

MODEL COMPARISONS AMONG TESTLET RESPONSE THEORIES (TRT) ON A
READING COMPREHENSION TEST

by

Kyungtae Kim

A Dissertation Submitted to the
Faculty of the College of Graduate Studies at
Middle Tennessee State University
in Partial Fulfillment
of the Requirements for the Degree of
Doctorate of Philosophy
in Literacy Studies

Middle Tennessee State University
May 2015

Dissertation Committee:

Dr. Jwa K. Kim, Chair

Dr. Amy M. Elleman

Dr. Cyrille L. Magne

I dedicate this project to my family.

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my advisor, Dr. Jwa K. Kim, for his incredible guidance, care, encouragement, and patience. Without his persistent help, this dissertation would not have been possible. I would also like to thank my committee members, Dr. Amy Elleman and Dr. Cyrille Magne, for their suggestions and advice throughout the dissertation process. I would not have finished my dissertation without the dedicated guidance of my committee members.

I would like to thank my wife, Shinsil, and my two sons, Don and Kyu. They were always cheering me up and walked by me through every situation. Finally, I would like to thank Philip Ahn and Sophie Lee who provided help in every way they could.

I dedicate this dissertation to the almighty GOD.

If I rise on the wings of the dawn, if I settle on the far side of the sea, even there your hand will guide me, your right hand will hold me fast. Psalm 139:9-10

ABSTRACT

The purpose of this study was to evaluate the strengths and weaknesses of psychometric models such as Classical Test Theory (CTT), Item Response Theory (IRT), and Testlet Response Theory (TRT) as well as test items of a fifth grade reading comprehension test with a large data set ($N = 10,897$). The reading comprehension test contained 22 items with 7 passages along with 4 areas of reading standards of literature (RL), reading standards of informational text (RI), reading standards of foundation skills (RF), and language standards (L) of Common Core State Standards (CCSS). The 22-item showed a good internal consistency reliability index with the Cronbach's alpha of .79. The exploratory factor analysis (EFA) confirmed that the data could be analyzed with the traditional IRT analyses because the data showed a unidimensional solution. The model comparison criteria (-2LL, AIC, and BIC) revealed that the 3PLM was the best-fitting model for the data when compared with 1PLM and 2PLM. Comparisons of the results from CTT and 3PLM addressed the advantages of IRT over CTT with more item information (a , b , c -parameter estimates) along with detailed understandings of the item parameters for specific students' ability levels. The -2LL, AIC, BIC illustrated that local item dependence (LID) among test items was minimal in the 5th grade reading comprehension test so unidimensional IRT was more appropriate than the TRT models. However, several testlet variances from the generalized TRT model indicated that the testlet effects were not negligible ($\hat{\sigma}^2_{\gamma_4} = 0.18$, $\hat{\sigma}^2_{\gamma_5} = 0.19$, and $\hat{\sigma}^2_{\gamma_6} = 0.06$). The 3PLM, constrained TRT, and generalized TRT models provided consistent ability estimations with a mean of 0.00 and standard deviation of 1.00. Two item parameter estimates (a and

c-parameters) except the item difficulty parameter (b_i) were highly correlated among 3PLM and two TRT models. The *b*-parameters were associated with the estimated testlet mean. In this study, comparisons of psychometric models and test item parameters among CTT, IRT and TRTs on a reading comprehension test are meaningful for both researchers and practitioners to achieve the precise evaluation of a reading comprehension test.

Keywords: Psychometric models, CTT, IRT, TRT, item discrimination, item difficulty, and pseudo-chance parameter

TABLE OF CONTENTS

LIST OF TABLES	viii
LIST OF FIGURES	ix
CHAPTER ONE: INTRODUCTION	1
Theoretical models of reading comprehension	2
Assessment of reading comprehension	5
Psychometric theories	7
Purpose of the study.....	13
CHAPTER TWO: REVIEW OF THE LITERATURE	15
Reading comprehension models	15
Common Core State Standards	19
Assessment of reading comprehension	22
Classical Test Theory (CTT)	23
Item Response Theory (IRT)	28
Testlet Response Theory (TRT)	35
Research questions	39
CHAPTER THREE: METHODS	41
Participants	41
Measurement	41
Procedure	44
CHAPTER FOUR: RESULTS	47
CTT and IRT analysis results	47

Model comparisons of 3PLM and TRT models	59
Testlet effects	63
Comparisons of item parameters	68
CHAPTER FIVE: DISCUSSION	72
Limitation and recommendations for future research	77
REFERENCES.....	79
APPENDICES	85
APPENDIX A: 1PLM	86
APPENDIX B: 2PLM	87
APPENDIX C: 3PLM	88
APPENDIX D: 3PL CONSTRAINED TRT	89
APPENDIX E: 3PL GENERALIZED TRT	90
APPENDIX F: IRB APPROVAL	91

LIST OF TABLES

Table 1 Descriptive statistics of test items from CTT and estimated item parameters of 3PLM	48
Table 2 Eigenvalues of the correlation matrix	50
Table 3 Model-fit indices of three traditional IRT models	51
Table 4 Model-fit indices of 3PLM and two TRT models	60
Table 5 Estimated item parameters and testlet effects of constrained TRT model	66
Table 6 Estimated item parameters and testlet effects of generalized TRT model	67

LIST OF FIGURES

Figure 1 Item characteristic curve (ICC) for 1PLM	32
Figure 2 Item characteristic curve (ICC) for 2PLM	33
Figure 3 Item characteristic curve (ICC) for 3PLM	34
Figure 4 Scree plot of eigenvalues	51
Figure 5 Each four items for good and poor item discrimination parameters	53
Figure 6 Illustration of posterior PDFs of item discrimination parameters	54
Figure 7 Probability density functions of item difficulty parameters	56
Figure 8 Probability density functions of the pseudo-chance parameters	57
Figure 9 Iteration histories of item parameter estimates of 3PLM	58
Figure 10 Acceptance rates of MCMC iterations for 3PLM	59
Figure 11 Comparison of the person ability estimates between 3PLM and TRT models	61
Figure 12 Distribution of the person ability estimates from the IRT and TRT models	62
Figure 13 Difference score distributions of ability estimates from IRT and TRT models	63
Figure 14 Comparisons of the item discrimination estimates (a_1 and a_2) between 3PLM and TRT models	69

