GDM) framework (von Davier, 2005) was applied to the data through the *mdltm* software (von Davier, 2009). To compare the model, I also fitted a Rasch model to the data assuming that only one dimension, CIVED achievement, was underlying the 38 test items. To ensure that outcomes from the 2-dimensional IRT model and the Rasch model were comparable, I set the person parameters (latent variables) of the two models on the same scale. More specifically, I did so by centering the scale on the item scale, meaning that the item difficulties set the scale. The *mdltm* does so by making the average difficulty the same (mean = 0), which means that the location of the person parameters is on the same scale with respect to the common difficulty location.

In addition, for both models, I set the skill levels of each latent dimension to 15 to approximate a continuous normal latent skill distribution, and the range of each latent skill distribution from -3.0 to 3.0. For both models, I constrained the slope parameter γ to be 1.0. As for the trait distribution, I set it as saturated (based on the *mdltm* manual, this means there are no constraints) (more details about the model settings are described in *mdltm user manual,* Seo, Xu, & von Davier, 2009).

Results. Table 5.2.2 presents indices of model-data fit for the two IRT models. Log-likelihood, Akaike information criterion (AIC), and Bayesian information criterion (BIC) suggest that there is no salient difference between these two models in terms of model fit. Descriptive statistics of the student-level outcomes from the two runs are shown in Table 5.2.3. On average, the standard error of estimate (SE) of the CIVED achievement scores from the 2-dimensionalmodel is smaller than the SE of the same attribute scores from the Rasch model, which suggests that the quality of domain achievement scores slightly improved when using the 2-dimensional model.

Table 5.2.2 *Model Fit for CIVED Data*

Model	# of parameters	Log-Likelihood	AIC	BIC
Rasch	50	-53473.623	107047.245	107343.863
2-dim Rasch	259	-53306.172	107130.344	108666.826

Note. AIC = Akaike information criterion
BIC = Bayesian information criterion
2-dim Rasch = 2-dimensional Rasch Model

Table 5.2.3 *Descriptive Statistics of the Standard Error of the Estimate from the Rash Model and Multidimensional IRT model*

	N	Mean	SD	Min	Max
1dim-SE of CIVED scores	2786	0.372	0.094	0.133	0.615
2dim-SE of CIVED scores	2786	0.367	0.092	0.071	0.664
2dim-SE of Linguistic scores	2786	0.34	0.128	0.235	1.394

Note. 1dim = 1 dimensional Rasch model
2dim = 2 dimensional Rasch model
S.E. = Standard error of estimate

In addition to the comparison between the standard error of estimates, I also compared students' achievement scores from the 2-dimensional model and 1-dimensional model using bivariate correlation. Results showed that the correlation coefficient was .983 (p < .01). It implies the MIRT model did not separate out much reading demand that is independent of domain achievement.

In conclusion, my results suggest that the multidimensional IRT methods separated out a small amount of variance associated with the reading demand from the total variance of the CIVED achievement measure, but this did not make a substantial difference when comparing with scores yielded from the one dimensional model. One possible explanation is that the reading demand on CIVED items was within the reading capabilities of almost all of the students, and there was not much variance associated with reading demand to be teased out. Future study should apply the MIRT method to other civics assessments that involve more reading (e.g., the IEA Civic and Citizenship Education Study—ICCS—2009) to explore if there is substantial amount of reading demand variance can be separated out.

Next, I applied the same methods to the TIMSS science assessment of 1999 to see whether there is a different pattern.

The TIMSS Science Test of 1999. At this step, I focused on the 104 multiple-choice items from the TIMSS science assessment of 1999. The data source is the item responses from 9072 U.S. students on 104 science multiple-choice items. The TIMSS test involved a booklet design (i.e., matrix sampling design, Gonzales & Miller, 2001). Therefore, each student was only administrated a small proportion of the 104 items. In the TIMSS student-level dataset, an item that was not assigned to a student was marked as “not administered”, and the *mdltm* software treats this item as if it was not administered to the student when the software estimates item parameters and person parameters through the one dimensional Rasch model and MIRT model.

House weights in the IEA TIMSS 1999 science dataset were applied when I conducted parameter estimations and statistical analyses at the student level.

For the predicted reading difficulty, I compared two cutoff scores by applying them to 2-dimensional MIRT model analyses. The two cutoff criteria are the 25 percentile (i.e., -.741) and -.440 (the cutoff point the same as in the CIVED test previously reported). Results from the two runs show that when I used the -.440 as the cutoff point, I obtained a better model fit from the 2-dim MIRT model. For that reason, I used -.440 as the cutoff criterion throughout the subsequent analyses.

IRT Model Specification. I applied a 2-dimensional Rasch Model to the data through the *mdltm* software. This model assumes one general science achievement attribute, and a linguistic (reading) attribute. The relation between the latent attributes and items was specified through the design matrix—Q matrix (See Table 5.2.4). Sixty-two items with predicted reading difficulty of -.440 or greater were set to depend only on the science achievement dimension, and other 42 items were set as depending on both science achievement and reading dimension.

To compare the outcomes from the 2-dimensional Rasch model, I also fit a 1-dimensional Rasch model to the TIMSS data using *mdltm*. To make sure that the outcomes are on the same scale and comparable, I applied similar model setting approach as I did for the CIVED data and set both models on the same scale. That is, I centered the scale on the item scale so that the item difficulties set the scale. For both models, I constrained the slope parameter γ to be 1.0. In addition, I set the skill levels of each latent dimension to 15, and the range of each latent skill distribution from -3.0 to 3.0. (more details about the model settings are described in *mdltm user manual*, Seo, Xu, & von Davier, 2009).

Table 5.2.4 *The Q-Matrix of 2-dimensional IRT Model of CIVED Test*

Item	Science Achievement	High Level Reading	Item	Science Achievement	High Level Reading	Item	Science Achievement	High Level Reading
s012005	1	1	**s022275**	1	1	**s012045**	1	0
s012009	1	1	**s022202**	1	1	**s012047**	1	0
s012010	1	1	**s022157**	1	1	**s012048**	1	0
s012011	1	1	**s022054**	1	1	**s022183**	1	0
s012014	1	1	**s022181**	1	1	**s022276**	1	0
s012017	1	1	**s022126**	1	1	**s022019**	1	0
s012020	1	1	**s012001**	1	0	**s022002**	1	0
s012021	1	1	**s012002**	1	0	**s022294**	1	0
s012025	1	1	**s012003**	1	0	**s022073**	1	0
s012030	1	1	**s012004**	1	0	**s022009**	1	0
s012032	1	1	**s012006**	1	0	**s022012**	1	0
s012036	1	1	**s012007**	1	0	**s022117**	1	0
s012039	1	1	**s012008**	1	0	**s022235**	1	0
s012042	1	1	**s012012**	1	0	**s022074**	1	0
s012044	1	1	**s012013**	1	0	**s022240**	1	0
s012046	1	1	**s012015**	1	0	**s022058**	1	0
s022115	1	1	**s012016**	1	0	**s022295**	1	0
s022106	1	1	**s012018**	1	0	**s022194**	1	0
s022150	1	1	**s012019**	1	0	**s022187**	1	0
s022042	1	1	**s012022**	1	0	**s022222**	1	0
s022099	1	1	**s012023**	1	0	**s022040**	1	0
s022082	1	1	**s012024**	1	0	**s022007**	1	0
s022094	1	1	**s012026**	1	0	**s022238**	1	0
s022278	1	1	**s012027**	1	0	**s022145**	1	0
s022225	1	1	**s012028**	1	0	**s022178**	1	0
s022188	1	1	**s012029**	1	0	**s022030**	1	0
s022206	1	1	**s012031**	1	0	**s022041**	1	0
s022014	1	1	**s012033**	1	0	**s022280**	1	0
s022131	1	1	**s012034**	1	0	**s022245**	1	0
s022132	1	1	**s012035**	1	0	**s022290**	1	0
s022118	1	1	**s012037**	1	0	**s022208**	1	0
s022123	1	1	**s012038**	1	0	**s022264**	1	0
s022293	1	1	**s012040**	1	0	**s022064**	1	0
s022137	1	1	**s012041**	1	0	**s022254**	1	0
s022198	1	1	**s012043**	1	0			

Results. Table 5.2.5 presents indices of model-data fit for the two IRT models. The model fits are similar as what I found from the CIVED data. That is, both 2-dimensional model and 1-dimensional model fitted the TIMSS science data equally well. Descriptive statistics of the outcomes from the two models (Table. 5.2.6) show that the average standard error of estimate (SE) of the science achievement scores from the 2-dimensional model was smaller than the average SE of the same attribute scores from the 1-dimensional model. Overall, outcomes from these analyses suggest that the quality of science achievement estimates slightly improves when using the 2-dimensional model because on average standard error of estimation is lower when using the MIRT model.

Bivariate correlation between the science scores from the two models is .997, meaning achievement scores from the MIRT model are highly correlated with scores from the 1-dimensional IRT model. This indicates that the MIRT model did not tease out much noise that was associated with reading demand and independent of domain knowledge. Given that the reading-related features only explain about 20% of variance in item difficulty of the TIMSS science items, it is likely that there was not much linguistic-related noise in the test for the MIRT model to partial out. The small standard deviation of the linguistic attribute (SD = .228) also suggest the variance associated with reading is limited. In the future, it would be interesting to look at other science tests (e.g., PISA science literacy tests) that involve linguistically complex items, and to use this study as a baseline for comparisons.

Table 5.2.5 *Model Fit for the TIMSS Science Data*

Model	# of parameters	Log- Likelihood	AIC	BIC
Rasch	116	-143662.925	287557.850	288382.952
2-dim Rasch	325	-143624.373	287248.746	290210.454

Note. AIC = Akaike information criterion
BIC = Bayesian information criterion
2-dim Rasch= 2-dimensional Rasch model

Table 5.2.6 *Descriptive Statistics of the Standard Error of the Estimate from the Rash Model and Multidimensional IRT model*

	N	Mean	SD	Min	Max
1dim-SE of Science scores	9072	0.404	0.055	0.295	0.761
2dim-SE of Science scores	9072	0.398	0.045	0.242	0.922
2dim-SE of Linguistic scores	9072	0.248	0.171	0.115	1.41

Note. 1dim = 1 dimensional Rasch model
2dim = 2 dimensional Rasch model
SE = Standard error of estimate

Summary. Overall, results from CIVED test and TIMSS science test suggest that it is possible to partial out the noise associated with reading demand using the multidimensional IRT model. For each subject-matter achievement variable, the average standard error of estimate decreased when compared with the average standard error of estimate from the 1-dimensional IRT. Future studies should apply this method to subject-matter assessments, such as PISA science literacy test and the ICCS which involve relatively higher reading demand, to find out more.

Third, for both TIMSS and CIVED assessments, I found the same achievement scores yield from Rasch model and MIRT model are highly and positively correlated. This suggests the MIRT model did not partial out a substantial amount of reading demand that was independent of domain-knowledge demand.

5.2.3 Demographic Groups

My research question 3 asks whether the relation between reading demand and domain achievement varies by gender and by students' language status.

According to test validity theories (e.g., Messick, 1989) and Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1999), reading involved in a subject-matter assessment can be construct-irrelevant variance that biases the test score interpretations, especially for demographic groups who have disadvantages in language, such as English Language Learners (ELL) and students with disabilities. The bias may cause or increase achievement gaps between demographic groups. In this case, reading involved in the subject-matter assessment becomes construct-irrelevant variance.

My analysis at this step aims to test whether the reading demand involved in the CIVED test and TIMSS test is construct-irrelevant in the sense that it contributed to the achievement discrepancies between boys and girls, ELLs and non-ELLs. If high level reading demand was one of the factors that contributed to the achievement gap between certain demographic groups (e.g., boys and ELLs), then once the MIRT model teased out noise associated with the excessive reading demand, the average domain achievement scores of the low-language groups would increase. Furthermore, the achievement discrepancy between low language ability group and high language ability group would decrease. I performed mixed ANOVA for each assessment to test this hypothesis.

CIVED Assessment of 1999. Two mixed ANOVAs (both being a one Between-Subject effect and One Within-Subjects effect Design) were conducted for the CIVED data. The first ANOVA had the repeated measures of civic achievement scores from the Rasch model and 2 dimensional IRT model as two within-subjects variables, and gender as a between subject variable. *F* tests from the mixed ANOVA show there are a statistically significant within subject effect (*F* (1, 2935) = 8604.789, *p* < .01, partial Eta squared = .746), and between subject effect (*F (*1, 2935) = 15.088, *p* < .01, partial Eta squared = .005). In addition, the interaction effect is statistically significant (*F (*1, 2935) = 46.447, *p* < .01, partial Eta squared = .016).

Partial Eta squared is an effect size index that measures the ratio of variance explained in the dependent variable by a predictor (independent variable) controlling for other predictors (independent variables or interactions). In the partial Eta square measure, the effects of other predictors (independent variables or interactions) are partialled out (Richardson, 2011). Partial Eta squared varies from 0 to 1, and takes value of 1 “when all of the independent variables and interactions in the design explain all of the variance in the dependent variable” (Richardson, 2011, p. 141). In the case of the repeated measures ANOVA (or Within-Subjects effect Design), the partial eta squared refers specifically to how much of the variation between occasions can be explained by occasion. In this case, a value near one means that a mean shift characterized by the occasion effect accounts for nearly all the differences between students' two scores.

Generally speaking, the results imply that there is a statistically significant change within the two measures, but the change varies by gender. That is, the discrepancy of mean scores between boys and girls slightly broadens when using the MIRT model to estimate the achievement scores. However, the partial eta square of the interaction indicates that the effect is small (Cohen, 1988) Figure 5.2.1 presents interaction effect, which is almost negligible substantively.

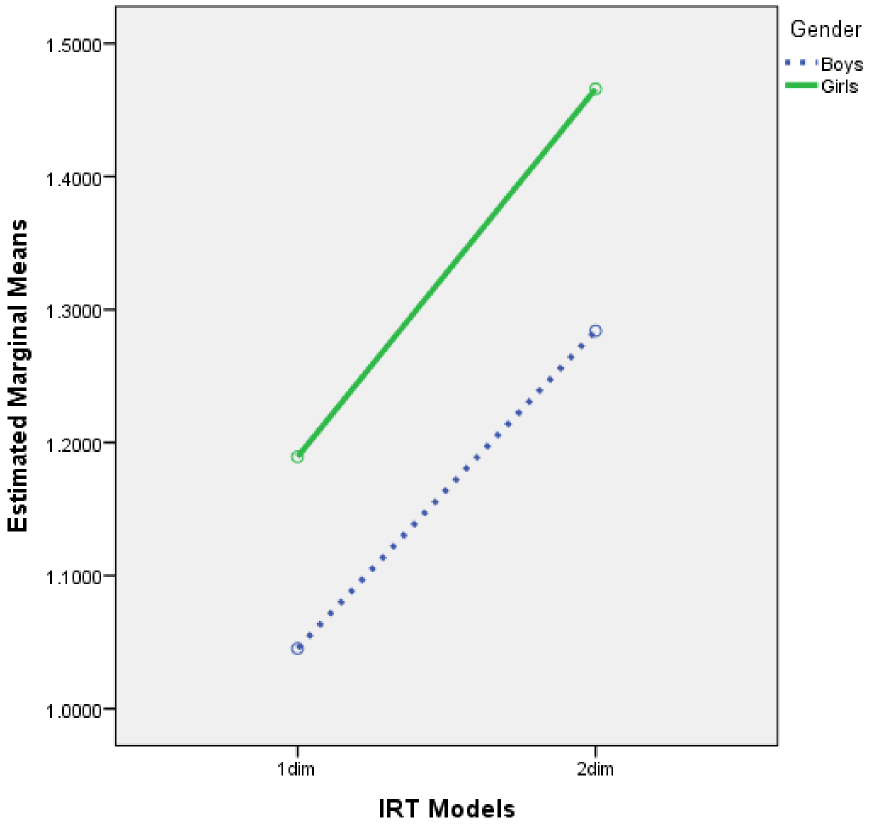


Figure 5.2.1. The interaction effect between gender and CIVED achievement scores from two IRT models

In the second mixed ANOVA, students' language status replaced gender as the between-subject variable. The ANOVA results indicate that on top of the significant within-subject effect (partial Eta squared = .448), there are a significant language status effect (*F (*1, 2868) = 83.909, *p* < .01, partial Eta squared = .028), and an interaction effect (*F (*1, 2868) = 19.106, *p* < .01, partial Eta squared = .007). Similar to the first mixed ANOVA, the interaction effect between civic scores and students' language status is statistically significant but counterintuitive. That is, the interaction indicates that the mean score discrepancy between ELLs and non-ELLs broadens when using the MIRT model. Non-ELLs who often have higher language proficiency benefit slightly more from methods of computing scores that attempt to separate out reading. Figure 5.2.2 shows the interaction effect. Overall, the results suggests that on average students' CIVED achievement scores increased when using the new method in which some noise associated with the reading demand was teased out. However, the non-ELLs seem to benefit more from the multidimensional model.

In summary, the mixed ANOVA results confirmed my findings previously. That is, reading demand had an influence on students' performance in CIVED test. When items with high level reading demand were taken into consideration, achievement scores from the MIRT model increased across students. Boys and ELLs benefited from using the advanced measurement model. However, groups (girls and non-ELLs) that have advantages in language benefited more from this approach. Next, I applied the same method to the TIMSS science data to explore whether the change of repeated science scores varies by gender and language background.

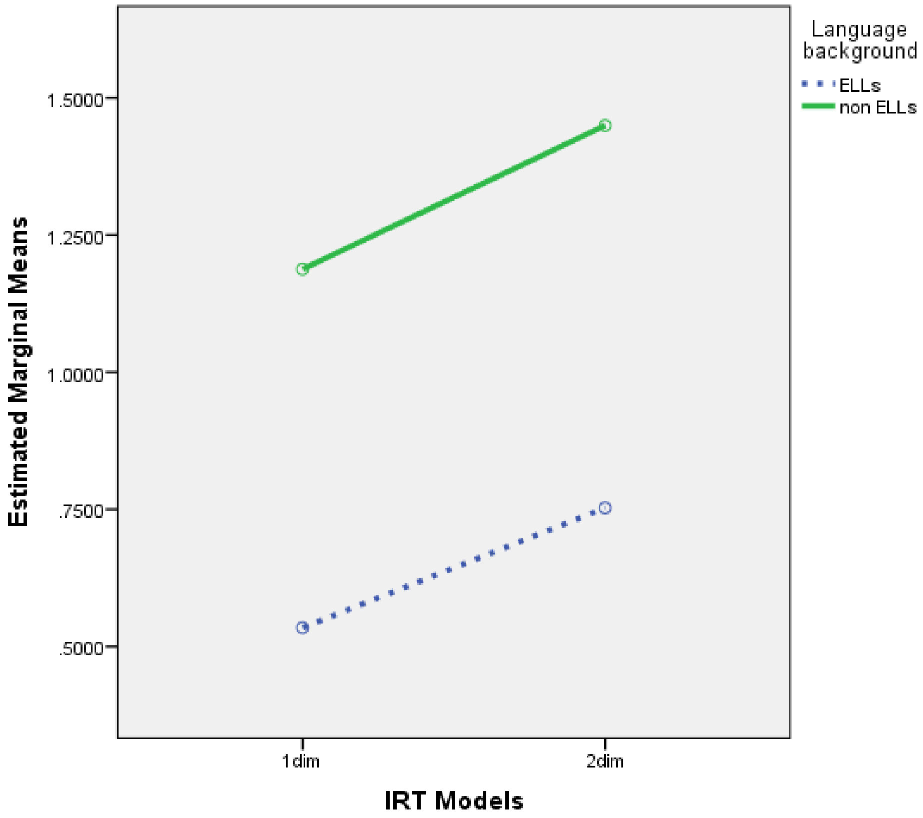


Figure 5.2.2. The interaction effect between language background and CIVED achievement scores from two IRT models

TIMSS Science Assessment of 1999. Two mixed ANOVA (One Between-Subject and One Within-Subjects Design) were applied to the TIMSS science assessment. The first mixed ANOVA model has gender as the between-subject factor, and the within-subjects variables are two repeated science achievement scores from Rasch model and 2-dimensional IRT model. Results reveal similar patterns to those from the CIVED assessment. That is, both the within-subjects effect (*F* (1, 10160) = 125862.701, *p* < .01, partial Eta squared = .925) and between-subject effect are statistically significant (*F* (1, 10160) = 115.790, *p* < .01, partial Eta squared = .011). The interaction is also significant (*F* (1, 10160) = 26.005, *p* < .01, partial Eta squared = .003). These indicate that both groups attain increased scores when reading demand is separated out by the MIRT model. However, girls, the group often thought of as disadvantaged in science and expository text, benefit more from the advanced IRT model (see Figure 5.2.3).

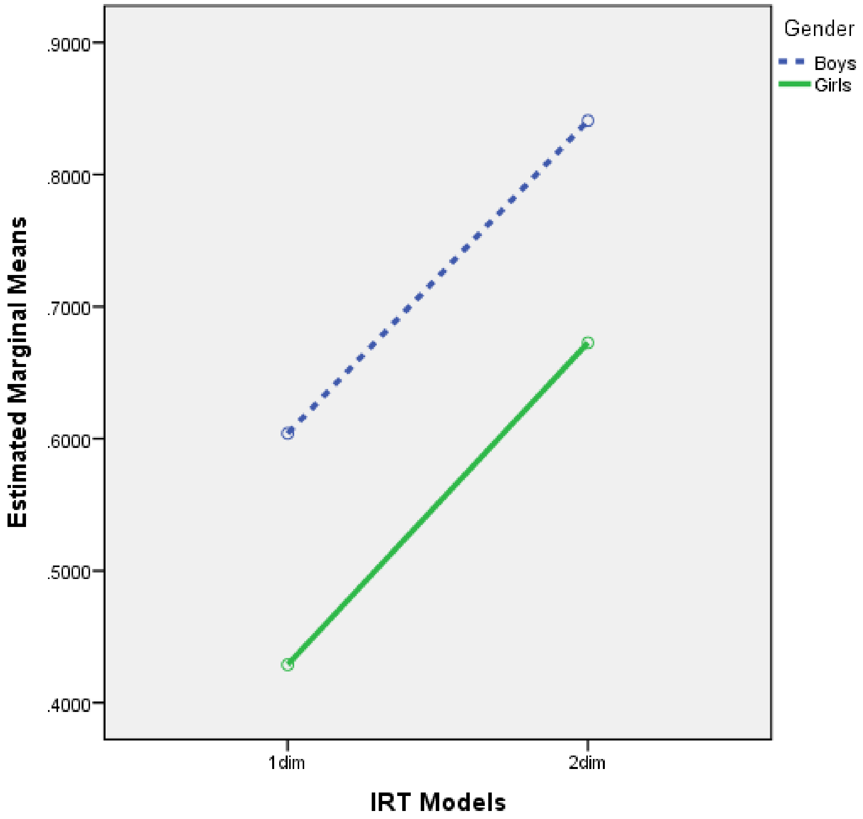


Figure 5.2.3. The interaction effect between gender and TIMSS Science achievement scores from two IRT models

The second mixed ANOVA replaced gender with language status as the between-subject variables. Results show that the within-subjects effect (*F* (1, 9806) = 42220.263, *p* < .01, partial Eta squared = .812) and between-subject effect (*F* (1, 9806) = 324.907, *p* < .01, partial Eta squared = .032) are statistically significant, but the interaction is not significant (*F* (1, 9806) = .069, *p* =.793, partial Eta squared = .000). It means that scores increase across students when using the MIRT model. ELLs and non-ELLs groups benefit from this approach equally (see Figure 5.2.4).

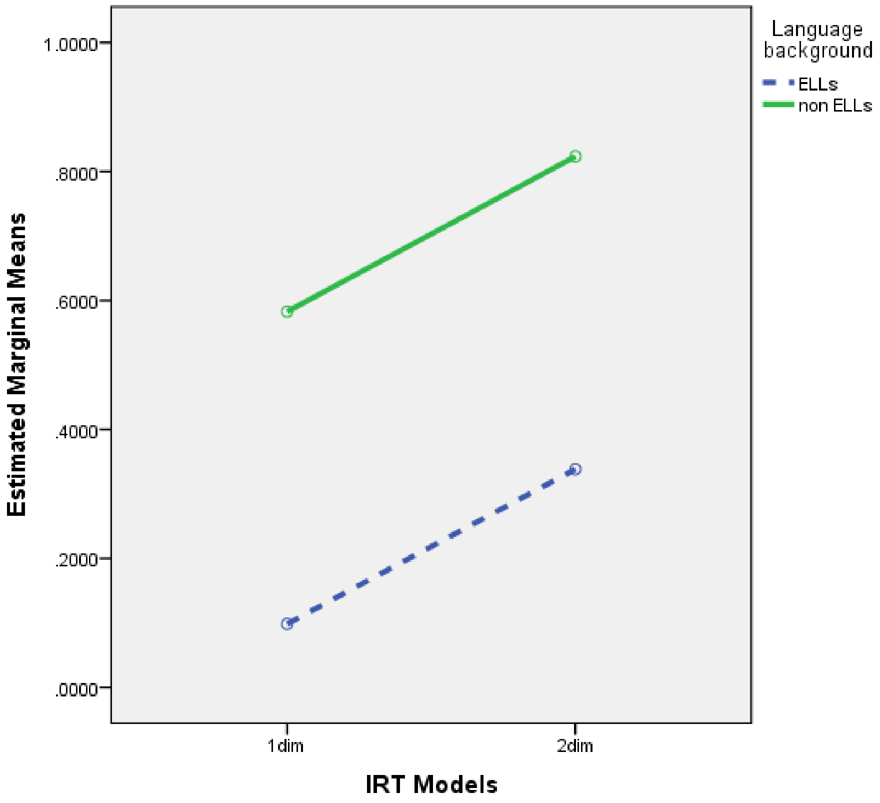


Figure 5.2.4. The interaction effect between language background and TIMSS Science achievement scores from two IRT models

Summary. Overall, results from both CIVED and TIMSS assessments show that after the high level of reading demand in some items was modeled by the MIRT model, domain achievement scores increased across students when compared with the same measures from traditional Rasch model. This suggests that the MIRT model successfully separated out a certain degree of noise associated with the high level of reading demand. However, the groups that might be thought disadvantaged by reading demand did not benefit from the adjusted scoring.

