APPLICATION OF THE IRT AND TRT MODELS TO A READING COMPREHENSION TEST

by

Weon H. Kim

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

Middle Tennessee State University
August 2017

Dissertation Committee:

Dr. Amy M. Elleman, Chair

Dr. Ying Jin

Dr. Mohammed Albakry

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Amy Elleman, for her persistent help and advice throughout the dissertation process. She advised me in many ways regardless of her extremely busy schedule. I would also like to thank my committee members, Dr. Ying Jin and Dr. Mohammed Albakry, for their suggestions and comments. With the dedicated guidance and help from the all committee members, I could complete my dissertation well. Finally, I would like to thank my husband, Jwa Kim, and my three children. They always supported and encouraged me with their deep love and sacrifice. Without my husband's great sacrifices, I would not have been able to finish this long journey.

ABSTRACT

The purpose of the present study is to apply the item response theory (IRT) and testlet response theory (TRT) models to a reading comprehension test. This study applied the TRT models and the traditional IRT model to a seventh-grade reading comprehension test (n = 8,815) with eight testlets. These three models were compared to determine the best model for a testlet-based reading comprehension assessment. The goodness-of-fit indices such as -2 log likelihood, Akaike information criterion, and Bayesian information criterion were utilized as model comparison indices. The standardized local dependence X^2 statistic was computed for a comparison of local dependence among the three different models. Scatter plots were obtained to evaluate parameter-estimation consistency among models. Correlations and mean differences between the estimated parameters were also examined to detect and quantify the magnitude of inaccuracy due to the use of a worse-fitting model. Finally, items were evaluated based on the item parameters from the TRT models and compared to the results from the Coh Metrix.

According to the three goodness-of-fit indices, the bi-factor model was the best-fitting model and the testlet-effects model was the second best-fitting model among the three models, showing a superiority of the TRT models for the testlet-based assessments compared to the more traditional IRT model. The standardized local dependence X^2 statistic revealed significant local dependencies when the traditional IRT model was applied to the data, while the application of the TRT models resolved the problematic local dependencies. Items requiring the comprehension ability of context-related vocabulary showed high local dependencies when the 3-plm was applied.

Comparisons of the item and person parameters and their standard error estimates showed a slight underestimation and less stable standard errors of the item discrimination parameters under the locally dependent condition. The results of the item parameters from the TRT models supported the advantages of the TRT models in selecting good items for accurate assessments. This study provided replicable evidence that the TRT models are crucial for testlet-based reading comprehension assessments. For future study, various comprehension tests containing more questions for each testlet as well as smaller sample sizes might be useful.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE: INTRODUCTION

According to Shaywitz (2003), reading is not a natural process regardless of individual differences, while the capacity for language is innate for human beings. Reading is an extremely complex process demanding interactions of visual, auditory, linguistic, cognitive, and reasoning skills (Bell & McCullum, 2008). The fundamental goal of reading is comprehension, which is a non-unitary construct that consists of multiple components and skills including various cognitive processes (Keenan, Betjeman, & Olson, 2008; Kintsch & Kintsch, 2005).

Reading Comprehension

Reading Comprehension Theory

Reading comprehension comes from two separate but highly correlated processes: word recognition and comprehension (Perfetti, Landi, & Oakhill, 2005). Perfetti and Stafura (2014) further noted that readers need to combine two processes for full reading comprehension; one process involves constructing a text-based model that represents the explicit meaning of the text (i.e., bottom-up process based on words). The other process requires building a situation model (Kintsch, 1988) that represents a broader meaning implied by the text (i.e., top-down process based on knowledge).

Comparatively, the situation model is a higher-level process for deeper comprehension that requires the integration of text information with related prior knowledge and the reader's ability to make inferences (Kintsch & Rawson, 2005). Background knowledge and inference making skills are crucial for situation model development (Compton, Miller, Elleman, & Steacy, 2014). Specifically, background

knowledge is an important component for deeper comprehension (Cain & Oakhill, 2007; Edmonds et al., 2009). Stahl and Nagy (2006) affirmed that rich knowledge of concepts including background knowledge drives comprehension.

Based on Perfetti's verbal efficiency theory, when lower-level processes that support lexical access are efficiently executed, the reader has cognitive resources available for higher-level processes such as comprehension (Perfetti et al., 2005). These two associated processes support the acquisition of reading comprehension reciprocally (Kendeou, Papadopoulos, & Spanoudis, 2012; Perfetti et al.). In general, word recognition is the major instructional emphasis in the early elementary school years and influences a higher-level process of deeper comprehension for older adolescent readers (Graesser, McNamara, & Kulikowich, 2011).

Adolescent Reading Comprehension

According to reports of the National Assessment of Educational Progress (NAEP), 31% of 4th grade students and 24% of 8th grade students read below the basic level, indicating that adolescent readers do not have sufficient reading comprehension abilities at their grade level (National Center for Education Statistics, 2015). Reading comprehension is crucial for achievement in various subjects including mathematics and science. Adolescents with reading difficulties first need to improve their reading comprehension abilities in order to understand content area instruction. The reality of student reading difficulties calls for higher-level instruction of reading comprehension for adolescent readers such as comprehension strategies (Duke & Carlisle, 2011)

Moreover, text levels of middle and high schools have decreased over the past 50 years. The decreased text levels have resulted in the problem that many current high

school graduates are not prepared to read the texts in college or in the workplace (Hiebert & Mesmer, 2013). It is critical to provide adolescent students with higher-level instruction of reading comprehension for the workplace and postsecondary educational settings.

Common Core State Standards (CCSS)

To provide identical standards across all states for students' college and career readiness, the Common Core State Standards (CCSS, 2014) were established as academic benchmarks. The CCSS from Kindergarten through 12th grade encompass grade-level expectations of complex skills and knowledge in English language arts (ELA) as well as other subjects such as history, social sciences, and science. Each grade level has standards that consist of four separate specific areas such as literature, information, language, and writing.

The CCSS for ELA focus on text complexity and recommend accelerated text levels beginning with the 2nd and 3rd grades. It was assumed that increasing the complexity of texts from the primary grades onward could close the gap between the levels of texts in high school and college (Hiebert & Mesmer, 2013). Those accelerated text levels are aligned to college and career readiness expectations (Hiebert & Mesmer).

Text Complexity

According to the CCSS for ELA (CCSS, 2014), text complexity is a key component for reading comprehension. It is crucial to analyze text complexity for the purpose of selecting the appropriate text levels for students (Graesser et al., 2011). The traditional approach to text analysis uses a unidimensional metric and provides a simple solution of grade-level text difficulty. The traditional approach includes the Flesch-

Kincaid Grade Level (Klare, 1974), Degrees of Reading Power (DRP; Koslin, Zeno, & Koslin, 1987), and Lexile scores (Stenner, 2006). Those three unidimensional metrics of text difficulty offer an indication of the overall complexity of texts through two text measures (i.e., a word factor or vocabulary index and a syntax factor). In the Lexile framework, the vocabulary index is the average frequency of words relative to words in a digital databank. At the syntactic level, Lexiles use average sentence length which is an index of a text's syntactic complexity (Hiebert & Mesmer, 2013).

The problem is that reading comprehension is complex and multi-faceted. Not all researchers agree that vocabulary rarity and sentence length capture the complexity of reading comprehension (McNamara, Graesser, McCarthy, & Cai, 2014). Therefore, a new approach considering various facets of reading comprehension may be needed for text analysis (McNamara et al., 2014). Coh-Metrix (McNamara, Louwerse, Cai, & Graesser, 2005) provides multilevel analyses of text complexity based on text cohesion (Graesser et al., 2011). According to McNamara et al. (2014), text cohesion denotes "the connectedness of concepts presented in a text" (p. 11). Cohesion gaps require the reader to make inferences, which can be challenging and even unsuccessful without sufficient prior knowledge.

Coh-Metrix focuses on text characteristics which are related with a higher-level of deeper comprehension. Those text characteristics called "cohesive cues" include overlapping key words across sentences, connectives (e.g., and, but), narrativity, syntactic simplicity, word concreteness, referential cohesion, and deep cohesion (McNamara et al., 2014, p. 20). In Coh-Metrix, a statistical technique called Latent Semantic Analysis (LSA; Landauer & Dumais, 1997) was adopted along with various

lexicons in order to compute text cohesion as well as world knowledge (McNamara et al., 2014).

Based on Coh-Metrix using LSA, narrativity is associated with word familiarity, world knowledge, and oral language. A text which is high in narrativity indicates that it is more story-like and may have more familiar words; more story-like texts are usually easier to understand. High syntactic simplicity means that the text has simple sentence structures. If a text has high word concreteness, it indicates that there are many words which are easier to visualize and comprehend. High referential cohesion may help the reader comprehend a text by providing a significant overlap in words and ideas between sentences. A text which is high in deep cohesion contains relatively more connecting words to help clarify the relationship between events, ideas, and information (McNamara et al., 2014).

Assessment in Education

It is through assessment that teachers are able to plan and implement their instruction based on students' diverse knowledge. Accurate assessments related to the school curriculum are crucial for teachers to know their students' needs. The improvement of instruction relies on information about instructional effectiveness, which results from valid, reliable, and sensitive assessments (Sweet, 2005). Thus, assessment and instruction are inseparable. Shepard (2000) asserted that the nature of the assessment is significantly influential to the nature of instruction in a real classroom setting. An adequate assessment system is necessary for the progress of all aspects of reading research (Sweet). Bell and McCullum (2008) defined assessment as the process of collecting data for progress monitoring and educational decision making. It is important

that the primary goal of any educational assessment should be to enhance student performance (Fletcher et al., 2002).

As discussed thus far, assessment is essential to education in that it helps teachers with instructional planning, monitors students' progress, determines educators' accountability, guides decisions for special education services, and provides feedback to parents and students (Bell & McCullum, 2008; Campbell, 2005). In particular, accurate assessment is imperative to decisions for special education services because inaccurate assessment may result in misidentification of children with learning disabilities or reading difficulties. Nation and Snowling (1997) emphasized the importance of accurate assessments of reading difficulty in order to implement the most appropriate type of intervention for struggling readers. In short, instructionally sensitive and accurate assessments are needed to evaluate the effects of different types of instruction and interventions, which provide teachers with information and direction for their students' learning needs (RAND Reading Study Group, 2002).

Reading Comprehension Assessment

As Kendeou et al. (2012) asserted, designing reading comprehension assessments is challenging due to the complex nature of reading comprehension. Hannon and Daneman (2001) discovered that even though there were plenty of standardized reading tests, the majority of them were not based on a theoretical understanding of reading processes. They claimed that it is necessary to take a measurement that directly addresses "theoretically important component processes and resources of reading" (p. 104). Compton et al. (2014) noted that reading theory is important for the implementation of reading interventions as well. Pearson and Hamm (2005) pointed out

that most reading comprehension tests were developed long before the theoretical framework of comprehension processes was established. Kintsch and Kintsch (2005) also put emphasis on the importance of theory in reading comprehension assessment, claiming that the current reading comprehension tests do not reflect reading processes consistent with the reading theory.

The problem is that there has not been one unified theory because reading comprehension theories have grown and evolved continuously based on research findings (Duke, 2005) and there is consensus among comprehension researchers that comprehension is not a unitary construct (Catts & Kamhi, 2017). Pearson and Hamm (2005) demonstrated the impact of theoretical perspectives on reading comprehension assessments through historical analyses. They contend that reading comprehension assessments have been developed over time on the basis of specific theoretical perspectives such as behaviorism, cognitive psychology (i.e., sociocultural and literary perspectives), schema theory, and constructivism.

In addition to the reading theory, psychometric qualities such as reliability and validity are critical to reading assessments. Duke (2005) stated that both theory and "psychometric soundness" are key to reading assessments (p. 95). Being consistent with a theory indicates psychometric soundness, which is related to the concept of validity (Duke). The RAND Reading Study Group (2002) also emphasized the importance of psychometric criteria including reliability and validity for reading assessments.

Reliability and Validity

Reliability and validity are fundamental necessities for psychometric measurement such as reading comprehension. Test results should be consistent under

similar conditions; this is the issue of reliability (McKenna & Stahl, 2008). In addition, the basic assumption for assessment is that any given measurement tool measures what it is supposed to measure (Thompson, 2004); this is the issue of validity. The value of validity cannot exceed the square root of the product of two reliability indices (Allen & Yen, 2002). Reliability is a necessary but not sufficient condition for validity. In other words, high reliability does not guarantee high validity.

Several types of reliability indices include alternate form, internal consistency using Cronbach's alpha, and test-retest. There are also several types of validity indices: content validity, construct validity, predictive validity, concurrent validity, and consequential validity. Construct validity is the degree to which a test measures the intended construct. Some tests are specifically designed to predict future performance or success, and therefore possess good predictive validity (e.g., the SAT and ACT). If a new measure and an established measure are administered at about the same time to the same students and the two scores turn out to be highly correlated, the result could be regarded as evidence of concurrent validity for the new test (Bell & McCullum, 2008).

It is critical that a test measures the intended construct. In general, it is assumed that many tests of reading comprehension available on the market are measuring the same construct—reading comprehension (Keenan et al., 2008). Therefore, they are supposed to be interchangeable, which means that an examinee's ability of reading comprehension should be consistent when different reading comprehension tests are used. Research in this area has been challenging. Many researchers have found it difficult to establish construct validity among well-known reading comprehension tests (e.g., Gray Oral Reading Test-3 [GORT-3]; Woodcock-Johnson Passage Comprehension-3 [WJPC-3];

Keenan et al., 2008; Keenan & Meenan, 2014; Nation & Snowling, 1997). Many of the comprehension measures studied have been shown not to be interchangeable with one another.

Literature Review on Reading Comprehension Assessment

Nation and Snowling (1997) raised the question about what the reading tests measure. In one study, they revealed the loadings of the individual tests on the two factors (i.e., decoding and comprehension skills) through a principal components analysis of factor analysis. The pattern of loadings across the two factors indicated that different reading tests measure different aspects of the reading process, showing the possible lack of validity. For example, the Suffolk Reading Scale (Hagley, 1987), a test of reading ability, loaded far more highly on the decoding factor than on the comprehension factor, while the Neale Reading Comprehension (Neale, 1989) loaded similarly both on the decoding factor and the comprehension factor. However, the internal reliability (i.e., Cronbach's alpha) of all of the tests they used exceeded .90 (i.e., high reliability), indicating that high reliability does not guarantee high validity.

In a second study, Nation and Snowling (1997) compared the performance of skilled and less skilled comprehenders on the same tests. The poor comprehenders had the greatest difficulty with tests which were heavily dependent on linguistic comprehension and had the least difficulty on decoding measures. According to Nation and Snowling, even though the reading tests had high reliability, they were not measuring the same component skills of comprehension.

In a study of the underlying factors important for performance on four reading comprehension tests (i.e., GORT-3; Peabody Individual Achievement Test [PIAT];

Qualitative Reading Inventory-3 [QRI-3]; WJPC), Keenan, Betjeman, and Olson's analyses (2008) also revealed two factors of reading comprehension (i.e., decoding and comprehension) using an exploratory principal components factor analysis. The PIAT and the WJPC loaded significantly more highly on decoding than on the comprehension factor. The result of hierarchical regressions revealed that the PIAT and the WJPC were more sensitive to individual differences in decoding skills than were the GORT-3 and the QRI-3.

Additionally, Keenan et al. (2008) assessed developmental differences both as a function of age and reading ability based on the regression analyses. As other research has shown (Catts, Hogan, & Adolf, 2005), Keenan et al. found that decoding skill accounts for more variance when children are younger or have low reading ability. The more important finding was that those developmental differences in variance accounted for by decoding were dramatically larger on the PIAT and WJPC than on the GORT-3 and QRI-3; that is, there were significant developmental differences across tests. Keenan et al. concluded that the reading comprehension tests examined in this study were not comparable and thus violated the assumption of validity.