Figure 15 Comparisons of the item difficulty estimates between 3PLM and TRT models	70
Figure 16 Comparisons of the pseudo-chance parameter estimates between 3PLM and TRT models	71

CHAPTER ONE: INTRODUCTION

According to reports of the National Assessment of Educational Progress (NAEP), a total of 32% of students in 4th grade and 22% in 8th grade read below the basic level (National Center for Education Statistics, 2013). It is crucial that students read at their grade-level or better because without adequate reading comprehension skills, students may struggle in many subject areas including science, social studies, and math (Neufeld, 2005). Problems related to a lack of reading comprehension will not stop after graduating from high school; it will continue even in college. College students are asked to comprehend what they read on scientific journals, textbooks, or magazines (Pritchard, Wilson, & Yamnitz, 2007). However, various reports have confirmed that many college students rarely have the ability to comprehend increasing complexity in texts (Common Core State Standards, 2014; Heller & Greenleaf, 2007). Lower reading proficiency may significantly impact students' academic success, careers, and life in general.

Reading comprehension has been a major research topic in literacy. Durkin (1993) described that comprehension was the ultimate goal of all activities related to reading, which was an intentional manner in order to construct meaning of a text. Researchers have used different definitions for reading comprehension to emphasize various skills and activities. De Corte, Verschaffel, and Van De Ven (2001) defined reading comprehension as activities of understanding, interpreting, and constructing meaning through a variety of student-related, text-related, and environmental factors. Comprehension may be considered as a complex process by consisting of various language skills and activities rather than by a single construct. Phonological processes,

orthographic awareness, and oral language proficiency including vocabulary and grammatical knowledge have been considered to be main factors that affect reading comprehension.

Hoover and Gough (1990) conceptualized the simple view of reading (SVR), a theory that focused on decoding and linguistic comprehension as the primary factors that comprised reading comprehension. The skill of decoding is the ability to “read isolated words quickly, accurately, and silently” in terms of word recognition in alphabetic orthography (Gough & Tunmer, 1986, p.7). Another constituent in SVR is linguistic comprehension which consists of discourse skills that construct meaning from texts and monitor comprehension (Oakhill & Cain, 2011). Syntactic (grammatical structure) and semantic (meaning of vocabulary) skills are required to build linguistic comprehension (Mutter, Hulme, Snowling, & Stevenson, 2004). The strength of the relationship between decoding and reading comprehension or between linguistic comprehension and reading comprehension changes over time. For early grade school students, decoding is a more important skill for comprehension than linguistic comprehension. However, linguistic comprehension becomes more important to reading comprehension than decoding later in development, when a student’s decoding ability is not significantly different among students (Adlof, Catts, & Little, 2006; Curtis, 1980).

Theoretical models of reading comprehension

There are other models in reading in addition to SVR that provide more details about the process of reading comprehension. The construction-integration model

(Kintsch, 1998) and the sociocultural context theory are two additional major reading models. The construction-integration (C-I) model highlights the importance of two cognitive processes to attain deeper understanding of the text: Construction and integration (Kintsch, 1998). Construction is a stage of building several possible interpretations of text-based information. Integration is a stage that selects the most plausible interpretations based on prior knowledge or the reader's experience. In order to build higher-order comprehension, three levels of processing are required. The first one is the surface level known as decoding in SVR that recognizes words from the text. The second is at a text-based level which allows for a reader to access possible meanings of the text using a linkage through propositions. In this step, syntactic and semantic knowledge helps a reader form the coherent understanding of the sentence. The last level is to build inferences based on prior knowledge or experience. Prior knowledge plays a critical role in enhancing a deeper comprehension (Fisher & Frey, 2009; Stahl, Hare, Sinatra, & Gregory, 1991).

While a cognitive perspective on reading comprehension highlights a process of constructing and integrating meaning, sociocultural theory emphasizes the importance of interactions among three dimensions (the reader, text, and activity) beyond the importance of internal cognitive process to build a meaning of text. The RAND Reading Study Group (RRSG) also depicts the importance of interaction among these three elements for defining reading comprehension as the extracting and constructing processes (Snow & Sweet, 2003). According to Vygotsky (1978), children learn through interactions with parents, siblings, teachers, and their environment. In sociocultural

theory, motivation for reading and discussion in the classroom plays a crucial role in reading comprehension.

With many different models and theories of reading comprehension, it is difficult to unify reading comprehension skills and activities from these various models. Only some shared reading comprehension skills can be presented. Reading comprehension involves interpreting information from a written text using prior knowledge and constructing a coherent mental representation (Duke, 2005). Skilled readers have an automatic process of interpreting written text without having to decode words, understand sentences, make inferences, draw connections between the text and prior knowledge, and identify themes (Kendeou, van den Broek, White, & Lynch, 2007).

In order to describe the skills and knowledge for academic success as well as college and career readiness, Common Core State Standards (CCSS) as academic benchmarks have been established in 45 states ([www. corestandards.org](http://www.corestandards.org), 2013). Formative assessment may be a key component of the CCSS from Kindergarten through 12th grade, which contains grade-level expectations of complex skills and knowledge in English language arts (ELA) and math. The CCSS was developed in order to provide uniform standards across all states for students' readiness in the areas for English language arts as well as other subjects such as history, social sciences, and science. The ELA of CCSS focuses on text complexity and the developmental growth of reading comprehension. In this framework, reading comprehension tests should contain factors such as: Understanding key ideas and details, integrating knowledge and ideas, and constructing an author's craft and structure (CCSS, 2014). CCSS can help educators

develop instructional goals and objectives for students in each grade level. Any reading comprehension tests developed to test students' reading abilities can follow the structure and guidelines of CCSS.

Assessment of reading comprehension

Assessing reading comprehension based on different comprehension models offers many challenges due to the number of cognitive processes involved in the models, such as recognizing individual words, constructing meaning, activating prior knowledge, and generating inferences (Paris & Stahl, 2005). The accurate assessment of a construct is a vital step in research, diagnosis, and prediction in any field of study. Reading assessment is not an exception. The appropriate assessment of reading comprehension plays a key role in a broad purpose of educational planning and evaluation. In addition, assessment may help increase our understanding about the construct of reading comprehension and may provide resources for instructional decision-making for administrators, teachers, and parents. However, it may not be possible to fully describe the construct of reading comprehension because the construct itself may not be unidimensional and cannot be measured with 100% accuracy. Since some common skills of reading are shared by different types of comprehension tests, it is worth investigating conjoint aspects of reading comprehension with diverse assessment tools (Duke, 2005).