5.3 Research Questions 4 and 5

With respect to the two additional subject-matter assessments, my research question 4 asks: Is there a relation between the general word knowledge and students' achievement scores in each subject-matter assessment? To further explore the relation, my research question 5 asks: Does the relation between the general word knowledge and students' achievement scores vary by demographic factors? I employed the 1970s Six Subject Surveys in Civic Education and Science to answer these two questions, where students' general word knowledge had been assessed in an independent scale.

5.3.1 1970s Six Subject Survey in Civic Education

I examined a total of 3207 U.S. 14-year-old students from the IEA Six Subject Survey in Civic Education. They were nationally representative sample of U.S. 14-year-old students. The stratum weights in the 1970s' Civic Education dataset were applied when I conducted statistical analysis and parameter estimation.

The cognitive Civics test (scale) contains 47 multiple-choice items that measure students' conceptual knowledge in Civic Education, and the general word knowledge test (scale) contains 40 items. Because this study was conducted in early 1970s, the data set only reported total raw scores of correct responses for each student in each scale (Torney, Oppenheim & Farnen, 1975). To take into consideration measurement errors and the item characteristics, I applied a Rasch model to students' item responses in each scale through the *mdltm* software in order to obtain IRT scores for each scale. The domain achievement scores estimated by the Rasch model from the cognitive Civics test were assumed to conflate Civic-related proficiencies including domain knowledge, civics skills, and basic reading competence. The word knowledge IRT scores from the general word knowledge test were assumed to reflect students' knowledge in general vocabulary. The scale settings of each Rasch model here were kept the same as the ones applied to CIVED and TIMSS tests.

The ability scores (IRT theta scores) of each scale are the focus of this study. They are theta values of civic test (M=.768, SD = .834, Min =-1.7, Max =2.78), and theta values of word knowledge (M= 1.216, SD = .603, Min = -.432, Max = 2.913). I standardized both scores to z scores for subsequent analyses.

Next, I performed a set of multiple regression to answer my research questions 4 and 5. The criterion variable is the standardized IRT scores of civics cognitive test (M =.00, SD = 1.00, Min=-2.977, Max= 2.416). The predictors are (1) the standardized IRT scores of word knowledge test (M =.00, SD = 1.00, Min=-2.732, Max= 2.814), (2) gender (boys (0) =1621, girls (1) =1493), and (3) the interaction between gender and the word knowledge (the product of the gender variable and standardized word knowledge IRT scores). The missing data contained in each of the criterion variable and predictors were less than 8% of total sample. Listwise deletion therefore was used for the subsequent analyses.

At the first step, I entered the standardized word knowledge scores in to the regression model to predict the standardized civics scores. Results show that the word knowledge scores explain 47.4 percent of variance in civics scores (R^2=.474, F (1, 3184) = 2871.395, p < .01). The standardized regression coefficient of word knowledge β was .689 (p < .01). Because both the criterion variable and predictor were standardized, the regression coefficient reflects the correlation of these two variables, that is, .689.

Gender. Next, I used gender to predict the standardized civics scores and standardized word knowledge scores. Results show a statistically significant gender difference in terms of civic scores (β = -.037, p < .05). On average, boys have high civic scores than girls. Gender explains 0.1 percent of variance of the civics scores. However, there is no gender difference in terms of word knowledge scores (β = -.001, p = .976).

Finally, I used word knowledge, gender, and the interaction between word knowledge and gender to predict the standardized civics scores (Table 5.3.1). Overall, the predictors explain 49.2 % of variance in civics scores (F (3, 3094) = 998.815, p < .01). Interaction is not statistically significant (β = -.026, p =.127) after taking main effects—the word knowledge scores and gender—into account. Figure 5.3.1 illustrates that non-interaction effect between boys and girls.

In summary, the multiple regression analyses reveal that students' word knowledge scores are highly related to their civics achievement scores, and this relation did not vary by gender. In other words, boys tend to have higher civics scores than girls regardless of their word knowledge. In this case, if word knowledge capability influenced boys' civics test performance, it also impacted girls' performance equally in the civics test.

Table 5.3.1 *Multiple Regression Results for Civics Achievement Scores of the Six Subject Survey in Civic Education*

Model	Unstandardized		Standardized	t	Sig.
	B	SE	Beta		
(Constant)	.057	.018		3.145	.002
Gender	-.074	.026	-.036	-2.823	.005
Word knowledge	.693	.017	.718	41.517	.000
Interaction	-.038	.025	-.026	-1.525	.127

a. Dependent Variable: Zscore (civics IRT scores)

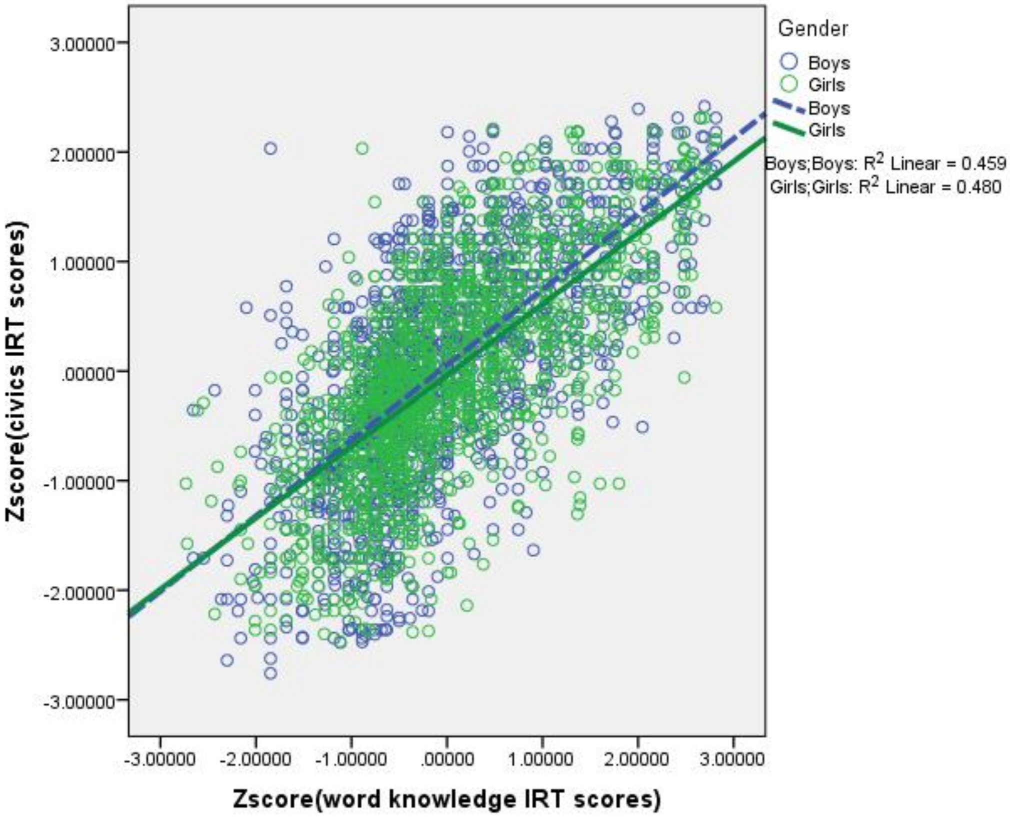


Figure 5.3.1. The scatterplot of civics and word knowledge scores by gender

Home Literacy Resources. Very few ELLs participated in the Six Subject Survey assessment in civic education (*n* = 78). Therefore, I added a categorical demographic variable—number of books at home to explore to what degree it is associated with civics achievement and word knowledge scores. Descriptive statistics of this variable are presented in Table 5.3.2.

I dichotomized the number of books into two levels using 50 as cutoff score, and renamed the variable as "home literacy resources". Students who had 50 or less books at home belonged to the low home resource group (coded as "0", *n* =854), students who had 50 or more books at home belong to the high home resource group (coded as "1", *n* = 2221).

Table 5.3.2 *Frequency Distribution of the Number of Books in Home*

	Frequency	Percent	Valid Percent	Cumulative Percent
A NONE	29	.9	.9	.9
B 1 - 10	114	3.6	3.6	4.5
C 11 - 25	224	7.0	7.0	11.4
D 26 - 50	487	15.2	15.2	26.6
E 51 OR MORE	2221	69.3	69.3	95.9
Missing	132	4.1	4.1	100.0
Total	3207	100.0	100.0	

Next, I used the home literacy resource to predict civics IRT scores and word knowledge IRT scores. Results show that home resources significantly predict the civics IRT scores (β = 268, *p* < .01), indicating that students from high resource home outperformed those from lower resource home in civic achievement. Home resources are also significantly associated with word knowledge IRT scores (β = .241, $p < .01$), suggesting that students who had 50 or more books at home had better word knowledge scores than those with fewer books. Overall, home resources explain 7.2% of variance in civics scores, and 5.8% of variance in word knowledge scores.

Finally, I entered home literacy resources, standardized word knowledge IRT scores, and the interaction of home resources and word knowledge scores into a multiple regression model to predict the standardized civics IRT scores. Overall, the combination of predictors explains 50 percent of variance in IRT scores of civics test. Results suggest that home resources significantly predict the civics scores (β = 102, $p < .01$). This suggests that the group with 50 or more books at home outperforms the group with lower number of books in the civics test after word knowledge is controlled. There is no statistically significant interaction effect found in this step of analysis. This suggests if word knowledge affected test performances of students from families with low literacy resources, it also equally impacted test performances of students from families with high literacy resources. Table 5.3.3 and Figure 5.3.2 illustrate the results.

Table 5.3.3 *Multiple Regression Results for Civics Achievement Scores of the Six Subject Survey in Civic Education*

Model	Unstandardized		Standardized	t	Sig.
	B	SE	Beta		
(Constant)	-.137	.027		-5.070	.000
Home Resources	.230	.031	.102	7.399	.000
Word knowledge	.670	.028	.698	23.649	.000
Interaction	-.028	.032	-.026	-.892	.372

a. Dependent Variable: Zscore (Civics IRT scores)

Interaction = Home resources (low vs. high) * standardized word knowledge IRT scores

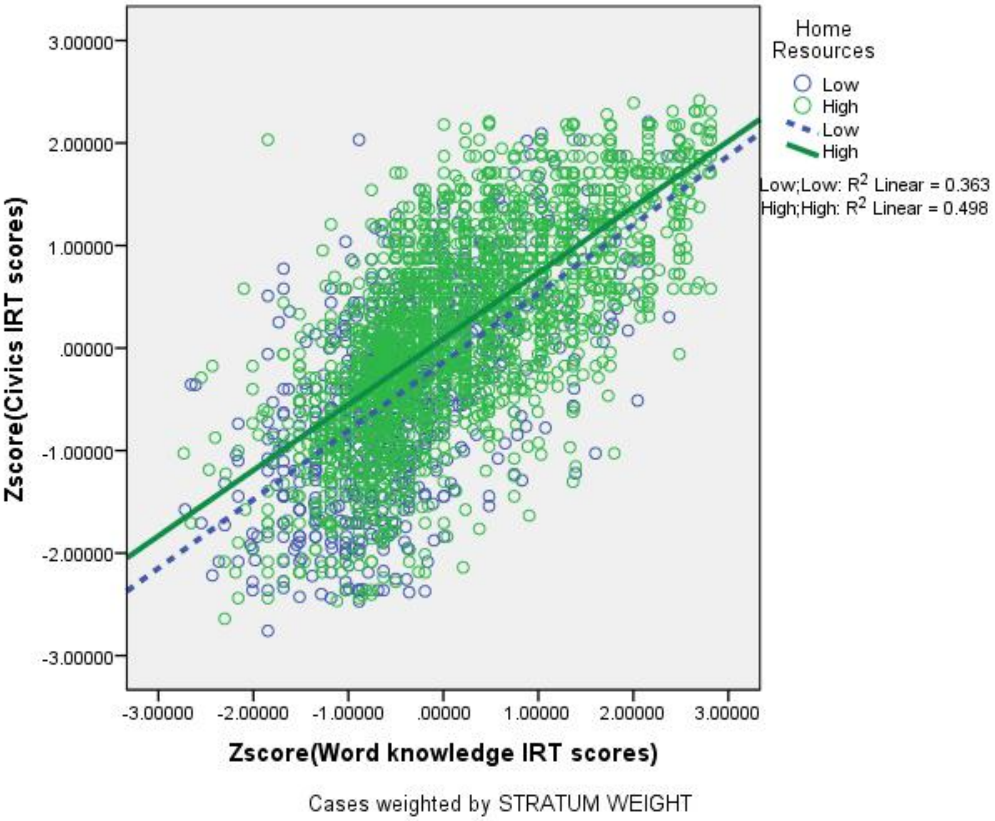


Figure 5.3.2. The scatterplot of civics and word knowledge scores by home literacy resources

In summary, this step of analyses looked at the relation between general word knowledge and civics achievement from a different perspective. Results revealed a similar pattern though. That is, 14-year-old students' word knowledge capability was highly related to their civic knowledge. Students who showed a high level of word knowledge also achieved high scores in the civics test. The relation did not vary by groups, suggesting that word knowledge affected test performances of students from different family background equally. Therefore, further inference is that the reading demand posed on the 70s civic cognitive test was not likely to influence students test performance in a way that would lead to different ordering of students or groups. To compare and contrast the results, next, I conducted analogous analyses on the 1970s Six Subject Survey in Science.

5.3.2 1970s Six Subject Survey in Science

This step of analysis focuses on a total of 3467 U.S. 14-year-old students who took the Form B of science test and the word knowledge test from the Six Subject Survey in Science. The Form B of science test contains 40 multiple-choice items measuring science achievement, and the word knowledge test contains another 40 multiple-choice items measuring students' knowledge of general vocabulary (Comber & Keeves, 1973; Thorndike, 1973). The dataset only provided total raw scores for each test. To obtain more precise measures, I applied the Rasch model to students' item responses through the *mdltm* software to obtain IRT scores for each test. The scales of the Rasch model were set the same as the ones applied to CIVED and TIMSS tests. My focus was the ability scores from the Rasch model. That is, theta values of science achievement (M= -.481, SD = .522, Min = -1.393, Max =1.564), and theta values of word knowledge (M= 1.226, SD = .512, Min = -1.292, Max = 2.810). I standardized both theta values to z scores for subsequent analyses. The stratum weight in the 1970s' Science dataset was applied for all statistical analyses and parameter estimations. Listwise deletion was used in analysis because less than 5 % of missing data were found in each variable of interest.

Bivariate correlation showed that the relation between science IRT scores and word knowledge IRT scores is .578 ($p < .01$), suggesting that students who performed well on science test tended to achieve good scores in vocabulary knowledge and vice versa.

Gender. Next, I examined how gender (boys (0) = 1567, girls (1) = 1760) was related to science achievement and word knowledge by regressing gender on the standardized science achievement IRT scores, and the standardized word knowledge IRT scores respectively. Results showed that gender explained 4% of variance of science achievement scores, and significantly predicted the science achievement scores (β = - .199, p < .01). Overall, boys achieved higher scores from the science test than girls did. On the other hand, there was no statistically significant gender difference in terms of students' word knowledge (β = .018, p =.300).

Third, I used multiple regression to explore whether the relation between science achievement and word knowledge varies by gender. Predictors are gender, the standardized word knowledge IRT scores, and the interaction between gender and word knowledge IRT scores. The criterion variable is the standardized science IRT scores. Table 5.3.4 presents the regression results. Overall, the combination of predictors explains 38.8 % of variance of science scores (F (3, 3323) = 701.991, p < .01). Interaction is statistically significant (β = -.77 p < .01) after taking main effects—the word knowledge and gender—into account. Figure 5.3.3 illustrates the interaction. Results show that the science achievement gap was most salient among students who performed well on both science and general word knowledge tests. Among students who achieved high scores in general vocabulary knowledge, boys outperformed girls in science tests.

In summary, in the 1970s science assessment, I found that students' science performance was highly associated with their knowledge of content-irrelevant vocabulary. Students who performed well on the science test tended to have a broad level of vocabulary knowledge. Because the correlational nature of the large-scale assessment, students' performance in science test, however, could not be attributed to their word knowledge assessed. Furthermore, the relation between students' word knowledge and science performance is more likely to be reciprocal. That is, students who performed well in the science test tended to read more. On the other hand, students who had better vocabularies knew more words in a sentence and were able to determine the meaning of an unknown word from the context in a science test.

One explanation for the multiple regression results is among the high achievers of the science test, boys were more likely to read scientific related books and / or materials than girls. Therefore, boys may obtain a relatively higher level of science knowledge than girls at the same level of reading proficiency through reading scientific-related materials. To better understand the relation between science achievement and word knowledge, it would be helpful to look at it from a different perspective. Because the 1970s science assessment did not measure students' language status, I looked at the home resource variable instead.

Table 5.3.4 *Multiple Regression Results for Science Achievement Scores of the Six Subject Survey in Science*

Model	Unstandardized		Standardized	t	Sig.
	B	SE	Beta		
(Constant)	.231	.020		11.798	.000
Gender	-.416	.027	-.210	-15.445	.000
Word knowledge	.639	.019	.642	33.750	.000
Interaction	-.110	.027	-.077	-4.060	.000

a. Dependent Variable: Zscore (Science IRT scores)

Note. Interaction = Gender * Standardized word knowledge IRT scores

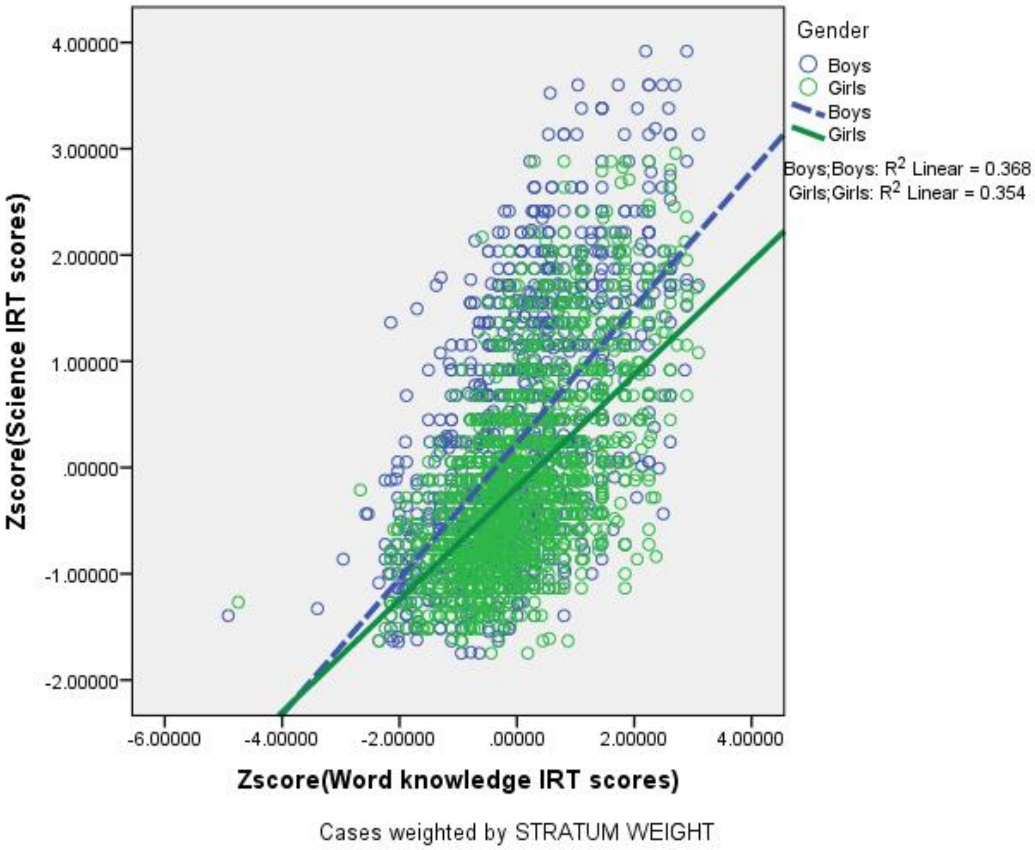


Figure 5.3.3. The scatterplot of science and word knowledge scores by gender

Home Literacy Resources. The number of books in the home is a categorical variable in the 1970s science data file. Table 5.3.5 presents the descriptive statistics of this variable. I dichotomized the variable using 50 as the cutoff score, and created a new variable of home literacy resource in which students from home with books less than 50 were coded as 0 (low home literacy resources, *n* = 992), and students with books of 51 or higher were coded as 1 (high home literacy resources, *n* = 2316).

Table 5.3.5 *Frequency Distribution of the Number of Books in Home*

	Frequency	Percent	Valid Percent	Cumulative Percent
NONE	36	1.1	1.1	1.1
1 - 10	97	2.9	2.9	4.0
11 - 25	271	7.9	8.2	12.2
26 - 50	588	17.3	17.8	30.0
51 OR MORE	2316	67.9	70.0	100.0
Total	3308	97.1	100.0	
Missing	100	2.9		
Total	3409	100.0		

First, the dichotomized home literacy resources variable was regressed on the standardized science IRT scores and standardized word knowledge scores. Results showed that home resources explained about 4% of variance of science scores, and 4% of the variance of word knowledge scores. In addition, home resource significantly predicted science scores (β = .206, p < .01) and word knowledge scores (β = .210, p < .01). This means students from families that had 50 books or higher achieved higher science scores as well as word knowledge scores.

Next, I used the home literacy resources, standardized word knowledge scores, and the interaction between the two variables as predictors to predict the standardized science scores. Results are presented in Table 5.3.6. Overall, the combination of predictors accounted for 34.8% of variance of science scores. When word knowledge and home resources were controlled, the interaction effect was statistically significant (β= .088, p < .01), indicating the relation between science achievement and word knowledge varied by home resources. Figure 5.3.4 demonstrates the interaction effect, and shows that the gap between students from low and high home resource families is most evident among high-performing students.

To summarize, in the science assessment, I found similar patterns to those I discovered from the 70s civics test. That is, students' word knowledge was highly related to their performance in the domain specific test. Students who mastered a broad range of vocabulary, also tended to possess a high level of domain knowledge. In addition, the significant interaction between home literacy resources and general word knowledge suggests that other factors associated with the high home resources might contribute to the difference. I discuss the possible factors and explanations in the next chapter.