Rimrodt, Lightman, Roberts, Denckla, and Cutting (2005) also demonstrated the inconsistencies across three tests (i.e., Gates-MacGinitie Reading Test [MacGinitie, MacGinitie, Maria, Dreyer, & Hughes, 2000]; GORT [Wiederholt & Bryant, 1992]; Wechsler Individual Achievement Test reading comprehension subtest [Wechsler, 1992]) they used to identify children with a reading comprehension deficit (CD). In this study, only 9.4% of the participants were identified as having a CD using all three tests, while 43.5% of the participants were identified as having a CD when using at least one of the

three tests. This study also revealed that different reading comprehension tests are not interchangeable.

In regard to CD diagnoses, Keenan and Meenan (2014) examined how test differences affect diagnoses of a CD. They explored whether a child diagnosed as having a CD with one test might not be diagnosed as having a deficit if a different test is used. They found that only 20 children among 100 (all 100 children were approximately in the 10^{th} percentile) were found to be consistently identified by all four tests. The consistency between tests in diagnosis, when measured as percentage overlap, ranged from a low of 35% between the PIAT and GORT-3 to a high of 56% between the PIAT and the WJPC-3. The average across all pairwise test comparisons was only 43%, indicating that the odds are less than half that a child diagnosed with a CD with one test would get that same diagnosis if a different test had been used.

Keenan and Meenan (2014) also found that there were inconsistencies across tests in identification of the top performers (those in the 90^{th} percentile and above). Their results showed that there was even less consistency across tests in identification of the top performers than in identification of the poorest performers, and the age group analyses revealed less consistency between tests for older children. They explained the reasons of the inconsistencies in diagnosis; the first reason for the inconsistencies was the reliabilities of the tests. Particularly, the GORT-3 showed the lowest reliability. Another reason was that the reading tests they used were measuring different component skills of comprehension, showing the possible lack of validity. In sum, reading comprehension tests used in this study were not reliable and possibly not valid.

As a study of construct validity using widely used reading tests, Cutting and Scarborough (2006) examined reading comprehension scores from three reading tests (Gates-MacGinitie Reading Test-Revised [G-M; MacGinitie et al., 2000], GORT-3 [Wiederholt & Bryant, 1992], WIAT [Wechsler, 1992]). Results of the study showed that the unique contributions of word-recognition skills varied across comprehension measures; nearly twice as much variance was accounted for in the WIAT scores than in the G-M and GORT scores. They concluded that "comprehension tests vary in their task demands and conceptual underpinnings," and the contributions of each factor may not be the same across tests (p. 281). The internal reliabilities of the G-M, GORT-3, and WIAT ranged from .87 to .93, indicating high reliability. Although the tests used in the study showed adequate reliability, their validity in measuring reading comprehension is questionable, indicating that high reliability does not guarantee high validity.

Betjemann, Keenan, Olson, and DeFries (2011) investigated whether a specific reading comprehension test results in different outcomes for behavior genetic analyses, measuring reading comprehension-decoding (RC-D; reading comprehension that loaded most highly on decoding), reading comprehension-listening comprehension (RC-LC; reading comprehension that loaded more strongly on listening comprehension), listening comprehension, and word reading. The standardized path coefficient between listening comprehension and RC-D was much smaller than that from listening comprehension to RC-LC. The path from the word reading factor was much larger to RC-D than to RC-LC.

According to their findings (Betjemann et al., 2011), the gene covariation between word decoding and comprehension depends on which test was used to assess reading comprehension; that is, the choice of test influences the outcomes of genetic analyses.

Their conclusion is that it is critical to choose the appropriate reading comprehension test that will be able to identify not only children who suffer from word decoding difficulties, but also those who suffer from comprehension deficits despite adequate word reading. Based on the research results of Betjemann et al., not all reading comprehension tests are the same, indicating a possible lack of validity.

From the same angle of construct validity, Bowyer-Crane and Snowling (2005) examined the relative performance of skilled and less-skilled comprehenders on questions of different types of inference using the Neale Analysis of Reading Ability (NARA II; Neale, 1989) and the Wechsler Objective Reading Dimensions Test of Reading Comprehension (WORD; Wechsler, 1990). They reported that children having comprehension difficulties identified by the NARA II performed normally on the WORD comprehension subtests, which means that the two different reading tests target different types of inferencing skills. This result indicates that those two tests are not interchangeable. The NARA II relied on generating knowledge-based inferences, while the WORD comprehension subtest was reliant on the retention of literal information. Their results also showed that less-skilled comprehenders have difficulties making knowledge-based inferences when reading.

Conclusively, most of the reading comprehension tests vary from test to test. Many reading comprehension tests are not interchangeable because of the differences between the tests in the underlying comprehension skills that they assess. They are reliable but may not be valid. The variations of test scores from different tests may show the violation of the general assumption of reading comprehension tests—that all tests

measure the same types of reading skills for the estimate of reading comprehension ability.

Purpose of the Study

The first thing to do to rectify this validity problem may be to construct tests on the basis of sound reading comprehension theories (Hannon & Daneman, 2001). However, it would be difficult to find the perfect theory-based assessment because the reading theory is still developing and changing (Pearson & Hamm, 2005). A second approach could be to test validity utilizing factor analysis for construct validity and correlations for predictive and concurrent validity (Allen & Yen, 2002; Bell & McCullum, 2008).

A third way to solve validity problems may be to select the appropriate test that measures the same components of comprehension ability. In particular, it may be crucial for a diagnostic test or for a research-project test to select reading comprehension tests from the market because it is directly linked to the appropriate type of intervention to be chosen (Keenan & Meenan, 2014). Caution should be exercised when selecting tests for the assessment of reading difficulties (Nation & Snowling, 1997). As Cain and Oakhill (2006) stated, no assessment tool is perfect. They contend that awareness of the strengths and weaknesses of each test will lead to the selection of the most appropriate assessment and right interpretation of test scores.

Finally, item response theory (IRT; Lord, 1952) models could be an alternative to the traditional classical test theory (CTT; Spearman, 1904) to address the issue of validity. The CTT has been the most popular test theory for a long time in psychometrics. When CTT is applied to a given test, validity and reliability are essential requirements because

an indirect measurement such as reading comprehension assessment is complex and difficult to assess (Allen & Yen, 2002; Bell & McCullum, 2008). On the contrary, IRT does not necessarily require validity and reliability because in IRT, the examinee's true ability is estimated using the likelihood function of the item response pattern based on a statistical model (Hambleton, Swaminathan, & Rogers, 1991; Kim & Nicewander, 1993).

Even if some test scores have high reliability indices when using CTT, it is still possible that item indices may show weak item characteristics. In other words, because CTT is a test- and sample-dependent theory, item indices such as item total correlation (item discrimination index) or *p*-value (item difficulty index) are variant depending on the individual tests, which may lead to variant results from test to test (Hambleton et al., 1991).

The IRT focuses on the responses of examinees on each item based on the probability of answering each item correctly. It is a set of statistical models which can be tested with empirical data. The IRT resolves the shortcomings of CTT and offers the possibility of computing invariant examinees' abilities and item indices (Hambleton & van der Linden, 1982; Kim & Nicewander, 1993). In brief, IRT is an item-oriented theory providing invariant item indices and examinees' abilities. This property of invariance indicates that IRT is not a test- and sample-dependent, which may contribute to more accurate assessments than CTT, regardless of differences in tests for psychometric measurement such as reading comprehension (Hambleton et al., 1991).

Specifically, the testlet response theory (TRT) model could be another alternative to the traditional CTT for reading comprehension tests with several testlets (Wainer, Bradlow, & Wang, 2007; Wainer & Wang, 2000). In reading comprehension measures, a

testlet can be defined as a cluster of items around a single passage when the measure is composed of multiple passage/question combinations (DeMars, 2012). Within a testlet, items are most likely correlated due to the fact that items are derived from the same content, which can give rise to the problem of local item dependence (Wainer et al.). The TRT model is a specific class of the IRT models designed to model local item dependence in a testlet-based test (Baldwin, 2007).

It is clear that more recent approaches to reading comprehension assessment such as IRT and TRT may be needed instead of the traditional CTT. The purpose of the present study is to apply the IRT and TRT models to a reading comprehension test for more accurate reading comprehension assessments. This study compares three models of IRT and TRT to determine the best model for a testlet-based reading comprehension test and examines the advantages of TRT over IRT for reading comprehension assessments.

CHAPTER TWO: REVIEW OF LITERATURE

The complex characteristics of reading comprehension assessment are directly related to the area of psychometrics concerning theories and practice of mental process measurement. In psychometrics, there are two major theories. They are CTT and IRT.

Psychometric Theories

Classical Test Theory (CTT)

The CTT has been the most popular test theory until the turn of the 21st century (Embretson & Reise, 2000). When CTT is applied to a given test, validity and reliability are essential requirements due to the challenging nature of measurement of mental abilities such as reading comprehension (Allen & Yen, 2002; Bell & McCullum, 2008). Theoretically, CTT is a tautology which cannot be proved or disproved. In the practical applications of CTT, the true test scores which express examinees' abilities are test-dependent. When the test is difficult, the examinee's true score will be low; when the test is easy, the examinee will appear to have a higher true score (Hambleton, Swaminathan, & Rogers, 1991). Assessments using CTT may bring about problems in comparing examinees who take different tests because the examinee's true score depends on the difficulty level of a given test.

In CTT, the item and test indices are sample-dependent. For example, the item difficulty index of CTT (i.e., *p*-value) is the proportion of examinees who answer the item correctly. Whether an item is difficult or easy depends on the examinees' abilities. The same test will be easy for good students but difficult for weak students, which may lead to the inappropriateness of comparing item difficulty indices. In the end, when CTT

is applied for assessments, it may be problematic to compare examinees' test scores from different tests and item indices from different groups of examinees because of their variant characteristics (Hambleton et al., 1991).

In CTT, two item indices are well-known and often-reported: item difficulty index and item-test correlation. Item-test correlation represents the item point-biserial correlation with the total score. It is a measure of differentiating strength of the item and considered as the best item index in CTT. Test indices in CTT include reliability and validity estimates (Allen & Yen, 2002; Bell & McCullum, 2008; Duke, 2005). As mentioned above, these item and test indices are sample-dependent, resulting in the incomparability of the tests.

The CTT model assumes the existence of the parallel test, which is a difficult requirement to meet in practice. Two tests are called parallel tests if $T_1 = T_2$ and $\sigma^2_{E1} = \sigma^2_{E2}$, where T_1 = the true score for test 1, T_2 = the true score for test 2, E_1 = error of measurement for test 1, E_2 = error of measurement for test 2, σ^2_{E1} = variance of E_1, and σ^2_{E2} = variance of E_2 (Hambleton & van der Linden, 1982).

The aforementioned testing problems spawned the need for alternative theories and models to measure reading comprehension ability. The RAND report (2002) recommended psychometric approaches that provide more accurate and invariant estimations. According to Hambleton et al. (1991), IRT offers a powerful method for comparing tests and examinees' abilities because it computes invariant item and person parameters.

Item Response Theory (IRT)

The IRT has been used with increasing frequency in recent decades as an alternative to CTT, because the IRT is a superior tool for mental measurements including reading comprehension. It focuses on the responses of examinees on each item while CTT focuses on the total test score (Hambleton et al., 1991). The IRT provides statistical models which can be proved or disproved based on empirical data (Hambleton & van der Linden, 1982). One benefit of IRT is that it has statistical models that characterize the examinee's response to individual items (Hambleton, Swaminathan, & Rogers, 1991; Wang, Bradlow, & Wainer, 2002).

The IRT has two major assumptions—unidimensionality and conditional independence (or local independence). Unidimensionality indicates that items in a test measure a single latent trait. Local independence assumes that given a fixed ability level on the part of the examinee, the probability of answering an item correctly is independent of the probability of answering another item correctly (Hambleton & Swaminathan, 1985).

The IRT has many theoretical advantages over CTT. First, it has falsifiable models while CTT is a tautology. Basically, IRT allows three different models: one-parameter logistic model (1-plm) with only item difficulty parameter (*b*-parameter), two-parameter logistic model (2-plm) with *b*-parameter and item discrimination parameter (*a*-parameter), and three-parameter logistic model (3-plm) with *a*-parameter, *b*-parameter, and pseudo-chance parameter (*c*-parameter). A given IRT model may or may not be appropriate for a particular data set. The advantages of the IRT models can be attained only when the model fits the given test data (Hambleton et al., 1991). Therefore, in IRT

applications, it is fundamentally necessary to compute the goodness-of-fit indices of particular IRT models in order to determine the best-fit model among the three models.

Second, IRT offers invariant person and item indices whereas CTT provides variant indices (Kim & Nicewander, 1993). The most important distinction of IRT from CTT is that estimates of an examinee's ability and item parameters are invariant (Reckase, 2009). This property of invariance indicates that the item parameters are not dependent on the examinees' abilities and the examinee's ability is independent of the test (Hambleton et al., 1991).

Third, no parallel test assumption is needed for IRT while CTT assumes the existence of a parallel test. Finally, IRT is an item-oriented theory whereas CTT is a test-oriented theory. The IRT provides "a model that is expressed at the item level rather than at the test level" (Hambleton et al., 1991, p. 5). The IRT handles an examinee's performance for each item. The information provided by answering the question, "What is the probability of an examinee answering a given item correctly?" is useful for the applications of testing in many ways. The IRT also has practical advantages; it is an individualized test (e.g., graduate record examination) and fewer items are needed to estimate examinees' abilities.

An examinee's ability in IRT is called the person parameter, person's ability, theta (θ) parameter, or person's latent trait (Hambleton & van der Linden, 1982). Psychometric calibration based on IRT provides at least seven indices for reporting: θ-parameter, *a*-parameter, *b*-parameter, *c*-parameter, item characteristic curve (ICC), item information function (IIF), and test information function (TIF). The θ-parameter is

estimated using the likelihood function of the response pattern. The θ and b are on the same scale with values of 0 and 1 for the mean and the standard deviation, respectively.

The value of *a*-parameter influences the amount of item information. An item with low discriminating power (i.e., low value of *a*-parameter) does not add useful information in a test. If the equal discrimination index assumption is violated, the 1-plm is not valid for the data. Only if the item-test correlation distribution is homogeneous, the *a*-parameter value is equal, which is almost impossible in a real test setting. The *c*-parameter represents the probability of answering an item correctly without any knowledge about the question. Only if the performance of the low-ability students on the most difficult items is close to zero, the 1-plm or 2-plm is valid because the *c*-parameter is minimal, which also does not happen often.

The IIFs display the contribution of items to person's ability estimation at each level of theta. An item with less information indicates that the item is less useful for assessing a person's ability. Poorly fitting items mislead IIFs. The IIFs are the most important index of IRT in evaluating and selecting items for test development. According to Hambleton et al. (1991), the IIFs "provide new directions for judging the utility of test items and constructing tests" (p. 93). The TIF shows that this test is more useful for assessing a given level of an examinee's ability. The TIF is calculated by the sum of the item information functions. The ICC shows that as the person's ability increases, the probability of a correct response to an item increases. The poor-fit items to the data show relatively large deviations from the ICC while the good-fit items show small deviations.

In sum, IRT has many theoretical and practical advantages over CTT for reading comprehension assessment. The item-oriented IRT theory resolves the limitation of CTT which provides no basis for an examinee's performance estimates for an item because CTT is a test-oriented theory. In IRT, validity and reliability are not necessary components because the examinee's true ability is estimated using the likelihood function of the item response pattern (Hambleton, Swaminathan, & Rogers, 1991; Kim & Nicewander, 1993).

Testlet Response Theory (TRT)

The IRT assumes that given a fixed ability level on the part of the examinee, the probability of answering an item correctly is independent of the probability of answering another item correctly, which is called local item independence (Hambleton & Swaminathan, 1985). In other words, local item independence is a probabilistic term where two items' covariance is assumed to be zero if the underlying ability factor is controlled. However, there are some situations in which the local item independence assumption of IRT is violated (Wang et al., 2002).