Since Binet (Binet & Simon, 1916; cited in Johnston, 1984) used reading comprehension test items in 1895 as a part of his intelligence quotient (IQ) battery, many researchers have dedicated themselves to develop reading comprehension assessments

with sound psychometric indices such as reliability, validity, and efficacy (Johnson & Pearson, 1975) as well as consistent assessments of reading comprehension constructs (Keenan, Betjemann, & Olson, 2008). However, there is still a lack of coherent and comprehensive research strategies investigating the shared skills of reading comprehension. Kintsch and Kintsch (2005) criticized different reading comprehension tests for their lack of understanding about theoretical process of reading comprehension. The RRSG also addressed several complaints in assessing reading comprehension; most comprehension assessments failed to evaluate the complexity of the target construct, reflect developmental sensitivity in reading comprehension, and address minimal criteria for reliability and validity (Sweet, 2005).

The demand for an accurate assessment of comprehension is high, especially due to various political and societal mandates such as the No Child Left Behind Act of 2001 (NCLB) and Response to Intervention (RTI). It is difficult to accurately and reliably measure students' reading skills with high standards in the assessment of reading comprehension (Pearson & Hamm, 2005; Sweet, 2005) because reading comprehension is a complex process with multidimensional constructs. Cutting and Scarborough (2006) found that unique contributions of word recognition and language proficiency varied across three reading comprehension subtests from the Gates-MacGinitie Reading Test-Revised (G-M; MacGinitie, MacGinitie, Maria, & Dreyer, 2000), the Gray Oral Reading Test-Third Edition (GORT-3; Wiederholt & Bryant, 1992), and Wechsler Individual Achievement Test (WIAT; Wechsler, 1992). Keenan, Betjemann, and Olson (2008) also demonstrated that the contributions of decoding and linguistic comprehension differed for

different types of reading comprehension tests such as the GORT, Qualitative Reading Inventory (QRI), Woodcock-Johnson Passage Comprehension subtest (WJPC), and Peabody Individual Achievement Test (PIAT). Various tests measure different aspects of reading comprehension skills even though most tests are developed to assess the same skill. Duke (2005) suggested that researchers and educators should make priorities in assessing reading comprehension because all students cannot be assessed by all domains of comprehension in practice.

Psychometric theories

In order to scientifically evaluate diverse reading comprehension assessment tools and prioritize different domains of reading comprehension constructs, one must utilize psychometric methods and indices such as reliability, validity, and other item and test statistics which may provide criteria to evaluate various reading comprehension tests. Most standardized assessments are considered valid and reliable measures because items are selected to maximize reliability and criterion validity (Carpenter & Paris, 2005). Although numerous reliability and validity indices have been proposed by different statisticians in educational assessments, these indices are based on the measurement of whole test following the Classical Test Theory (CTT). Thus, it is possible that a test with high reliability and validity index may contain several items with poor item characteristics. Although CTT has a few item indices such as the p -value (item difficulty index) and d -value (item discrimination index), the main purpose of CTT is test-oriented indices (e.g., reliability and validity, Hambleton & van der Linden, 1982).

CTT, developed by Spearman (1904), has dominated psychometrics for more than half the 20th century. However, CTT has both theoretical and practical issues.

Theoretically, observed scores are decomposed by two independent components, true-score and error-score. In this framework, the true-score component for an individual is defined simply in terms of the expected value of an individual's scores on the assessment that can be estimated through repeated assessments on the same instrument under identical conditions, which is impossible in practice. Thus, the CTT paradigm is a tautology which cannot be proved or disproved with empirical data (Hambleton & van der Linden, 1982). In addition, there are several practical shortcomings with CTT. Person statistics (i.e., ability or observed scores) are test-dependent, and item statistics (i.e., item difficulty and item discrimination) are dependent on sample groups. The parallel-test assumption and equal standard error of measurement assumption are other shortcomings of CTT, which are almost impossible to meet in practice (Lord, 1984). The final shortcoming is that CTT is test-oriented rather than item-oriented. A true score from a test does not provide any information on how examinees answer a given item.

Proposed by Lord (1952) and Birnbaum (1968) among others, item response theory (IRT) offered the possibility of resolving the shortcomings of CTT. One theoretical advantage of IRT over CTT is that falsifiable mathematical models can be determined through empirical data for various situations. IRT makes it possible to investigate the model-data fit with diverse stochastic distributions. Once the model-data fit is evaluated and an appropriate mathematical model is selected, then estimations for person and item parameters can be obtained through different estimation methods. The

IRT estimations will yield invariant person and item parameters across different tests and populations. IRT mainly focuses on the item-level estimation instead of test-level indices.

According to Hambleton, Swaminathan, and Rogers (1991):

IRT rests on two basic postulates: a) the performance of an examinee on a test item can be predicted (or explained) by a set of factors called traits, latent traits, or abilities; and b) the relationship between examinee's item performance and the set of traits underlying item performance can be described by a monotonically increasing function called an item characteristic function or item characteristic curve. This function specifies that as the level of the trait increases, the probability of a correct response increases. (p. 7)

There are two main streams in IRT models: Normal ogive models and logistic models.

Lord (1952) proposed three normal ogive models with 1, 2, and 3 parameters. Later, Birnbaum (1968) demonstrated that logistic models could compute very similar probabilities compared with normal ogive models. The logistic model is simpler in mathematical forms than the normal ogive models, so it is often preferred. The most general three-parameter logistic model (3PLM) can be expressed as,

$$P(Y_{ij} = 1) = c_j + (1 - c_j) \frac{1}{1 + e^{-1.7a_j(\theta_i - b_j)}}, \quad (1)$$

where $P(Y_{ij} = 1)$ is the probability that an examinee (i) with ability (θ) answers item j correctly, c_j is the lower-asymptote parameter known as the pseudo-chance parameter of item j , a_j is the item discrimination parameter, b_j is the item difficulty parameter, and 1.7 is a constant to make the 2- and 3-logistic models similar to the normal ogive models. The lower-asymptote parameter allows for examinees with low ability (θ) to answer a correct response by guessing only. The 2PLM can be said

to be a special case of 3PLM where the c -parameter is released, and the 1PLM is a special case of 2PLM where the a -parameter is fixed to 1. The 2PLM can be used in a situation where guessing is minimized due to no alternatives in each item (short-answer items). The 1PLM may be appropriate for a set of items with the same item discrimination value.