Table 5.3.6 *Multiple Regression Results for Science Achievement Scores of Six Subject Survey in Science*

Model	Unstandardized		Standardized	t	Sig.
	B	SE	Beta		
(Constant)	-.145	.027		-5.382	.000
Home resources	.210	.032	.097	6.637	.000
Word knowledge	.484	.028	.486	17.357	.000
Ineraction	.104	.032	.088	3.209	.001

a. Dependent Variable: Zscore(Science IRT scores)

Note. Interaction = Home resources (low vs. high) * Standardized word knowledge IRT scores

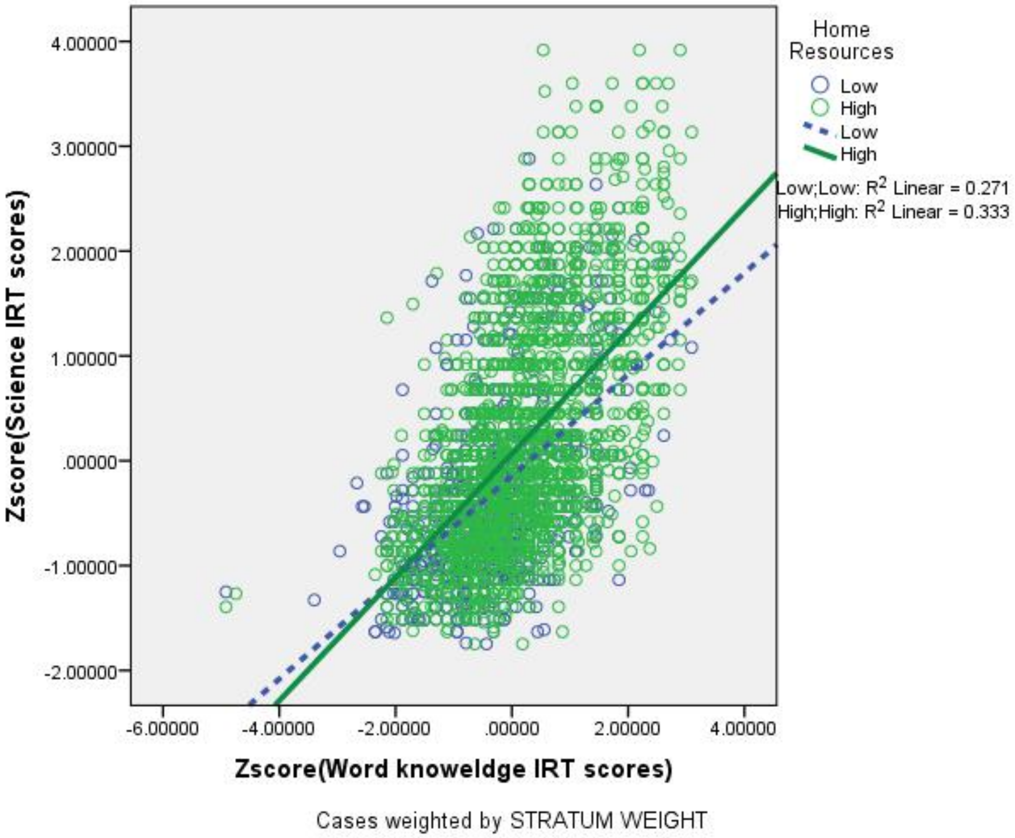


Figure 5.3.4. The scatterplot of science and word knowledge scores by home literacy resources

Chapter 6: Discussion

Large-scale educational assessments such as IEA TIMSS and IEA CIVED have been used to document what students know and can do in subject-matter domains. Outcomes from these large-scale assessments provide a base for policy makers, curriculum specialists, and researchers to better understand the quality of our educational systems. Therefore, ensuring accurate and valid information about student achievement in content areas is critical. The present study investigated a potential source of construct-irrelevant variance—reading comprehension—that might affect the scores that students obtain (e.g. Abedi, 2002; Haladyna & Downing, 2004; Messick 1989). The current study extended previous research (conducted on reading and literacy tests, and sometimes mathematics) by examining the role of reading comprehension on science and civics assessments through a cognition-centered approach based on the Evidence-Centered Design (ECD) framework.

This chapter begins with a synopsis of the ECD Framework and how it guides the current research design and analysis. Next I summarize specific findings and offer potential explanations based on the ECD framework. The chapter concludes with a discussion of limitations and suggestions for future research.

6.1 Evidence-Centered Design Framework

National Research Council (2001) advocates an interdisciplinary approach to assessment design and validation. At the heart of this approach is making use of advances in cognitive theories and cutting edge statistical models to acquire the best evidence from structured theoretical frameworks with the intention of enhancing our understanding of students' achievement and the process of their learning. Evidence-Centered Design (Mislevy, Steinberg, & Almond, 2003) provides such a theoretical framework that integrates cognitive theories and measurement models into assessment design and validation. The framework facilitates researchers and test designers in reasoning from the best available evidence with respect to what students know and can do. The present study utilized the theoretical framework of ECD to guide the research design, analyses, and interpretations.

The ECD framework conceptualizes assessment development and validation as closely paralleling the process of hypothesis testing in social science. That is, researchers first hypothesize a model describing students' knowledge or proficiencies, provide operational definitions, and then build measures to collect data, and finally evaluate the evidence for and against their hypotheses (synthesized by Gorin, 2007, in a review). Psychometricians including Kane (1992, 2006) and Mislevy (2009) refer this process as constructing the assessment argument. They suggest that one way to strengthen the assessment argument is to include the testing of alternative explanations for a person's high or low test scores. Significant and credible alternative explanations might indicate that test validity is threatened. Ruling out the plausibility of alternative explanations can help ensure that assessments do measure what they were intended measure.

Grounded in this theoretical framework, the present study explored the role of reading comprehension in association with the domain specific achievement that a subject-matter assessment was intended to measure. Particularly, this study examined whether reading difficulty can be an alternative explanation for 14-year-old U. S. students' high scores or low scores in standardized science tests and civics tests.

The current study started with defining reading comprehension based on the conceptual frameworks developed by Kintsch (1998). Factors (at the personal level, text level, and item level) that contribute to reading difficulty were explicitly discussed.

In the second step, the present study identified task features including individual-item characteristics and text features in each large-scale assessment based on Kintsch's reading comprehension theory (1998) and previous empirical research on reading comprehension. In particular, text features were identified through an advanced technical tool--Coh-Metrix, which was developed in alignment with Kintsch's theory.

In the third step, the reading demands of the subject-matter assessments were predicted using multiple regression, and relations between reading demands and students' test performance were investigated.

Finally, the presence of a high level of reading demand along with the subject-matter achievement was modeled through a multidimensional IRT model. This approach allows for an estimate of domain achievement while taking reading comprehension components and items with high level of reading demand into account. The hypothesis is that estimates of domain achievement are more accurate because the noise associated with high reading demand is partialled out. Results were compared across demographic groups to test whether the high level of reading demand biased test performances of students, especially those who belonged to the group with language deficiency (those individuals who do not speak English at home).

I present the summaries and interpretations in the next section.

6.2 Summary and Interpretation of Findings

I summarize findings based on my research questions. Within each section, I address how the research question pertaining to each subject-matter assessment was answered, describe how results are consistent or conflict with previous research, consider practical and theoretical explanations for the findings, and discuss theoretical and practical implications. Recognizing that this study was not experimental, I cannot reach causal conclusions or identify explicit causes. However, I can speculate about specific mechanism that could explain the relations that I have found. Some of these mechanisms could be examined in subsequent research.

6.2.1 Reading-Related Task Features Contribute to Difficulty in Answering Subject-Matter Items

Advances in cognitive theories such as Kintsch's construction-integration (CI) theory have provided a feasible framework to understand the nature and characteristics of comprehension processes when students read assessment tasks. According to the CI theory, reading comprehension is constructed and built on integrated mental models derived from and activated by the text. During the comprehension process, the reader activates concepts expressed in the text and forms connections between activated concepts and relevant prior knowledge of words, concepts, ideas and personal experience. The networks of concepts that are compatible with the context enhance the activation of one another, while concepts that are not compatible with the context lose activation. In summary, the comprehension processes are regulated by mental models and constrained by contexts (Kintsch & van Dijk, 1978; Stahl & Hiebert, 2006). Kintsch (1998) suggests that a successful comprehension of the text depends on a variety of factors including reader factors, text factors, and context factors such as the tasks and reading activities that a reader engages in.

In Research Question 1, I examined the degree to which text factors and item characteristics contribute to the difficult level of multiple-choice items in subject-matter assessments within the theoretical framework of reading comprehension theory. The subject-matter assessments include the IEA CIVED test of 1999 and IEA TIMSS Science test of 1999.

Even though these two domain-specific assessments have different emphases in term of what they were intended to assess, the language of items across tests has some common characteristics. First, all items were designed to present a high level of demand on subject-matter knowledge or skills that the assessment was intended to measure. Second, they were written in the form of short texts. A typical item in these two assessments starts with a short sentence followed by a question or an incomplete statement which calls for the answer to complete it. Four alternatives are presented after the stem, and each alternative usually encompasses one complete or incomplete sentence. Third, items in these two assessments were not necessarily designed as highly cohesive texts, since the purpose was eliciting students' domain knowledge or skills (in contrast to instructional texts). Fourth, the language of test items is expository.

In addition, two facts about the studies from 1999 should be noted. First, the items were written to be translated into 20 plus languages, and that is one reason the reading difficulty was kept at a relatively simple level. Second, the preliminary set of items was examined for DIF, and thus a few items difficult for certain groups or in certain languages may have been left out.

IEA CIVED Test of 1999. My results from the IEA CIVED data suggested that linguistic features that measure different levels of reading comprehension significantly predicted item difficulty of the CIVED test items. More than 1/3 of variance in the difficulty level of test items was accounted for by linguistic features originally identified through a software package called Coh-Metrix.

At the *surface level* of reading comprehension, vocabulary predicted item difficulty. First, the inclusion of lengthy words was associated with difficult items. This is consistent with reading theories (e.g., Kintsch, 1998, Perfetti, 2010, RAND, 2002) which suggest that lengthy words usually take more working memory space when the reader processes text information. Therefore lengthy words increase reading difficulty.

At the *textbase level*, syntactical indices predicted item difficulty. First, the number of noun phrases in a stem was associated with item difficulty. That is, when an item stem encompassed one or more noun phrases, the item appeared to be easier than other items which contain no noun phrases. Example of the noun phrases embedded in the CIVED item stems include “which of the following”, and “evidence of government corruption”. One possible explanation is that noun phrases may aid readers in chunking the information into fewer units so that they can use less effort to process the information in their working memory. Therefore, when the text lengths are relatively short and similar (about two or three sentences per item), the more noun phrases embedded in a stem, the easier the item is to process.

Second, negative expressions in the distractors were related to easy items. For instance, “The United Nations has its own flag even though it is not a country.” Another example is “People with very low incomes should not pay any tax”. It seems students were less likely to choose a distractor if it contains negative expression. One explanation is that negative expression increases the reading difficulty of a sentence. In fact, Abedi and his colleagues (e.g., Abedi & Lord, 2001) found that the passive voice of verb phrases contributes to reading difficulty of standardized reading items (NAEP reading), and math word problem items (NAEP math). Abedi (2009) points out that it is difficult for students, especially for ELL students, to understand test items that are complex in their linguistic structure. To conclude, my study results imply that students were less likely to spend time and make effort to process a distractor if it involves complex linguistic structure. Therefore, students were less likely to select this distractor as the correct answer. This increased their likelihood of choosing correct answers.

Finally, I found that the difficulty of the 38 CIVED items was associated with a Coh-Metrix feature pertaining to the higher level of reading comprehension: the *situation model*. This has to do with word meanings and students' knowledge of vocabulary. First, the inclusion of concrete words is associated with easy items. My results based on Coh-Metrix indicated that the mean concreteness value of all content words in a stem predicted the item difficulty. The more concrete content words (including nouns, adverbs, adjectives, main verbs) in an item stem, the easier the item to read. Reading comprehension theories (e.g., Perfetti, 1985; 2011; RAND Reading Study Group, 2002, Snow, 2010) suggest that vocabulary knowledge is the major component of reading ability and plays an important role in reading comprehension. This has been confirmed by empirical studies (e.g., Holmes, 2009; Simmons, et. al., 2010) which found strong correlations among reading comprehension (specific word recognition) and knowledge of word meaning in adolescents and adults. More specifically, Schwanenflugel and Akin

(1994) revealed that words that are meaningful to a reader are identified faster and more accurately than words that are abstract. In general, findings from the present study converge with previous reading theories and empirical studies.

Second, intentional verbs embedded in item distractors were related to item difficulty of the civics test. Intentional verbs were identified based on WordNet (Fellbaum, 1998), a lexical database that comprises a large number of semantic characteristics of words. Coh-Metrix 2.0 classifies a verb as "intentional" if it belongs to particular WordNet categories. The higher the occurrence of intentional actions in a text, the more the text is assumed to convey goal-driven content. My analysis revealed that items tended to be more difficult if any of their distractors had high intentional verb values. Below are few examples of distractors that have high values of intentional verbs.

- To ask for public debates about a political issue.
- Increase citizens' interest in government.
- He makes statements supporting other leaders in his party.

One possible explanation is that intentional verbs were likely to activate students' prior knowledge. However, when distractors contained these verbs, the activated prior knowledge did not match with the purpose of these items. Therefore, the difficulty level of item increased.

One limitation of using the Coh-Metrix 2.0 is that the online software only provides users numerical values about the frequency of the intentional verbs. Users are not informed as to which verbs are identified by Coh-Metrix 2.0 as high in intentional actions, and which are low. To find out more, future studies should obtain access to the WordNet database. Empirical research should be conducted to manipulate the main verbs in the test items like those used here based on the information from WordNet to examine whether certain verbs are more likely to contribute to item difficulty.

Overall, my results suggested that when students took this standardized test in civic education, difficulty levels of items were related to linguistic features pertaining to vocabulary and syntactical structures. Some features such as short words / frequent words, words with concrete meanings, and certain noun phrases were likely to facilitate students' comprehension of civics multiple-choice items. Contrariwise, lengthy words, words that carried abstract meanings and complex syntactic structure appear to increase item difficulty by hindering students' comprehension of civics items.

In addition, intentional verbs can facilitate comprehension because they may assist the progress of activating students' knowledge schemata. However, the inclusion of intentional verbs in distractors may lead to a reverse effect. That is, they may trigger students' prior knowledge that does not match with what the item calls for.

In terms of syntactical structure, negative expressions are likely to hinder comprehension because they increase the syntactic complexity of a sentence or sentences. The inclusion of negative expression in distractors, however, can make a test item easier.

Constructing items with noun phrases that students are familiar with is likely to facilitate comprehension. Noun phrases can aid students in chunking the item information into fewer units.

I was also interested in how language use in science test contributes to difficulty levels of science item. Next, I describe and discuss findings from the TIMSS science test of 1999.

IEA TIMSS Science Test of 1999. I applied similar analysis procedures to 104 TIMSS science multiple-choice items. Overall, I found some similar results from the TIMSS science test. That is, only linguistic features identified by Coh-Metrix 2.0 predicted item difficulty values. Item characteristics including graphic features and the abstractness of item questions judged by raters were not related to item difficulty of the science test. Compared with CIVED items, variance of science item difficulty explained by linguistic features dropped to 20 percent, suggesting that linguistic features played a less important role in the TIMSS science test.

In terms of linguistic features identified through Coh-Metrix 2.0, results from TIMSS science 104 multiple-choice items showed that item difficulty values were associated with linguistic features pertaining to all three levels of reading comprehension processes.

At the *surface level*, vocabulary factors including word length and word frequency contributed to difficulty levels of science items. First, I found that lengthy words embedded in stems were associated with difficult items. Additionally, infrequent words in distractors were related to difficult items. This is in fact contrary to previous findings from standardized reading comprehension assessments (e.g., Embretson & Wetzel, 1987; Gorin & Embretson, 2006; Sheehan & Ginther, 2001). For example, Sheehan and Ginther examined test items in the TOEFL reading test. They found that rather than frequent words, infrequent words in the distractors made the item easier. Their explanation was similar to what Embretson and Wetzel (1987) provided. That is, test takers were less likely to expend the time and effort to process the distractors in a reading comprehension test if the distractor contained rare words. The divergent finding from the present study suggested that students may have used different cognitive strategies when they read science test items of TIMSS. They appear to have been likely to spend equivalent time and effort in weighing every item alternative before they made their choice unless the syntactical structure of an alternative overburdened them.

At the *textbase level* and *situation level*, difficulty levels of science items were associated with logical connectives in stems, and negative logical connectives in distractors. First, logical connectives in item stems were related to easy items. Students were more likely to answer science items correctly if descriptions in item stem were connected by a logical operator such as *and*, *or*, *not*, *if*, *then*. Below are examples of TIMSS science items that have high values in logical connectives.

- The Moon produces no light, and yet it shines at night. Why is this?
- Immediately before and after running a 50 meter race, your pulse and breathing rates are taken. What changes would you expect to find?

One possible explanation is that adding logical operators in item texts facilitated students constructing a coherent textbase level model, and therefore enhanced their comprehension. This finding is consistent with previous research conducted by Embretson and Wetzel (1986) on GRE reading comprehension items. They modeled the difficulty of items as a function of text features, and their findings revealed that logical connectives facilitated comprehension.

Second, I found that negative logical connectives in distractors were related to item easiness. Below are two examples of distractors with negative logical connectives.

- No change in pulse but a decrease in breathing rate.
- From either his father or his mother, but not from both.

One explanation is that these negative connectives were likely to increase cognitive load when students processed the option. Students, especially those who were not highly motivated to take the test, were likely to ignore such alternatives of an item. Instead, they focused on other alternatives (including the correct responses) that were easier to process. As a result, the item was relatively easy.

When examining GRE reading assessment, Embretson and Wetzel (1986) found that test takers were less likely to choose a distractor as a correct answer if it encompassed difficult vocabulary because many test takers were not willing to extend their time and effort to process the alternative. As a result, item with distractors that contain structures that are difficult to process may be easier than item with distractors simple to process. My finding is analogous to their discovery.

In my analyses of TIMSS science items, I did not find statistically significant relations between item difficulty and the inclusion of graphic features. This suggests that the graphs, diagrams, and tables did not substantially facilitate students' comprehension process, nor hinder it in TIMSS Science assessment. These features were likely to be necessary parts of students' cognitive processes that were within most students' range of cognitive capacities.

Summary and Implications. Traditionally, difficulty of text passages has been gauged through the word length and sentence length. Recent studies have shown that other factors pertaining to different levels of reading comprehension also relate to reading difficulty (e.g. Kintsch & Kintsch, 2005; McNamara & Kintsch, 1996). My results from the CIVED and TIMSS science tests were consistent with contemporary research, indicating that traditional readability methods may not be the best (or only) way to detect problems pertaining to reading difficulty in assessments such as the CIVED and TIMSS science tests. Linguistic features related to higher levels of reading comprehension processes should also be taken into account when designing and validating items in a subject-matter assessment. Particularly, test designer and researchers should be cautious about the following linguistic features when constructing standardized science or civics multiple-choice items. These features are likely to increase unnecessary reading demand of subject-matter test items:

- Word length—average syllables per words.
- Word frequency—familiarity or frequency of content words. Content words are nouns, adverbs, adjectives, main verbs, and other categories with rich conceptual content.
- Word abstractness—abstractness of content words.
- Negation expressions, e.g., “People with very low income should not pay any tax.”
- Negative logical connectives, e.g., “from either his father or his mother, but not from both”.

On the other hand, other linguistic features are likely to facilitate students’ comprehension. For example, constructing an item with noun phrase that students are familiar with can boost their working memory process. Adding logical connectives in item stem may aid students in constructing a coherent textbase model. Using appropriate intentional verbs in item stems and correct responses may help to activate students’ domain knowledge or skills that the item is intended to measure.

Finally, because all TIMSS science and CIVED items were identified as written in expository texts, the current study could not compare the expository texts with narrative texts in terms of their influence of genre on item difficulty. However, previous research implies that test takers may benefit from using expository language to construct science and civics tests. For example, Wolfe and Woodwyk (2010) conducted an experimental study on 61 undergraduate students, and examined the impact of text genre on students' memory. In their study, participants were asked to read to-be-learned content that was embedded in narrative or expository texts. A sentence recognition task was then used to assess their memory. Their results showed that students tended to make more associations to prior knowledge when reading expository texts. The implication is that expository texts are more likely to prompt students to use relevant prior knowledge than narrative texts.

Another experimental study conducted more recently (Adams, Mayer, MacNamara, Koenig, & Weiness, 2012) confirmed this finding. In this study, researchers examined the impact of a computer-based narrative discovery learning game on college students' learning outcomes. Students who learned by playing the narrative game performed worse in a posttest than those who learned from a matched slideshow presentation that was more expository. Their results suggested that narrative tasks are often not effective in facilitating learning.

6.2.2 Reading Demand and Subject-Matter Test Performance

My item-level results discussed in the previous section contribute important evidence to understanding the role of reading comprehension in subject-matter tests including CIVED and TIMSS science by providing a list of linguistic features and estimates of their impact on item difficulty. Next, I made use of advances in statistical models to further explore the role of reading comprehension in CIVED and TIMSS Science tests. Particularly, my analyses focused on answering following questions dealing with the student level of analysis:

(1) To what extent will an advanced statistical model—a multidimensional IRT model--partial out the noise associated with high reading demand?

(2) To what degree do the average standard error of estimated scores of the domain-specific proficiency change after taking into account the reading demand of test items?

(3) Does the relation between the reading and students' domain proficiency vary by gender and language status in each subject-matter assessment?

First, it is possible to partial out the noise associated with reading demand using multidimensional IRT model.

This rationale behind the modeling is that all test items of subject-matter required some reading proficiency. In standardized subject-matter assessments such as TIMSS science and CIVED, students have to be able to read in order to understand what a standardized subject-matter test item asks. Therefore, a threshold amount of reading is a necessary part of the construct of the subject-matter assessment. Once the level of reading demand of an item exceeds a reasonable range of students' linguistic ability, however, students (e.g., ELLs, and students with reading difficulties) who have deficiencies in English language competency may not be able to demonstrate their domain-specific proficiencies because they misunderstood or could not process what the test item asks for. In this case, the item should be modified or rewritten to reduce the amount of reading demand. Nevertheless, this process can take a long time, and may have to go through trials or experimental studies. In the present study, I proposed an alternative way to accomplish this by utilizing advances in model-based approach (i.e. multidimensional IRT model) to separate out the noise associated with high level of reading demand.

Based on this rationale, I used a 2-dimensional IRT model to model a domain achievement variable and linguistic variable. In this MIRT model, items with high level of reading demand were assumed to call for students' domain-specific proficiencies (the construct) as well as high level of linguistic capability which was not what the test was intended to measure. On the other hand, items with relatively low reading demand were assumed to call for predominantly domain-specific proficiencies. The combination of domain-specific proficiencies reflects the domain-specific achievement that a standardized subject-matter assessment often assesses. Generally speaking, the model specification had to do with the belief about the role of reading comprehension in subject-matter assessments. That is, we have to decide: (1) is reading is a part of construct that the subject-matter test is intended to measure? (2) If it is, what is the appropriate degree of reading demand in a standardized subject-matter test?

I applied a 2-dimensional IRT model to both CIVED 38 items and TIMSS Science 104 items to test the assumption about the role of reading comprehension in a subject-matter test. To compare the results, I also applied a 1-dimensional Rasch model to the datasets. The Rasch model (i.e. 1 PL IRT model) is one of the most common measurement models that standardized subject-matter assessments such as TIMSS, CIVED, PISA and NAEP employ to estimate students' domain specific achievement. When using the traditional 1-dimensional Rasch model to estimate students' domain achievement, estimated scores reflect an overall achievement competency that conflates domain proficiencies, reading proficiency and probably other attributes. The assumption underlying common practice is that it is the domain achievement proficiency that accounts for the majority of variance in students' scores.

After model comparison for each test, my results showed that the average standard error of estimates from the MIRT model was smaller than that of the Rasch model in terms of domain score estimates. Even though the difference between standard error of estimates was not substantial, evidence still indicated that MIRT model separated out a small amount of variance that was associated with high level reading demand, and produced a slightly "purer" estimate of domain achievement.

Third, for both TIMSS and CIVED assessments, I found the same achievement scores yield from Rasch model and MIRT model were highly and positively correlated. This suggests what the MIRT model partialled out is not substantial in terms of the amount of reading demand that was independent of domain-knowledge demand. Potential explanations for this result are that (a) the linguistic demands of all items, including the linguistically complex ones, may be within the reading capabilities of almost all of the students, and (b) domain-knowledge and reading abilities are highly correlated, so that students who have increasing difficulty with linguistic aspects of items are also likely to have commensurately difficulty with respect to domain knowledge.

Summary and Implications. Previously, my study identified a list of linguistic features that contributed to difficulty levels of items in CIVED test, and TIMSS science test. Next, making use of the information, I obtained a gauge of item reading demand as a function of these salient linguistic features at the item level in each test.

Next, I used a multidimensional IRT model to model the high level of reading demand while estimating students' domain proficiency scores. The purpose was to separate out noise variance associated with the high level of reading demand that some items had, so as to attain a "purer" estimate of the domain achievement that the subject-matter test was intended to assess. My results indicated that the MIRT model partialled out a very small amount of variance that was associated with high reading demand and independent of the domain achievement. By separating out the noise variance, the standard error of mean decreased to a small extent.

Overall, the current study contributes to understanding the role of reading comprehension in subject-matter tests including CIVED and TIMSS science by providing feasible methods (1) to estimate reading demand of test items, and (2) to partial out variance related to high level of reading demand that is independent of the domain proficiencies that the subject-matter assessment was intended to measure.

These two methods were based on the assumption that items in the standardized subject-matter assessment tapped two attributes. One is an excessive-linguistic proficiency attribute which was not what the test had been intended to assess (i.e., construct-irrelevant), and the other is students' achievement in the subject domain. The domain specific achievement attribute conflates multiple proficiencies some of which are construct-relevant, including students' domain knowledge, procedural skills, problems-solving strategies, and basic reading capacity (and perhaps other factors such as test taking strategies). Future studies should apply more advanced statistical models to the TIMSS science and CIVED datasets so as to obtain finer-grained information. For example, a 3-dimensional IRT model could be applied to the datasets to model three attributes: construct-irrelevant linguistic capacity, domain knowledge, and procedural skills. Information from this model could help us learn more about how reading comprehension interacts with domain knowledge and skills. Some of the experimental studies suggested by the item-level analysis could also be useful here.