An example of when the local item independence assumption may be violated is when a test consists of testlets (Wang et al., 2002). In reading comprehension measures, a testlet can be defined as a collection of several questions on one reading passage so that each question can measure a different aspect of the examinee's comprehension of the passage (DeMars, 2012). Reading comprehension measures in particular are often composed of several testlets. This is a typical case where the local item independence assumption of IRT is violated (Wang et al.).

Items of reading comprehension tests are most likely correlated with each other because of testlets that comprise questions about the same passage, which may result in the problem of local dependence (Wainer, Bradlow, & Wang, 2007). This source of local item dependence, called response dependence (Marais & Andrich, 2008), may lead the examinee's response for a given item to impact the response for a subsequent item within the same testlet. Additionally, responses to items within a testlet could be correlated with a secondary trait (i.e., testlet trait) relevant to the passage as well as an original trait measured by the test (Min & He, 2014). This kind of local item dependence is called trait dependence (Marais & Andrich).

In a reading comprehension test, items under the same reading passage are dependent on a testlet, or "a common stimulus" (Li, Bolt, & Fu, 2006, p. 3). Therefore, the secondary trait related to the common stimulus may result in the problem of trait dependence. The secondary trait in a reading comprehension test may include background knowledge about the passage contents or the levels of text complexity for the passages (DeMars, 2006).

The consideration of local item dependence related to testlets is required for this particular structure, because the use of the traditional IRT underlies the assumption of local item independence. Therefore, testlet traits must be accounted for in order to correct the problematic local item dependence (DeMars, 2006). It is challenging for the test developers to appropriately model those local item dependencies (Zenisky, Hambleton, & Sireci, 2002).

The TRT model was designed to resolve the problem of local item dependence in a testlet-based assessment (Baldwin, 2007; Ip, 2010; Li et al., 2006). The most

important concept of the TRT model is that the testlet model uses "statistical random effects to capture the residual correlation" after controlling for the primary latent trait that is supposed to be measured (Ip, p. 468). To model the examinee's responses to testlet items, a random testlet-effect parameter is added to the traditional IRT model (Li et al.). The TRT model takes into account testlet effects by modeling local item dependence explicitly (Schroeders, Robitzsch, & Schipolowski, 2014) and incorporates secondary traits related to testlets into the IRT model (Min & He, 2014; Rijmen, 2010).

Researchers have proposed several TRT models. Wang and Wilson (2005) proposed the modified Rasch model by adding a random testlet-effect parameter to a Rasch model. Bradlow, Wainer, and Wang (1999) modified the two-parameter logistic model (2-plm). Additionally, the three-parameter logistic model (3-plm) was modified for a testlet effect by Wainer, Bradlow, and Du (2000). All of those three models can be categorized as the testlet-effects model (DeMars, 2006).

Among the three testlet-effects models, the modified 3-plm (i.e., 3-pl testlet-effects model) is as follows:

$$p(\theta) = c_i + (1-c_i)\frac{e^{1.7a_i(\theta-b_i-\gamma_{g(i)})}}{1+e^{1.7a_i(\theta-b_i-\gamma_{g(i)})}}, \quad (1)$$

where $p(\theta)$ is the probability that an examinee answers item i correctly, θ = latent trait (person ability), a_i = item discrimination parameter, b_i = item difficulty parameter, c_i = pseudo-chance parameter, and $\gamma_{g(i)}$ = random testlet effect. This model is almost the same as the 3-plm of IRT except for one parameter— $\gamma_{g(i)}$. In the testlet-effects model, an item is expected to have the same discrimination parameter a_i for $\gamma_{g(i)}$ and θ (Min & He, 2014).). It is critical to apply this testlet-effects model for reading comprehension

assessments instead of the traditional IRT model, because the testlet-effects model incorporates a secondary trait such as background knowledge into the IRT model without ignoring the effects of the secondary trait, which may lead to more accurate estimations for the primary trait—reading comprehension ability (DeMars, 2006).

In addition to the testlet-effects models, the bi-factor model has been proposed for testlet-based assessments (DeMars, 2006; Li et al., 2006; Min & He, 2014; Rijmen, 2010). In the bi-factor model, each item response is a function of both the primary trait and one of the second testlet traits. The bi-factor model assigns separate discrimination parameters to the primary and secondary dimensions, while the testlet-effects models assign the same discrimination parameters to the secondary dimensions associated with testlets as the primary dimension.

The bi-factor model is as follows:

$$p(\theta) = c_i + (1-c_i)\frac{e^{1.7(a_i\theta_j + a_{Ti}\theta_T - d_i)}}{1+e^{1.7(a_i\theta_j + a_{Ti}\theta_T - d_i)}}, \quad (2)$$

where $P(\theta)$ is the probability of correct response on item i, θ is composed of θ_j and the vector uT of the testlet traits, c_i is the pseudo-chance parameter, a_i is the item discrimination on the primary trait, a_{Ti} is a vector of testlet discrimination parameters for item i, and d_i is the item difficulty. For any item i, all but one of the testlet a's is equal to zero, and thus only one element of uT has an impact on the function. For brevity, a_T will refer to the nonzero element of a_{Ti} (for item i within a testlet). For reading comprehension assessments, the bi-factor model would be also better than the traditional IRT model because the bi-factor model incorporates a secondary trait into the IRT model,

allowing more parameters than the testlet-effects model due to the separate assignment of discrimination parameters to the primary and secondary dimensions.

According to TRT studies, ignoring the testlet effects results in inaccurate estimations of item and person parameters as well as item misfit (DeMars, 2012; Eckes, 2014; Min & He, 2014; Schroeders et al., 2014; Yao, Rich, & McGraw-Hill, 2008; Zenisky et al, 2002). Other studies also found the biased item discrimination parameters when used the traditional IRT for the testlet-based assessments (Bradlow et al., 1999; Wainer et al., 2000; Wainer & Wang, 2000). The TRT model may be an alternative to the traditional IRT model for the testlet-based assessments such as reading comprehension measures. Figure 1 shows graphical representation of the IRT and TRT models.

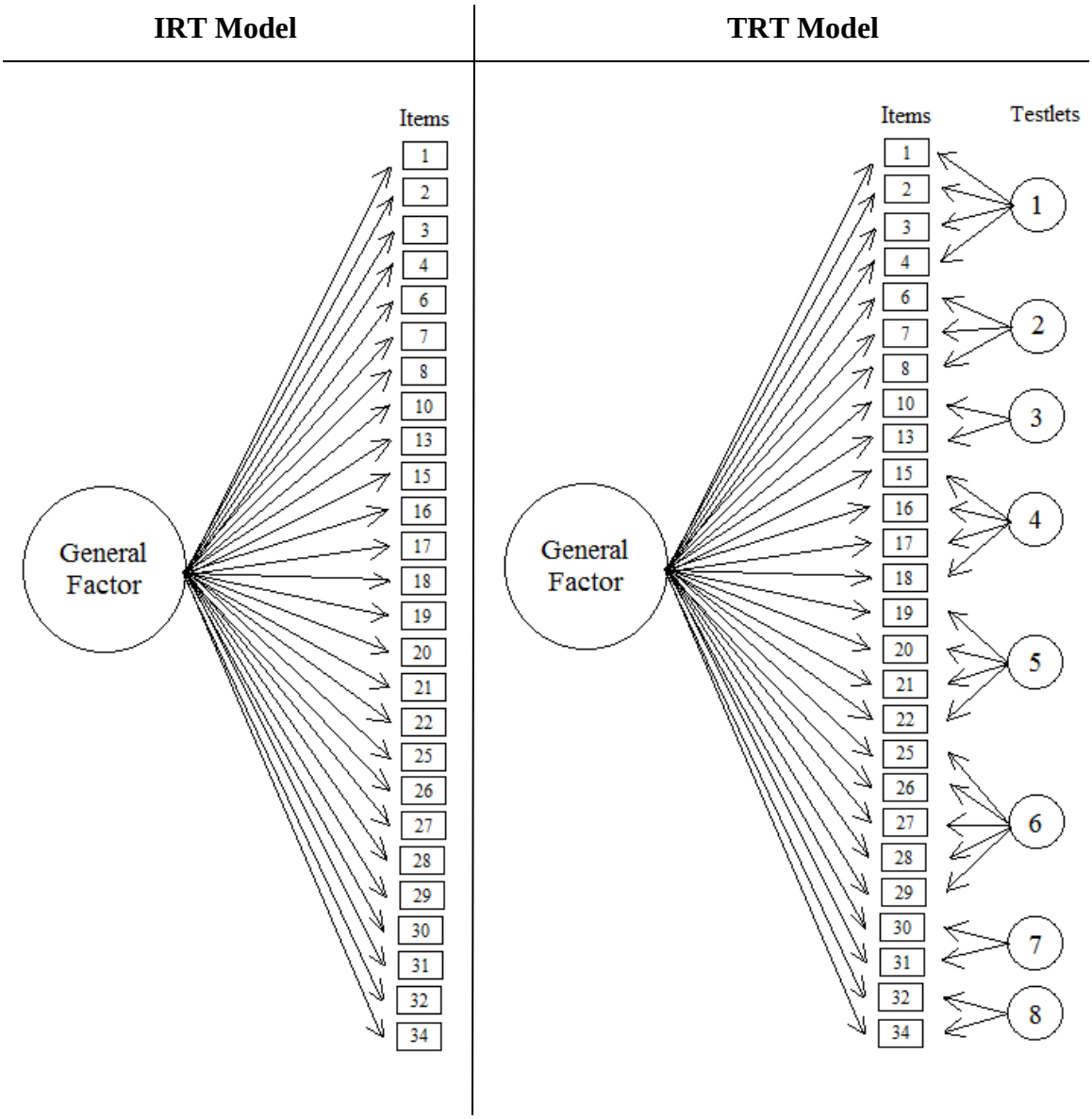

Figure 1. Graphical representation of the IRT and TRT models.

Current Literature on TRT with Reading Comprehension

An electronic literature search was conducted to find studies that address the application of the IRT and TRT models to a reading comprehension test. The PsycINFO and ERIC databases were searched based on the following category limiters: *reading comprehension AND testlet response theory*. The PsycINFO and ERIC databases yielded only six and one studies, respectively. A total of four studies were selected for the abstract examination. These abstracts were examined on the basis of the inclusion criteria. The inclusion criteria were as follows: (1) methodological studies using the TRT models, (2) studies using reading comprehension tests as an instrument, and (3) studies published in peer-reviewed journals. Only two studies met the inclusion criteria. The Google Scholar database was searched additionally. Eight more studies were selected based on the inclusion criteria. Thus, 10 studies were selected and coded to extract important information to be utilized for the literature review.

Among 10 studies, six illustrated the application of both IRT models and TRT models to determine the best model for a testlet-based assessment. Two of them (Baldonado, Svetina, & Gorin, 2015; Li et al., 2006) applied only 3-plm and only TRT models, respectively. Li et al. concluded that the bi-factor model was the best fitting model among the four TRT models. DeMars (2012) conducted a simulation only study comparing two pairs of testlet models to confirm testlet effects. The other study (Rijmen, 2010) compared TRT models with a second-order model which belongs to hierarchical models, concluding that the bi-factor model was the preferred model for a reading comprehension assessment.

Min and He (2014) applied three different IRT models, the 2-pl testlet-effects model, and the bi-factor model to a reading comprehension test. The study revealed that the bi-factor model was the best-fitting model for testlet-based assessments. DeMars (2006) used the bi-factor model, testlet-effects model, polytomous model, and independent-item model to choose the best model for reading tests. The study showed that the testlet-effects model was the best. Wainer and Wang (2000) applied both the testlet-effects model and 3-plm of IRT to reading comprehension sections of 86 TOEFL. They demonstrated that the testlet-effects model was better than the IRT model.

The testlet-effects model was compared with two IRT models in Eckes' study (2014). The study also revealed that the testlet-effects model is needed for a testlet-based assessment. Tao, Xu, Shi, and Jiao (2013) proposed the 2-pl normal ogive TRT model (2PNOTRT) to compare with the 2-pl normal ogive IRT model (2PNOIRT), revealing that the 2PNOTRT is a better-fitting model for testlet-based assessments.

Uniquely, Jiao, Kamata, Wang, and Jin (2012) proposed a four-level IRT model in order to account for local person dependence as well as local item dependence. According to them, when a cluster sampling method is used, local person dependence is likely to be introduced. They compared their proposed model with three other models (i.e., Rasch model, Rasch testlet model, and three-level Rasch model) for testlet-based assessments.

As for an index of local item dependence, Min and He (2014) utilized the standardized local dependence (LD) X^2 statistic. The result revealed significant differences in local item dependence when applying the IRT models and testlet-effects models to testlet-based tests. The IRT models showed severe local item dependence

while the testlet-effects models showed no local item dependence. Baldonado et al. (2015) estimated Q_3, LD X^2, and G^2 as indices of local item dependence. None of the item pairs in their study met the commonly used criterion of $Q_3 = 0.2$ to be classified as locally dependent.

According to Min and He (2014), estimates of the *a*-parameter and person's ability parameter were inaccurate when the traditional IRT model was applied to a testlet-based test. The testlet-effects model yielded slightly lower standard errors of item parameter estimates than the bi-factor model. The study of DeMars (2006) revealed that applying the traditional IRT model to a testlet-based test yielded the underestimated *a*-parameter and greater standard error for the *b*-parameter compared to the results using the TRT model.

Wainer and Wang (2000) demonstrated that when local item independence was incorrectly assumed, estimates of both *a*-parameter and *c*-parameter were biased. Eckes' study (2014) also revealed inaccurate estimations of test reliability and standard error of person's ability parameter. However, estimates of *a*- and *b*-parameters remained largely unaffected.

Tao et al. (2013) found that the mean square error of the TRT model (i.e., 2PNOTRT) is generally smaller than that of the traditional IRT model (i.e., 2PNOIRT), which means that the TRT model better fits testlet-based assessments. Jiao et al. (2012) showed that the proposed four-level IRT model "recovered the item difficulty and person ability parameters with the least total error" (p. 82), accounting for both local item dependence and local person dependence.

As for model comparison indices, overall likelihood ratio tests (-2 log-likelihood [-2LL] difference tests), Akaike information criterion (AIC), and Bayesian information criterion (BIC) were utilized. Li et al. (2006) estimated deviance information criterion (DIC) in addition to AIC and BIC. Jiao et al. (2012) used only DIC to choose the best fit model under various simulation conditions; neither AIC nor BIC could properly identify their simulated true model. DeMars (2012) also proposed a new index, sample-size adjusted Bayesian information criterion (SSA-BIC), as well as -2LL, AIC, and BIC.

In short, all 10 studies demonstrated that when the local item independence assumption of IRT is violated, the TRT is the alternative to the IRT methods to accommodate a possible dependence of items in real test settings (Baldwin, 2007). The TRT may offer an applicable assessment tool for testlet-based assessments such as reading comprehension tests, resolving the problem of item correlation. When local item dependencies are neglected, data analyses using the traditional IRT model may be misleading (DeMars, 2012). The correlated nature of items in the testlet structure raises an issue of overestimation or underestimation for item parameters in IRT (DeMars, 2006). Even though there has not been any consistent trend for differences in item and person's ability parameters, and standard errors of those parameters, it is clear that parameter estimates may be biased if testlet effects are not treated in a proper way (Eckes, 2014; Wainer & Wang, 2000).

To conclude, the majority of the 10 studies revealed that the TRT models such as the testlet-effects model and the bi-factor model yielded the best model-data fit for the testlet-based assessments (DeMars, 2006; Li et al., 2006; Min & He, 2014; Rijmen, 2010;

Tao et al., 2013). Many studies used -2LL, AIC, and BIC for model comparisons (DeMars, 2006; DeMars, 2012; Li et al., 2006; Min & He, 2014; Rijmen, 2010).