When estimating both person and item parameters, two groups of estimation methods can be utilized: Maximum Likelihood Estimation (MLE) and Bayesian Estimation. The Maximum Likelihood method is employed to find the highest point of the likelihood function of the binomial distribution for both the person and item parameters (Kim & Nicewander, 1993). The Bayesian methods can apply the prior distributions of the ability and item characteristics to the computation of the posterior probabilities of the estimated parameters (Bock & Aiken, 1981). The Maximum-a-Posteriori (MAP) and Expected-a-Posteriori (EAP) are some examples of the Bayesian methods (Bock & Mislevy, 1982, Samejima, 1969).

In order to estimate item parameters using the IRT models, two major assumptions should be satisfied: Local independence and unidimensionality. The unidimensionality assumption stipulates that the items in a test should measure only one dimension of a trait. This assumption can be easily met if items are carefully constructed with one specific trait in mind. The local independence assumption states that the probabilities of answering items correctly are independent of each other given the examinee's ability. However, most reading comprehension tests have items which are grouped with a common passage. These items in a common passage may be locally

dependent. Local item dependence (LID) may occur by “item interaction” (Tuerlinckx & De Boeck, 2001). It may also occur by many other factors including learning, cheating, fatigue, carelessness, etc. (Hambleton, Swaminathan, & Rogers, 1991). Tuerlinckx and De Boeck found that a positive correlation between items produced overestimated item parameters, and a negative correlation yielded underestimated item parameters.

Generally, LID shows a propensity to provide an overestimate of the measurement’s precision (Bradlow, Wainer, & Wang, 1999) as well as inflated item discrimination and item difficulty parameter estimates (Yen, 1984). However, LID affects test reliability in CTT to make it overestimated (DeMars, 2006). The different strength of LID makes the differences in reliability estimates (Gessaroli & Folske, 2002).

A set of test items grouped in a common passage is named as testlet by Wainer and Kiely (1987). Items within a testlet have been a prevalent and useful element of reading comprehension tests even though the use of IRT estimation methods is limitedly applied due to the presence of LID (Wainer, Bradlow, & Wang, 2007). In order to resolve the violation of the local independence assumption, several psychometricians have proposed diverse testlet response theory (TRT) models (e.g., Bradlow, Wainer, & Wang, 1999; Wainer, Bradlow, & Du, 2000). In all of these models, the testlet effect parameter is inserted as an adjustment to the basic models of IRT. The 3PL random-effect TRT model which is constrained by item discrimination parameter (Bradlow, Wainer, & Wang; Demars, 2012; Wainer, Bradlow, & Wang) for dichotomous responses can be described as,

$$P(Y_{ij} = 1) = c_j + (1 - c_j) \frac{1}{1 + e^{-1.7a_j(\theta_i - b_j - \gamma_{id(j)})}}, \quad (2)$$

where $\gamma_{id(j)}$ represents the testlet effect of item j with examinee i within $d(j)$ which indicates j^{th} passage. In this constrained TRT model, the testlet parameter ($\gamma_{id(j)}$) contains the covariance effect among items within a testlet. As one can see from Equation (2), a single item discrimination parameter (a -parameter) affects all of the item difficulty parameter, examinee ability, and testlet parameter. In this model, item discrimination parameter cannot be interpreted independently from examinee ability, item difficulty, and testlet parameter (Li, Bolt, & Fu, 2006). However, in reality, the effect of item discrimination parameter may not be constant to all other parameters. For example, an item with high item discrimination parameter value on an examinee's ability may show poor item discrimination parameter value on the testlet parameter in a reading passage. In order to address this limitation, the generalized 3PL TRT model (Li, Bolt, & Fu) was proposed which formulated two separate item discrimination parameters on person and testlet parameter, respectively as,

$$P(Y_{ij} = 1) = c_j + (1 - c_j) \frac{1}{1 + e^{-1.7(a_{1j}\theta_i - b_j - a_{2j}\gamma_{id(j)})}}, \quad (3)$$

where a_{1j} is the item discrimination parameter for examinees' abilities (θ_i) and a_{2j} is the item discrimination for testlet effects ($\gamma_{id(j)}$). Notice that the b -parameter is not associated with any item discrimination parameter. In order to fully utilize the advantages of IRT models over CTT, these two testlet models warrant further investigations for empirical data, especially with the reading comprehension assessment

where items stem from the same passage. The advantages of TRT over the traditional IRT can be tested using a few criteria including the χ^2 -test with -2LL (Log-Likelihood), AIC (Akaike information criterion; Akaike, 1973), and BIC (Bayesian information criterion; Schwarz, 1978). Estimated item parameters along with item and test information were also used to compare different IRT and TRT models.

Purpose of the study

The purpose of this study was to evaluate psychometric models such as CTT, traditional IRT, and TRT models on a reading comprehension test constructed on the basis of CCSS. The test contained the items about key ideas, meaning of words, integration of knowledge, and comprehension of text complexity that was developed by a for-profit educational assessment company as a benchmark measurement. For this research goal, firstly, the CTT model was applied to the data for both item and test indices including the *p*-value, item discrimination index, item-test correlation, Cronbach's alpha, and a construct validity estimate utilizing exploratory factor analysis (EFA). The traditional IRT models (1PL, 2PL, and 3PL) were applied to obtain item parameters such as item discrimination, item difficulty, and pseudo-chance parameters. Secondly, the model-data fit for each IRT model was investigated and compared among the three IRT models with the model comparison criteria (-2LL, AIC, BIC). Once the best-fitting IRT model is selected, then two TRT models (constrained TRT and generalized TRT) would be employed for discerning efficiency of each TRT model. The estimated item parameters and the testlet parameters along with item and test information

were computed and compared. The similarities and dissimilarities between CTT and IRT, and between IRT and TRT were discussed in conjunction with previous findings.

CHAPTER TWO: REVIEW OF LITERATURE

This chapter will review various theoretical reading comprehension models as a foundation for the current investigation of psychometric validation of a reading assessment in order to benchmark reading comprehension test among fifth graders.