6.2.3 Reading Demand and Demographic Groups

Results from the previous section indicate that high level of reading demand had a negative association overall with students' test performance in the CIVED and TIMSS Science tests. My next question was whether the relation between reading demand and domain achievement varies by gender and by students' language background.

Test validity theories (e.g., Messick, 1989; Mislevy, 2009) and the common test standards (AERA, APA, NCME, 1999) suggest that reading involved in a subject-matter assessment can be construct-irrelevant variance that biases the test score interpretations, especially for demographic groups who have disadvantages in language, such as English Language Learners (ELL) and students with reading difficulties. This bias may increase achievement gaps between demographic groups. My analysis at this step focused on examining whether the reading demand involved in the CIVED test and TIMSS test biased students' test performance and increased discrepancies between boys and girls, ELLs and non-ELLs.

The assumption was if high level of reading demand biased students' test score by causing more error variance in the estimation of achievement scores, one should expect that the discrepancy between mean scores of boys and girls, as well as ELLs and non-ELLs would decrease after the noise variance of reading demand was separated out.

For both CIVED and TIMSS assessments, the analysis contradicted the assumption, however. My results showed that the discrepancy of mean scores between boys and girls unexpectedly increased when the scores came from the MIRT model. In CIVED assessment, the mean scores discrepancy between ELLs and non-ELLs also unpredictably broadened. All effects were small, but statistically significant, and in a direction that I would not expect. One possible explanation is that students who had difficulty with the high linguistic demands had low domain knowledge that was revealed when linguistic difficulty was separated out. In other words, if the high level reading demand was removed (for example by rewriting the items to be simpler linguistically), one would have seen the low reading capability students performing better, but not as much better as the groups who had higher reading capability as well as higher domain proficiency that was being masked by linguistic requirements.

There are two ways to find out whether this hypothesis is likely to be correct. The first method is to rewrite and modify these linguistic complex items in CIVED and TIMSS to be simpler linguistically, and to conduct an experimental study using the original items and linguistically simple items. The hypothesis would be supported if the linguistic simple items benefit all students especially students who had high scores from original items and belong to groups that have high proficiency in language. The second method is to look at the statistical correlation between an independent measure of reading comprehension and subject-matter proficiency. Highly positive correlation between the two attributes would support this hypothesis.

International large-scale subject-matter tests such as the CIVED and TIMSS were designed to assess students' domain specific proficiencies. Little attention has been given to other influences such as students' reading comprehension. Due to lack of direct measures of reading comprehension or reading-related cognitive proficiencies, I used an indirect way to approximate the amount of reading demand posed on each test item. In order to obtain more information with respect to the relation between reading and subject-matter achievement, I added two additional large-scale assessments into my study: the Six Subject Survey in Science and in Civic Education conducted by IEA in the 1970s. In addition to measures of domain achievement, each subject-matter assessment contains an independent measure of students' knowledge of vocabulary (content-irrelevant vocabulary) with 40 items. I discuss the results from these two additional subject-matter assessments in the section following the summary.

Summary and Implications. Mislevy (2009) suggests that validation of standardized assessment can be summarized as an argument that encompasses: (a) a claim about a person possessing at a given level a certain targeted proficiency, (b) the data (e.g., test scores) that would likely result if the person possessed a certain level of the targeted proficiency, (c) the warrant (or rationale, based on theory and experience) that explains why the person's level in the targeted proficiency would lead to occurrence of the data, and (d) "alternative explanations" for a person's high or low test scores (explanations that can potentially be tested).

Especially, he emphasizes that significant and credible alternative explanations might indicate that test validity is threatened. By ruling out potential alternative explanations, we can be more confident that the test assessed what was intended to assessment. Reading demand in the subject-matter assessment, especially high level of reading demand in some items, may impede some students' comprehension of test items, and become an alternative explanation for these students' low test scores. The current study revealed that even though reading demand showed an impact on students' performances in the CIVED and TIMSS science tests, the impact was almost equivalent across all test takers. Therefore, for both tests, students' low scores cannot be attributed to excessive reading demand on some test items. This evidence strengthened the argument for the validity of these two subject-matter assessments and suggested that these test items measured what was intended to measure. Researchers who are interested in the achievement outcomes from these two subject-matter assessments should feel more confident about using the test scores yielded from these test items for their research.

6.2.4 Vocabulary Knowledge and Subject-Matter Achievement

The main purpose of the current study is to understand the role of reading comprehension in standardized science and civics assessments. In the previous sections, I explored this topic utilizing the CIVED assessment of 1999 and TIMSS Science test of 1999. My results provided insights into the functional and meaningful relation between reading demands at the item level and students' test performance in these tests. Due to the lack of independent measures of students' reading comprehension competency, I was not able to directly examine reading comprehension or reading-related cognitive proficiencies at the student level, or how students' reading comprehension and reading-related proficiencies were related to their performance in the standardized science test and civics test administered in 1999. To fill in the gap, I analyzed two more large-scale subject-matter assessments: the IEA Six Subject surveys in Science (administered in 1969), and in Civic Education (administrated in 1971). Both of these assessments included a separate test that had an independent measure of students' knowledge of general vocabulary. I examined the relation between students' knowledge of general words (synonyms/antonyms) and achievement scores in science and civics. Furthermore, my study investigated whether the relations between the students' general word knowledge and achievement scores vary by demographic factors in each subject-matter assessment.

Overall, my results revealed that students' knowledge in general vocabulary was highly related to their achievement scores in the science test and civics test. This suggested that students who had wide vocabularies also achieved high scores in science and civics tests, and students who had poor word knowledge did not received high domain achievement scores. This evidence supports my previous hypothesis with respect to students' reading-related proficiencies and their domain achievement, and explains why the MIRT model did not partial out a great amount of variance that is associated with excessive reading demand and independent of domain achievement in the TIMSS and CIVED tests.

Based on this information, however, it is hard to conclude whether students' word knowledge actually facilitated or hindered their performance in the science and civic tests. Based on research on reading and students' general academic achievement (e.g., Alexander & The Disciplined Reading and Learning Research, 2012; Pearson, Moje & Greenleaf, 2010; Snow, 2010), it is likely that there are reciprocal association among students' word knowledge, domain knowledge and skills, as well as other domain specific proficiencies. Each is in service of the others and all contribute to students' academic competence developed through years of learning and schooling. For example, students who knew a broad range of vocabulary were likely to have more exposure to a variety of literacy resources and engage in reading. By reading and learning through academic-related literacy, they could obtain more domain knowledge which would help them to understand academic content better and learn more. Consequently, they would end up being knowledgeable in both general vocabulary and content areas of science and civics. This hypothesis is examined in the next section (to the extent possible given the existing datasets).

Home Literacy Resources. In the absence of information about the home language background of students in the studies conducted in forty years ago, my study also explored the group differences with respect to word knowledge and domain achievement by more general home literacy background. In the present study, home literacy resources were measured by asking students the numbers of books at their homes. My results were consistent with previous research such as Leseman and de Jong (1998), Senechal and LeFevre (2002), and Lugo-Gil and Tamis-LeMonda (2008). That is, on average students from homes that had many literacy resources (i.e. more than 50 books) outperformed those from homes with lower literacy resources in general word knowledge, civics and science tests. This implies that students' exposure to literacy at home and their word knowledge are important factors associated with their achievement in science and civics.

Moreover, in the six subject survey in science, the present research found a significant interaction effect between home literacy and word knowledge (as illustrated in Figure 5.3.4). Results suggested that on average students from families with many literacy resources achieved higher scores than those from families with fewer literacy recourses. The gap was most evident among high performers on the word knowledge test.

That is, among test takers who showed high level of word knowledge, students from high literacy homes outperformed those from low literacy home in science. This suggests that factors other than reading experience and word knowledge contribute to students' achievement in science. They can be factors associated students' informal learning experience such as visiting science museums, or summer camps.

Gender. In terms of gender, the current study did not find gender differences in word knowledge. My results showed that 14-year-old boys performed as well as the girls in the word knowledge tests in the two assessments: six subject surveys in science and in civics. The outcome was not consistent with findings from other more recent studies using large-scale reading assessments (e.g., Lynn and Mikk, 2009) which indicated that girls in general had reading advantage over boys in standardized reading achievement assessments; this was not the case for the 14-year-olds in 1970s in terms of word knowledge.

Multiple regression results showed gender differences in science and civics in the 1970s. Overall boy's outperformed girls in both civics and science subject areas. In the science test, the present research also found a statistically significant interaction effect (as illustrated in Figure 5.3.3), which indicates that the gender difference was most substantial among students who achieved high scores in the science and general word knowledge tests. Among this group of students who showed high level of word knowledge, boys showed higher science competence than girls did. One possible explanation is that this group of students, who mastered a broad range of vocabulary and had better knowledge in science, also had richer experience including reading. However, boys had different preferences in reading materials compared with girls. These boys might read materials related to science and political or civic matters, whereas girls probably read more narrative materials such as popular fictions. This could explain some of the gender differences in terms of the correlation between the general word measure and domain achievement in science. Research on gender difference in reading achievement supports this hypothesis (e.g., Logan & Johnston, 2010), and suggests that gender difference in reading achievement can be attributed to many factors including motivation. A number of studies indicate that boys enjoy a wider range of genres and topic including news, science fiction, and special-interest books, whereas girls generally like narrative texts such as modern or classic fiction, romance stories, or song lyrics (e.g., Baker & Wigfield, 1999; Canadian Council on Learning, 2009; Logan & Johnston, 2009; Young & Brozo 2001).

Summary and Implications. In summary, in the six subject surveys in science and civics, the current study revealed that students' word knowledge was highly related to their achievement of subject-matter competence. This finding was aligned with my preceding results from the CIVED assessment of 1999 and TIMSS science assessment of 1999, which showed that vocabulary at the item level, especially the infrequent and abstractness of words, contributed to item difficulty of science and civics tests. The combination of evidence from all four subject-matter assessments indicates that in designing of large-scale science and social studies tests in which test items are usually composed of short texts, vocabulary is one of the most important factors to consider.

In addition, I found that boys outperformed girls in both science and civics, students from home with ample literacy resources outperformed students from homes low literacy resources. The group differences in subject-matter achievement were likely to contribute to students' reading experiences and preferences.

One appropriate line of future research is to employ latent class analysis methods (e.g., cognitive diagnostic model) to classify students into different cognitive profiles based on their mastery status of word knowledge and domain knowledge. This method may provide more information about characteristics of these students. Students' informal learning experience should be investigated too.

6.3 Limitations of the Study and Future Directions

Researchers have extensively examined the association between reading comprehension and students' subject-matter achievement including their domain knowledge and problem-solving skills. However, few have explored the sources of reading comprehension difficulties and their influence on students' performance in large-scale subject matter assessment. Furthermore, few have attempted to partial out the influence of reading demand in large-scale subject-matter assessments in science and civics. The current study extends previous research by examining the role of reading comprehension in large-scale subject-matter assessments using advances in cognitive and measurement theories, cutting-edge technological tools and statistical models. Additionally, the nationally-representative sample enables findings to be generalized to the national population of 14-year-olds. However, there are some limitations that are important to note.

First, the current study put emphasis on reading difficulty and the influence of reading comprehension on students' test performance in large-scale science and civics tests. I conceptualized students' domain achievement (what a large-scale subject-matter assessment is usually intended to assess) as a compound construct that conflates manifold proficiencies including domain knowledge, procedural skills, problem-solving strategies, basic reading capacity, motivation, and interest. Based on this assumption, I modeled the domain achievement in each test as a unidimensional latent variable in IRT models. The conceptual assumption and relevant statistical procedures were mostly aligned with some common practices of large-scale subject-matter assessments such as TIMSS and CIVED. However, it did not fully reflect current cognitive and educational psychologists' view about domain knowledge. For example, Alexander and her colleague point out that domain knowledge is a complex and multidimensional construct that entails many types of knowledge, such as knowledge about content, content-specific vocabulary, knowledge in language syntactic, the domain, the world, and cultures (e.g., Alexander, Kulikowich, & Schulze, 2004).

Future studies should implement finer-grained inspections of domain knowledge based on current cognitive and learning theories (e.g., Murphy, Alexander, & Muis, 2011; Webb, 2006). In addition, advanced research and measurement methods should be employed to disentangle compound domain achievement variables, and treat them as multidimensional, multilevel, and/or dynamic constructs. For example, Shavelson and his team at Stanford University developed a conceptual framework to understand adolescent's science achievement in large-scale assessments including TIMSS Science test (see details in Shavelson & Ruiz-Primo, 1999; Li *et al*., 2011). This framework addresses the connections among instruction, student learning, educational measurement, standards, and science curriculum, and conceptualizes science achievement as four types of knowledge. Future study should draw on frameworks such as this to explore issues pertaining to reading comprehension and science achievement in a more explicit way.

Second, a significant proportion of the current study was devoted to investigating linguistic features at the item level, and the degree to which these features were associated with item difficulty of subject-matter items. Originally, I proposed to use the advanced tool – Coh-Metrix 2.0 to identify various text features pertaining to different levels of Kintsch's reading comprehension model. However, I learned through experience that in order to obtain most of text features that Coh-Metrix 2.0 manual promises to provide, the to-be-analyzed text has to be more than 200 words in length. This was not the case for the items of my study. Eventually, most of the variables (text features) that I obtained from Coh-Metrix 2.0 and used for statistical analysis were counting variables such as average syllables per word, the frequency of logical operators in a text, and the number of noun phrases per text. I could not obtain most of text cohesive features and features pertaining to the situation model of reading comprehension due to the length of each written item. Future research should apply this approach to examine subject-matter assessments that involve more reading material (e.g., PISA scientific literacy tests). By obtaining more text features from the Coh-Metrix or relevant text analysis software, researchers can learn more about functions of various linguistic features in relation to difficulty levels of subject-matter test items.

Third, this study adapted coding schemes from previous research such as Ozuru, et al., (2008) and Mosenthal (1996), and had two reading experts to identify item characteristics pertaining to reading difficulty. Raters reached good agreement when they evaluated TIMSS Science items. However, the inter-rater reliabilities of ratings for IEA

CIVED test items were not very high. The strength of the agreement between raters on CIVED may be attenuated by several factors. First of all, the CIVED test items on average involve more reading than the TIMSS Science items, and a main part of the reading difficulty comes from vocabularies used in the CIVED items. However, the line between the content-related and content-irrelevant words was not distinct for the raters involved in the present study. They were likely to perceive bias related to vocabularies differently. For example, during the coding process, one rater indicated that items which contain infrequent words such as “democracy” can be difficult for some 8th graders to comprehend. Conversely, another rater insisted that “democracy” and some other infrequent words in the CIVED test are content-related, and therefore should not be considered as a bias toward construct validity.

Furthermore, agreement may be underestimated because the TIMSS and CIVED test items have been extensively screened for bias at test development stages, and many obvious biases have already been eliminated.

Another factor may lower the reliability is the variance of ratings provided by the raters. This is especially true for rating *item 5* which asks raters to identify the level of abstractness of each item question. TIMSS Science items and CIVED items were designed to elicit specific content-related knowledge or skills. It is not surprising that few test item questions were identified by the raters as level 1—“identification of persons, animals, or things”, or level 5—“identification of equivalence, difference, or theme”. The lack of variability in the ratings may attenuate the inter-rater reliability of *item 5*.

Finally, after their coding, these two raters reflected that the language of a few rating scales (e.g., *item 5*—the scale for abstractness of questions) were abstract so that they had to interpret the rating item first before they used it to rate TIMSS science or CIVED test items. This may contribute to some inconsistence between raters' ratings especially on the abstractness of questions. Nevertheless, researchers who intend to adapt these existing schemes to their study should work with the language used in the rating scheme, and provide more training and examples to raters. By doing so, the researchers may be able to find significant associations between item characteristics and difficulty level of items in a standardized subject-matter test. However, it is not the case that holistic ratings such as this should always be preferred (Engelhard, Hansche, & Rutledge, 1990).

Fourth, the current study found that vocabulary plays a predominant role among item-level features (including text features and item characteristics) in TIMSS Science and CIVED test, in which items were written in short texts. Students' knowledge of general vocabulary predicted their science and civic achievement in Six Subject Surveys in Science and Civics. When evaluating construct-irrelevant variance that poses threats to validity of a subject-matter test, it is critical to differentiate content-irrelevant words from content-based vocabulary. Content-related vocabulary can be an essential part of the construct that a subject-matter test is intended to measure, whereas content-irrelevant words that hinder some students' comprehension of test items should be subject to modification. The present research used Coh-Metrix 2.0 to identify vocabulary features such as word length, word frequency, word abstractness, and intentional verbs. However, this version of Coh-Metrix does not have the function distinguish between content-related and content-irrelevant vocabularies. Item evaluations from raters did not provide much information either. Future studies could have human raters familiar with the content evaluate subject-matter test items based on existing criteria or standards (e.g., the Common Core Sate Standards) and determine what vocabulary words in items are content-related. Some State Departments of Education (e.g., the Oklahoma State Department of Education) provide lists of academic vocabularies on their websites. Each list is core to a subject-matter domain and corresponds to one grade level (e.g., Algebra, Biology, or Economics).

Fifth, the present study provided a list of linguistic features and estimated their associations with students' performance in science and civics tests. However, due to lack of experimental methods, I could not rule out other confounding factors that had potential influence students' test performances. Future studies should make use of the information from the present study to modify or simplify standardized test items in science and civic education. Experimental studies should be conducted to further examine the effect of these linguistic features on test performances of students. Special attention should be given to those have deficiencies in language (e.g., ELLs and students with learning difficulties). Detailed experimental design ideas were described in the results section.

Sixth, using existing large-scale assessment data such as TIMSS Science and CIVED of 1999, I attempted to separate out noise associated with high level of reading demand on some test items through multidimensional IRT model. However, the MIRT model only partialled out a very small amount of variance that was associated with high reading demand and independent of the domain achievement. One possibility is the reading demands of the test items were within the reading capabilities of almost all of the students who took the TIMSS science or CIVED test.

To find out more, further studies could replicate this study and apply the current methods to subject-matter assessments that have higher degree of reading demands and are more up to date, such as, PISA Scientific Literacy Assessments (2006, 2009, 2012), and International Civic and Citizenship Education Study of 2009. Furthermore, experimental approaches could be used and they may be more effective in teasing out noise and error variance associated with excessive reading demands on some test items. Additional complexity would be involved to look at reading demand of these tests in other languages but could be explored in the future.

Seventh, like many studies using cross-sectional large-scale assessment data, the current study provided snapshots with respect to the role of reading comprehension in standardized science and civics tests for 14-years-old students. Previous research on reading comprehension suggests that linguistic features pertaining to lower level language skills (e.g., word recognition, fluency, and oral language abilities) are likely to affect students' reading comprehension in the early elementary years. As students get to higher grades, these features become less associated with their reading comprehension. Linguistic features that call for higher level language skills (e.g., as semantic skill, and the use of comprehension strategies) are more important determinants of reading comprehension by 5th or 6th grade (Duke & Carlisle, 2011). Another appropriate line of research should be conducted with students at lower grades. Additionally, longitudinal studies should be conducted to explore how the relation between reading comprehension and domain knowledge develops over the school years. Ways in which Kintsch's theory can help in this process will be discussed later in this section.

Eighth, large-scale data for six subject surveys in science and civic education have potential to be used for secondary analysis. The current study only focused on independent measures of students' word knowledge and domain achievement when using these two additional datasets. Coh-Metrix approach was not applied to these two assessments. Therefore, linguistic demands of test items were not taken into account. Another potential analysis for future researchers is making use of advantages of the technique from the earlier analysis of CIVED and TIMSS Science items and investigating the degree to which the linguistic demands on test items are associated with students' word knowledge and domain achievement.

Finally, Kintsch (1998) suggests that reading comprehension involves complex cognitive processes that integrate information from the text with the readers' background knowledge and experiences and is subject to contextual constraints. The current study mainly focused on context factors including item-level linguistic features and students' demographic information such as home literacy resources that were available in the large-scale subject-matter assessments that I employed. Further research can complement the current study by looking at other contextual features in which students build and use their knowledge and competence, such as testing environment (e.g., paper-pencil vs. computer-based), curriculum and instruction, students' classroom experience, school environment, and the cultural contexts.

6.4 Conclusion

Published research on international large-scale assessments such as TIMSS and CIVED examining the influence of reading comprehension on performance is generally lacking. Utilizing data from the IEA international large-scale assessments, the study contributes to the literature by employing modern cognitive models and psychometric methods to enhance the current state of knowledge regarding the role of reading comprehension in large-scale subject-matter assessments. This study identified which types of comprehension-related item features contribute most to the difficulty of items. This information can afford researchers, educators, and test designers greater insight into types of cognitive proficiencies and processing that are tapped by assessment items in subject-matter areas. The current research also provides feasible theoretical and methodological frameworks for test makers and psychometricians who want to reduce excessive reading demands in a domain specific assessment without compromising the theoretical validity of the assessment.

Finally, there are three overall points to be made based on this study.

First, almost all test items of subject-matter required some degree of reading proficiency in addition to domain proficiencies. The big issue is whether those items that demand a great deal were hard for students because of their reading difficulty level rather than their subject-matter demands. The ideal situation is not that items demand no reading proficiency, but rather that the level of construct-irrelevant reading proficiency they demand is within the capabilities of the testing population.

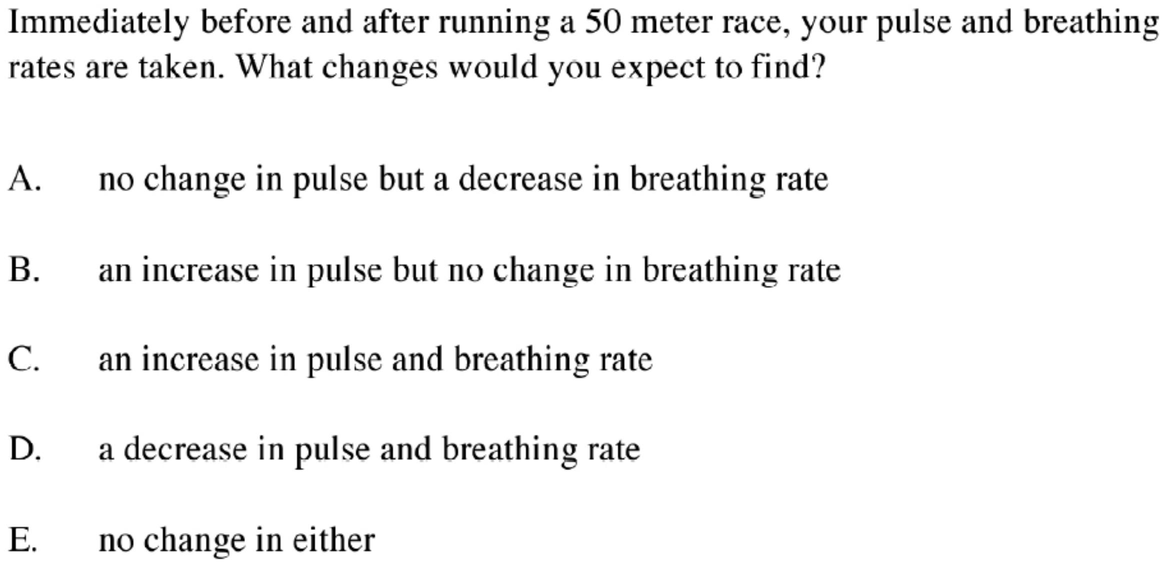
Second, there are differences between subject matters in the particular features that appear to be important in the comprehension of items and texts. This corroborates the work of scholars such as Stodolsky (1988) and Torney-Purta and Amadeo (2012). This suggests that one cannot generalize about reading and cognitive processes without examining specific characteristics of a subject matter domain.

Further, the study advances the understanding of the contributions of Evidence-Centered Design (Mislevy, 2006; 2008) and suggests a variety of ways to combine data from large-scales assessments with more targeted experimental studies.

Appendices

Appendix A

Sample Items from the TIMSS Science Test of 1999



Immediately before and after running a 50 meter race, your pulse and breathing rates are taken. What changes would you expect to find?