As for participants and instruments, the majority of the studies used entrance exams or international assessments such as the Graduate School Entrance English Exam in China, the Test of German as a Foreign Language, TOEFL, Law School Admission Test, or Program for International Student Assessment 2000 (PISA). Their instruments were not the typical reading comprehension tests which were administered under the regular school systems in the United States (e.g., middle schools or high schools). Only the study of Jiao et al. (2012) used a reading comprehension test for high school graduation in a southern state in the United States. Thus, participants of the majority of the studies came from a specific demographic such as Chinese students applying to master's programs, not from students who are attending middle schools or high schools in the United States. DeMars (2012) conducted a simulation study without real participants. Table 1 presents methodological characteristics of 10 TRT studies.

Table 1

Methodological Characteristics of 10 TRT Studies

Study	Participant description	Instrument	Item types	Used models	Model comparison indices	Results
Min & He, 2014	14,089 candidates applying to master's programs	Graduate School Entrance English Exam: reading comprehension section	Four 20-item testlets	3 IRT models, TRT, & bi-factor model	-2LL, AIC, BIC, local-dependence X^2 statistic, test reliability, test information, & item parameters	The bi-factor model fits best.
DeMars, 2006	5,000 examinees who completed the selected test booklet 7	Program for International Student Assessment 2000 (PISA 2000): reading tests	Eight 2-item, three 3-item, three 5-item testlets	Bi-factor, testlet-effects, Polytomous, & IRT models	-2LL, AIC, & item parameters	The testlet-effects model is the best one
Rijmen, 2010	13,508 examinees	International English Assessment Test: reading comprehension section	Four 5-item testlets	Bi-factor, TRT, & second-Order models	AIC& BIC	The bi-factor model is the best one.
Wainer & Wang, 2000	26,977 examinees	86 TOEFL: reading and listening comprehension sections	Fifty 13-item reading comprehension testlets, 36 listening comprehension testlets	TRT model, & 3-plm	Item parameters, & test information	The testlet-effects model is the best.
Baldonado, Svetina, & Gorin, 2015	5,734 high school students	High-stakes test of U.S. high school students for college preparedness and scholarship purposes	Four 26-item testlets	3-plm	Item parameters & Q_3, X^2, G^2 for local-dependence identification	Cognitively similar item pairs had higher local dependence values

Table 1 (cont.)

Eckes, 2014	1st sample: 2,859 examinees 2nd sample: 2,214 examinees	Test of German as a Foreign Language: listening section	Three 8-, 10-, 7-item testlets	TRT, independent-items IRT, & graded-response model	Item parameters, person ability parameter, & test reliability	The testlet-effects model is the best one.
Li, Bolt, & Fu, 2006	2,000 examinees	Law School Admission Test (LSAT), & English Placement Test (EPT)	Four 5-8-item, Ten 4-6-item, eight 4-6-item testlets	2-pl testlet, bi-factor, bi-factor with constraints, bi-factor with constant *a*-parameters for testlet effects	AIC, BIC, DIC	The bi-factor model best fits the given data.
DeMars, 2012	Confirmation of testlet effects (simulation study)	Simulated 3 test forms	Five 5-item, ten 5-item, five 10-item testlets	Comparison of unidimensional vs. single testlet model, & comparison of all-but-one vs. complete model	-2LL, AIC, BIC, SSA-BIC	The second comparison was most useful for detecting testlet effects.
Tao, Xu, Shi, & Jiao (2013)	Simulation & 1,289 Japanese students	Reading comprehension test	4 testlets with 8, 7, 8, and 5 items per testlet, respectively	2-pl normal ogive IRT & 2-pl normal ogive TRT model	Item parameters & mean square error	2-pl normal ogive TRT model better fits.
Jiao, Kamata, Wang, & Jin (2012)	1,644 students in 424 school districts	Reading comprehension test for high school graduation	4 testlets with 8, 8, 9, & 7 items	Rasch , Rasch testlet, three-level Rasch , & four-level IRT model	DIC	A four-level IRT model is the best one.

Note. DIC: deviance information criterion; SSA-BIC: sample-size adjusted Bayesian information criterion; Q_3, X^2, and G^2: indices of local-dependence.

Research Questions

Based on the information from the literature review, it is necessary to utilize a typical reading comprehension test that was administered to students attending regular schools in the United States for an application of the TRT and IRT models. Specifically, it is crucial to analyze an adolescent reading comprehension test for more accurate assessments. It is notable that none of the TRT studies took text complexity metrics into account even though text complexity is crucial for reading comprehension assessments, especially for older adolescent readers. Therefore, text complexity metrics were added to the present study for data analyses. The purpose of this study was to apply the testlet-effects model, the bi-factor model, and the traditional IRT model to a seventh-grade reading comprehension test administered in the regular schools in the United States. The three models were compared to determine the best model for a testlet-based reading comprehension test. Examining a reading comprehension measure, four research questions were addressed:

1. Which model best fits the given reading comprehension test based on goodness-of-fit indices?
2. To what extent do the IRT model and the TRT models show differences in local item dependence?
3. To what extent do the IRT model and the TRT models show differences in item and person parameters and their standard errors?
4. What are the strengths and weaknesses of each item based on item analyses from the TRT models and text complexity metrics from Coh-Metrix?

CHAPTER THREE: METHOD

The data set used in the present study was obtained from a testing company. This 34-item reading comprehension test was administered in the spring semester of the 2014-2015 academic year in 18 states in the United States. The test was developed to monitor student growth in states using the CCSS and provided subtest scores in the following areas: literature, information, language, and writing questions. The test was administered online.

Participants

Participants (n = 8,815) were seventh-grade adolescent students. Based on the reported data, the distribution of gender was almost equal for both female and male (49.9% female). The majority of the population was Caucasian students, with minority populations of Hispanics (24.9%), African Americans (7.5%), Asians (2.2%), and Native Americans (3.4%). Approximately 0.8% of the population received English as a Second Language (ESL) services. Students receiving special education were about 2.6% of the population.

Instrument

The assessment comprised multiple choice questions with four answer choices. The test consisted of 34 items that can be divided into nine testlets. Some testlets were longer with more than two items (i.e., 3-5 items), and some were shorter with only two items. All 34 of these items were scored either correct or incorrect.

According to item-level analysis based on comprehension skills, four items (i.e., items 5, 9, 14, and 24) were constructed separately without any comprehension passage.

Additionally, the sixth testlet contained only one item (i.e., item 23). Because the present study investigated local item dependencies between two items within a testlet under the same comprehension passage, those five items (i.e., items 5, 9, 14, 23, and 24) and the sixth testlet were excluded for further data analyses. Three more items (i.e., items 11, 12, and 33) were also excluded because they were independent questions which were not connected with the passage. Therefore, a total of eight testlets and 26 items were utilized for data analysis. Table 2 provides item level analysis based on comprehension skills.

Table 2

Item Level Analysis Based on Comprehension Skills

Item Number	Comprehension Skills
1	Inferential
2	Context-related Vocabulary
3	Context-related Vocabulary
4	Summarizing
6	Author's Craft & Theme
7	Setting Inferential
8	Point of View & Knowledge
10	Inferential Organization & Semantic
13	Main Idea
15	Context-related Vocabulary
16	Literal Detail
17	Knowledge & Inferential
18	Inferential
19	Point of View, Text Purpose, & Author's Craft
20	Text Structure
21	Author's Craft & Main Idea
22	Main Idea
25	Text Structure
26	Context-related Vocabulary
27	Literal
28	Main idea & Summarizing
29	Context-related Vocabulary
30	Context-related Vocabulary

Table 2 (cont.)

31	Text Structure
32	Organization
34	Context-related Vocabulary

Based on CCSS analysis, literature questions used narrative/literature passages and asked questions related to understanding text structure and key points as well as determining themes. Information questions used expository text, asking explicit questions including vocabulary knowledge. Language questions were related to understanding proper language use such as grammar and word meaning in addition to understanding figurative language. Writing questions pertained to text organization and planning, information gathering, arguments, and conducting research.

In addition, Coh-Metrix was used for multilevel analyses of text complexity on the basis of text cohesion. It provided information about eight comprehension passages. Table 3 describes each passage including text type, content, number of questions, and *M*s and *SD*s of paragraph-, sentence-, and word-lengths. Table 4 presents text easability principle component scores for each passage. The principle components for each passage include narrativity, syntactic simplicity, word concreteness, referential cohesion, deep cohesion, verb cohesion (the degree of overlapping verbs in the text), and connectivity.

Narrativity is important to reading comprehension because it is associated with word familiarity and story-likeness which influence the text difficulty. Narrative text is easier to understand than non-narrative texts. Syntactic simplicity refers to sentence structures. The simplicity of sentence structures has effects on a reader's comprehension. When a text contains "shorter sentences, few words before the main verb of the main

clause, and few words per noun-phase," a text is easier to comprehend (p. 70). Word concreteness is also crucial to comprehension, because concrete words help a reader comprehend a text easily by providing concepts which are easy to represent visually. Referential cohesion implies overlapping of content words with other sentences in the text, which has effects on reading comprehension. Low-cohesion text is difficult to comprehend due to fewer connections of ideas. Deep cohesion refers to "the degree to which the text contains causal and intentional connectives" (p. 85) when the text involves causal and logical relationships. The reader may need the process of inference to comprehend a text with low deep cohesion (McNamara et al., 2014).

Five of eight passages fell within the fifth grade level, and three of them the eighth grade level; the average grade level was six. Passages contained around 417 average number of words and 34 average number of sentences. Passage 7 was a poem which had the fewest number of words and sentences. They varied in the amount of narrativity (range of *z* score = from -0.96 to 1.03), referential cohesion (range of *z* score = from -1.54 to 0.55), and deep cohesion (range of *z* score = from -0.12 to 3.58), but showed comparatively less variances on dimensions of syntactic simplicity (range of *z* score = from -0.35 to 0.98) and word concreteness (range of *z* score = from -0.4 to 1.14).

Table 3

Descriptive Statistics, Text Type, and Content for Each Passage

Passages	1	2	3	4	5	7	8	9
Text Type	Speech	Narrative Story	Student Draft	Drama	Speech	Article	Poem	Student Draft
Text Content	City council election	Love story	Helping the community	Loving the library	Science	Science	In the woods	Organi-zing a space
Number of Paragraphs	5	21	5	20	9	7	5	7
Number of Sentences	21	55	31	50	36	29	16	33
Number of Words	226	634	332	597	545	543	28	430
Paragraph Length: *M* (*SD*)	4.2 (1.1)	2.62 (1.8)	6.2 (3.9)	2.5 (1.32)	4 (1)	4.14 (2.34)	3.2 (1.1)	4.71 (2.75)
Sentence Length: *M* (*SD*)	10.76 (5.62)	11.53 (8.67)	10.71 (5.07)	11.94 (6.15)	15.14 (6.03)	18.72 (6.25)	13.63 (6.14)	13.03 (10.03)
Word Length: *M* (*SD*)	1.66 (0.87)	1.39 (0.77)	1.4 (0.74)	1.38 (0.74)	1.45 (0.79)	1.57 (0.86)	1.38 (0.68)	1.28 (0.55)
Number of Questions	4	3	2	4	4	5	2	2

Note. Paragraph length: number of sentences, Sentence length: number of words, Word length: number of syllables, *M*: Mean, *SD*: Standard Deviation.

Table 4

Text Easability Principle Component Scores for Each Passage

Passages	1	2	3	4	5	7	8	9
Z score of Narrativity	-0.96	0.77	-0.09	0.56	-0.32	-0.61	1.03	0.01
Z score of syntactic Simplicity	0.8	0.45	0.98	0.35	0.36	-0.35	0.62	0.66
Z score of Word Concreteness	0.29	-0.4	0.72	-0.03	0.66	1.02	1.14	0.94
Z score of Referential Cohesion	-1.54	-1.16	-1.4	-0.3	0.55	-0.69	-1.21	-0.23
Z score of Deep Cohesion	1.73	0.24	0.48	-0.12	0.17	3.58	1.22	0.1
Argument Overlap (*M*)	0.35	0.28	0.3	0.53	0.77	0.54	0.33	0.47
Content Word Overlap (*M*)	0.06	0.08	0.05	0.12	0.16	0.08	0.06	0.09
LSA Overlap (*M*)	0.15	0.11	0.19	0.12	0.27	0.21	0.1	0.26
All Connectives Incidence	106.2	89.91	69.28	67.00	84.4	104.97	96.33	86.05
Pronoun Incidence	70.8	113.57	75.3	110.55	58.72	42.36	169.73	81.4

Table 4 (cont.)

Age of Acquisition for Content Words	424.32	317.69	294.88	308.39	349.34	340.13	315.55	249.73
Concreteness for Content Words (*M*)	377.6	366.71	413.26	388.6	422.05	405.5	405.03	418.13
Readability (Flesch-Kincaid Grade Level)	8.24	5.34	5.05	5.35	7.44	7.97	4.88	5.1

Data Analyses

The IRTPRO software was utilized to analyze the test and items. This program implements the Expected A Posteriori (EAP) method to estimate the person parameter. The prior distributions were set for each parameter following the default of the program. A normal prior distribution was set for the item discrimination parameter. The prior distribution for the item difficulty parameter was set as the standard normal distribution. As for the pseudo-chance parameter, the logit-form normal distribution was set, which resulted in the inclusion of its negative values.

Exploratory Factor Analysis (EFA) was performed to test dimensionality. The extraction method of principal component analysis was applied for factor analysis. For the IRT model, the three models (1-plm, 2-plm, and 3-plm) were compared to select the best model for the given data based on goodness-of-fit indices. Goodness-of-fit indices include -2LL difference tests, AIC, and BIC. The -2LL differences (i.e., 1-plm vs. 2-plm, 1-plm vs. 3-plm, and 2-plm vs. 3-plm) were compared on the basis of chi-square

test. Statistically significant differences validate the move to the more complicated model. The information criteria (i.e., AIC and BIC) reduce the tendency toward model over-parameterization (i.e., models with more parameters tend to fit a data set better). The model with the smallest AIC or BIC indicates the model with the best comparative fit. Then, the selected best IRT model, the testlet-effects model, and the bi-factor model were applied to the data and compared for further data analyses of the present study.

To answer the first research question regarding the determination of the best-fitting model among the three different models (i.e., the IRT model, the testlet-effects model, and the bi-factor model), three goodness-of-fit indices were utilized as model comparison indices. First, -2LL difference tests were performed for model comparison. In addition, AIC and BIC were computed for further model comparison.

To compare local item dependencies among the three different models for the second research question, the standardized local dependence (LD) X^2 statistic was computed. The standardized LD X^2 statistic was proposed by Chen and Thissen (1997). It shows the extent of item dependence for each item pair. Because the test consisted of one five-item testlet, three four-item testlets, one three-item testlet, and three two-item testlets (a total of eight testlets and 26 items), 34 item pairs using the combination formula ($1 \times {}_5C_2 + 3 \times {}_4C_2 + 1 \times {}_3C_2 + 3 \times {}_2C_2 = 34$) were examined to reveal local item dependencies within a testlet. Values of the standardized LD X^2 statistic exceeding four show clear local dependence, and values exceeding 10 represent extreme local dependence between items (Chen & Thissen, 1997; Min & He, 2014).

For the third research question, item parameters (i.e., item discrimination, item difficulty, and pseudo-chance parameters), the person parameter, and their standard errors

were computed to compare differences among the three models. Specifically, scatter plots were obtained to evaluate parameter-estimation consistency among models. Correlations and mean differences between the estimated parameters were also obtained to detect and quantify the magnitude of inaccuracy due to the use of a worse-fitting model. High correlations provide evidence for no differences in estimation from the models. The larger mean difference indicates the smaller correlation. For the fourth research question, the strengths and weaknesses of each item were examined based on item analyses from the TRT models. Item parameters for each item including *a*-, *b*-, and *c*-parameters were computed from the TRT models and compared with the results from Coh-Metrix.