Reading comprehension models

Diverse reading comprehension models contain different multidimensional cognitive and linguistic skills in readings based on different perspectives. Reading comprehension models have been developed with the perceptual processing of “bottom-up” or “top-down.” The bottom-up process suggests reading comprehension to start with word recognition from a text and moves upward to comprehension. On the other hand, the top-down method starts with prior knowledge and experience in reading and proceeding downward to word recognition. Word identification and linguistic comprehension are two main reading processes.

Gough and Tunmer (1986) proposed that reading comprehension is a product of decoding and linguistic comprehension. Decoding, defined as word recognition from a written text, is measured by the accuracy of word and non-word reading. The skill of decoding is central in reading. If children have the ability to decode automatized, they are enabled to use higher level of cognitive resources without efforts to comprehend a text (Cain, Oakhill, & Bryant, 2004). In other words, reading comprehension will be compromised by inefficient decoding skills when texts are longer or more complex. At the primary level of learning to read, decoding is considered to be the foundation of

reading. For later reading comprehension development, or periods of reading to learn, linguistic comprehension is more emphasized with a wide range of texts. In the research by Gough and Tunmer (1986), decoding skill was described as the aspect of word recognition. Word recognition is influenced by phonological and orthographical skills (Henderson, 1982; Plaut, 2005). Phonological awareness refers to the ability to recognize, distinguish, and manipulate separate sounds. Phonological skills are strongly associated with word-reading development (e.g., Bradely & Bryant, 1983; Wagner & Torgesen, 1987). Phonological processing deficits can impair the ability to retain verbal information in working memory (Shankweiler, 1989). Siegel (1993) demonstrated that phonological knowledge would enable children to divide whole sounds into its smallest units. The knowledge of phoneme and grapheme aids to associate those units of sound with a letter. The orthographic knowledge helps to access or recognize a word directly in lexical memory (Cunningham, Perry, & Stanovich, 2001).

Linguistic comprehension, defined as the ability to interpret phrases and texts from lexical information, is measured using comprehension questions administered after listening to or reading a printed text. This definition is similar to reading comprehension which is measured by only printed texts. According to the SVR, linguistic comprehension is measured with parallel methods with reading comprehension. When assessments are not parallel, differences between written and oral texts, such as complexity of grammatical structures could produce differences in results of reading and linguistic comprehension (Kershaw & Schatschneider, 2012).

Curtis (1980) found that decoding and linguistic comprehension were significant predictors of reading comprehension. Poor readers at kindergarten through 4th grade were struggling with both decoding skills and linguistic comprehension, or with one of these components (Catts, Hogan, & Fey, 2003). Thus, it is evident that decoding and linguistic comprehension are strongly correlated to reading comprehension (e.g., Juel, Griffith, & Gough, 1986). Decoding allows readers to access the meaning of words. This semantic knowledge is essential for comprehension. Syntactic knowledge enables readers to assign a grammatical function to words within sentences. Semantics have been found to be one of the best predictors of reading comprehension (e.g., Carroll, 1993, Thorndike, 1973). Semantic skills include one's knowledge of word meaning as well as the efficiency of retrieving the meaning of a word. Comprehension of written and spoken language is dependent on an individual's knowledge of vocabulary (McGregor, 2004). Thorndike (1973) found a strong relationship between comprehension and vocabulary knowledge. Seigneuric and Ehrlich (2005) also revealed a reciprocal association between vocabulary and comprehension skills. In their study, the reading comprehension of first-grade students was found to be accounting for 10% of variability in second-grade vocabulary and 15% of the variability in third-grade vocabulary. Syntactic knowledge helps children detect reading errors and enhance comprehension monitoring as well as aid in word recognition (Tunmer & Hoover, 1992). Semantic and syntactic knowledge serves as cue for the construction of meaning with certain predictions about sentence structures, which will enhance children's reading comprehension.

A number of other theories about reading comprehension exist in which different parts of the reading process are described. In the view point of cognitive theories, reading comprehension is the outcome in terms of mental representations. One of the most influential cognitive theories is the C-I model by Kintsch (1998). This model describes the reading process from recognizing words to constructing a representation of the meaning of the text. Comprehension processes result in three levels of mental representation. The first is the surface level of representation which is a word-for-word representation of the text. The second is the proposition level of representation, in which the reader extracts the core ideas from the literal text. In this level, the reader builds the text base by linking together the propositions. The third is the situation model, also known as the mental model (Kintsch, 1998), which is the highest level of representation of the text's meaning and represents the integrated situation described in a text. Situation models describe the representation constructed when readers integrate and update what they already know about the topic into a more complex and holistic conceptualization of it. Kintsch (1998) proposed that comprehension resulted from the process of construction and selection of meaning from a text. The construction phase involves the formation of diverse meanings from a text followed by the selection (integration) of constructed meanings utilizing prior knowledge.

Another influential model in reading comprehension is the sociocultural model. According to the model, reading comprehension process occurs between a reader, text, and the reading activity, within a range of sociocultural factors. The RRSB (2002) emphasized that these elements are actively interrelated. In the sociocultural context,

environmental factors include “economic resources, class membership, ethnicity, neighborhood, and school culture, can be seen in oral language practices, in students self-concepts, in the types of literacy activities in which individuals engage, in instructional history, and of course in the likelihood of successful outcomes (p.7)” (RRSG, 2002). The RRSG addressed three categories as the outcomes of reading comprehension: Knowledge, application, and engagement (Sweet, 2005). Knowledge indicates the process of successful comprehension of a text with integration and evaluation using prior knowledge. Application is the act of applying practical tasks. Engagement is the manner of reflecting with knowledge, experience, ideas, and information of a text.

These various reading comprehension models have been proposed as an effort to enhance our understanding of the construct and process of reading comprehension. In order to have a better understanding the constituents, well-developed test items which represent target domains of reading comprehension are necessary. However, many current comprehension tests fail to reflect the nature and characteristics of reading comprehension due to the interaction of a variety of component processes and skills (Kintsch & Kintsch, 2005). A practical guideline was developed by state leaders and the Council of Chief State School Officers (CCSSO) as CCSS to direct all grade level students in both ELA and mathematics (<http://www.corestandards.org>).