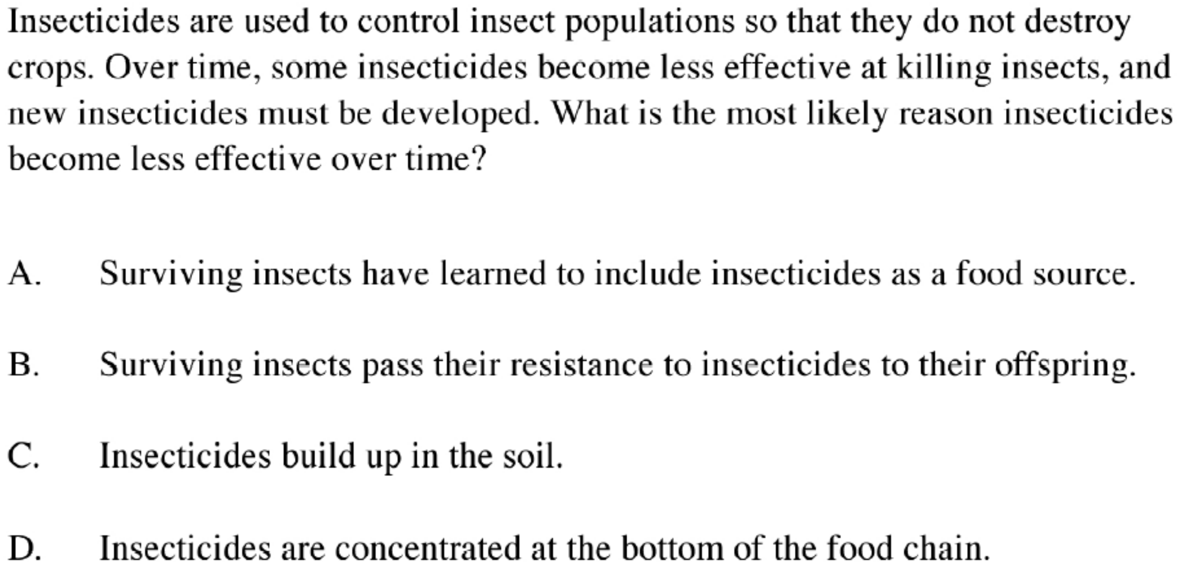
A. no change in pulse but a decrease in breathing rate

B. an increase in pulse but no change in breathing rate

C. an increase in pulse and breathing rate

D. a decrease in pulse and breathing rate

E. no change in either



Insecticides are used to control insect populations so that they do not destroy crops. Over time, some insecticides become less effective at killing insects, and new insecticides must be developed. What is the most likely reason insecticides become less effective over time?

A. Surviving insects have learned to include insecticides as a food source.

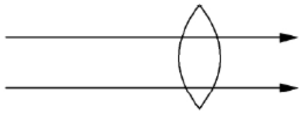
B. Surviving insects pass their resistance to insecticides to their offspring.

C. Insecticides build up in the soil.

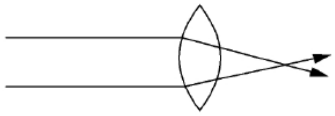
D. Insecticides are concentrated at the bottom of the food chain.

Which diagram best shows what happens when light passes through a magnifying glass?

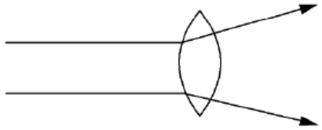
A



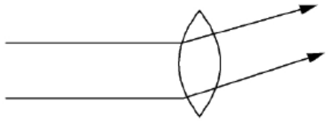
B



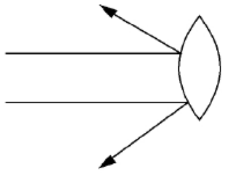
C



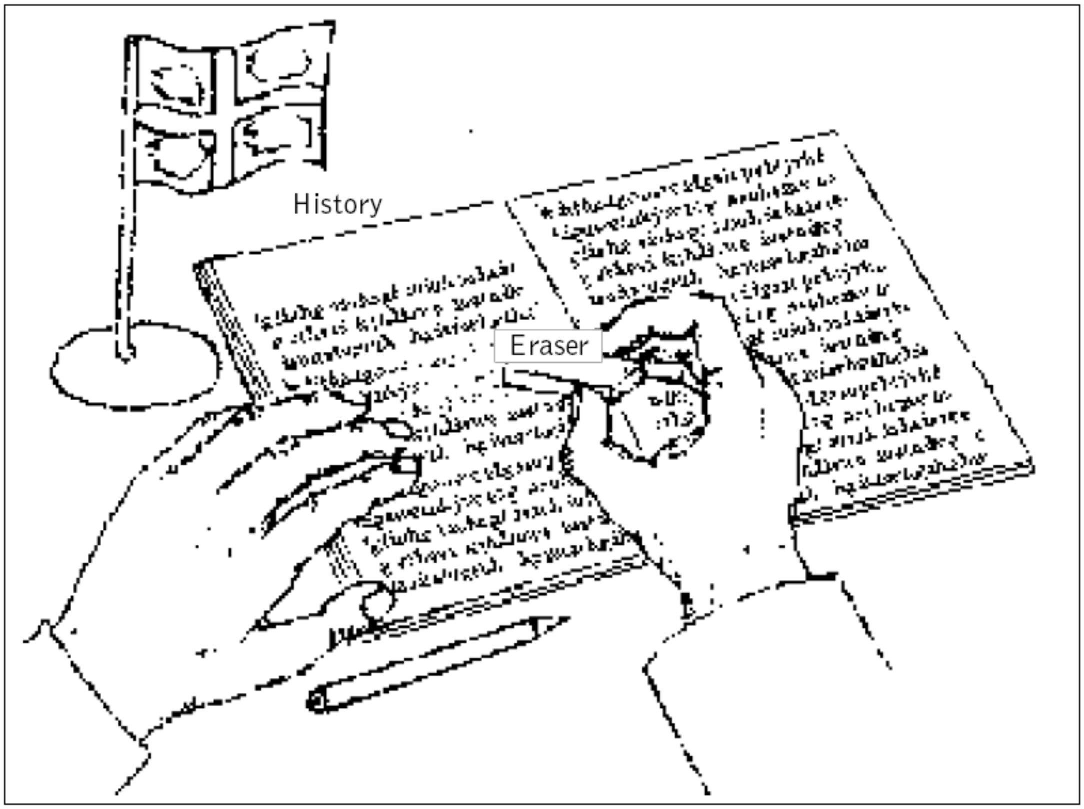
D



E



Sample Items from the CIVED Test of 1999



36. What is the message or main point of this cartoon? History textbooks ...

   A. ☒ are sometimes changed to avoid mentioning problematic events from the past.

   B. □ for children must be shorter than books written for adults.

   C. □ are full of information that is not interesting.

   D. □ should be written using a computer and not a pencil.

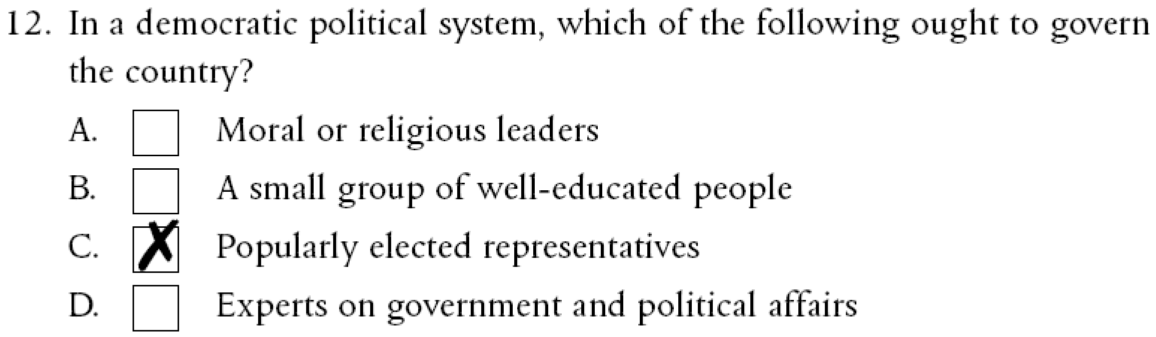
17. Which of the following is most likely to cause a government to be called non-democratic?

   A. ☒ People are prevented from criticising [not allowed to criticise] the government.

   B. □ The political parties criticise each other often.

   C. □ People must pay very high taxes.

   D. □ Every citizen has the right to a job.



12. In a democratic political system, which of the following ought to govern the country?

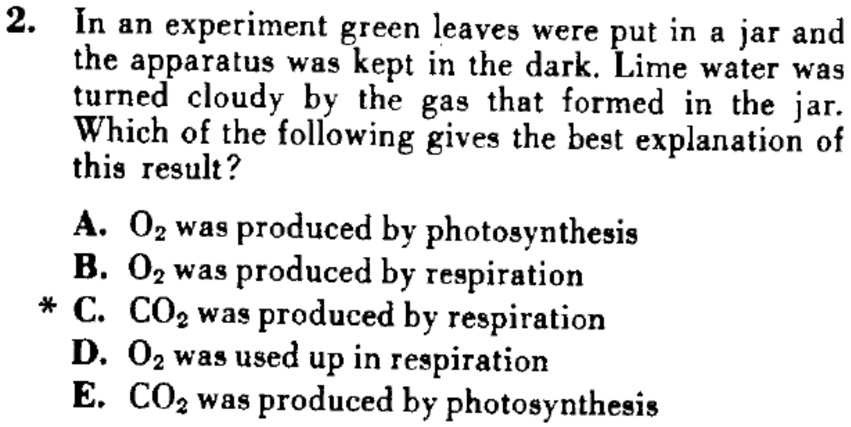
A. ☐ Moral or religious leaders

B. ☐ A small group of well-educated people

C. ☒ Popularly elected representatives

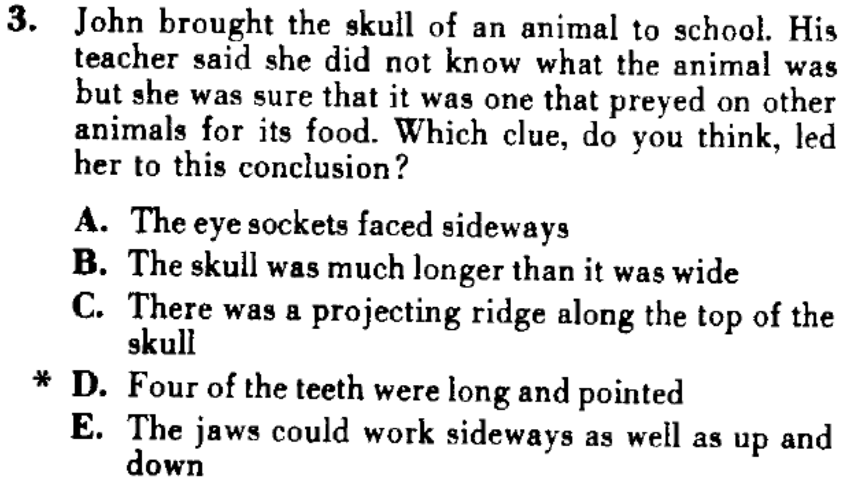
D. ☐ Experts on government and political affairs

Sample Items from the Six Subject Survey in Science



2. In an experiment green leaves were put in a jar and the apparatus was kept in the dark. Lime water was turned cloudy by the gas that formed in the jar. Which of the following gives the best explanation of this result?

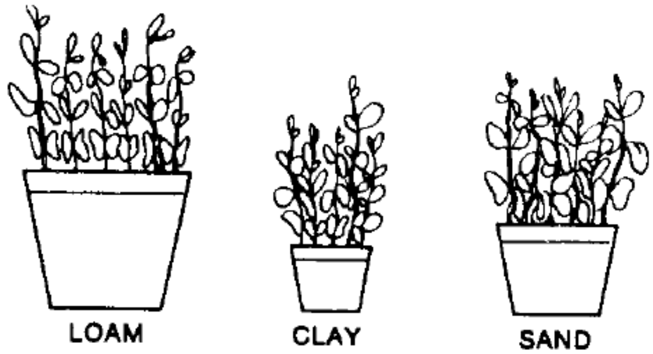
A. O_2 was produced by photosynthesis
B. O_2 was produced by respiration
* **C.** CO_2 was produced by respiration
D. O_2 was used up in respiration
E. CO_2 was produced by photosynthesis



3. John brought the skull of an animal to school. His teacher said she did not know what the animal was but she was sure that it was one that preyed on other animals for its food. Which clue, do you think, led her to this conclusion?

A. The eye sockets faced sideways
B. The skull was much longer than it was wide
C. There was a projecting ridge along the top of the skull
* **D.** Four of the teeth were long and pointed
E. The jaws could work sideways as well as up and down

4. Tom wanted to learn which of three types of soil—clay, sand, or loam—would be best for growing beans. He found three flowerpots, put a different type of soil in each pot, and planted the same number of beans in each, as shown in the drawing. He placed them side by side on the window sill and gave each pot the same amount of water.



Why was Tom's experiment *NOT* a good one for his purpose?

A. The plants in one pot got more sunlight than the plants in the other pots

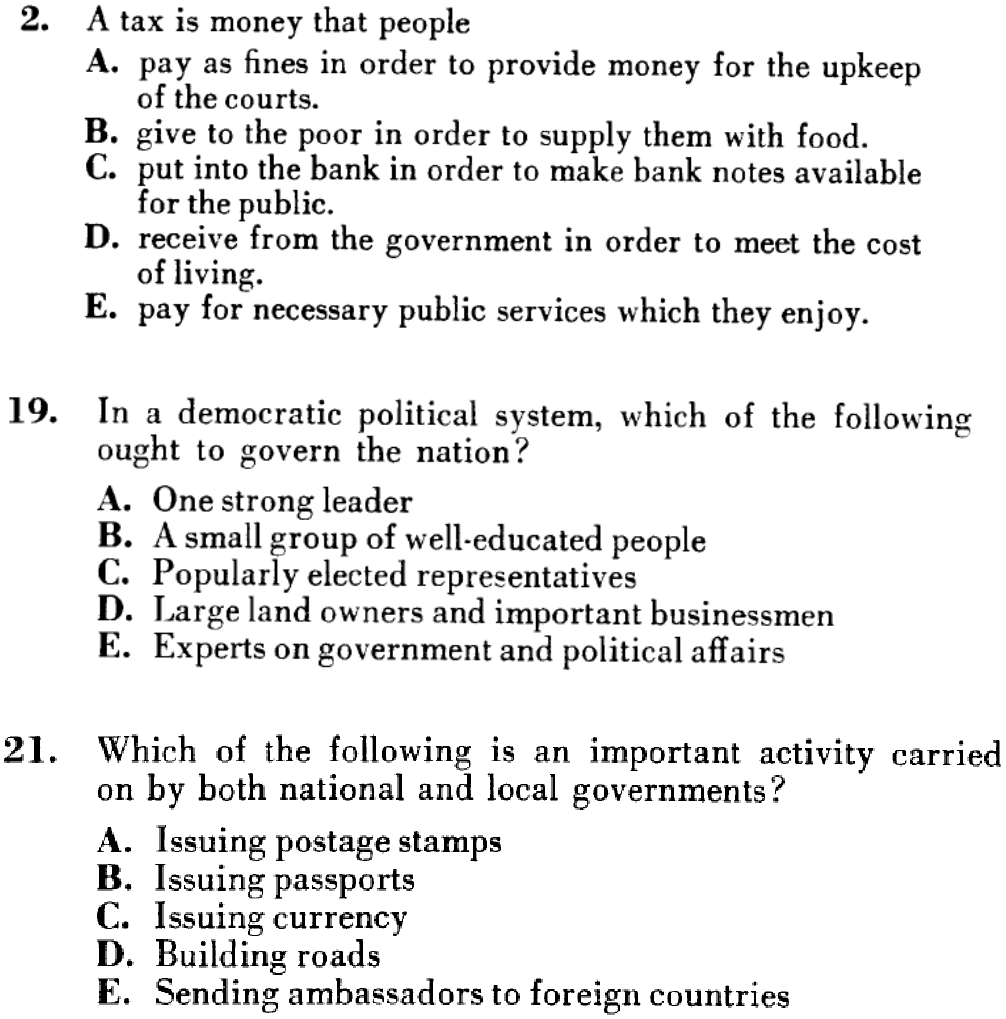
* **B.** The amount of soil in each pot was not the same

C. One pot should have been placed in the dark

D. Tom should have used different amounts of water

E. The plants would get too hot on the window sill

Sample Items from the Six Subject Survey in Civic Education



2. A tax is money that people

A. pay as fines in order to provide money for the upkeep of the courts.
B. give to the poor in order to supply them with food.
C. put into the bank in order to make bank notes available for the public.
D. receive from the government in order to meet the cost of living.
E. pay for necessary public services which they enjoy.

19. In a democratic political system, which of the following ought to govern the nation?

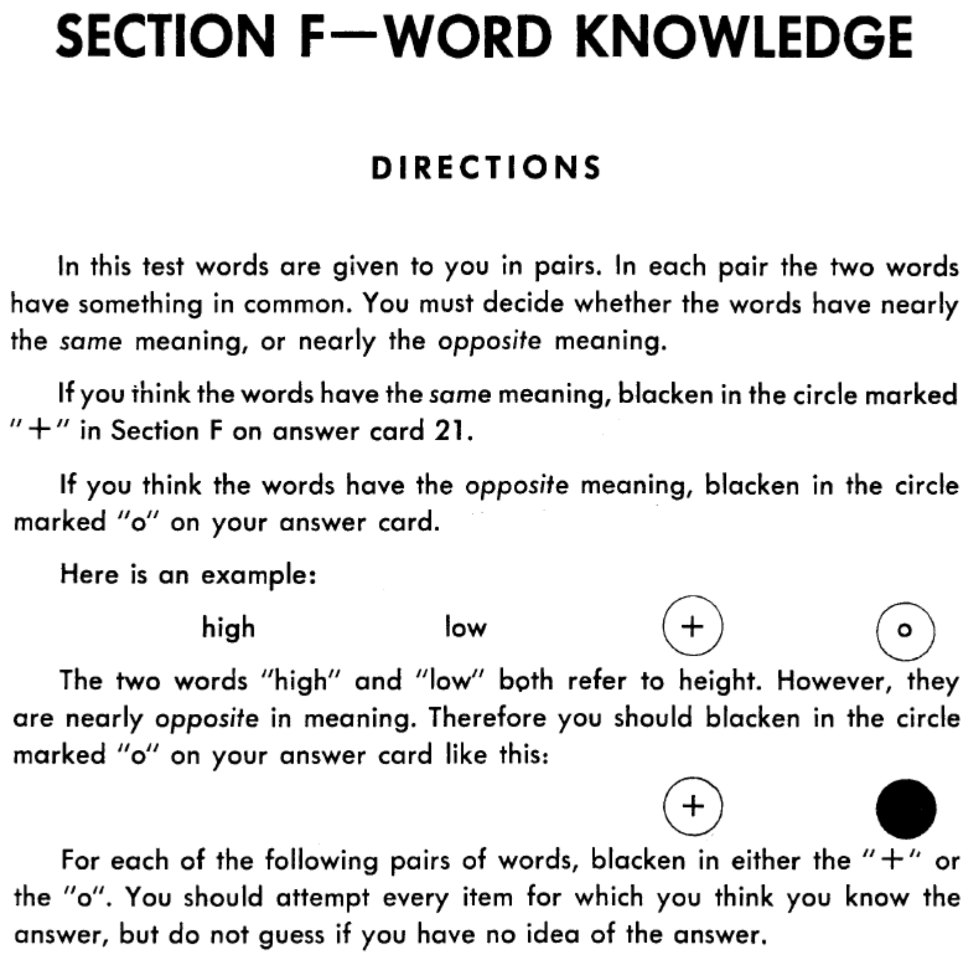
A. One strong leader
B. A small group of well-educated people
C. Popularly elected representatives
D. Large land owners and important businessmen
E. Experts on government and political affairs

21. Which of the following is an important activity carried on by both national and local governments?

A. Issuing postage stamps
B. Issuing passports
C. Issuing currency
D. Building roads
E. Sending ambassadors to foreign countries

Appendix B

Six Subject Surveys in Science and Civic Education—Word Knowledge Test



SECTION F—WORD KNOWLEDGE

DIRECTIONS

In this test words are given to you in pairs. In each pair the two words have something in common. You must decide whether the words have nearly the *same* meaning, or nearly the *opposite* meaning.

If you think the words have the *same* meaning, blacken in the circle marked "+" in Section F on answer card 21.

If you think the words have the *opposite* meaning, blacken in the circle marked "o" on your answer card.

Here is an example:

high low ⊕ ⓞ

The two words "high" and "low" both refer to height. However, they are nearly *opposite* in meaning. Therefore you should blacken in the circle marked "o" on your answer card like this:

⊕ ●

For each of the following pairs of words, blacken in either the "+" or the "o". You should attempt every item for which you think you know the answer, but do not guess if you have no idea of the answer.

SECTION F—WORD KNOWLEDGE TEST

1.	savory	insipid
2.	informed	unaware
3.	precarious	stable
4.	rapid	sluggish
5.	supple	malleable
6.	associate	partner
7.	decoration	ornamentation
8.	mute	voluble
9.	prosperity	opulence
10.	ordered	confused
11.	prohibited	forbidden
12.	boastfulness	modesty
13.	wealthy	impoverished
14.	adjacent	contiguous
15.	create	originate
16.	garrulous	taciturn
17.	expatiate	harangue
18.	rare	habitual
19.	benevolent	intolerant
20.	vague	precise
21.	wise	judicious
22.	acquire	dispel
23.	ancient	antique
24.	abstruse	explicit
25.	loosen	relax
26.	despise	scorn
27.	flagrant	obvious
28.	gauge	measure
29.	paltry	exorbitant
30.	absolute	relative
31.	everlasting	permanent
32.	conformity	dissimilarity
33.	converge	approach
34.	consecrate	dedicate
35.	deny	repeal
36.	variable	inconstant
37.	bounty	generosity
38.	delicate	tactful
39.	repudiate	disavow
40.	obvious	indisputable

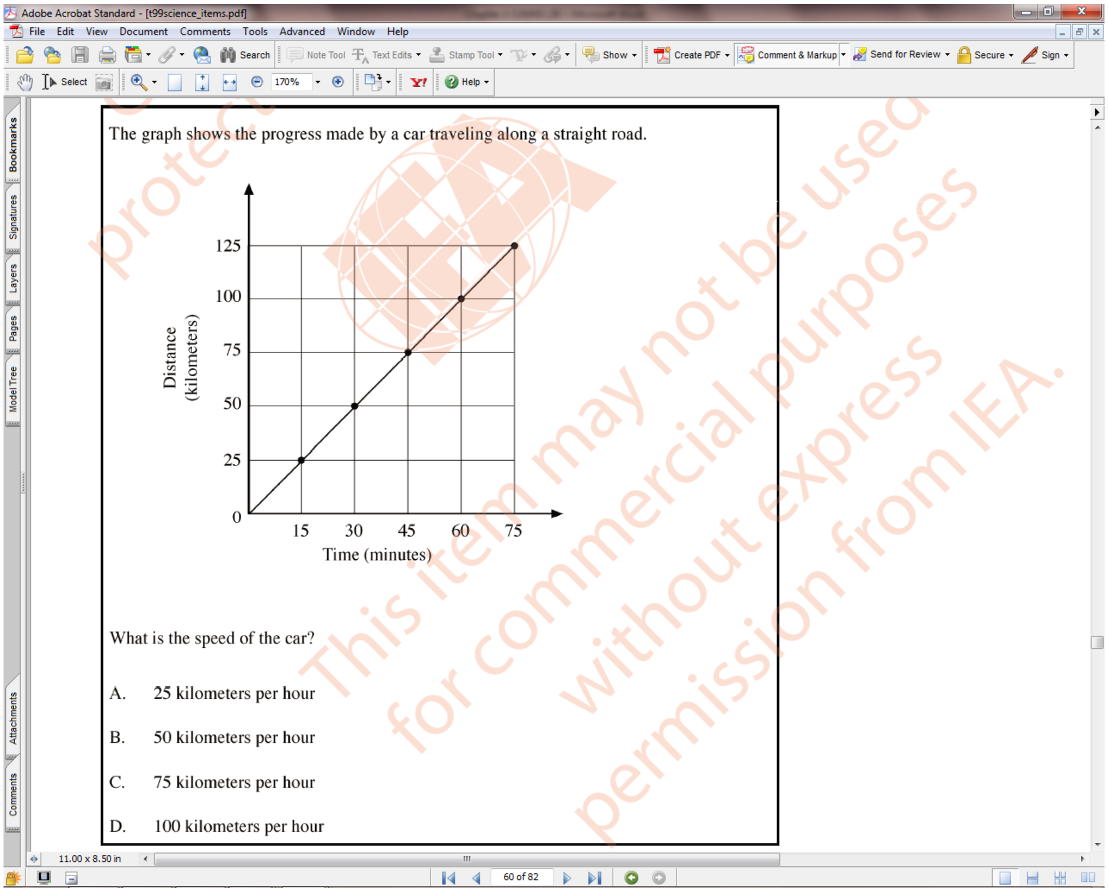
Appendix C

Sample Coh-Metrix Outputs and Corresponding Items

Figure C.1 presents a typical multiple-choice item from the TIMSS Science test of 1999, and Figure C.2 illustrates some partial outcomes from the Coh-Metrix based on the stem of the item. This example also can represent a typical CIVED multiple-choice item because the language and sentence structure of the CIVED multiple-choice items are very similar to the typical TIMSS Science item except that most of CIVED items don't contain graphic features.

Figure C.3 shows a typical PISA science item stem which demands much more reading than a typical TIMSS science item does, and Figure C.4 demonstrates partial outcomes of Coh-Metrix based on the item stem. I present the outcome from a PISA science item to illustrate the difference between test items with higher degree of reading demand and lower degree of reading demand (e.g., a TIMSS science item).