CHAPTER FOUR: RESULTS

EFA revealed a one-factor solution with an eigenvalue of 6.14 and 23.6% variance explained. However, when eigenvalues over one were included, a total of 31.9% variance was accounted for a multiple-factor solution. For the IRT model, the 3-plm was selected as the best model for the given data among the three IRT models (i.e., 1-plm, 2-plm, and 3-plm). The 3-plm revealed the smallest values of AIC and BIC, indicating the best fit to the data. The -2LL difference tests also supported that the 3-plm was the best fitting model among the three models. Table 5 summarizes the fit indices of the three IRT models, including -2LL, AIC, BIC, and -2LL difference tests. Thus, the 3-plm for the IRT model, the testlet-effects model, and the bi-factor model for the TRT model were applied to the data and compared for further investigation of the present study.

Table 5

Summary of Fit Indices of Three IRT Models

IRT models	-2LL	AIC	BIC	-2LL difference tests
1-plm	264575.83	264629.83	264821.10	ΔG^2=1452.01* 1plm vs 2plm
2-plm	263123.82	263227.82	263596.19	ΔG^2=947.68* 2plm vs 3plm
3-plm	262176.14	262332.14	262884.71	ΔG^2=2399.69* 1plm vs 3plm

Note. * p < .05.

Goodness-of-Fit Indices for Model Comparisons

The likelihood ratio tests were performed for three nested models; the 3-plm is nested within the testlet-effects model, and the testlet-effects model is nested within the bi-factor model. When models are compared, the more constrained model is considered to be nested within the other complex model. The 3-plm is nested within the testlet-effects (as well as the bi-factor) model because the random testlet-effect parameter is constrained to zero in the 3-plm. The testlet-effects model is the more constrained model compared to the bi-factor model due to the constrained item discrimination parameters (i.e., the item discrimination parameters for testlets are the same as the primary item discrimination parameters).

It was tested that the more complex model (i.e., the testlet-effects model or the bi-factor model) would significantly improve the model-data fit over the simpler model (i.e., the 3-plm or the testlet-effects model). The results revealed that the -2LL differences were statistically significant based on the chi-square test: ΔG^2 = 194.23 from the 3-plm to the testlet-effects model, ΔG^2 = 463.19 from the 3-plm to the bi-factor model, and ΔG^2 = 268.96 from the testlet-effects model to the bi-factor model. Those results indicated that the more complicated testlet-effects model and the bi-factor model fit significantly better than the traditional 3-plm for the present data. Furthermore, the bi-factor model was the best fitting model when compared to the testlet-effects model. Critics point out that the bi-factor model best fits the data because it has more parameters than the other two models. However, "in general, an information-criterion-index-based model selection strategy does not always select a model with a large

number of parameters since the model complexity is compensated for in the index" (Jiao et al., 2012, p. 96).

Therefore, the information criteria (i.e., AIC and BIC) were also computed for additional indices of model comparisons. Comparisons of information criteria also demonstrated that the testlet-effects model and the bi-factor model better fit than the 3-plm and that the bi-factor model was the better fitting model than the testlet-effects model; the values of AIC and BIC for the testlet-effects model and the bi-factor model were smaller than those for the 3-plm. The values of two information criteria for the bi-factor model were smaller than those for the testlet-effects model, indicating that the bi-factor model was the best fitting model among the three models. Based on the goodness-of-fit indices, the IRT model (i.e., the 3-plm) was the worst model among the three different models, while the bi-factor model was the best one for the given reading comprehension data. The testlet-effects model was the second best model for the reading comprehension test regarding the model-data fit. Table 6 presents comparisons of goodness-of-fit indices for the three different models.

Table 6

Comparisons of Goodness-of-Fit Indices of the 3-plm, the Testlet-Effects Model, and the Bi-Factor Model

Models	-2LL	AIC	BIC	-2LL difference tests
3-plm	262176.14	262332.14	262884.71	ΔG^2=194.24* 3plm vs Testlet-effects model
Testlet-effects model	261981.91	262153.91	262763.16	ΔG^2=268.96* Testlet-effects model vs Bi-factor model
Bi-factor model	261712.95	261914.95	262630.45	ΔG^2=463.19* 3plm vs Bi-factor model

Note. * p < .05.

Comparisons of Standardized LD X^2 Statistics

The results of the standardized LD X^2 statistic revealed that when the 3-plm was applied to the data, seven out of 34 item pairs were extremely dependent (i.e., LD $X^2 \geq$ 10; Chen & Thissen, 1997), and nine out of 34 item pairs were clearly dependent (i.e., 4 $\leq$ LD X^2 < 10; Chen & Thissen). A total of 16 item pairs showed the problem of local item dependence. Particularly, the magnitude of local item dependence between items 15 and 16 was the highest (LD X^2 = 41.4). Item dependence between items 2 and 3 was the second highest (LD X^2 = 22.7). Those three items (i.e., items 2, 3, and 15) were vocabulary questions. Item 16 was a question requiring the comprehension ability for literal details. However, when the testlet-effects model was applied to the same data, three out of 34 item pairs showed extreme local dependence, and three other item pairs

were clearly dependent. Specifically, when the bi-factor model was applied, none of the item pairs revealed local item dependencies, showing the superiority of the bi-factor model in resolving the problem of local item dependence. This result indicated that the TRT models including the testlet-effects model and the bi-factor model reduced local item dependence dramatically because the TRT model incorporated the local dependence among items in testlets into its mathematical equation. The bi-factor model was the best one among the three different models based on the comparisons of Standardized LD X^2 Statistics. Table 7 provides a comparison of the standardized LD X^2 statistic among the 3-plm, the testlet-effects model, and the bi-factor model.

Table 7

Comparisons of Standardized LD X^2 Statistics among the 3-plm, the Testlet-Effects Model, and the Bi-Factor Model

	3-plm				Testlet-effects model				Bi-factor model			
Item	1	2	3		1	2	3		1	2	3	
2	2.0				0.9				-0.6			
3	18.6	22.7			0.9	2.8			-0.5	-0.6		
4	0.0	0.3	3.8		4.8	-0.3	-0.3		0.9	-0.5	-0.4	
	6	7			6	7			6	7		
7	9.4				1.6				0.4			
8	1.8	8.8			1.1	-0.7			-0.6	-0.5		
	10				10				10			
13	9.0				0.9				-0.5			
	15	16	17		15	16	17		15	16	17	
16	41.4				14.5				-0.7			
17	2.3	9.6			-0.5	0.0			-0.4	1.4		
18	0.9	5.2	1.6		5.7	11.8	-0.1		-0.4	-0.7	2.1	
	19	20	21		19	20	21		19	20	21	
20	13.7				12.5				0.3			
21	0.0	7.0			-0.3	5.3			-0.6	-0.6		
22	-0.5	3.4	1.6		-0.6	2.9	1.3		-0.6	-0.6	0.2	
	25	26	27	28	25	26	27	28	25	26	27	28
26	11.9				0.9				-0.6			
27	4.8	14.4			-0.4	2.5			-0.6	2.6		
28	0.7	3.8	1.0		-0.6	-0.2	-0.6		-0.5	-0.2	-0.5	
29	3.9	8.2	0.8	4.0	-0.3	-0.3	-0.4	0.8	-0.7	-0.5	-0.6	1.0
	30				30				30			
31	21.6				-0.5				0.5			
	32				32				32			
34	0.5				3.4				-0.7			

Note. LD X^2 statistic $\geq$10: extreme local dependence, 4 $\leq$ LD X^2 statistic < 10: clear local dependence, and LD X^2 statistic < 4: no local dependence.

Comparisons of the Person and Item Parameters

For the parameter estimates, the estimated person parameters from the three different models were found to be highly correlated, indicating no relative differences among the models; all correlation coefficients were equal to .999 (r = .999). The mean differences of the person parameter estimates from the three models were very small, showing that there were no differences in estimation from the models. The estimated standard errors of the person parameters from the 3-plm, the testlet-effects model, and bi-factor model ranged from 0.23 to 0.74, from 0.28 to 0.78, and from 0.27 to 0.7, respectively, showing no significant differences. Figure 2, 3, and 4 present scatter plots of the estimated person parameters from the three models. Correlations and mean differences among the three models are reported in Table 8. Table 9 shows the estimated standard errors of the parameters from the three models.

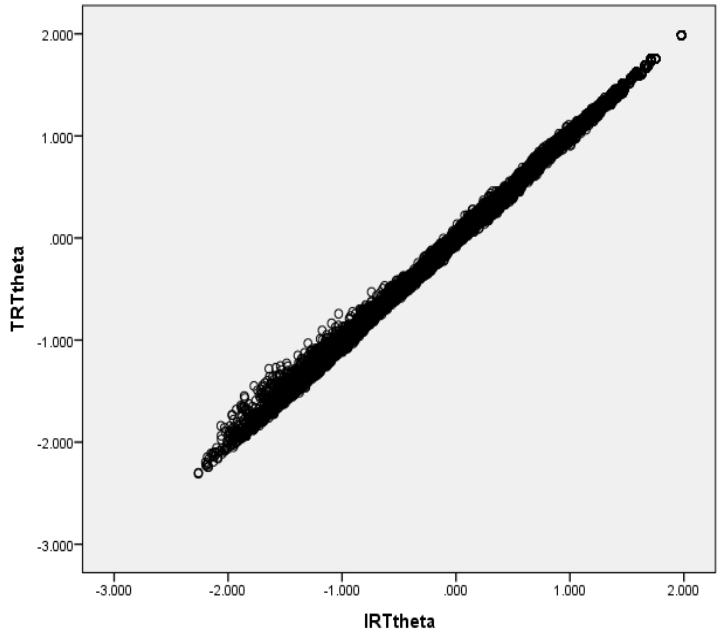

Figure 2. Scatter plot of the person parameter estimates from the 3-plm and the testlet-effects model.

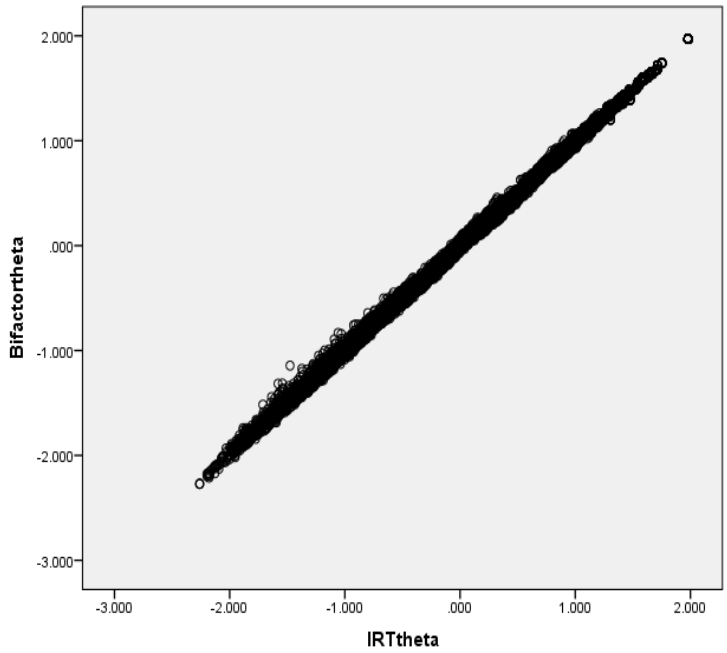

Figure 3. Scatter plot of the person parameter estimates from the 3-plm and the bi-factor model.

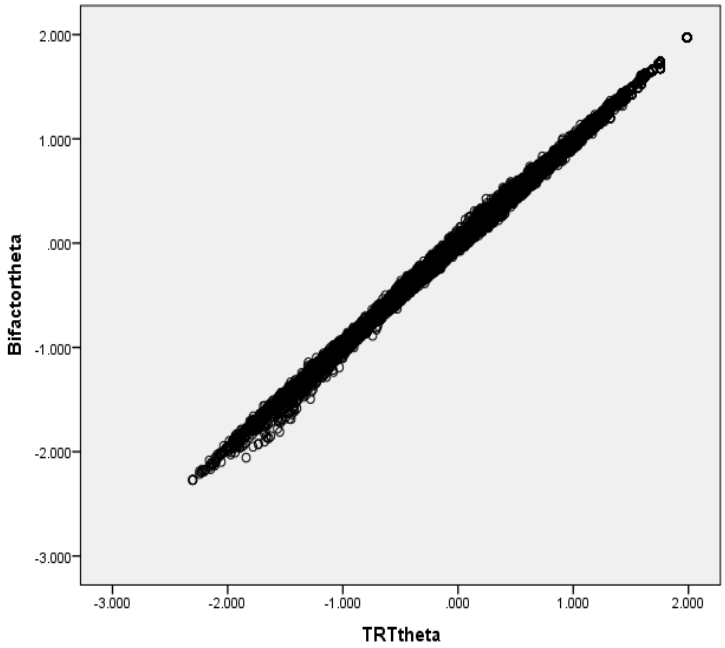

Figure 4. Scatter plot of the person parameter estimates from the testlet-effects model and the bi-factor model.

Table 8

Correlations and Mean Differences among the 3-plm, the Testlet-Effects Model, and the Bi-Factor Model

Parameter	Correlation	Mean Difference
Person parameter		
3-plm & Testlet-effects model	.999	0.005
3-plm & Bi-factor model	.999	0.004
Testlet-effects model & Bi-factor model	.999	-0.0002
Item discrimination parameter		
3-plm & Testlet-effects model	.97	-0.036
3-plm & Bi-factor model	.93	-0.125
Testlet-effects model & Bi-factor model	.92	-0.089
Item difficulty parameter		
3-plm & Testlet-effects model	.99	0.051
3-plm & Bi-factor model	.999	0.031
Testlet-effects model & Bi-factor model	.99	-0.02
Pseudo-chance parameter		
3-plm & Testlet-effects model	.992	-0.065
3-plm & Bi-factor model	.995	-0.06
Testlet-effects model & Bi-factor model	.99	0.005

Table 9

Ranges of the Estimated Standard Errors of the Parameters from the 3-plm, the Testlet-Effects Model, and the Bi-Factor Model

Parameters	3-plm	Testlet-Effects Model	Bi-Factor Model
Person Parameters	0.23 – 0.74	0.28 – 0.78	0.27 – 0.7
Item Discrimination Parameters	0.03 – 0.41	0.03 – 0.12	0.04 – 0.4

The correlation between the estimated item discrimination parameters from the 3-plm and the testlet-effects model was .97. The mean difference from the 3-plm to the testlet-effects model was -0.036. These two results indicated no biased item discrimination parameter estimates. As for the standard error, which is an index of the accuracy of the parameter estimation, it was revealed that the estimated standard errors of the 3-plm and the testlet-effects model ranged from 0.03 to 0.41 and from 0.03 to 0.12, respectively. This result showed that the estimated standard errors of the item discrimination parameters from the testlet-effects model were more stable than those from the 3-plm, indicating the accuracy of the testlet-effects model.

The correlation between the 3-plm and the bi-factor model for the estimated item discrimination parameters was .93, indicating no differences in item discrimination parameter estimates. The mean difference from the 3-plm to the bi-factor model was -0.125, which showed a slight underestimation of the item discrimination parameters when the 3-plm was applied to the data. The estimated standard errors of the item discrimination parameters from the bi-factor model ranged from 0.04 to 0.4. The estimated item discrimination parameters from the testlet-effects model and the bi-

factor model were highly correlated (r =.92), revealing no differences between the two models. The mean difference between the two models for item discrimination estimates also revealed no differences between the two models (mean difference = -0.089). Table 8 describes correlations and mean differences of the item discrimination parameter estimates among the three models. Scatter plots for the estimated item discrimination parameters are presented in Figure 5, 6, and 7.