Common Core State Standards

The CCSS for ELA provides guidance and structure for reading curriculum for all grade levels. The CCSS (2014) were established with a focus on defining general and

cross-disciplinary goals that students must meet in order to prepare for college and career readiness. Each grade level has standards that are broken into specific areas of focus in order to achieve various goals. Elementary school goals are anchor standards, foundational skills, informational text, writing, speaking, and listening (CCSS, 2014). Secondary school standards also include a focus on history and social studies, science and technical skills, and a deeper look at writing.

Policymakers concluded that creating common educational standards and increasing rigor in schools to prepare all students for college or career readiness in the 21st century were vital if the United States was to strive for and surpass educational excellence (<http://www.corestandards.org>, 2014). Since the initial discussion for developing common standards in 2008, the Common Core has been adopted by forty-three states and the District of Columbia, four territories, and the Department of Defense Education Activity (<http://www.corestandards.org/about-the-standards/development-process>). The effort to draft common standards was launched in 2009. The National Governors Association (NGA) and the Council of Chief State School Officers (CCSSO) led to the initiative with guidance from an advisory group to help states raise academic standards. In 2010, NGA and CCSSO released the CCSS as an academic benchmark to define the knowledge and skills for college and career readiness. The CCSS contains mainly two categories ([www. corestandards.org](http://www.corestandards.org)). The first is for college and career readiness that students are required to understand and know by their graduation from high school. The second category comprises standards that K-12 students are expected to acquire literacy skills through high school.

The guideline by the CCSS for ELA highlights formative assessments on the continuum of developmental progress in reading. The key endeavor of the CCSS is to provide reliable and consistent standards in literacy complexity for all grade levels. The CCSS for ELA and other subjects define the expectations by each grade level. The standards for reading describe four key dimensions and ten sub-skills (CCSS, 2014):

Key Ideas and Details

1. Read closely to determine what the text says explicitly and to make logical inferences from it; cite specific textual evidence when writing or speaking to support conclusions drawn from the text.
2. Determine central ideas or themes of a text and analyze their development; summarize the key supporting details and ideas.
3. Analyze how and why individuals, events, and ideas develop and interact over the course of a text.

Craft and Structure

4. Interpret words and phrases as they are used in a text, including determining technical, connotative, and figurative meanings, and analyze how specific word choices shape meaning or tone.
5. Analyze the structure of texts, including how specific sentences, paragraphs, and larger portions of the text (e.g., a section, chapter, scene, or stanza) relate to each other and the whole.
6. Assess how point of view or purpose shapes the content and style of a text.

Integration of Knowledge and Ideas

7. Integrate and evaluate content presented in diverse media and formats, including visually and quantitatively, as well as in words.
8. Delineate and evaluate the argument and specific claims in a text, including the validity of the reasoning as well as the relevance and sufficiency of the evidence.
9. Analyze how two or more texts address similar themes or topics in order to build knowledge or to compare the approaches the authors take.

Range of Reading and Level of Text Complexity

10. Read and comprehend complex literary and informational texts independently and proficiently (p. 10).

According to Appendix A (http://www.corestandards.org/assets/Appendix_A.pdf, p. 5-6), the CCSS have multiple categories and steps in terms of structure, language conventionality, and diverse knowledge demands for both qualitative and quantitative measures. The structure can develop from a simple structure to a complex one, from explicit to implicit, from conventional to unconventional, from chronological to non-chronological, from common genre to particular discipline, from simple graphics to sophisticated graphics, and from supplementary to essential graphics. Language conventionality can change from literal to figurative or ironic, from clear to ambiguous, from contemporary or familiar to archaic or unfamiliar, from conversational to academic and domain-specific. Diverse knowledge demands also have different categories. Life experiences can progress from simple to complex themes, single to multiple themes, from common to unique experiences, from single to multiple perspectives and from one's own experience to another's. Cultural and literacy knowledge demands may grow from everyday knowledge to useful cultural and literacy knowledge and from low intertextuality to high intertextuality for both literal texts and informational texts. These categories and steps can be used as assessment guidelines to gauge a students' performance.

Assessment of reading comprehension

Researchers and educators are increasingly calling for reliable and valid assessments that reflect children's progress towards reading comprehension benchmarks.

Large-scale assessments of reading comprehension have been used for monitoring large numbers of students and evaluating educational programs. However, large-scale assessments are often criticized due to a lack of theoretical underpinnings (Snow, 2003). Snow addressed that it is hard to achieve construct validity because the target domain of reading comprehension is complex and multidimensional. However, most comprehension assessments which are commonly used for research and diagnosis reflect a single dimension. A unidimensional test does not fully represent the construct the test is trying to measure.

In order to establish a clear understanding of reading comprehension, one must recognize the process and outcome of reading comprehension assessment because without accurate assessments of the process, other predictions and diagnoses will be inaccurate. The assessment of reading comprehension should be understood through the theories of educational measurement. The area of psychometrics describes both theoretical and practical principles and issues. Thus, one must understand the basic test theories of psychometrics. There are two main streams in psychometrics: Classical test theory (CTT) and item response theory (IRT). The two theories of psychometrics warrant more detailed descriptions.

Classical Test Theory (CTT)

Measurement theory was originally developed in the early 20th century from the work of Spearman (1904). Spearman measured individual differences in mental abilities. Since then, CTT has been widely used in many research areas. The fundamental feature

of CTT is the formulation of an observed outcome (X_{ip}) as a composite of two independent components, an underlying true-score component (T_{ip}) and measurement error (E_{ip}):

$$X_{ip} = T_{ip} + E_{ip}. \quad (4)$$

In this framework, the true score (T) for item i and person p is defined as the expected value of an individual's observed scores (X) on the repeated assessments with the same instrument to the same examinee under an identical condition. There are several assumptions in CTT (Allen & Yen, 2002). First, the expected value of observed scores is the true score. The expected value of error scores in the population is zero and the error scores are normally distributed. Second, there is no correlation between the true and error scores. Third, the error scores from two different tests are not correlated. Fourth, there is no correlation between the true score from Test 1 and the error score from the Test 2 in the population. The fifth assumption is that parallel tests exist. The conditions for parallel tests are that the two tests have the same true scores ($T_1 = T_2$) and that the two error score variances are identical ($\sigma_{E_1}^2 = \sigma_{E_2}^2$). The last assumption of CTT is the existence of τ -equivalent tests. The τ -equivalent tests assumption requires $T_1 = T_2 + C$ (constant). The equal error variance condition does not apply to the τ -equivalent tests assumption.