Please note that I used Coh-Metrix 2.0 for all my text analysis. Only these two examples presented in Figure C. 2 and C.4 were from Coh-Metrix 3.0. I finished all my text analysis using Coh-Metrix 2.0 before November, 2012. The Coh-Metrix team took down the Coh-Metrix 2.0 in late 2012 and replaced it with Coh-Metrix 3.0. When I created these two Figures on Nov. 28, 2012, I found I had no choice but to use the Coh-Metrix 3.0.



The graph shows the progress made by a car traveling along a straight road.

What is the speed of the car?

A. 25 kilometers per hour

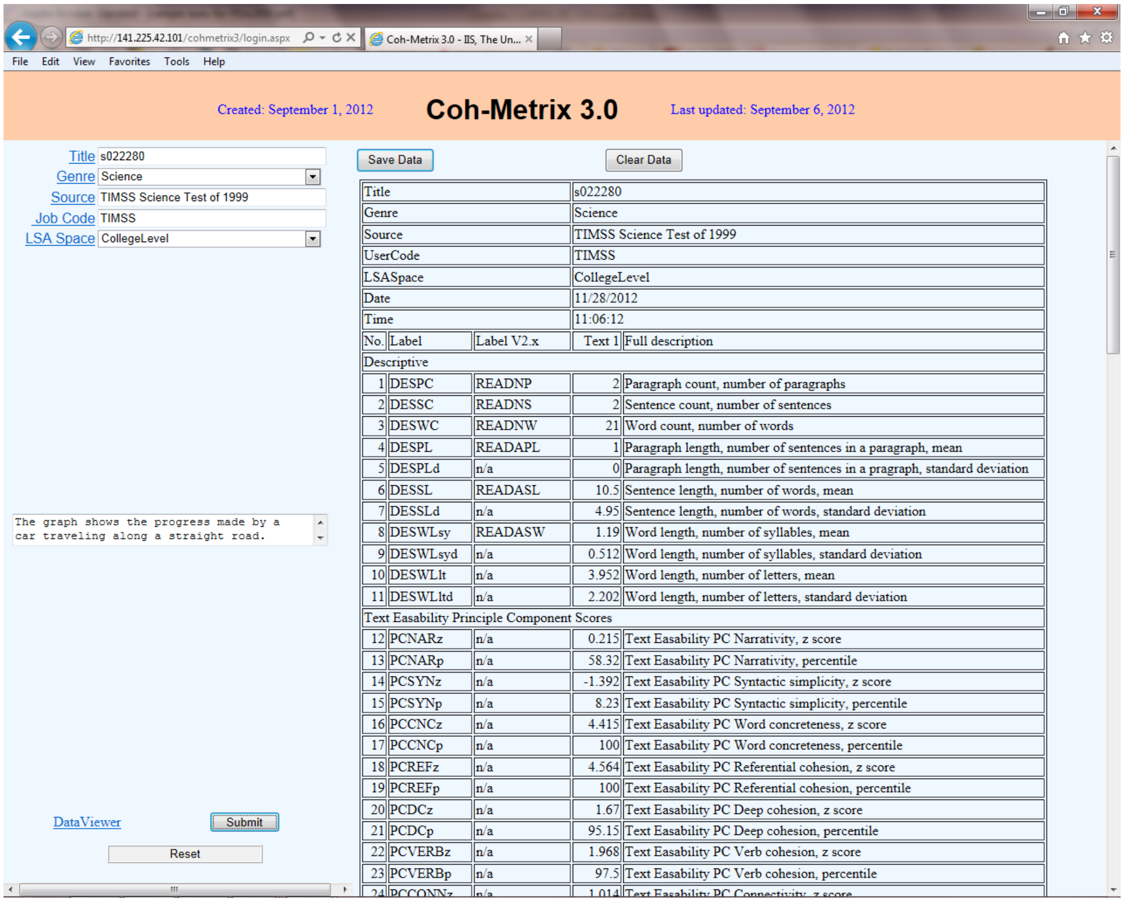
B. 50 kilometers per hour

C. 75 kilometers per hour

D. 100 kilometers per hour

This item may not be used for commercial purposes without express permission from IEA.

Figure C.1. IEA TIMSS Science Test of 1999--*Item s022280*


http://141.225.42.101/cohmetrix3/login.aspx
Coh-Metrix 3.0 - IIS, The Un...
File
Edit
View
Favorites
Tools
Help
Save Data
Clear Data
DataViewer
Submit
Reset

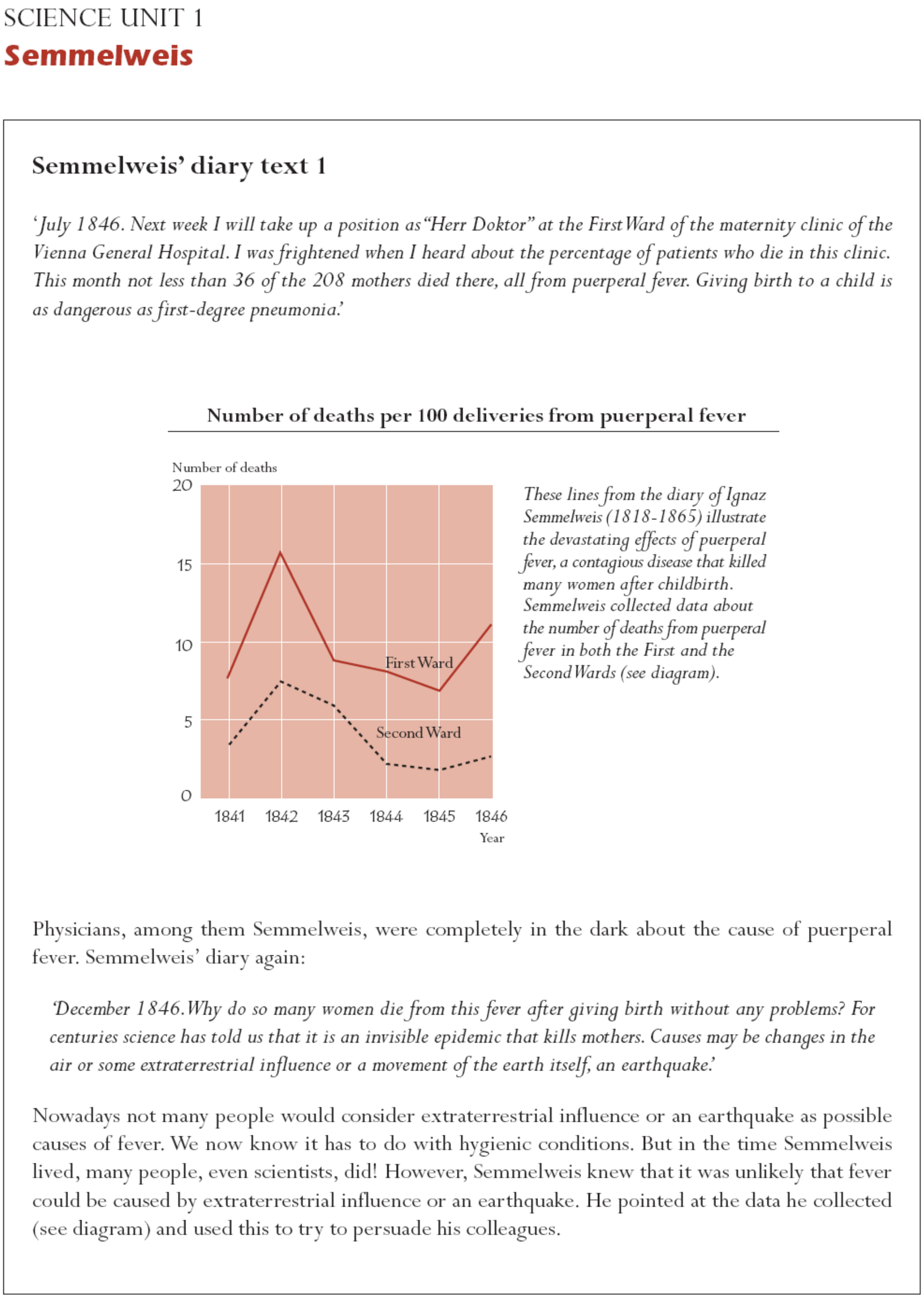
Created: September 1, 2012 **Coh-Metrix 3.0** Last updated: September 6, 2012

Title s022280
Genre Science
Source TIMSS Science Test of 1999
Job Code TIMSS
LSA Space CollegeLevel

```
The graph shows the progress made by a
car traveling along a straight road.
```

Title		s022280		
Genre		Science		
Source		TIMSS Science Test of 1999		
UserCode		TIMSS		
LSASpace		CollegeLevel		
Date		11/28/2012		
Time		11:06:12		
No.	Label	Label V2.x	Text 1	Full description
Descriptive				
1	DESPC	READNP	2	Paragraph count, number of paragraphs
2	DESSC	READNS	2	Sentence count, number of sentences
3	DESWC	READNW	21	Word count, number of words
4	DESPL	READAPL	1	Paragraph length, number of sentences in a paragraph, mean
5	DESPLd	n/a	0	Paragraph length, number of sentences in a pragraph, standard deviation
6	DESSL	READASL	10.5	Sentence length, number of words, mean
7	DESSLd	n/a	4.95	Sentence length, number of words, standard deviation
8	DESWLsy	READASW	1.19	Word length, number of syllables, mean
9	DESWLsyd	n/a	0.512	Word length, number of syllables, standard deviation
10	DESWLlt	n/a	3.952	Word length, number of letters, mean
11	DESWLltd	n/a	2.202	Word length, number of letters, standard deviation
Text Easability Principle Component Scores				
12	PCNARz	n/a	0.215	Text Easability PC Narrativity, z score
13	PCNARp	n/a	58.32	Text Easability PC Narrativity, percentile
14	PCSYNz	n/a	-1.392	Text Easability PC Syntactic simplicity, z score
15	PCSYNp	n/a	8.23	Text Easability PC Syntactic simplicity, percentile
16	PCCNCz	n/a	4.415	Text Easability PC Word concreteness, z score
17	PCCNCp	n/a	100	Text Easability PC Word concreteness, percentile
18	PCREFz	n/a	4.564	Text Easability PC Referential cohesion, z score
19	PCREFp	n/a	100	Text Easability PC Referential cohesion, percentile
20	PCDCz	n/a	1.67	Text Easability PC Deep cohesion, z score
21	PCDCp	n/a	95.15	Text Easability PC Deep cohesion, percentile
22	PCVERBz	n/a	1.968	Text Easability PC Verb cohesion, z score
23	PCVERBp	n/a	97.5	Text Easability PC Verb cohesion, percentile
24	PCCONNz	n/a	1.014	Text Easability PC Connectivity, z score

Figure C.2. Coh-Metrix 3.0 outcomes based on the stem of *Item s022280* from IEA TIMSS Science test of 1999



SCIENCE UNIT 1

Semmelweis

Semmelweis' diary text 1

'July 1846. Next week I will take up a position as "Herr Doktor" at the First Ward of the maternity clinic of the Vienna General Hospital. I was frightened when I heard about the percentage of patients who die in this clinic. This month not less than 36 of the 208 mothers died there, all from puerperal fever. Giving birth to a child is as dangerous as first-degree pneumonia.'

Number of deaths per 100 deliveries from puerperal fever

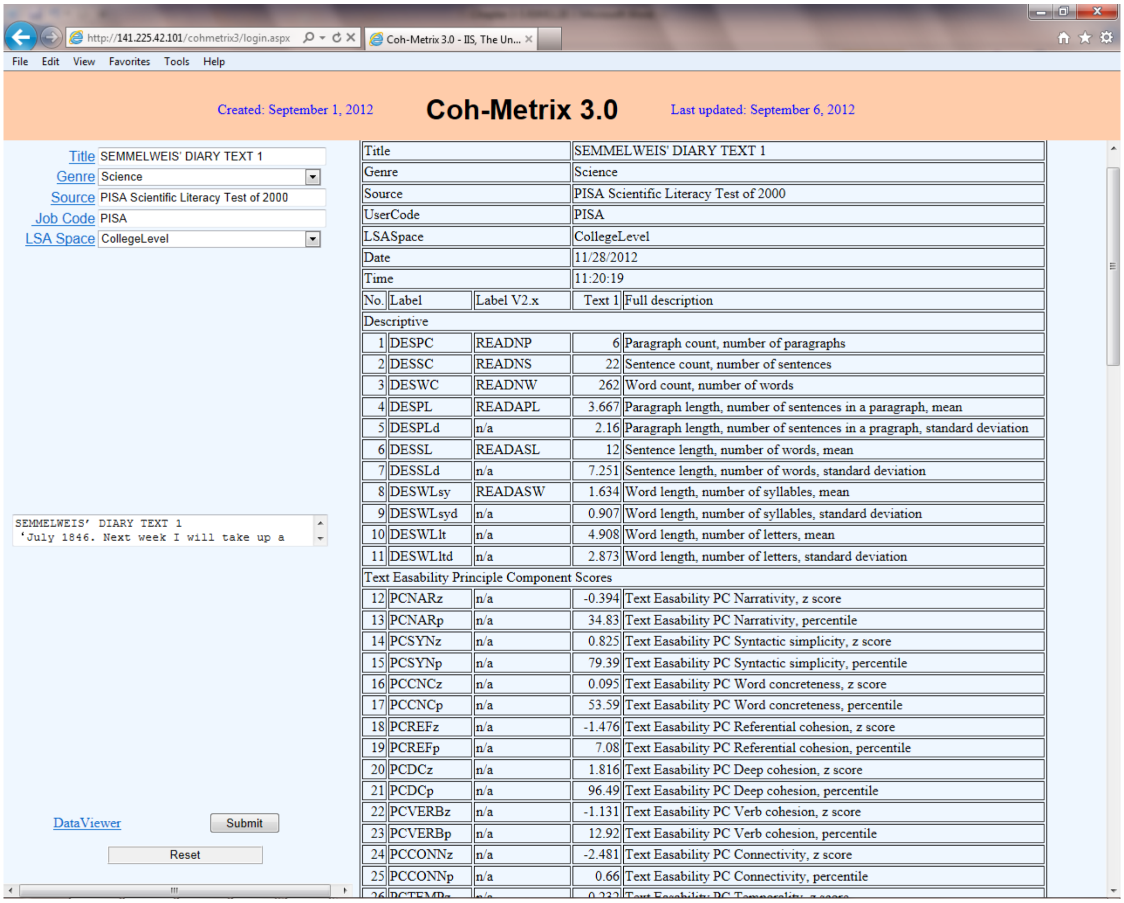
These lines from the diary of Ignaz Semmelweis (1818-1865) illustrate the devastating effects of puerperal fever, a contagious disease that killed many women after childbirth. Semmelweis collected data about the number of deaths from puerperal fever in both the First and the Second Wards (see diagram).

Physicians, among them Semmelweis, were completely in the dark about the cause of puerperal fever. Semmelweis' diary again:

> *'December 1846. Why do so many women die from this fever after giving birth without any problems? For centuries science has told us that it is an invisible epidemic that kills mothers. Causes may be changes in the air or some extraterrestrial influence or a movement of the earth itself, an earthquake.'*

Nowadays not many people would consider extraterrestrial influence or an earthquake as possible causes of fever. We now know it has to do with hygienic conditions. But in the time Semmelweis lived, many people, even scientists, did! However, Semmelweis knew that it was unlikely that fever could be caused by extraterrestrial influence or an earthquake. He pointed at the data he collected (see diagram) and used this to try to persuade his colleagues.

Figure C.3. A sample task (item stem only) from PISA Scientific Literacy Assessment of 2000 (OECD, 2002, p. 108)



http://141.225.42.101/cohmetrix3/login.aspx
Coh-Metrix 3.0 - IIS, The Un...
File Edit View Favorites Tools Help
DataViewer
Submit
Reset

Created: September 1, 2012 **Coh-Metrix 3.0** Last updated: September 6, 2012

Title	SEMMELWEIS' DIARY TEXT 1
Genre	Science
Source	PISA Scientific Literacy Test of 2000
Job Code	PISA
LSA Space	CollegeLevel

```
SEMMELWEIS' DIARY TEXT 1
'July 1846. Next week I will take up a
```

Title	SEMMELWEIS' DIARY TEXT 1
Genre	Science
Source	PISA Scientific Literacy Test of 2000
UserCode	PISA
LSASpace	CollegeLevel
Date	11/28/2012
Time	11:20:19

No.	Label	Label V2.x	Text 1	Full description
Descriptive				
1	DESPC	READNP	6	Paragraph count, number of paragraphs
2	DESSC	READNS	22	Sentence count, number of sentences
3	DESWC	READNW	262	Word count, number of words
4	DESPL	READAPL	3.667	Paragraph length, number of sentences in a paragraph, mean
5	DESPLd	n/a	2.16	Paragraph length, number of sentences in a pragraph, standard deviation
6	DESSL	READASL	12	Sentence length, number of words, mean
7	DESSLd	n/a	7.251	Sentence length, number of words, standard deviation
8	DESWLsy	READASW	1.634	Word length, number of syllables, mean
9	DESWLsyd	n/a	0.907	Word length, number of syllables, standard deviation
10	DESWLlt	n/a	4.908	Word length, number of letters, mean
11	DESWLltd	n/a	2.873	Word length, number of letters, standard deviation
Text Easability Principle Component Scores				
12	PCNARz	n/a	-0.394	Text Easability PC Narrativity, z score
13	PCNARp	n/a	34.83	Text Easability PC Narrativity, percentile
14	PCSYNz	n/a	0.825	Text Easability PC Syntactic simplicity, z score
15	PCSYNp	n/a	79.39	Text Easability PC Syntactic simplicity, percentile
16	PCCNCz	n/a	0.095	Text Easability PC Word concreteness, z score
17	PCCNCp	n/a	53.59	Text Easability PC Word concreteness, percentile
18	PCREFz	n/a	-1.476	Text Easability PC Referential cohesion, z score
19	PCREFp	n/a	7.08	Text Easability PC Referential cohesion, percentile
20	PCDCz	n/a	1.816	Text Easability PC Deep cohesion, z score
21	PCDCp	n/a	96.49	Text Easability PC Deep cohesion, percentile
22	PCVERBz	n/a	-1.131	Text Easability PC Verb cohesion, z score
23	PCVERBp	n/a	12.92	Text Easability PC Verb cohesion, percentile
24	PCCONNz	n/a	-2.481	Text Easability PC Connectivity, z score
25	PCCONNp	n/a	0.66	Text Easability PC Connectivity, percentile
26	PCTEMPz	n/a	0.232	Text Easability PC Temporality, z score

Figure C.4. *Coh-Metrix 3.0 outcomes based on the stem of the sample task from PISA Scientific Literacy Test of 2000*

Appendix D

Descriptions of Selected Text Features from the Coh-Metrix Version 2.0 (McNamara, et al., 2005)

Text Features	Description	The level of reading comprehension	Full description
CONLGni	Negative logical connectives	Textbase and/or situation model. Text cohesion	This is the incidence of negative logical connectives, including but, until, and although.
CONLGpi	Positive logical connectives	Textbase and/or situation model Text cohesion	This is the incidence of positive logical connectives.
DENLOGi	Logical Operators	Textbase Syntactic complexity	Logical operators express logical reasoning, and are a type of metric that assesses syntactic complexity in a text. They include operators such as and, or, not, if, then, and other similar conditionals.
DENNEGi	Negation	Textbase. Syntactic complexity	This is an incidence score for negation expressions. Negation is a process that turns an affirmative statement (I am American) into its opposite denial (I am not American). Negation can be adjective (there is no computer), or pronoun (Nobody is American here), or adverb (I never was American).
DENSPR2	Pronoun ratio	Textbase. Syntactic complexity	This is the ratio of pronouns to the noun phrases in a text. A high density of pronouns can increase text syntactic complexity, and create comprehension problems when the reader does not know what the pronoun refers to. For example, “The fourth stage of mitosis is called telophase, because telo- means ‘end,’ and it begins when all the daughter chromosomes reach the two cell poles.” The word “it” is tagged as a pronoun, whereas phrases such as “the fourth stage”

			are tagged as noun phrases. If there is one pronoun and 8 total noun-phrases (the pronoun itself being a noun phrase) then the ratio would be 0.125.
DENSNP	Noun phrase incidence	Textbase Syntactic complexity	The noun phrase incidence is a type of syntactic index that assess syntactic complexity in a text. It is the frequency of noun-phrase constituents per 1000 words. The higher the score, the more noun-phrases in the text. For example, consider the sentence "Cell division occurs to reproduce and replace cells." There are two main DENSNPs in the sentence: cell division and cells. There are a total of eight words, hence the incidence score for this sentence is 2/8*1000 = 250.
FRQCRmcs	Min. raw frequency of content words		In Coh-Metrix, this index initially computes the lowest frequency score among all of the content words in each sentence. The frequency scores vary between 0 to 1,000,000. A word with the lowest frequency score is the rarest word in the sentence.
FRQCRaw	Raw frequency of content words	Surface Word frequency	This is the average raw frequency of all the content words in the text. In Coh-Metrix, content words are nouns, adverbs, adjectives, main verbs, and other categories with rich conceptual content.
FRQCLacw	Log frequency of content words		This is the log frequency of all content words in the text. In Coh-Metrix, content words are nouns, adverbs, adjectives, main verbs, and other categories with rich conceptual content. Previous research suggests that taking the log of the frequencies instead of the raw scores is consistent with

			research on reading time (Haberlandt & Graesser, 1985; Just & Carpenter, 1980).
FRQCLmcs	Log min. raw frequency of content words		According to the Coh-Metrix version 2.0 indices online manual, this index “initially computes the lowest log frequency score among all of the content words in each sentence. A mean of these minimum log frequency scores is then computed. The logarithm is to the base 10. Content words are nouns, adverbs, adjectives, main verbs, and other categories with rich conceptual content. The word with the lowest log frequency score is the rarest word in the sentence. (Scores range from 0-6).” (p. 9)
HYNOUNaw	Average hypernym values of nouns	Surface/situation model. Vocabulary	This is the mean hypernym value of nouns in the text. Hypernymy measure is one way of assessing the abstractness of a word based on WordNet (Fellbaum, 1998; Miller, et al., 1990). An abstract word is one with few distinctive features and few attributes that can be pictured in the mind. A word with high hypernym levels lean toward concrete, and low hypernym levels to abstract.
HYVERBaw	Average hypernym values of verbs	Surface/situation model. Vocabulary	This is the mean hypernym value of main verbs in the text.

INTEi	Intentional content	Situation model	Text comprehension researches have suggested at least five situational dimensions that can contribute to the situation model (Zwaan & Radvansky, 1988), and intentional content belongs to one of the dimensions, namely, intentional dimension. The intentional content reflects the extent to which sentences are related by intentional particles (e.g., in order to, so that, for the purpose of, by means of, by, wanted to), actions, and events. Coh-Metrix estimates intentional actions and events by counting the number of main verbs that are intentional (actions which are performed in pursuit of goals) based on WordNet (Fellbaum, 1998; Miller, et al., 1990). The higher the counts in a text, the more the text is assumed to carry goal-driven content.
READASL	Average words per sentence	Surface level/ Readability index	This is the mean number of words per sentence.
READASW	Average syllables per word	Surface level/ Readability index	This is the mean number of syllables per word.
READNS	Number of sentences	Surface level/ Readability index	This is the number of sentences in the entire text.
SYNHw	Higher level constituents	Textbase. Syntactic complexity	Structurally dense sentences tend to have more high order syntactic constitutes per word.
SYNNP	Mean number of modifiers per noun-phrase	Textbase. Syntactic complexity	This is the mean number of modifiers per noun-phrase. "A modifier is an optional element that describes the property of a head of a phrase. Modifiers per NP refer to adjectives, adverbs, or determiners that modify the head noun. For example, the noun-phrase *the lovely, little girl* has three

			modifiers: *the*, *lovely* and *little*. A second metric is mean number of higher level constituents per sentence, controlling for number of words.”
WORDCacw	Mean concreteness of words in a text	Surface/situation model. Vocabulary	This is the mean concreteness value of all content words in a text that match a word in the MRC Psycholinguistics Database (Coltheart, 1981). Concreteness measures how concrete a word is based MRC concreteness ratings. The higher the score, the more concrete the word is. The scores range from 100 to 700.
WORDCmcs	Mean concreteness of words across sentences	Surface/situation model. Vocabulary	This is the mean of low-concreteness of words across sentence The scores range from 100 to 700 with high values leaning toward concrete.

Appendix E

Reading Demand Coding Scheme

The purpose of the rating scheme is to understand the reading difficulty of test items. Please read the item first and evaluate the difficulty of the text of item stem and of multiple-choice options. For example, below is an item from the IEA civic education test. The highlighted part is the item stem and the four options listed below are the multiple-choice options.

Which of these statements best describes the role of the citizen in democratic countries? The citizen …

A. can vote on the national budget.

B. can vote for representatives who then vote for laws.

C. must always vote for the same political party.

D. must obey leader without question.

Key: B

1. **Please rate the level of reading difficulty for the item as a whole for an 8th grader in a public school in the United States.**

1	2	3	4	5
Very easy		Moderate difficulty		Very difficult

2. **If you have rated the item at a 3 or higher on difficulty, where is the reading difficulty? Circle all that apply.**

(1) The difficulty of vocabulary in the item stem

(2) The difficulty of vocabulary in the multiple-choice options. Please specify which option(s).