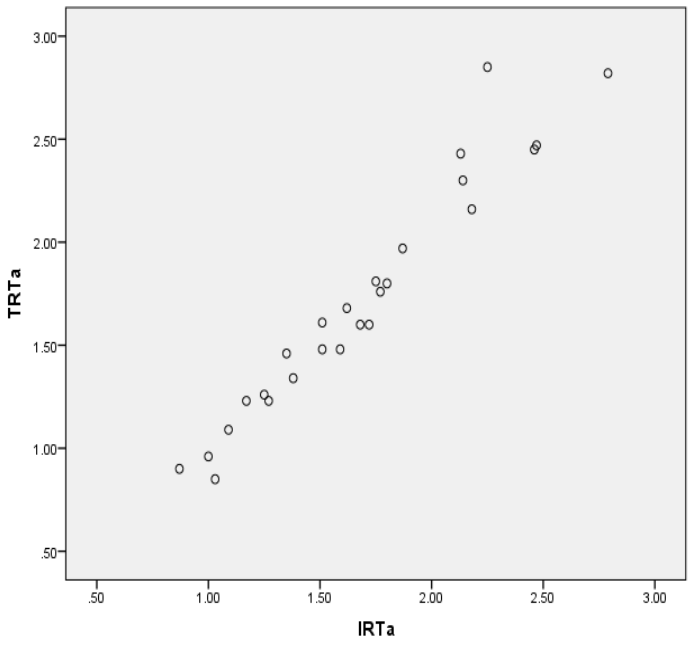

Figure 5. Scatter plot of the estimated item discrimination parameters from the 3-plm and the testlet-effects model.

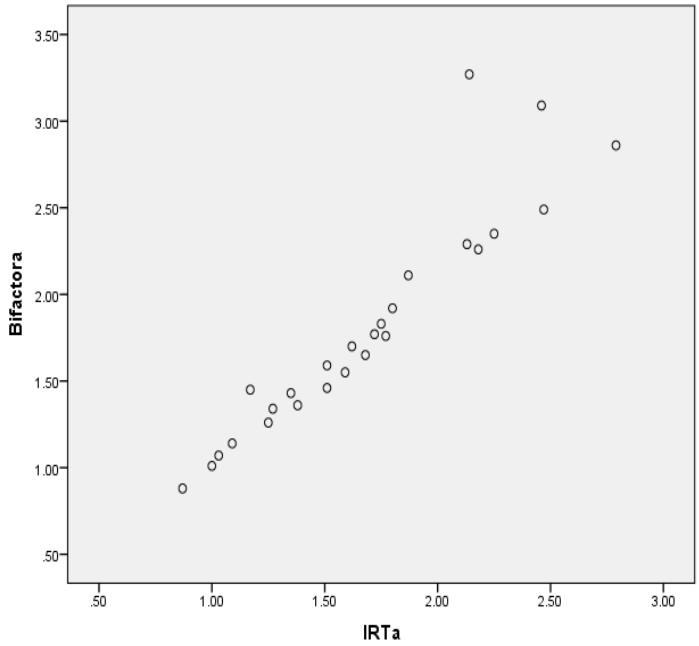

Figure 6. Scatter plot of the estimated item discrimination parameters from the 3-plm and the bi-factor model.

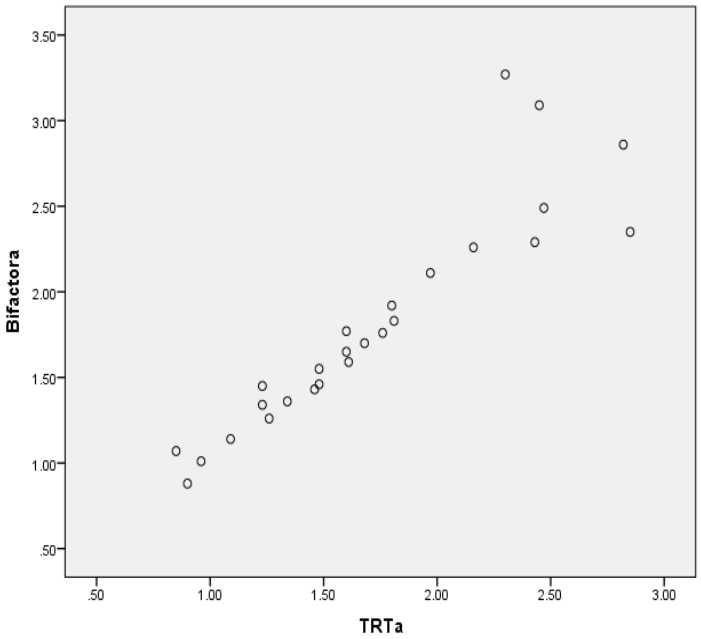

Figure 7. Scatter plot of the estimated item discrimination parameters from the testlet-effects model and the bi-factor model.

The item difficulty parameter estimates were very similar for the three models; correlation coefficients among the three models were .99, and mean differences were very small. These results indicated that there were no relative differences in estimating item difficulty parameters. Mean differences showed the same results as correlations (See Table 8). Figure 8, 9, and 10 present scatter plots for the item difficulty parameters estimates. Correlations and mean differences of the pseudo-chance parameter estimates revealed almost the same results as those of the item difficulty parameters (See Table 8). Scatter plots of the pseudo-chance parameters are displayed in Figure 11, 12, and 13.

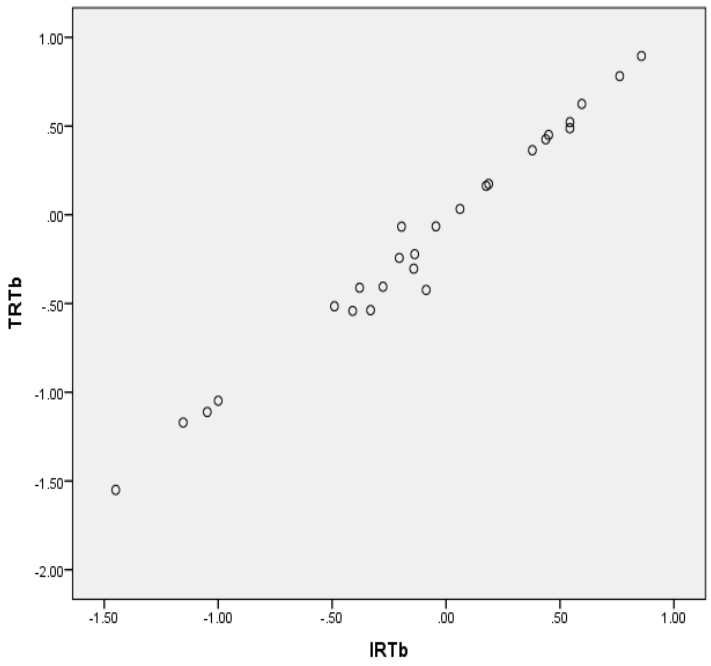

Figure 8. Scatter plot of the estimated item difficulty parameters from the 3-plm and the testlet-effects model.

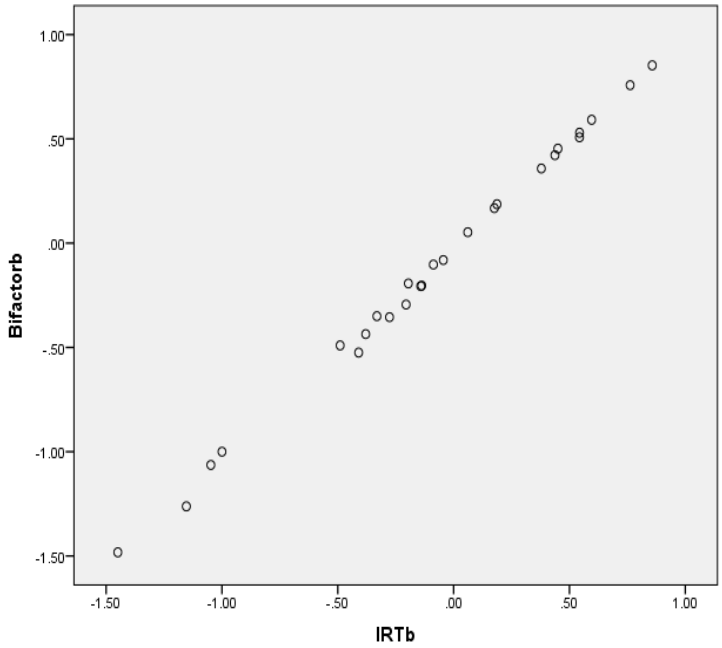

Figure 9. Scatter plot of the estimated item difficulty parameters from the 3-plm and the bi-factor model.

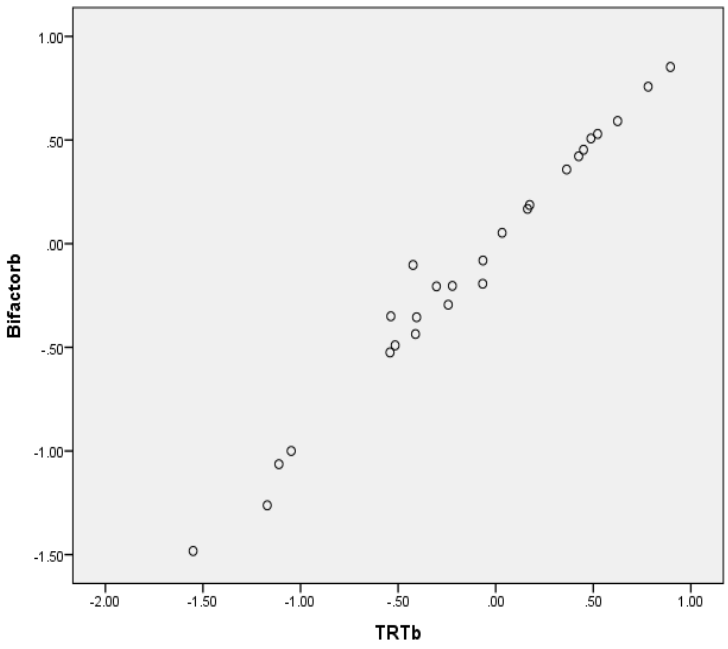

Figure 10. Scatter plot of the estimated item difficulty parameters from the testlet-effects model and the bi-factor model.

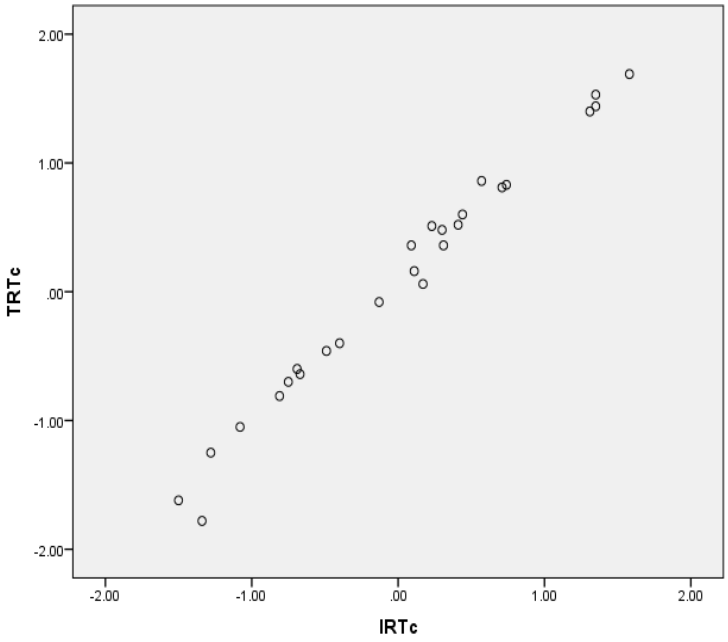

Figure 11. Scatter plot of the pseudo-chance parameter estimates from the 3-plm and the testlet-effects model.

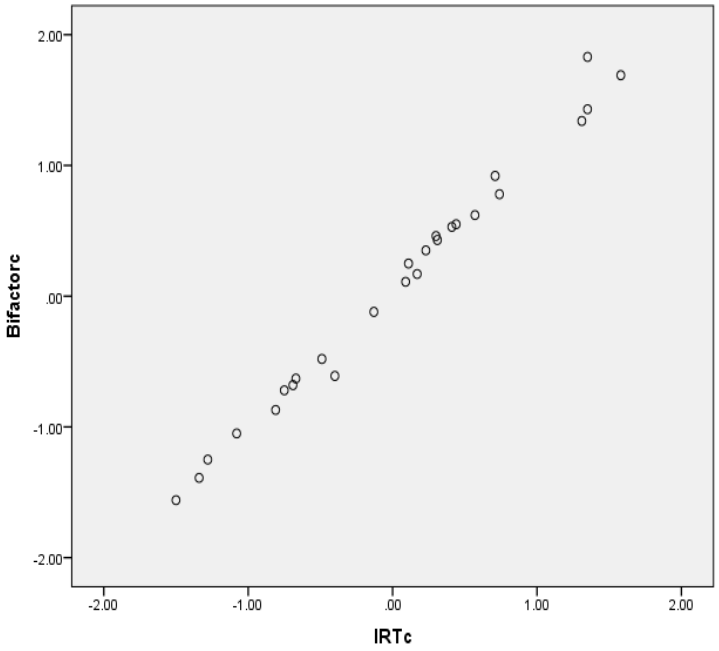

Figure 12. Scatter plot of the pseudo-chance parameter estimates from the 3-plm and the bi-factor model.

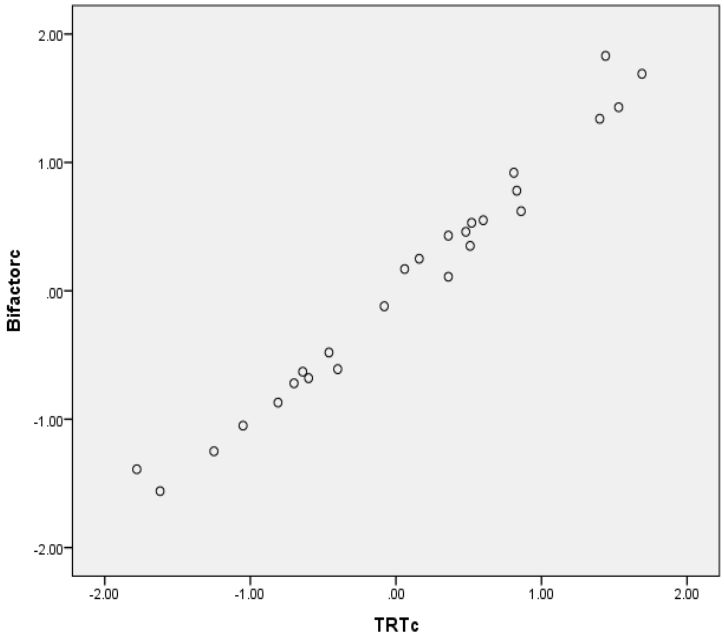

Figure 13. Scatter plot of the pseudo-chance parameter estimates from the testlet-effects model and the bi-factor model.

In short, according to the comparisons of the item and person parameters from the three different models, there were no significant differences in parameter estimations. The reason might be that the test for the present study had fewer numbers of items for each passage than those from the previous TRT studies (Wainer, Bradlow, & Wang, 2007). The test used for the present study had only two to five items per passage, while other TRT studies (e.g., Jiao et al., 2012) consisted of nine items for each passage. In addition, the large sample size of the present study (n = 8,815) might lead to the results of less significant differences in standard errors of item and person parameters. Nevertheless, the testlet-effects model revealed more stable standard errors of the item discrimination parameters than the 3-plm. Moreover, the 3-plm showed a

slight underestimation of the item discrimination parameters when compared to the bi-factor model.