As indicated in the introduction, CTT is a measurement model on the observed scores, which cannot be proved or disproved with actual data. The true score (T_{ip}) for item i and person p as the expected value of observed scores, $E(X_{ip})$, for item i and person

p is impossible to define in practice and should administer the same test to the same person infinitely many times in order to compute the expected value from infinitely many observed scores of the person. Thus, testing to see if whether the CTT model is true with data in practice is impossible. If the same test is given to the same person for the second time, there will be numerous factors which may affect the observed scores other than measurement error (Embretson & Reise, 2000).

In addition to this theoretical weakness, CTT also has several practical issues. The person's true score (person parameter) is test-dependent. For an easy test, the person's true score is high; for a difficult test, the same person's true score is low. At the same time, item and test indices are sample-dependent. For example, the item difficulty index (p -value) for item i will be high (easy item) if the item is given to a very high-ability group. The same item's p -value will be low (difficult item) if the item is given to a very low-ability group. The assumption of parallel-test in CTT is extremely difficult, if not impossible, to meet in a real test construction setting. Also, CTT is a test-oriented model instead of item-oriented. Although some item indices are computed in the CTT analyses such as the p -value, d -value, and item-test correlation, the two main features of CTT are validity and reliability (Hambleton & van der Linden, 1982).

There are several different types of validity: Content validity, criterion-related validity, and construct validity (Allen & Yen, 2002). Content validity can be established through face validity and logical validity. Bollen (1989) defined content validity as “a qualitative type of validity where the domain of the concept is made clear and the analyst judges whether the measures fully represent the domain” (p.185). Criterion-related

validity, including predictive validity and concurrent validity, is usually measured by the degree of relation between the test and the external criteria. The Multitrait-Multimethod (MTMM) method and factor analysis are two types of construct validity estimation methods. MTMM can provide the degree of convergent and discriminant validations (Campbell & Fiske, 1951). Factor analysis is another often-used method for construct validity (Allen & Yen). Construct validity is a validity of theoretical construct that can be established by psychometric methods such as exploratory factor analysis and MTMM.

Different methods for defining and estimating reliability also exist. Within CTT, reliability is defined as the ratio of the true score variance to the observed score variance. Test-retest, parallel-forms, alternate-forms, and internal-consistency (including the Spearman-Brown formula and Cronbach's alpha) are ways to estimate reliability (Allen & Yen, 2002). Test-retest reliability refers to the strength or weakness of correlation between two test results which are administered to the same examinees with the same test at different times. Test-retest reliability estimate is affected by serious problems such as carry-over effect and the length of time interval between the two tests (Allen & Yen). Similar to the test-retest reliability, parallel-forms or alternate-forms reliability is measured by the correlation between two parallel tests or two τ -equivalent tests. The criterion for parallel tests is that true scores and error variances of two tests should be identical, which is almost impossible to meet in practice. The τ -equivalent tests require that the true score of one test should be a linear function of the true score of the other test and the error variances do not need to be identical. Internal-consistency estimates are methods to test reliability with two divided parts of items from the same test. Internal-

consistency estimates have an advantage to avoid the problems with the repeated testing from the procedures of test-retest and parallel-forms reliability. In the case of two subtests being parallel, the Spearman-Brown formula can be utilized to estimate internal consistency while Cronbach's alpha can be used when two tests are τ -equivalent.

There are several item indices in CTT including the p -value, d -value, and item-test correlation (Allen & Yen, 2002). The p -value, known as the item difficulty, can be computed by the proportion of the number of people who have the correct answer for item i and the total sample size. The p -value ranges from 0 to 1. If an item has a p -value of either 1 or 0, this item is not very useful because these values indicate that all students got the item correct or incorrect. It is desirable if the range of the p -value is between .30 and .70. The second item index in CTT is the d -value (item discrimination), which can be computed by:

$$d_i = U_i/n_{iU} - L_i/n_{iL}, \quad (5)$$

where the U_i/n_{iU} represents the ratio of the proportion of examinees in the upper group who have the right answer for item i to the total number of examinees in the group. The L_i/n_{iL} is the proportion of examinees in the lower group who have the right answer for item i to the total number of examinees in the same group. The sample size of the upper and lower group is the same or similar in many cases and ranges from 10% to 33% of the total sample. The item-test correlation as an alternative method of item discrimination (d) can be computed as:

$$r_{iX} = \frac{\bar{X}_i - \bar{X}}{S_x} \sqrt{\frac{P_i}{1 - P_i}}, \quad (6)$$

where \bar{X}_i is the mean of item i and \bar{X} and S_x are the mean and standard deviation of the test. P_i is the item difficulty index. The point-biserial correlation (r_{iX}) provides the index of the association between each item and the test (Allen & Yen, 2002).

Although CTT has several item-level indices, the main focus of CTT lies in test indices such as reliability and validity. Theoretical and practical shortcomings and difficulties for interpreting item and test indices in CTT have been criticized in psychological and educational measurements (Hambleton & van der Linden, 1982). Due to the issues and shortcomings of CTT, there was a call for a better psychometric theory.

Item Response Theory (IRT)

Unlike CTT, which has theoretical and practical problems with test-dependent person parameters, sample-dependent item parameters, and the parallel test assumption, a modern measurement theory known as IRT developed by Lord (1952) and Birnbaum (1968) offers many important advantages over CTT. Embretson and Reise (2000) described benefits of utilizing IRT rather than CTT (p. 15);

<i>The old rules (CTT)</i>	
Rule 1	The standard error of measurement applies to all scores in a particular population.
Rule 2	Longer tests are more reliable than shorter tests.
Rule 3	Comparing test scores across multiple forms is optimal when the forms are parallel.
Rule 4	Unbiased estimates of item properties depend on having representative samples.
Rule 5	Test scores obtain meaning by comparing their position in a norm group.

Rule 6	Interval scale properties are achieved by obtaining normal score distributions.
Rule 7	Mixed item formats leads to unbalanced impact on test total scores.
Rule 8	Change scores cannot be meaningfully compared when initial score levels differ.
Rule 9	Factor analysis on binary items produces artifacts rather than factors.
Rule 10	Item stimulus features are unimportant compared to psychometric properties.