(3) Complexity of grammar or syntax in the item stem.

(4) Complexity of grammar or syntax in the multiple-choice options. Please specify which option(s).

(5) Other. Please specify.

3. If you have rated the item at a 3 or higher on difficulty, do you think the reason(s) you selected as causing the item difficulty is/are relevant to the content which the item assesses?

1	2	3
Yes	Somewhat	No

4. Do you think this item could be rewritten to reduce the reading difficulty, but still assess the relevant content?

1	2	3
Yes	Somewhat	No

5. When you read the item stem, how abstract do you think the item question is? Please circle one.

(Provide raters with this descriptive sheet, but on the form for the actual rating format the response similar to your other rating scales from left to right

1	2	3	4	5
Most concrete	Highly concrete	Intermediate	Highly abstract	Most abstract

(1) The first level, most concrete, asks for the "identification of persons, animals, or things."

(2) The second level, the highly concrete class of questions, asks for the "identification of amounts, times, or attributes."

(3) The third level, intermediate questions, asks for the “identification of manner, goal, purpose, alternative, attempt, or condition.”

(4) The fourth level, highly abstract, asks for the “identification of cause, effect, reason, or result.”

(5) The highest level, the most abstract questions, asks for the “identification of equivalence, difference, or theme”.

6. **When you read the item stem, what genre do you think the text belongs to? Please circle one.**

- *Narrative.* Narrative passages tend to describe relatively mundane events with which most people have some familiarity from a personal perspective.
- *Expository.* Expository passages tend to describe historical, social, and/or scientific facts from a nonpersonal, objective perspective.

1	2	3	4	5
Narrative		Mixed/both		Expository

7. **Other Comments:**

Illustrating Examples for Rating Item 5

5. **When you read the item stem, how abstract do you think the item question is?**

   Please circle one.

(Provide raters with this descriptive sheet, but on the form for the actual rating format the response similar to your other rating scales from left to right

1	2	3	4	5
Most concrete	Highly concrete	Intermediate	Highly abstract	Most abstract

E. **The first level, *most concrete*, asks for the "identification of persons, animals, or things."**

Examples: D06. Seeds develop from which part of a plant?

J08. Sunscreen is used to protect the skin from exposure to which type of solar radiation?

F. **The second level, the *highly concrete* class of questions, asks for the "identification of amounts, times, or attributes."**

Examples: L02. What is the primary function of the large leaves found on seedlings growing in a forest?

L03. Which one of the following characteristics is most likely to be found in mammals that are preyed on by other mammals for food?

N08. Which statement best explains why mammals are found in very cold region of the world but lizards are not?

G. **The third level, intermediate questions, asks for the "identification of manner, goal, purpose, alternative, attempt, or condition."**

Examples: J04. A student put 100ml of water in each of the open containers and let them stand in the sun for one day. Which container would probably lose the most water due to evaporation?

B04. Immediately before and after running a 50 meter race, your pulse and breathing rates are taken. What changes would you expect to find?

H. The fourth level, highly abstract, asks for the "identification of cause, effect, reason, or result."

Examples: J06. Which of the following is an important factor in explaining why seasons occur on Earth?

J07. The BEST reason for including protein in a healthy diet is because it is the main source of

I. The highest level, the most abstract questions, asks for the "identification of equivalence, difference, or theme".

Another Example (Mosenthal, 1996, p. 323**)**

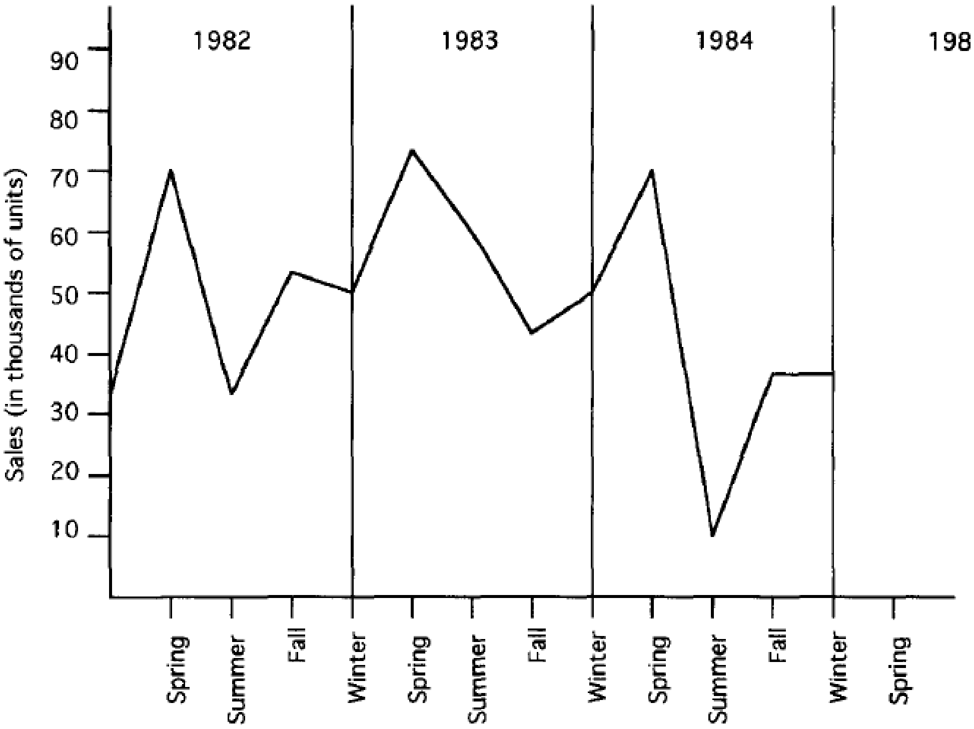


Figure 3. A sales graph by season from 1982 to 1984.

The resulting scale forms a continuum of difficulty depending on how concrete or abstract different types of requested information are. This continuum was as follows.

1. Questions requesting information regarding the identification of persons, animals, or things were scored the highest in concreteness and therefore received a score of 1 and were hypothesized to be the easiest to answer.

2. Questions requesting information regarding the identification of amounts, times, attributes, types, actions, and locations (e.g., for Figure 3, "How much was the sales [in thousands] for the winter1984?" [answer: "38"]) were assigned a concreteness score of 2 and were hypothesized to be the next easiest to answer.

3. Questions requesting information regarding the identification of manner, goal, purpose, alternative, attempt, condition, pronominal reference, and predicate adjectives (e.g., "What is the purpose of the sales graph shown in Figure 3?" [answer: "To show a company's sales over a 3-year period, from 1982 to 1984"]) were assigned a concreteness score of 3 and were hypothesized to be of moderate difficulty to answer.

4. Questions requesting information regarding the identification of cause, effect, reason, result, evidence, similarity, and explanation (e.g., "Given the seasonal pattern shown on the graph, what similar pattern appears for spring in 1982, 1983, and 1984?" [answer: "This is the month that sales tend to be the highest"]) were assigned a concreteness score of 4 and were hypothesized to be difficult to answer.

5. Finally, questions requesting information regarding the identification of equivalent, difference, and theme were assigned a concreteness score of 5 (the

term *equivalence* in this case refers to highly unfamiliar or low-frequency vocabulary items for which respondents must provide a definition). Questions requesting these types of information were hypothesized to be the most difficult to answer. (An example of a level 5 type-of-information question as applied to Figure 3 would be "What is the major difference between sales between spring and summer and sales between winter and spring?" [answer: "Sales tend to fall between spring and summer but climb between winter and spring."])

Reference:

Mosenthal, P. (1996). Understanding the strategies of document literacy and their conditions of use. *Journal of Educational Psychology, 88*, 314-332.

Bibliography

Abedi, J. (2002). Standardized achievement tests and English language learners: Psychometrics issues. *Educational Assessment, 8*(3), 231-257. doi:10.1207/S15326977EA0803_02

Abedi, J. (2009). Validity of assessments for English language learning students in a national/international context. *Estudios Sobre Educacion, (16*), 167-183. Retrieved from EBSCOhost.

Abedi, J., Bayley, R., Ewers, N., Mundhenk, K., Leon, S., Kao, J., & Herman, J. (2012). Accessible reading assessments for students with disabilities. *International Journal of Disability, Development and Education, 59*(1), 81-95. doi:10.1080/1034912X.2012.654965

Abedi, J., Courtney, M., & Leon, S. (2003).*Effectiveness and validity of accommodations for English language learners in large-scale assessments*. Los Angeles: University of California, Center for the Study of Evaluation/National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., & Gándara, P. (2006). Performance of English language learners as a subgroup in large-scale assessment: Interaction of research and policy. *Educational Measurement: Issues and Practice, 25*(4), 36-46. doi:10.1111/j.1745-3992.2006.00077.x

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education, 14*(3), 219-234.

Abedi, J., Lord, C., Hofstetter, C., & Baker, E. (2000). Impact of accommodation strategies on English language learners' test performance. *Educational Measurement: Issues and Practice*, *19*(3), 16-26. doi:10.1111/j.1745-3992.2000.tb00034.x

Adams, M. J. (1990). *Beginning to Read*. Cambridge, MA: MIT Press.

Adams, D. M., Mayer, R. E., MacNamara, A., Koenig, A., & Wainess, R. (2012). Narrative games for learning: Testing the discovery and narrative hypotheses. *Journal of Educational Psychology*, *104*(1), 235-249. doi:10.1037/a0025595

Alexander, P. A. (1997). 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., & The Disciplined Reading and Learning Research, L. (2012). Reading into the future: Competence for the 21st century. *Educational Psychologist*, *47*(4), 259-280. doi:10.1080/00461520.2012.722511

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. 3, pp. 285–310). New Mahwah, NJ: Lawrence Erlbaum.

Alexander, P.A. & Judy, J.E. (1988). The interaction of domain-specific and strategic knowledge in academic performance. *Review of Educational Research*, 58, 375–404.

Alexander, P. A., & Kulikowich, J. M. (1991). Domain knowledge and analogic reasoning ability as predictors of expository text comprehension. *Journal of Reading Behavior, 23*,165-190.

Alexander, P. A., Kulikowich, J. M., & Schulze, S. K. (1994). How subject-matter knowledge affects recall and interest. *American Educational Research Journal,* 31(2), 313-337. doi:10.2307/1163312.

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*, 315-343.

Allington, R. L. & Cunningham, P. (2006). *School that work: All children read and write.* Pearson: Allyn & Bacon.

Allington, R. L., & Weber, R. (1993). Questioning questions in teaching and learning from texts. In B. K. Britton, A. Woodward, M. R. Binkley, B. K. Britton, A. Woodward, M. R. Binkley (Eds.), *Learning from textbooks: Theory and practice* (pp. 47-68). Hillsdale, NJ England: Lawrence Erlbaum Associates, Inc. Retrieved from EBSCO*host*.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association

Anastasi, A, & Urbina, S. (1997). *Psychological testing: Seventh edition.* Upper Saddle River, NJ: Prentice Hall.

Anderson, R. C. (1982). How to construct achievement tests to assess comprehension. *Review of Educational Research,* 42,145-170.

Anderson, R. C., & Freebody, P. (1981). Vocabulary knowledge. In J. T. Guthrie (Ed.), *Comprehension and teaching: Research reviews* (pp. 77-117). Newark, DE: International Reading Association.

Anderson, R. C., & Pichert, J. W. (1978). Recall of previously unrecallable information following a shift in perspective. *Journal of Verbal Learning and Verbal Behavior, 17,* 1-12.

Anderson, R. C., & Biddle, W. B. (1975). On asking people questions about what they are reading. In G. Bower (Ed.), *The psychology of learning and motivation* (vol. 9, pp. 89-132). New York, NY: Academic Press.

Baker, L., & Brown, A. L. (1984). Metacognitive skills and reading. In P. D. Pearson (Ed.)*, Handbook of reading research* (pp. 353 – 394). New York: Longman.

Baker, L., & Wigfield, A. (1999). Dimensions of children's motivation for reading and their relations to reading activity and reading achievement. *Reading Research Quarterly, 34 (*4), 452-477.

Baldi, S., Peries, M., Skidmore, D., Greenberg, E. & Hahn, C. (Ed.). (2001). *What democracy means to ninth graders: U.S. results from the IEA Civic Education Study*. Washington, D.C.: National Center for Education Statistics.

Beck, I., McKeown, M., & Omanson, R, (1987). The effect and uses of diverse vocabulary instruction techniques. In M. McKeown & M. E. Curtis (Eds.)*, The nature of vocabulary acquisition* (pp. 147-163). Hillsdale NJ: Erlbaum.

Bennett, R. E., & Ward, W. C. (Eds.). (1993). *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment.* Hillsdale NJ: Erlbaum.

Bertam, B. & Newman, S. (1981). *Why readability formulas fail. Reading education report No. 28*. Illinois University, Urbana, Center for the Study of Reading (Eric document service number ED205915). Retrieved from EBSCO*host*.

Best, R., Ozuru, Y., Floyd., R., & McNamara, D.S. (2006). Children's text comprehension. Effects of genre, knowledge, and text cohesion. In S. A. Barab, K. E. Hay, D. T. Hickey (Eds.), *Proceedings of the seventh international conference of the learning sciences* (pp. 37-42). Mahwah, NJ: Erlbaum.

Bland, J.M., Altman, D.G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet, i.* 307-310.

Bower, G.H., Black, J.B., & Turner, T.J. (1979). Scripts in memory for text. *Cognitive Psychology, 11*, 177-220.

Bransford, J. D., Brown, A. L., & Cocking, R.R. (Eds.). (1999). *How people learn: Brain, mind, experience, and school.* Washington, D.C.: National Academy Press.

Bransford, J. D., & Johnson, M. K. (1972). Contextual prerequisites for understanding: Some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior, 11*, 717-726.

Bransford, J. D., & Johnson, M. K. (1973). Considerations of some problems of comprehension. In W. G. Chase (Ed.), *Verbal information processing* (pp. 383-438). New York: Academic.

Breese, J. S., Goldman, R. P., & Wellman, M. P. (1994). Introduction to the special section on knowledge-based construction of probabilistic and decision models. *IEEE Transactions on Systems, Man, and Cybernetics, 24*, 1577-1579.

Bridgeman, B., & Rock, D. A. (1993). Relationships among multiple-choice and open-ended analytical questions. *Journal of Educational Measurement*, 30(4), 313-29. Retrieved from EBSCO*host*.

Brozo, W. G. (2002). *To be a boy, to be a reader: Engaging teen and preteen boys in active literacy.* Newark, DE: International Reading Association.

Burton, S., Sudweeks, R., Merrill, P. & Wood, B., (1991), *How to prepare better multiple-choice test items: Guidelines for university faculty*, Brigham Young University Testing Services and The Department of Instructional Science.

Bybee, R.W. (1997) *Achieving scientific literacy: from purpose to practice*. Portsmouth, NH: Heinemann.

Campbell, J. R. (1999). Cognitive processes elicited by multiple-choice and constructed-response questions on an assessment of reading comprehension. Ph.D. dissertation, Temple University, United States -- Pennsylvania. Retrieved September 21, 2011, from Dissertations & Theses: Full Text. (Publication No. AAT 9938651).

Canadian Council on Learning (CCL) (2009).*Why boys don't like to read: Gender differences in reading achievement*. Retrieved November 27, 2012, from http://www.nald.ca/library/research/ccl/lessons_learning/why_boys/why_boys.pdf

Carroll, J.M., Snowling, M.J., Hulme, C., & Stevenson, J. (2003). The development of phonological awareness in preschool children. *Developmental Psychology, 39*(5), 913-923.

Carver. R., P. (1990). Intelligence and reading ability in Grades 2–12. *Intelligence* 14: 449–55.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284-290.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37-46.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Academic Press.

Collins, J.(2007). Are our children reading proficiently and how would we know? An examination of state and national elementary reading assessments. Ph.D. dissertation, The University of Oklahoma, United States -- Oklahoma. Retrieved September 21, 2011, from Dissertations & Theses: Full Text.(Publication No. AAT 3271227).

Coltheart, M. (1981). The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology*, *33*(a), 497–505

Comber, L.C., & Keeves, J.P. (1973). *Science education in nineteen countries: An empirical study*. Stockholm: Almqvist & Wiksell.

Cordón, L. A., & Day, J. D. (1996). Strategy use on standardized reading comprehension tests. *Journal of Educational Psychology*, *88*(2), 288-295. doi:10.1037/0022-0663.88.2.288.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Harcourt Brace Jovanovich College Publishers: Philadelphia.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.). *Test validity* (pp. 3-17). Hillsdale, NJ: Lawrence Erlbaum.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281-302.

Crooks, T. J., Kane, M. T., & Cohen, A. S. (1996). Threats to valid use of assessment. *Assessment in Education,* 9(3), 265-285.

Cummins, D. D., Kintsch,W., Reusser, K., &Weimer, R. (1988). The role of understanding in solving word problems. *Cognitive Psychology, 20,* 405–438.

Dale, E., & Chall, J. (1948). A formula for predicting readability. *Educational Research Bulletin, 27*, 37-54.

Dansereau, D. F. (1985). Learning strategy research. In S. C. J. Segal & R. Glaser (Eds.), *Thinking and learning skills: Relating instruction to research* (Vol. 1, pp. 209-240). Hillsdale, NJ: Lawrence Erlbaum Associates.

D'Arcy, R. N., Service, E., Connolly, J. F., & Hawco, C. S. (2005). The influence of increased working memory load on semantic neural systems: a high-resolution event-related brain potential study. *Cognitive Brain Research, 22*(2), 177-191. doi:10.1016/j.cogbrainres.2004.08.007.

De Corte, E., & Verschaffel, L. (1993). Some factors influencing the solution of addition and subtractionword problems. In K. Durkin & B. Shire (Eds.), *Language in mathematical education* (pp. 118-130). Philadelphia: Open University Press.

DeMars, C. E. (1998). Gender differences in mathematics and science on a high school proficiency exam: The role of response format. *Applied Measurement in Education, 11*(3), 279-299.

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. (2000). 3.6 minutes per day: The scarcity of informational texts in first grade. *Reading Research Quarterly, 35*(2). 202-224.

Duke, N., & Carlisle, J. F. (2011). Comprehension development. In M. L. Kamil, P. D. Pearson, P. A. Afflerbach, & E. B. Moje (Eds.), *Handbook of reading research, Vol. 4 (pp. 199-228).* NY: Routledge.

Duran, R. P. (2011). Ensuring valid educational assessments for ELL students: Scores, score interpretation, and assessment uses. In M. Bastera, E. Trumbull, & G. Solano-Flores (eds.). *Cultural validity in assessment: Addressing linguistic and cultural Diversity*. New York: Routledge/Taylor & Francis Group.

Eason, S. H., Goldberg, L. F., Young, K. M., Geist, M. C., & Cutting, L. E. (2012). Reader-text interactions: How differential text and question types influence cognitive skills needed for reading comprehension. *Journal of Educational Psychology, 104(*3), 515-528.

Educational Testing Service (1998b). *Learning to write reading and comprehension materials for GRE & GMAT verbal skills tests.* International ETS publication.

Elfenbein, A. (2011). Research in text and the uses of Coh-Metrix. *Educational Researcher*, *40*(5), 246-248. doi:10.3102/0013189X11414181.

Embretson, S. E. (1994). Application of cognitive design systems to test development. In C. R. Reynolds (Ed.), *Cognitive assessment: A multidisciplinary perspective* (pp. 107-135). New York: Plenum.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3,* 380–396.

Embretson, S. E., & Gorin, J. S. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement, 38*(4), 343-368.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah,New Jersey: Lawrence Erlbaum Associates.

Embretson, S. E., & Wetzel, C. D. (1987). Component latent trait models for paragraph comprehension. *Applied Psychological Measurement, 11,* 175-193.

Engelhard, J. G., Hansche, L., & Rutledge, K. E. (1990). Accuracy of Bias Review Judges in Identifying Differential Item Functioning on Teacher Certification Tests. *Applied Measurement in Education, 3(*4), 347-360.

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.

Flesch, R. (1951). *How to test readability*. New York: Harper and Brothers.

Fry, E. (1968). A readability formula that saves time. *Journal of Reading, 11*, 513-516.

Gamma, E., Helm, R., Johnson, R., & Vlissides, J. (1994). *Design patterns*. Reading, MA: Addison-Wesley.

Gaskins, I.W. (2003). Taking charge of reader, text, activity, and context variables. In A. Sweet & C. Snow (Eds.). *Rethinking reading comprehension* (141-165). NY: Guilford Press.

Gernsbacher, M. A. (1990). *Language comprehension as structure building.* Hillsdale, NJ: Erlbaum.

Gernsbacher, M. A. (1997). Two decades of structure building. *Discourse Processes, 23*, 265–304.

Gordon, R. M. (1980). The readability of an unreadable text. *English Journal, 69*(3), 60-61.

Gorin, J. S. (2002). Cognitive and psychometric modeling of text-based reading-comprehension GRE-V items. Ph.D. dissertation, University of Kansas, United States -- Kansas. Retrieved July 6, 2011, from Dissertations & Theses: Full Text. (Publication No. AAT 3083176).

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.

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher, 40*(5), 223-234. doi:10.3102/0013189x11413260

Graesser, A. C., McNamara, D. C., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, and Computer, 36*, 193-202.

Graesser, A. C., & Person, N. K. (1994). Question asking during tutoring. *American Educational Research Journal*, 31(1), 104-37. Retrieved from EBSCO*host*.

Graesser, A.C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review, 101*, 371-395.

Gunning, R. (1968). *The technique of clear writing*. New York: McGraw-Hill.

Guthrie, J.T., & Mosenthal, P. (1987). Literacy as multidimensional: Locating information and reading comprehension. *Educational Psychologist, 22*(3/4), 279–297.

Guthrie, J. T., & Wigfield, A. (1999). *How motivation fits into the science of reading*. Special issue, *Scientific Studies of Reading, 3* No. 3.

Guthrie, J. T., Wigfield, A., & Klauda, S. L. (2012). *Adolescents' engagement in academic literacy* (Report No. 7). Retrieved November 27, 2012, from www.cori.umd.edu/.../2012_adolescents_engagement_ebook.pdf

Haberlandt, K. F., & Graesser, A. C. (1985). Component processes in text comprehension and some of their interactions. *Journal of Experimental Psychology: General, 114,* 357-374.

Haertel, E. H., & Wiley, D. E. (1993). Representations of ability structures: Implications for testing. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 359-384). Hillsdale, NJ: Erlbaum.

Haertel, G., Haydel DeBarger, A., Cheng, B., Blackorby, J., Javitz, H., Ructtinger, L., Snow, E., Mislevy, R. J., Zhang, T., Murray, E., Gravel, J., Rose, D., Mitman Colker, A., & Hansen, E. G. (2010). *Using evidence-centered design and universal design for learning to design science assessment tasks for students with disabilities* (Assessment for Students with Disabilities Technical Report 1). Menlo Park, CA: SRI International.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items (3rd ed.)*. Mahwah, NJ US: Lawrence Erlbaum Associates Publishers. Retrieved from EBSCO*host*.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17-27.

Hallgen, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology, 8*(1), 23-34.

Helwig, R., Almond, P. J., Rozek-Tedesco, M. A., Tindal, G., & Heath, B. (1999). Reading as an access to mathematics problem solving on multiple-choice tests for sixth-grade students. *Journal of Educational Research, 93*(2), 113. Retrieved from EBSCOhost.

Hansen, E. G., Mislevy, R. J., Steinberg, L. S., Lee, M. J., & Forer, D. C. (2005). Accessibility of tests within a validity framework. *System: An International Journal of Educational Technology and Applied Linguistics, 33,* 107-133.

Harp, S. F., & Mayer, R. E. (1998). How seductive details do their damage: A theory of cognitive interest in science learning. *Journal of Educational Psychology, 90*(3), 414-434. doi:10.1037/0022-0663.90.3.414

Hewitt, M.A., & Homan, S.P. (2004). Readability level of standardized test items and student performance: The forgotten validity variable. *Reading Research and Instruction, 43*(2), 1-16.

Heubert, J. & Hauser, R.M. (1999). *High stakes testing for tracking, promotion, and graduation*. Washington, D.C.: National Academy Press.

Holmes, V. M. (2009). Bottom-up processing and reading comprehension in experienced adult readers. *Journal of Research in Reading, 32*(3), 309-326. doi:10.1111/j.1467-9817.2009.01396.x.

Homan, S., & Hewitt, M. (1994). The development and validation of a formula for measuring single-sentence test item readability. *Journal of Educational Measurement*, 31(4), 349. Retrieved from EBSCO*host*.

Hyde, J. S. & Linn, M. C.(1988) Gender differences in verbal ability: a meta-analysis. *Psychological Bulletin* 104, 1, 53–69.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112,* 527–535.

Karabenick, S. A., Woolley, M. E., Friedel, J. M., Ammon, B. V., Blazevski, J., Bonney, C., De Groot, E., Gilbert, M., Musu, L. Kempler, T. M., & Kelly, K. L. (2007). Cognitive processing of self-report items in educational research: Do they think what we mean? *Educational Psychologist*, *42*(3), 139-151.