Item Analyses from the TRT Models

The strengths and weaknesses of each item were evaluated based on the item discrimination, item difficulty, and pseudo-chance parameters (i.e., *a*-, *b*-, and *c*-parameters) from the TRT models (i.e., the testlet-effects model and the bi-factor model) because one of the advantages of IRT including TRT is that IRT provides invariant item indices (Hambleton & van der Linden, 1982; Kim & Nicewander, 1993). According to the values of *a*-parameter from the testlet-effects model and the bi-factor model, items 34 and 16 showed the highest discriminating power (a = 2.85 and a_1 = 3.27, respectively). Item 34 was a vocabulary question, and item 16 was a question requiring comprehension ability for literal details. Based on the values of *a*-parameter from the two TRT models, items 20 and 21 were evaluated as overall good items with high discriminating power (a = 2.45 and 2.82 from the testlet-effects model; a_1 = 3.09 and 2.86 from the bi-factor model, respectively), indicating that the two items added useful information to the test. The value of *a*-parameter influences the amount of item information, which contributes to the person parameter estimation at each level of theta. Item 20 was a question about text structure, and item 21 asked about the main ideas of the passage. These two items shared the same passage 5. The format of passage 5 was an informational science speech about Caulder woods. Items 8, 26, and 27 also revealed comparatively high values of *a*-parameter. Item 8 asked point of view and knowledge about passage 2 (narrative story). Items 26 and 27 shared the same passage 7 of an informational science article, requiring comprehension ability of context-related vocabulary and literal information.

On the other hand, item 4 revealed the lowest discriminating power from the testlet-effects model and the bi-factor model (a = 0.90 and a_1 = 0.88, respectively). The low value of a-parameter indicates that this item does not add useful information to the test. Moreover, an item with less information such as item 4 indicates that the item is less useful for assessing a person's ability. Item 4 from passage 1 was a question requiring comprehension ability for summarizing. Passage 1 was an informational speech about election. Items 7 and 18 also showed low discriminating power from the testlet-effects model and the bi-factor model (a = 0.96 and a_1 = 1.01, and a = 0.85 and a_1 = 1.07, respectively). Both of the two items required inferential ability for reading comprehension. Passages for items 7 and 18 were a narrative story and a narrative drama, respectively.

According to the b-parameters from the testlet-effects model and the bi-factor model, item 10 showed the lowest value (b = -1.55 and -1.48, respectively), indicating that this item was the easiest one among the 26 items of the test. Item 10 was a question requiring reading comprehension abilities for inference and organization. This item came from passage 3, which talked about how to help in the community. The Flesch Kincaid Grade Level for passage 3 was 5, which was much lower than the original level of the test (i.e., seventh grade). Based on the text characteristics from Coh-Metrix, the text of passage 3 was high in syntactic simplicity (z = 0.98), indicating that this text has simple sentence structures which lead to an easier process for reading comprehension. This text also had high word concreteness (z = 0.72), meaning that it had easier words to visualize and comprehend. However, this text had low referential cohesion (z = -1.4), showing little overlap in words and ideas between sentences.

Item 6 revealed the highest value of the *b*-parameter from the testlet-effects model and the bi-factor model (b = 0.90 and 0.85, respectively), meaning that this item was the most difficult one among the 26 items. Item 6 required comprehension ability for the author's craft and theme regarding passage 2. This passage was a narrative story with a Flesch Kincaid grade level of 5.3, indicating that item 6 was difficult even though it came from an easy passage. Item 28 showed the second highest value of the *b*-parameter from the testlet-effects model and the bi-factor model (b = 0.78 and 0.76, respectively). According to the Coh-Metrix result, passage 7 for item 28 was difficult with a Flesch Kincaid Grade level of 8.5. The text of passage 7 had low referential cohesion (z = -0.69), requiring reader's ability for making inferences. However, it was high in deep cohesion (z = 3.58) with more connecting words to help clarify the relationships between events, ideas, and information. Table 10 reports item parameters from the TRT models.

Table 10

Item Parameters from the TRT Models

	Testlet-effects model			Bi-factor model			
Item	*a*	*b*	*c*	a_1	a_2	*b*	*c*
1	1.26	-1.11	1.40	1.26	0.38	-1.06	1.34
2	1.61	-0.52	0.83	1.59	0.60	-0.49	0.78
3	1.23	-1.17	1.44	1.45	1.42	-1.26	1.83
4	0.90	-0.07	0.06	0.88	0.18	-0.19	0.17
6	1.81	0.90	-1.62	1.83	0.68	0.85	-1.56
7	0.96	-0.54	0.52	1.01	0.56	-0.52	0.53
8	2.43	0.03	-0.08	2.29	0.65	0.05	-0.12
10	1.09	-1.55	1.69	1.14	0.55	-1.48	1.69
13	1.46	-1.05	1.53	1.43	0.55	-1.00	1.43
15	1.97	-0.41	0.81	2.11	1.07	-0.44	0.92
16	2.30	0.17	-0.40	3.27	1.96	0.19	-0.61
17	1.76	0.36	-0.64	1.76	0.24	0.36	-0.63
18	0.85	-0.42	0.36	1.07	0.28	-0.10	0.11
19	1.48	-0.24	0.36	1.46	0.41	-0.29	0.43
20	2.45	-0.07	0.16	3.09	1.53	-0.08	0.25
21	2.82	0.16	-0.46	2.86	0.54	0.17	-0.48
22	1.34	0.52	-0.70	1.36	0.28	0.53	-0.72
25	1.80	0.45	-0.81	1.92	0.85	0.45	-0.87
26	2.16	-0.22	0.48	2.26	0.91	-0.20	0.46
27	2.47	0.43	-1.05	2.49	0.75	0.42	-1.05
28	1.60	0.78	-1.25	1.65	0.46	0.76	-1.25
29	1.48	-0.41	0.60	1.55	0.51	-0.35	0.55
30	1.23	0.49	-0.60	1.34	0.73	0.51	-0.68
31	1.68	-0.30	0.51	1.70	0.73	-0.21	0.35
32	1.60	-0.54	0.86	1.77	0.40	-0.35	0.62
34	2.85	0.62	-1.78	2.35	0.40	0.59	-1.39

Note. *a*: *a*-parameter, *b*: *b*-parameter, *c*: *c*-parameter, a_1: *a*-parameter for the primary latent trait, a_2: *a*-parameter for the secondary latent trait.

CHAPTER FIVE: DISCUSSION

As the nature of reading comprehension is extremely complex, reading comprehension assessments are also complicated (Kendeou et al., 2012). The general assumption of reading comprehension assessments is that reading comprehension tests being used in real school settings are measuring the same construct of reading comprehension (Keenan et al., 2008). However, the reality is that test scores from different tests show variations, indicating a validity problem in reading comprehension assessments (Keenan & Meenan, 2014; Nation & Snowling, 1997). The IRT may be an alternative to the traditional CTT to rectify the validity problem. Specifically, the application of the TRT model may be required for testlet-based reading comprehension assessments because the TRT model was designed to resolve the problem of local item dependence due to testlets (Baldwin, 2007; Ip, 2010).

The present study applied a more recent psychometric theory, the TRT, along with the traditional IRT to a large data set from a testlet-based reading comprehension assessment for seventh graders. It demonstrated the importance of the TRT model for a testlet-based reading comprehension assessment. First, the fit of the models was evaluated via -2LL, AIC, and BIC. Those three goodness-of-fit indices demonstrated the best model-data fit was the bi-factor model rather than the 3-plm and the testlet-effects model for this 8-testlet reading comprehension test. The second best fitting model was the testlet-effects model. These results indicate the possible advantage of the TRT models for the testlet-based assessments compared to the more traditional 3-plm IRT models. The advantages of the IRT models can be attained only when the model fits the

given test data (Hambleton et al., 1991). Therefore, in IRT applications, it is fundamentally necessary to compute the goodness-of-fit indices of particular IRT models in order to determine the best-fit model among the various models. The application of the TRT model to a testlet-based comprehension assessment is crucial for constructing tests with good test items (Hambleton et al., 1991).

Second, local item dependencies within an 8-testlet test were investigated, and high dependencies were found within the test when the 3-plm was applied. Although initial dependencies indicated problematic item structure, the application of the TRT models including the testlet-effects model and the bi-factor model resolved the severe problem of item dependence. It demonstrated the usefulness of the TRT model for testlet-based reading comprehension assessments. Specifically, when the bi-factor model was applied, there no longer appeared to be any local dependence.

According to the results of local item dependencies, items 15 and 16 had the highest dependencies and were based on the same passage of drama. The items required the comprehension skills for vocabulary and literal details. The second highest dependent pair was items 2 and 3, which also required the comprehension skill of vocabulary around the same passage. Students were required to select the best meaning which was related to the context of the passage, which is one important component of the reading comprehension skills. That they were highly dependent makes sense given they were evoking the same skill set from the same passage. It is notable that a total of 10 items among 16 item pairs which showed local item dependence were vocabulary questions, which were associated with the same passage contexts. Those results based on the second research question may be aligned with research results claiming that context-

related vocabulary is a significant predictor of higher levels of reading comprehension ability (Cain & Oakhill, 2007; Duke & Carlisle, 2011).

Third, the parameter estimates of the three models were compared. Even though the results did not show significant differences in parameter estimations, they showed more stable standard errors of the item discrimination parameters when the testlet-effects model was applied instead of the 3-plm. This result is aligned with that of Jiao et al. (2012)'s study. Jiao et al. found differences in the estimated standard errors of the parameters, but not in item and person parameter estimations when they compared models for testlet-based assessments. Additionally, there appeared a slight underestimation of the item discrimination parameters under the locally dependent condition, demonstrating the usefulness of the TRT models for a testlet-based reading comprehension test.

Finally, based on the item parameters from the TRT models (i.e., the testlet-effects model and the bi-factor model), strengths and weaknesses of each item were evaluated. Items 34, 16, 20, 21, 8, 26, and 27 were good items with high values of *a*-parameter. Those items have great power for discriminating a person's ability because the value of *a*-parameter influences the amount of item information. Items 4, 7, and 18 were evaluated as bad items with low values of *a*-parameter, showing less power for assessing a person's ability. Two of the three bad items were questions which tested inferential skills for reading comprehension, indicating that discriminating a person's ability for making inferences may be complex and difficult. These results of *a*-parameter from the TRT models contribute to item and test development of reading comprehension.

Regarding *b*-parameters, when compared to the Coh-Metrix results for each passage, the results were not consistent in that some items were difficult even though they came from an easy passage and some items from a difficult passage were evaluated as easy items based on the TRT results. According to the CCSS which were established to provide adolescent students with higher-level instruction of reading comprehension for college and career readiness, text complexity is crucial for selecting the appropriate text levels for students. The problem is that the results of text complexity based on both the unidimensional (i.e., the Flesch-Kincaid Grade Level) and the multidimensional (i.e., narrativity, syntactic simplicity, word concreteness, referential cohesion, and deep cohesion) analyses were different from those of the TRT models (i.e., the values of item difficulty parameters): easy items from a difficult passage or difficult items from an easy passage.

Those results support the reality that designing reading comprehension assessments is a challenging process because the nature of reading comprehension is extremely complex (Kendeou et al., 2012). Text complexity is only one feature that could impact why items for a passage may exhibit a lack of independence. Other passage-level features may also be important to study. For example, background knowledge is known to impact a person's understanding of a text. As mentioned earlier, background knowledge is a crucial component for deeper comprehension (Cain & Oakhill, 2007; Edmonds et al., 2009) and might be one of the testlet effects in a reading comprehension test which contribute to the local item dependencies (DeMars, 2006). If some passages have more familiar background knowledge for students, the items for that passage may be easier, resulting in local item dependence.

For future study, further work using TRT may need to be done because the TRT models take into account the testlet effects such as background knowledge, which speaks to how comprehension construct is operationalized and should be considered in comprehension test development. The application of the TRT model to comprehension assessments provides test developers with item-based information such as the values of *b*-parameters. Therefore, the TRT models may be crucial for constructing reading comprehension tests because they offer item-based information for selecting appropriate text levels for students in addition to information regarding text complexity from Coh-Metrix.

The testlet format is often necessary for reading comprehension tests that use passages. Choosing the best model for the given data is crucial for more accurate assessment. The IRT including TRT has many theoretical and practical advantages over CTT *only* if there is a good model-data fit. Nevertheless, a proper analysis of testlet effects has not been widely performed (Eckes, 2014). Testlet-based assessments, particularly in the area of language and literacy studies, have been analyzed based on the traditional CTT or IRT models which neglect local dependencies within testlets.

The present study may provide replicable evidence that the TRT model better fits the testlet-based reading comprehension assessment than does the traditional IRT model. It revealed the same results as previous TRT studies (DeMars, 2006, 2012; Min & He, 2014) in terms of model comparisons using various indices, preferring the TRT models to the traditional IRT model. However, estimations of the item and person parameters and their standard errors showed variations. Nonetheless, this study revealed more stable standard errors when the TRT model was applied and

underestimation of the item discrimination parameters when the 3-plm was applied, which aligns with the studies of Jiao et al. (2012) and DeMars (2006), respectively.

To conclude, the TRT model may be highly recommended for reading comprehension assessment so that test evaluators/developers can minimize biases in analyzing the test and individual test items. Because language assessment in general and reading assessment specifically are so complicated and in many cases testlet based, it is critical to have accurate assessment to utilize analysis procedures that honor the structure of the data and the construct being measured. Specifically, accurate estimations of item discrimination and item difficulty parameters in addition to the analyses of text complexity from Coh-Metrix may prevent misinforming practitioners, educators, and decision makers on item and test selections. Additionally, it may be recommended to adopt the method of the Computer-Administered Test (CAT) using item banks such as graduate record examination (GRE), which could be another practical advantage of IRT and TRT because the CAT is an individualized test and fewer items are needed to estimate examinees' abilities when IRT is applied. The TRT model may be a better alternative method for testlet-based assessments. Commonly available IRT software should incorporate analysis features in order to make it possible to implement the TRT models in measures that are inherently testlet-based (e.g., reading comprehension tests).

Limitations and Recommendations for Future Study

The present study, however, has several limitations that warrant further investigation. It used only one reading comprehension test for data analyses. It might be useful to compare the TRT model with the traditional IRT model using various reading comprehension tests such as the Gates-MacGinitie Reading Tests (MacGinitie,

MacGinitie, Maria, Dreyer, & Hughes, 2000) and the Woodcock-Johnson Reading Tests (Woodcock, McGrew, & Mather, 2001) for future study.

In this study, each passage contained only two to five items, which might affect the algorithm performance of IRTPRO software. Data analyses using a test with more items per passage might be required for future study. In addition, a smaller sample size than that of the present study (i.e., between 200 and 300) might lead to more solid results for the future testlet-based assessments; at least 200 examinees are needed to examine the fit of a 1-plm, 2-plm, or 3-plm (Hambleton & Swaminathan, 1985), and too large of a sample size like the present study (n = 8,815) may influence the stability of the estimated standard errors of parameters, resulting in no significant differences in the estimated standard errors among different models. For a smaller sample size, it could be recommended to select a random sample from an original large sample (n = 8,815). The final limitation of the present study is that the IRTPRO software does not provide values of the random testlet-effect parameter, which is crucial for investigating the testlet effects of each testlet.

Regardless of these limitations, this study contributed to reading comprehension assessment by showing advantages of the TRT model based on an adolescent reading comprehension test of the United States. Minimized biases of item analyses on the basis of invariant item parameters from the TRT model may lead to more accurate selections and construction for test items. It is recommended for test developers to apply the TRT models in constructing the reading comprehension passages and items with the appropriate level of text complexity and item difficulty for test-takers. They could construct good items based on the values of the a-parameters. They could also construct

the same levels of items as the passage levels using the values of the *b*-parameters in addition to the information from Coh-Metrix.

REFERENCES

Allen, M. J., & Yen, W. M. (2002). *Introduction to measurement theory.* Long Grove, IL: Waveland Press, Inc.

Baldonado, A. A., Svetina, D., & Gorin, J. (2015). Using necessary information to identify item dependence in passage-based reading comprehension tests. *Applied Measurement in Education, 28*(3), 202-218. doi:10.1080/08957347.2015.1042154

Baldwin, S. G. (2007). A review of testlet response theory and its applications. *Journal of Educational and Behavioral Statistics, 32*(3), 333-336. doi:10.3102/1076998607305834

Bell, S. M., & McCallum, R. S. (2008). *Handbook of reading assessment.* Boston, MA: Pearson Inc.