Katz, I. R., Bennet, R. E., & Berger, A. E. (2000). Effects of response format on difficulty of SAT-mathematics items: It‘s not the strategy. *Journal of Educational Measurement*, 3*7*(1), 39-57.

Kintsch, E. (2005). Comprehension theory as a guide for the design of thoughtful questions. *Topics in Language Disorders*, *25*(1), 51.

Kintsch, W. (1974). *The representation of meaning is memory*. Hillsdale, NJ: Erlbaum.

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

Kintsch, W., & Kintsch, E. (2005). Comprehension. In S. G. Paris., & S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp. 73-92). Mahwah, NJ: IEA Publishers.

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

Kirsch, I.S. & Mosenthal, P.B. (1990), Exploring document literacy: Variables underlying the performance of young adults. *Reading Research Quarterly 25*(1), 5-30.

Klare, G.R. (1984). Readability. In P.D. Pearson (Ed.), *Handbook of reading research* (pp. 681-744). New York, Longman.

Landis, J. R., & Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174.

Leighton, J.P. & Gierl, M.J. (2011). *The learning sciences in educational assessment.* Cambridge, MA: Cambridge University Press.

Leseman, P. P. M., & de, J. P. F. (1998). Home literacy: Opportunity, instruction, cooperation and social-emotional quality predicting early reading achievement. *Reading Research Quarterly, 33(*3), 294-318.

Li, M., Ruiz-Primo, M.A., & Shavelson, R.J. (2006). Towards a science achievement framework: The case of TIMSS 1999. In S. Howie & T. Plomp (Eds.), *Contexts of learning mathematics and science: Lessons learned from TIMSS.* London: Routledge.

Lietz, P. (2006). Issues in the change in gender differences in reading achievement in cross-national research studies since 1992: A meta-analytic view. *International Education Journal*, 7(2), 127-149.

Linn, R. L., & Miller, M. D. (2005). *Measurement and assessment in teaching 9th ed.* Upper Saddle River, NJ: Pearson.

Liu, O. L., Bridgeman, B., & Adler, R. M. (2012). Measuring learning outcomes in higher education: Motivation matters. *Educational Researcher, 41,* 9, 352-362.

Logan, S.,& Johnston, R. (2009). Gender differences in reading ability and attitudes: Examining where these differences lie. *Journal of Research in Reading 32* (2): 199–214.

Logan, S., & Johnston, R. (2010). Investigating gender differences in reading. *Educational Review, 62 (*2), 175-187.

Loftus, E. F., & Suppes, P. (1972). Structural variables that determine problem solving difficulty in computer-assisted instruction. *Journal of Educational Psychology, 6,* 531–542.

Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading MA: Addison-Wesley.

Lugo-Gil, J., & Tamis-LeMonda, C. S. (2008). Family resources and parenting quality: links to children's cognitive development across the first 3 Years. *Child Development*, *79(*4), 1065-1085.

Lynn, R., & Mikk, J. (2009). Sex differences in reading achievement. *TRAMES: A Journal of the Humanities & Social Sciences*, 13(1), 3-13. doi:10.3176/tr.2009.1.01.

Magno, C. (2009). Demonstrating the difference between classical test theory and item response theory using derived test data. *The International Journal of Educational and Psychological Assessment*, 1(1), 1-11.

Marshall, S. P. (1995). *Schemas for problem solving*. Cambridge: Cambridge University Press.

Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist, 34*(4), 207.

Martiniello, M. (2008). Language and the performance of English language learners in math word problems. *Harvard Educational Review*, *78*, 333-368.

Martin, M. O., Gregory, K. D. & Stemler, S. E. (2000). *TIMSS 1999 tech report.* Chestnut Hill, MA: International study center Boston College.

Martin, M.O., Mullis, I.V.S., Gonzalez, E.J., Gregory, K.D., Smith, T.A., Chrostowski, S.J., Garden, R.A., & O'Connor, K.M. (2001). *TIMSS 1999 science benchmarking report, eighth grade achievement results for U.S. states and districts in an international context.* Chestnut Hill, MA: Boston College.

Mayer, R. E. (1993). Comprehension of graphics in texts: An overview. *Learning and Instruction, 3,* 3, 239-245.

McGraw, L. L. (1992). Historical text and secondary student comprehension. Ph.D. dissertation, Stanford University, United States -- California. Retrieved September 21, 2011, from Dissertations & Theses: Full Text. (Publication No. AAT 9221643).

McGraw, K. O, & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1,* 30-46.

McNamara, D. S., & Kintsch, W. (1996). Learning from texts: Effects of prior knowledge and text coherence. *Discourse Processes*, *22*(3), 247.

McNamara, D. S., Kintsch, E., Songer, N., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition & Instruction,* 14(1)*,* 1.

McNamara, D.S., Graesser, A.C., & Louwerse, M.M. (in press). Sources of text difficulty: Across the ages and genres. In J.P. Sabatini & E. Albro (Eds.), *Assessing reading in the 21st century: Aligning and applying advances in the reading and measurement sciences.* Lanham, MD: R&L Education.

McNamara, D.S., Louwerse, M.M., Cai, Z., & Graesser, A. (2005, January 1). Coh-Metrix version 1.4. Retrieved [October, 2011], from http//:cohmetrix.memphis.edu.

McNamara, D. S., & Magliano, J. (2009). Toward a comprehensive model of comprehension. In B. H. Ross, B. H. Ross (Eds.) , *The psychology of learning and motivation (Vol 51)* (pp. 297-384). San Diego, CA US: Elsevier Academic Press. doi:10.1016/S0079-7421(09)51009-2.

McKeown, M. G., & Curtis, M. E. (1987). *The nature of vocabulary acquisition*. Hillsdale, NJ: Erlbaum.

Memory, D. M. (1982). Written questions as reading aids in the middle grades: A review of research. In J. A. Niles & L. A. Harris (Eds.). *New inquiries in reading research and instruction* (pp.71-76). (Thirty-first Yearbook of the National Reading Conference). Washington, DC: The National Reading Conference, Inc.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education/Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 23(2), 13-23.

Mikk, J. (2008). Sentence length for revealing the cognitive load reversal effect in text Ccomprehension. *Educational Studies, 34*(2), 119-127. Retrieved from EBSCOhost

Mislevy, R.J. (1994). Evidence and inference in educational assessment. *Psychometrika, 59*, 439-483.

Mislevy, R.J. (1995). Probability-based inference in cognitive diagnosis. In P. D. Ncholes, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 43-71). Hillsdale, NJ; Erlbaum.

Mislevy, R.J. (2006). Cognitive psychology and educational assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.). Westport, CT: American Council on Education/Praeger.

Mislevy, R.J. (2008). How cognitive science challenges the educational measurement tradition. *Measurement: Interdisciplinary Research and Perspectives.*

Mislevy, R. J. (2009). Validity from the perspective of model-based reasoning. In R.L. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications.* Charlotte, NC: Information Age Publishing.

Mislevy, R.J. (2012). Modeling language for assessment. In C. Chapelle (Ed.), *The encyclopedia of applied linguistics.* Hoboken, NJ: Wiley-Blackwell.

Mislevy, R. J., & Riconscente, M. (2005). *Evidence-centered assessment design: Layers, structures, and terminology*. Menlo Park, CA: SRI International.

Mislevy, R.J., Steinberg, L.S., & Almond, R.G. (1999). *On the roles of task model variables in assessment design. CSE technical report 500.* Los Angeles: The National Center for Research on Evaluation, Standards, Student Testing (CRESST), Center for Studies in Education, UCLA.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1,* 3-67. (focus article for inaugural issue)

Mislevy, R. J., Steinberg, L. S., Breyer, F. J., Almond, R. G., & Johnson, L. (1999). A cognitive task analysis with implications for designing simulation-based performance assessment. *Computers in Human Behavior, 15(*3), 335.

Mislevy, R. J. & Wu, P. (1996). *Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing*, Educational Testing Service report no. RR-96-30-ONR.

Mislevy, R.J., & Yin, C. (2009). If language is a complex adaptive system, what is language testing? *Language Learning,* 59, Supplement 1, 249-267.

Mosenthal, P. (1996). Understanding the strategies of document literacy and their conditions of use. *Journal of Educational Psychology, 88,* 314-332.

Murphy, P. K., Alexander, P. A., & Muis, K. R. (2011). Knowledge and knowing: The journey from philosophy and psychology to human learning. In K. R. Harris, S. Graham, & T. Urdan (Eds.), *Educational psychology handbook: Vol. 1. Theories,* *constructs, and critical issues*. Washington, DC: American Psychological Association.

Myers, J. L., O’Brien, E. J., Albrecht, J. E., & Mason, R. A. (1994). Maintaining global coherence during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 876–886.

Oakland, T., & Lane, H. B. (2004). Language, reading, and readability formula: Implications for developing and adapting tests. *International Journal of Testing, 4,* 239-252.

OECD (2002), *Sample tasks from the PISA 2000 assessment: Reading, mathematical and scientific literacy*, PISA, OECD Publishing. doi: 10.1787/9789264194274-en

O’Neil, Jr. H. F., Sugrue, B., & Baker, E. L. (1995). Effects of motivational interventions on the National Assessment of Educational Progress mathematics performance. *Educational Assessment, 3*, 135-157.

O'Reilly, T., & McNamara, D. (2007). The impact of science knowledge, reading skill, and reading strategy knowledge on more traditional "high-stakes" measures of high school students' science achievement. *American Educational Research Journal, 44,* 161-196.

O'Reilly, T., & McNamara, D. S. (2007b). Reversing the reverse cohesion effect: Good texts can be better for strategic, high-Knowledge readers. *Discourse Processes, 43*(2), 121-152. doi: 10.1207/s15326950dp4302_2.

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.

Ozuru, Y., Rowe, M., O'Reilly, T., & McNamara, D. (2008). Where's the difficulty in standardized reading tests: The passage or the question? *Behavior Research Methods, 40,* 1001-1015.

Ozuru, Y., Best, R., Bell, C., Witherspoon, A., & McNamara, D. S. (2007a). Influence of question format and text availability on the assessment of expository text comprehension. *Cognition & Instruction*, *25*(4), 399-438.

Pappas, C. C., & Pettegrew, B. S. (1998). The role of genre in the psycholinguistic guessing game of reading. *Language Arts*, 75(1), 36. Retrieved from EBSCO*host*.

Pearson, P. D., Garavaglia, D., Lycke, K., Roberts. E., Danridge, J., & Hamrn, D. (1999). *The impact of item format on the depth of students' cognitive engagement.* Technical Report, American Institute for Research, Washington, DC.

Pearson, P. D., Moje, E., & Greenleaf, C. (2010). Literacy and science: Each in the service of the other. *Science, 328(* 5977), 459-463.

Pearson, P. D., & Valencia, S. W. (1987). Assessment, accountability, and professional prerogative. In J. E. Readence & R. S. Baldwin (Eds.), *Research in literacy: Merging perspectives, Thirty-Sixth Yearbook of the National Reading Conference* (pp. 3-16). Rochester, NY: National Reading Conference.

Pellegrino, J. W., Chudowsky, N., Glaser, R., & National Research Council (U.S.). (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

Perfetti, C. A. (1985). *Reading ability*. New York: Oxford Press.

Perfetti, C. (2010). Decoding, vocabulary, and comprehension: The golden triangle of reading skill. In M. G. McKeown, L. Kucan, M. G. McKeown, L. Kucan (Eds.), *Bringing reading research to life* (pp. 291-303). New York, NY US: Guilford Press. Retrieved from EBSCO*host*.

Pressley, M., & Forrest-Pressley, D. (1985). Questions and children's cognitive processing. In A. C. Graesser & J. B. Black (Eds.), *The psychology of questions* (pp. 277-295). Hillsdale, NJ: Erlbaum.

RAND Reading Study Group (2002). *Reading for understanding: Toward an R&D program in reading comprehension*. Santa Monica: CA. Retrieved August 1, 2011, from www.rand.org/pubs/monograph_reports/MR1465/MR1465.pdf.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen Danish Institution for Educational Research.

Ravid, D., & Tolchinsky, L. (2002). Developing linguistic literacy: A comprehensive model. *Journal of Child Language*, 29(2), 417-447. doi:10.1017/S0305000902009169.

Reckase, M. D. (2009). *Multidimensional item response theory*. New York: Springer.

Reich, G. (2009). Testing historical knowledge: Standards, multiple choice questions and student reasoning. *Theory and Research in Social Education, 37*, 325-360.

Richardson, J. T. E. (2011). Eta squared and partial eta squared as measures of effect size in educational research. *Educational Research Review, 6,* 2, 135-147.

Rodriguez, M. (2002). Choosing an item format, in G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment program for all students: Validity, technical adequacy, and implementation issues* (pp. 211-229). Mahwah, NJ: Lawrence Erlbaum.

Rubin, D.B. (1987) *Multiple imputation for nonresponse in surveys*. J. Wiley & Sons, New York.

Rupp, A. A., Garcia, P., & Jamieson, J. (2001). Combining multiple regression and CART to understand difficulty in second language reading and listening comprehension test items. *International Journal of Testing, 1 (3 & 4),* 185-216.

Ryan, J.M., & Demark, S. (2002). Variation in achievement scores related to gender, item format, and content area tested. In G. Tindal & T.M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 77-117). Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers

Ryan, K. E., Ryan, A. M., Arbuthnot, K., & Samuels, M. (2007). Students' motivation for standardized math exams. *Educational Researcher,* 36(1), 5-13. doi:10.3102/0013189X06298001.

Rygiel, M. A. (1982). Readability formulas: Pluses and minuses. *Teaching English in the Two-Year College, 9*(1), 45-49.

Sadler, P.M. (1998) Psychometric models of student conceptions in science: Reconciling qualitative studies and distractor-driven assessment instruments. *Journal of Research in Science Teaching,* 35(3), 265–96.

Scheuneman, J., Gerritz, K., & Embretson, S. (1989, March). *Effects of prose complexity on achievement test item difficulty.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Schoultz, J., Säljö, R., & Wyndhamn, J. (2001). Conceptual knowledge in talk and text: What does it take to understand a science question? *Instructional Science, 29*(3), 213-236. Retrieved from EBSCOhost.

Schulz, W. & Sibberns, H. (2002). *IEA civic education study technical report.* Amsterdam: IEA.

Schuman, J., & Presser, S. (1981). *Questions and answers in attitude surveys: Experiments in question form, wording, and context.* New York: Academic Press.

Schwanenflugel, P. J., & Akin, C. E. (1994). Developmental trends in lexical decisions for abstract and concrete words. *Reading Research Quarterly, 29*,251-263.

Schwarz, N. (2008). Self-reports: How the questions shape the answers. In R. H. Fazio, R. E. Petty, R. H. Fazio, R. E. Petty (Eds.) , *Attitudes: Their structure, function, and consequences* (pp. 49-67). New York, NY US: Psychology Press. Retrieved from EBSCO*host*.

Schwille, J. R. (1975). Predictors of between-student differences in civic education cognitive achievement. In J. V. Torney, A. N. Oppenheim, & R. F. Farnen (Eds.), *Civic education in ten countries: An empirical study*. New York: Halsted Press.

Sénéchal, M., & LeFevre, J.-A. (2002). Parental Involvement in the Development of Children's Reading Skill: A Five-Year Longitudinal Study. *Child Development, 73(* 2), 445-460.

Seo, M. H., Xu, X, & von Davier, M. (2009). *Mdltm software user manual*. Unpublished manuscript, Educational Testing Service.

Shaftel, J., Belton-Kocher, E., Glasnapp, D., & Poggio, J. (2006). The impact of language characteristics in mathematics test items on the performance of English language learners and students with disabilities. *Educational Assessment, 11*(2), 105-126. doi:10.1207/s15326977ea1102_2.

Shah, P., Mayer, R. E., & Hegarty, M. (1999).Graphs as aids to knowledge construction: Signaling techniques for guiding the process of graph construction. *Journal of Educational Psychology, 91*,690–702.

Shavelson, R.J. & Ruiz-Primo, M.A. (1999). On the assessment of science achievement. (English version) *Unterrichts wissenschaft*, 27(2), 102–27.

Sheehan, K. M. (1997). A tree-based approach to proficiency scaling and diagnostic assessment. *Journal of Educational Measurement*, *34*(4), 333-52. Retrieved from EBSCO*host*.

Sheehan, K. M., & Ginther, A. (2001). *What do passage-based multiple-choice verbal reasoning items really measure? An analysis of the cognitive skills underlying performance on the current TOEFL reading section*. Paper presented at the 2000 Annual Meeting of the National Council of Measurement in Education, New Orleans, LA.

Shrout, P.E. & Fleiss, J.L. (1979) Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 2,* 420-428.

Siegel, S., & Castellan, N.J. (1988). Nonparametric statistics for the behavioral sciences (2nd ed.) New York: McGraw-Hill.

Sigurd, B., Eeg-Olofsson, M., & van de Weijer, J. (2004). Word length, sentence length and frequency – Zipf revisited. *Studia Linguistica, 58*(1), 37-52. doi:10.1111/j.0039-3193.2004.00109.x.

Simmons, D., Hairrell, A., Edmonds, M., Vaughn, S., Larsen, R., Willson, V., & Byrns, G. (2010). A comparison of multiple-strategy methods: Effects on fourth-grade students' general and content-specific reading comprehension and vocabulary development. *Journal of Research on Educational Effectiveness, 3*(2), 121-156.

Snow, C. E. (2003). Assessment of reading comprehension. In A. P. Sweet & C. E. Snow (Eds.), *Rethinking reading comprehension (*pp. 192-218). New York: Guilford.

Snow, R.E., Corno, L., & Jackson, D., III (1996). Individual differences in affective and conative functions. In D.C. Berliner & R.C. Calfee (Eds.), *Handbook of educational psychology* (pp. 243-310). New York: Macmillan.

Snow, C. E. (2010). Academic language and the challenge of reading for learning about science. *Science, 328,* 450–452.

Solano-Flores, G., & Trumbull, E. (2003). Examining language in context: The need for new research and practice paradigms in the testing of English-language learners. *Educational Researcher, 32*(2), 3-13.

Spache, G. D. (1953). A new readability formula for primary-grade reading materials. *Elementary School Journal, 53*, 410-413.

Spearman, C. (1904). "General intelligence" objectively determined and measured. *American Journal of Psychology, 15,* 201-293.

Stahl, S.A. & Fairbanks, M.M. (1986). The effects of vocabulary instruction: A model-based meta-analysis. *Review of Educational Research, 56*(1), 72-110.

Stahl, S. A., & Hiebert, E. H. (2006). The "word factors". In K. A. Stahl & M. McKenna (Eds*.), Reading research at work* (pp. 403-424). New York: Guilford.

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly, 16*, 32–71.

Sternberg, R. J. (1998). Abilities are forms of developing expertise. *Educational Researcher,* 27(*3),* 11-20.

Stodolsky, S. (1998). *The subject matters: Classroom activity in math and social studies.* Troy, NY: Educator's International Press.

Storch, S. A., & Whitehurst, G. J. (2002). Oral language and code-related precursors to reading: Evidence from a longitudinal structural model. *Developmental Psychology*, *38*(6), 1.

Tatsuoka, K. K. (2009). *Cognitive assessment: An introduction to rule space method*. New York: Routledge.

Templeton, S., Cain, C. T., & Miller, J. O. (1981). Reconceptualizing readability: The relationship between surface and underlying structure analyses in predicting the difficulty of basal reader stories. *Journal of Educational Research, 74*, 382-387.

Thompson, R. A., & Zamboanga, B. L. (2004). Academic aptitude and prior knowledge as predictors of student achievement in Introduction to Psychology. *Journal of Educational Psychology*, 96, 778-784.

Thorndike, R.L. (1973). *Reading comprehension education in fifteen countries III: An empirical study*. New York: Wiley.

Tillers, P., & Schum, D.A. (1991). A theory of preliminary fact investigation. *U.C. Davis Law Review, 24,* 907-966.

Torney-Purta, J. (1990). International comparative research in education: Its role in educational improvement in the U.S. *Educational Researcher, 19,* 32-35.

Torney-Purta, J. (2009). International psychological research that matters for policy and practice. *American Psychologist, 64(8),* 825-837.

Torney-Purta, J. & Amadeo, J. (2012). The contribution of international large-scale studies in civic education and engagement. In M. von Davier, E. Gonzalez, I. Kirsch, & K. Yamamoto (Eds.), *The role of international large scale assessments.* New York: Springer (pp. 87-114).

Torney-Purta, J., Lehmann, R., Oswald, H., & Schulz, W. (2001). *Citizenship and education in twenty-eight countries: Civic knowledge and engagement at age fourteen.* Amsterdam, Netherlands: IEA.

Torney, J, Oppenheim, A. N., & Farnen, R. (1975). *Civic education in ten countries: An empirical study*. New York: Halsted Press.

Tourangeau, R. (2003). Cognitive Aspects of Survey Measurement and Mismeasurement. *International Journal of Public Opinion Research, 15,* 1, 3-7.

Tourangeau, R., & Rasinski, K. A. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin, 103*(3), 299-314. doi:10.1037/0033-2909.103.3.299.

Trabasso, T., van den Broek, P., & Suh, S. (1989). Logical necessity and transitivity of causal relations in stories. *Discourse Processes, 12,* 1–25.

van den Broek, P., Lorch, R. F., Linderholm, T., & Gustafson, M. (2001). The effect of readers' goals on inference generation and memory for texts. *Memory and Cognition, 29,* 1081-1087.

van den Broek, P., Young, M., Tzeng, Y., & Linderholm, T. (1999). The landscape model of reading. In H. van Oostendorp & S. R. Goldman (Eds.), *The construction of mental representations during reading* (pp. 71–98). Mahwah, NJ: Erlbaum.

Verhoeven, L., & Perfetti, C. (2008). Advances in text comprehension: Model, process and development. *Applied Cognitive Psychology, 22*, 293-301.

Vilenius-Tuohimaa, P. M., Aunola, K., & Nurmi, J.-E. (2008). The association between mathematical word problems and reading comprehension. *Educational Psychology*, *28*(4), 409-426. doi: 10.1080/01443410701708228.

von Davier, M. (2005). *A general diagnostic model applied to language testing data* (ETS Research Rep. No. RR-05-16). Princeton, NJ: Educational Testing Service.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology, 61*(2), 287-307. doi:10.1348/000711007X193957

Webb, N. (2006). Identifying content for student achievement tests. In S. Downing & T. Haladyna (Eds.), *Handbook of test development* (pp. 155-180). Mahwah, NJ: Erlbaum.

Wheeler, G., & Sherman, T. F. (1983). Readability formulas revisited. *Science and Children, 20*(7), 38-40.

Wiley, J. & Voss, J. F. (1999). Constructing arguments from multiple source: Tasks that promote understanding and not just memory for text. *Journal of Educational Psychology, 91*(2), 301-311.

Wilson, P. T., & Anderson, R. C. (1986). What they don't know will hurt them: The role of prior knowledge in comprehension. In J. Orasanu (Ed.), *Reading comprehension: From research to practice* (pp. 31-48). Hillsdale, NJ: Erlbaum.

Wilson, V. L., & Rupley, W. H. (1997). A structural equation model for reading comprehension based on background, phonemic, and strategy knowledge. *Scientific Studies of Reading*, *1*(1), 45.

Willoughby, T., Wood, E., & Khan, M. (1994). Isolating variables that impact or detract from the effectiveness of elaboration strategies. *Journal of Educational Psychology, 86*, 279–289.

Willoughby, T., Waller, T.G., Wood, E., & MacKinnon, G.E. (1993). The effect of prior knowledge on an immediate and delayed associative learning task following elaborative interrogation. *Contemporary Educational Psychology, 18*, 36–46.

Wise, S. L., & DeMars, C. D. (2005). Low examinee effort in lowstakes assessment: Problems and potential solutions. *Educational Assessment, 10,* 117.

Wolfe, M. W., & Woodwyk, J. M. (2010). Processing and memory of information presented in narrative or expository texts. *British Journal of Educational Psychology*, *80*(3), 341-362.

Xu, X., & von Davier, M. (2006). *General diagnosis for NAEP proficiency data* (ETS Research Rep. No. RR-06-08). Princeton, NJ: Educational Testing Service.

Xu, X., & von Davier, M. (2008a). *Fitting the structured general diagnostic model to NAEP data* (Research Report No. 08-27). Princeton, NJ: Educational Testing Service.

Young, J.P., & Brozo, W.G. (2001). Boys will be boys, or will they? Literacy and masculinities. *Reading Research Quarterly, 36*, 316-325.

Zhang, T., Torney-Purta, J. V., & Barber, C. (2012). Students' conceptual knowledge and process skills in civic education: Identification of profiles and classroom correlates. *Theory and Research in Social Education, 40*, 1-34. doi:10.1080/00933104.2012.649467

Zhang, T., Mislevy, R., Haertel, G., Javitz, H., Murray, E., & Gravel, J. (2010). *A design pattern for a spelling assessment for students with disabilities* (Assessment for students with disabilities technical report 2). Menlo Park, CA: SRI International.

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