Betjemann, R. S., Keenan, J. M., Olson, R. K., & DeFries, J. C. (2011). Choice of reading comprehension test influences the outcomes of genetic analyses. *Scientific Studies of Reading, 15*(4), 363-382.

Bowyer-Crane, C., & Snowling, M. J. (2005). Assessing children's inference generation: What do tests of reading comprehension measure? *British Journal of Educational Psychology, 75*, 189-201. doi:10.1348/000709904X22674

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64*, 153-168.

Cain, K., & Oakhill, J. (2006). Assessment matters: Issues in the measurement of reading comprehension. *British Journal of Educational Psychology, 76*, 697-708. doi:10.1348/000709905X69807

Cain, K., & Oakhill, J. (2007). Reading comprehension difficulties: Correlates, causes, and consequences. In K. Cain & J. Oakhill (Eds.), *Children's comprehension problems in oral and written language: A cognitive perspective* (pp. 41–75). New York: Guilford Press.

Campbell, J. R. (2005). Single instrument, multiple measures: Considering the use of multiple item formats to assess reading comprehension. In S. G. Paris & S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp. 347-368). Mahwah, NJ: Lawrence Erlbaum Associates.

Catts, H. W., Hogan, T. P., & Adolf, S. M. (2005). Developmental changes in reading and reading disabilities. In H. W. Catts & A. G. Kamhi (Eds.), *The connections between language and reading disabilities.* Mahwah, NJ: Erlbaum.

Chen, W., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*(3), 265-289.

Common Core State Standards (2014). *English language arts & literacy in history/social studies, science, and technical subjects*. Retrieved from http://www.corestandards.org/assets/CCSSI_ELA%20Standards.pdf.

Compton, D. L., Miller, A. C., Elleman, A. M., & Steacy, L. M. (2014). Have we forsaken reading theory in the name of "quick fix" interventions for children with reading disability? *Scientific Studies of Reading, 18*(1), 55-73.

Cutting, L. E., & Scarborough, H. S. (2006). Prediction of reading comprehension: Relative contributions of word recognition, language proficiency, and other cognitive skills can depend on how comprehension is measured. *Scientific Studies of Reading, 10*(3), 277-299.

DeMars, C. E. (2006). Application of the bi-factor multidimensional item response theory model to testlet-based tests. *Journal of Educational Measurement, 43,* 145-168. doi:10.1111/j.1745-3984.2006.00010.x

DeMars, C. E. (2012). Confirming testlet effects. *Applied Psychological Measurement, 36*(2), 104-121. doi:10.1177/0146621612437403

Duke, N. K. (2005). Comprehension of what for what: Comprehension as a nonunitary construct. In S. G. Paris & S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp. 93-104). Mahwah, NJ: Lawrence Erlbaum Associates, Inc. doi:10.4324/9781410612762

Duke, N., & Carlisle, J. (2011). The development of comprehension. In M. L. Kamil, P. D. Pearson, E. B. Moje, & P. P. Afflerbach (Eds.), *Handbook of reading research* (pp. 199-228). New York, NY: Routledge.

Eckes, T. (2014). Examining testlet effects in the TestDaF listening section: A testlet response theory modeling approach. *Language Testing, 31*(1), 39-61. doi:10.177/0265532213492969

Edmonds, M. S., Vaughn, S., Wexler, J., Reutebuch, C., Cable, A., Tackett, K. K., & Schnakenberg, J. W. (2009). A synthesis of reading interventions and effects on reading comprehension outcomes for older struggling readers. *Review of Educational Research, 79*(1), 262-300. doi:10.3102/0034654308325998

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

Fletcher, J. M., Foorman, B. R., Boudousquie, A., Barnes, M. A., Schatschneider, C. S., & Francis, D. J. (2002). Assessment of reading and learning disabilities a research-based intervention-oriented approach. *Journal of School Psychology, 40*(1), 27-63.

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

Hagley, F. (1987). *Suffolk reading scale*. Windsor: NFER-Nelson.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer Nijhoff.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory, 2*. Newbury Park, CA: Sage.

Hambleton, R. K., & van der Linden, W. J. (1982). Advances in item response theory and applications: An introduction. *Applied Psychological Measurement, 6,* 373-378. doi:10.1177/014662168200600401

Hannon, B., & Daneman, M. (2001). A new tool for measuring and understanding individual differences in the component processes of reading comprehension. *Journal of Educational Psychology, 93*(1), 103-128. doi:10.1037/0022-0663.93.1.103

Hiebert, E. H., & Mesmer, H. A. (2013). Upping the ante of text complexity in the Common Core State Standards: Examining its potential impact on young readers. *Educational Researcher, 42*(1), 44-51.

Ip, E. H. (2010). Interpretation of the three-parameter testlet response model and information function. *Applied Psychological Measurement, 34*(7), 467-482.

Jiao, H., Kamata, A., Wang, S., & Jin, Y. (2012). A multilevel testlet model for dual local dependence. *Journal of Educational Measurement, 49*(1), 82-100.

Kamhi, A. G., & Catts, H. W. (2017). Epilogue: Reading comprehension is not a single ability-Implications for assessment and instruction, *Language, Speech, and Hearing Services in Schools, 48*(2), 104-107. doi:10.1044/2017_LSHSS-16-0049

Keenan, J. M., Betjemann, R. S., & Olson, R. K. (2008). Reading comprehension tests vary in the skills they assess: Differential dependence on decoding and oral comprehension, *Scientific Studies of Reading, 12*(3), 281-300. doi:10.1080/10888430802132279

Keenan, J. M., & Meenan, C. E. (2014). Test differences in diagnosing reading comprehension deficits. *Journal of Learning Disabilities, 47*(2), 125-135. doi:10.1177/0022219412439326

Kendeou, P., Papadopoulos, T. C., & Spanoudis, G. (2012). Processing demands of reading comprehension tests in young readers. *Learning and Instruction, 22*(5), 354-367. doi:10.1016/j.learninstruc.2012.02.001

Kim, J. K., & Nicewander, W. A. (1993). Ability estimation for conventional tests. *Psychometrika, 58*(4), 587-599. doi:10.1007/BF02294829

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

Kintsch, W., & Kintsch, E. (2005). Comprehension. In S. G. Paris & S. A. Stahl (Eds.), *Current issues in reading comprehension and assessment* (pp. 71-92). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Kintsch, W., & Rawson, K. A. (2005). Comprehension. In M. J. Snowling & C. Hulme (Eds.), *The science of reading: A handbook* (pp. 209-226). Malden, MA: Blackwell. doi:10.1002/9780470757642.ch12

Klare, G. R. (1974). Assessing readability. *Reading Research Quarterly, 10*, 62-102.

Koslin, B. I., Zeno, S., & Koslin, S. (1987). *The DRP: An effective measure in reading.* New York, NY: College Entrance Examination Board.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211-240.

Li, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement, 30*(1), 3-21. doi:10.1177/0146621605275414

Lord, F. (1952). *A theory of test scores.* Richmond, VA: Psychometric Corporation.

MacGinitie, W. H., MacGinitie, R. K., Maria, K., Dreyer, L. G., & Hughes, K. E. (2000). *Gates-MacGinitie reading tests fourth edition.* Itasca, IL: Riverside.

Marais, I., & Andrich, D. (2008). Formalizing dimension and response violations of local independence in the unidimensional Rasch model. *Journal of Applied Measurement, 9,* 200-215.

McKenna, M. C., & Stahl, K. A. (2008). *Assessment for reading instruction.* New York, NY: Guilford Press.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix.* New York, NY: Cambridge University Press.

McNamara, D. S., Louwerse, M. M., Cai, Z., & Graesser, A. (2005). *Coh-Metrix version 1.4.* Retrieved from http//:cohmetrix.memphis.edu.

Min, S., & He, L. (2014). Applying unidimensional and multidimensional item response theory models in testlet-based reading assessment. *Language Testing, 31*, 453-477. doi:10.1177/0265532214527277

Nation, K., & Snowling, M. (1997). Assessing reading difficulties: The validity and utility of current measures of reading skill. *British Journal of Educational Psychology, 67*, 359-370.

National Center for Education Statistics. (2015). *The nation's report card: 2015 mathematics and reading (NCES 2015-144)*. Washington, DC: Institute of Education Sciences, U. S. Department of Education.

Neale, M. D. (1989). *The Neale analysis of reading ability-revised.* Windsor: NFER.

Pearson, P. D., & Hamm, D. N. (2005). The assessment of reading comprehension: A review of practices—past, present, and future. In S. G. Paris & S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp. 13-69). Mahwah, NJ: Erlbaum.

Perfetti, C., Landi, & Oakhill. (2005). The acquisition of reading comprehension skill. In M. J. Snowling & C. Hulme (Eds.), *The science of reading: A handbook* (pp. 227-247). Malden, MA: Blackwell. doi:10.1002/9780470757642.ch13

Perfetti, C., & Stafura, J. (2014). Word knowledge in a theory of reading comprehension. *Scientific Studies of Reading, 18,* 22-37. doi:10.1080/10888438.2013.827687

RAND Reading Study Group. (2002). *Reading for understanding: Toward an R & D program in reading comprehension.* Santa Monica, CA: RAND.

Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer. doi:10.1007/978-0-387-89976-3

Rijmen, F. (2010). Formal relations and an empirical comparison among the bi-factor, the testlet, and a second-order multidimensional IRT model. *Journal of Educational Measurement, 47*(3), 361-372.

Rimrodt, S., Lightman, A., Roberts, L., Denckla, M. B., & Cutting, L. E. (2005). *Are all tests of reading comprehension the same?* Poster presentation at the annual meeting of the International Neuropsychological Society, St. Louis, MO.

Schroeders, U., Robitzsch, A., & Schipolowski, S. (2014). A comparison of different psychometric approaches to modeling testlet structures: An example with C-tests. *Journal of Educational Measurement, 51*(4), 400-418.

Shaywitz, S. E. (2003). *Overcoming dyslexia: A new and complete science-based program for reading problems at any level.* New York, NY: Alfred A. Knopf.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*(7), 4-14.

Spearman, C. (1904). "General intelligence," objectively determined and measured. *American Journal of Psychology, 15*, 201-293. doi:10.2307/1412107

Stahl, S., & Nagy, W. (2006). *Teaching word meanings*. Mahwah, NJ: Lawrence Erlbaum Associates.

Stenner, A. J. (2006). *Measuring reading comprehension with the Lexile framework.* Dufham, NC: Metametrics, Inc.

Sweet, A. P. (2005). Assessment of reading: The RAND Reading Study Group vision. In S. G. Paris & S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp. 3-12). Mahwah, NJ: Erlbaum.

Tao, J., Xu, B., Shi, N., & Jiao, H. (2013). Refining the two-parameter testlet response model by introducing testlet discrimination parameters. *Japanese Psychological Research, 55*(3), 284-291. doi:10.1111/jpr.12002

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. Washington, DC: American Psychological Association.

Wainer, H., Bradlow, E. T., & Du, Z. (2000). Testlet response theory: An analog for the 3PL model useful in testlet-based adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 245-269). Dordrecht, Netherlands: Kluwer.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications.* New York, NY: Cambridge University Press. doi:10.1017/CBO9780511618765

Wainer, H., & Wang, X. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement, 37*(3), 203-220.

Wang, X., Bradlow, E. T., & Wainer, H. (2002). A general Bayesian model for testlets: Theory and applications. GRE board professional report No. 98-01P.

Wang, W. C. & Wilson, M. (2005). Exploring local item dependence using a random-effects facet model. *Applied Psychological Measurement, 26*, 109-128.

Wechsler, D. L. (1990). *Wechsler objective reading dimensions.* London: Psychological Press.

Wechsler, D. L. (1992). *Wechsler individual achievement test.* San Antonio, TX: Psychological Corporation.

Wiederholt, L., & Bryant, B. (1992). *Examiner's manual: Gray oral reading test-3.* Austin, TX: Pro-Ed.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III tests of achievement.* Itasca, IL: Riverside.

Yao, L., Rich, C., & McGraw-Hill, C. (2008). *Application of testlet-effect models to scaling performance assessments of mixed item types with multiple-criteria scoring rubrics.* Paper presented at the annual meetings of the National Council on Measurement in Education, March 23-27, New York.

Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (2002). Identification and evaluation of local item dependencies in the medical college admission test. *Journal of Educational Measurement, 39*(4), 291-309.

APPENDIX

APPENDIX A: IRB APPROVAL LETTER

IRB
INSTITUTIONAL REVIEW BOARD
Office of Research Compliance,
010A Sam Ingram Building,
2269 Middle Tennessee Blvd
Murfreesboro, TN 37129

IRBN007 – EXEMPTION DETERMINATION NOTICE

Friday, February 17, 2017

Investigator(s): Weon Kim; Dr. Amy Elleman
Investigator(s') Email(s): Weon.Kim@mtsu.edu; Amy.Elleman@mtsu.edu
Department: Literacy Studies

Study Title: Application of the IRT and TRT Models to a Reading Comprehension Test
Protocol ID: **17-1159**

Dear Investigator(s),

The above identified research proposal has been reviewed by the MTSU Institutional Review Board (IRB) through the **EXEMPT** review mechanism under 45 CFR 46.101(b)(2) within the research category *(4) Study involving existing data* A summary of the IRB action and other particulars in regard to this protocol application is tabulated as shown below:

IRB Action	EXEMPT from furhter IRB review***
Date of expiration	**NOT APPLICABLE**
Participant Size	Exisiting Data
Participant Pool	**Existing data provided by Discovery**

Mandatory Restrictions	Annonymous, de-identified; randomized data covered under provided permission letter on file with the Research Compliance Office	
Additional Restrictions	**None**	
Comments	None	
Amendments	**Date**	**Post-Approval Amendments** None at this time

***This exemption determination only allows above defined protocol from further IRB review such as continuing review. However, the following post-approval requirements still apply:

- Addition/removal of subject population should not be implemented without IRB approval
- Change in investigators must be notified and approved
- Modifications to procedures must be clearly articulated in an addendum request and the proposed changes must not be incorporated without an approval
- Be advised that the proposed change must comply within the requirements for exemption
- Changes to the research location must be approved – appropriate permission letter(s) from external institutions must accompany the addendum request form
- Changes to funding source must be notified via email (irb_submissions@mtsu.edu)
- The exemption does not expire as long as the protocol is in good standing

- Project completion must be reported via email (irb_submissions@mtsu.edu)
- Research-related injuries to the participants and other events must be reported within 48 hours of such events to compliance@mtsu.edu

The current MTSU IRB policies allow the investigators to make the following types of changes to this protocol without the need to report to the Office of Compliance, as long as the proposed changes do not result in the cancellation of the protocols eligibility for exemption:

- Editorial and minor administrative revisions to the consent form or other study documents
- Increasing/decreasing the participant size

The investigator(s) indicated in this notification should read and abide by all applicable postapproval conditions imposed with this approval. Refer to the post-approval guidelines posted in the MTSU IRB's website. Any unanticipated harms to participants or adverse events must be reported to the Office of Compliance at (615) 494-8918 within 48 hours of the incident.

All of the research-related records, which include signed consent forms, current & past investigator information, training certificates, survey instruments and other documents related to the study, must be retained by the PI or the faculty advisor (if the PI is a student) at the sacure location mentioned in the protocol application. The data storage must be maintained for at least three (3) years after study completion. Subsequently, the researcher may destroy the data in a manner that maintains confidentiality and anonymity. IRB reserves the right to modify, change or cancel the terms of this letter without prior notice. Be advised that IRB also reserves the right to inspect or audit your records if needed.

Sincerely,

Institutional Review Board

Middle Tennessee State University

Quick Links:

Click here for a detailed list of the post-approval responsibilities.
More information on exmpt procedures can be found here.