The new rules (IRT)

Rule 1	The standard error of measurement differs across scores (or response patterns), but generalizes across populations.
Rule 2	Shorter tests can be more reliable than longer tests.
Rule 3	Comparing test scores across multiple forms is optimal when test difficulty levels vary between persons.
Rule 4	Unbiased estimates of item properties may be obtained from unrepresentative samples.
Rule 5	Test scores have meaning when they are compared for distance from items.
Rule 6	Interval scale properties are achieved by applying justifiable measurement models.
Rule 7	Mixed item formats can yield optimal test scores.
Rule 8	Change scores can be meaningfully compared when initial score levels differ.
Rule 9	Factor analysis on raw item data yields a full information factor analysis.
Rule 10	Item stimulus features can be directly related to psychometric properties.

IRT focuses on the association between person parameter and item parameters in a test. Lord (1980) also described IRT;

We need to describe the items by item parameters and the examinees by examinee parameters in such a way that we can predict probabilistically the response of any examinee to any item. (p. 11)

Traditional IRT models were developed with two underlying assumptions of unidimensionality and local independence. These assumptions are required for test administrators and substantive educators to examine when using IRT. The unidimensionality assumption refers to a single latent trait (θ) which is observed by

items in a test. The local independence states that observed responses to each item are uncorrelated. Lord and Novick (1968) addressed the local independence assumption in their book,

Local independence means that within any group of examinees all characterized by the same values $\theta_1, \theta_2, \dots, \theta_k$, the (conditional) distribution of the item scores are all independent of each other. (p. 361)

Two mathematical forms in IRT models have been developed: Normal ogive models and logistic models. Both models yield similar stochastic results in IRT but the logistic models contribute to more simplified mathematical and computational forms than normal ogive models. In normal ogive models, item characteristics curve (ICC) is derived from the cumulative density function of a normal distribution. A mathematical form of the one parameter normal ogive model (Lord, 1952) is as follows:

$$P_i(\theta) = \int_{-\infty}^{\theta - b_i} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz, \quad (7)$$

where $P_i(\theta)$ is the probability of answering item i correctly for a given ability level θ , and b_i is the item difficulty parameter which ranges from $-\infty$ to $+\infty$, theoretically, but is used from -3 to $+3$, practically. The z is a standardized score ($z = \frac{X - \mu}{\sigma}$) of the examinee for item i .

The mathematical expression of 1PLM is,

$$P(Y_{ij} = 1) = \frac{1}{1 + e^{-D(\theta_i - b_j)}} = \frac{1}{1 + e^{-DL}}, \quad (8)$$

where the logistic deviate (L) is $\theta_i - b_j$. The θ_i represents the ability level, b_j is the difficulty parameter, e is the constant of 2.718, and D is a scaling factor and set 1 for the 1PLM. One parameter normal ogive and logistic models demonstrated the relationship

between only one item parameter (item difficulty) and the person parameter. Using the 1PLM of a given item, the difficulty parameter on the examinee's latent trait (θ) dimension, for instance, ICCs can be drawn as in Figure 1. The difficulty parameter is defined as the value with a 50% likelihood of a correct item response on the latent ability scale. The vertical axis represents the probability of a correct response. Figure 1 illustrated three ICCs for the probabilities of correct item responses based on the examinee's ability levels (moderately easier item ($b = -1$), medium ($b = 0$), and moderately harder item ($b = 1$)). All items of the 1PLM share identically shaped ICC only with different location parameter values of b_j .

The 2PLM is a generalized model of the 1PLM by adding the item discrimination parameter (a -parameter). The 2PLM can be expressed as,

$$P(Y_{ij} = 1) = \frac{1}{1 + e^{-Da_j(\theta_i - b_j)}} = \frac{1}{1 + e^{-DL}}, \quad (9)$$

where a_j is the item discrimination parameter for item j and D is a scaling factor ($D = 1.7$). The D yields similarly equivalent models and interpretations between the normal ogive and logistic models. The logistic deviate (L) for 2PLM is $a_j(\theta_i - b_j)$. Items with higher item discrimination parameter values provide more information about the examinee's ability at a specific location in the ability distribution than other items with lower values of item discrimination parameter. For example, in Figure 2, Item 3 ($a = 1.5$ and $b = 1.0$) delivers more item information at the person ability level of 1.0 than others. The graphs represent ICCs for three items: Item 1 contains item difficulties (b) value of -1 and item discrimination (a) value of 2.0. Item 2 has the b -value of 0.0 and the a -value of 1.0. Item 3 represents the b -value of 1 and the a -value of 1.5.

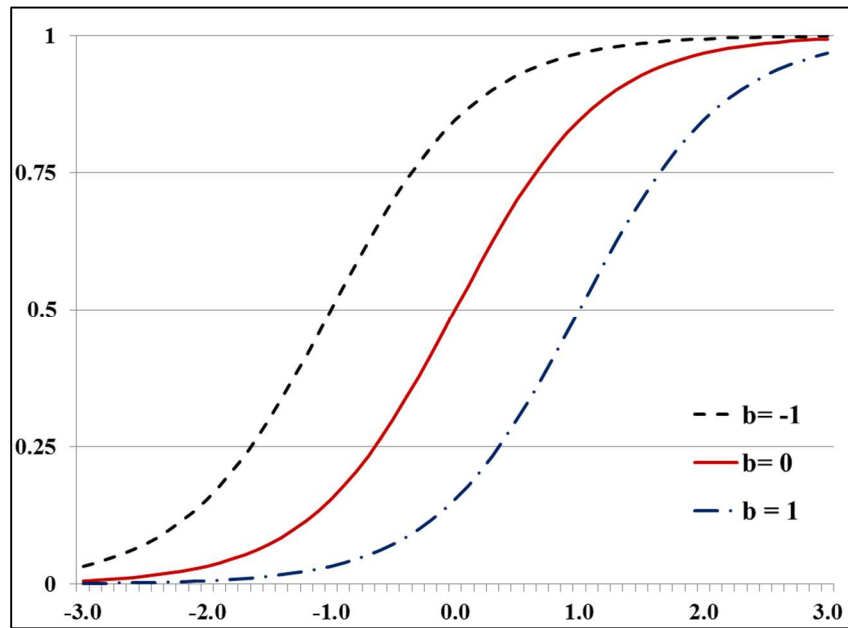


Figure 1. Item characteristic curve (ICC) for 1PLM

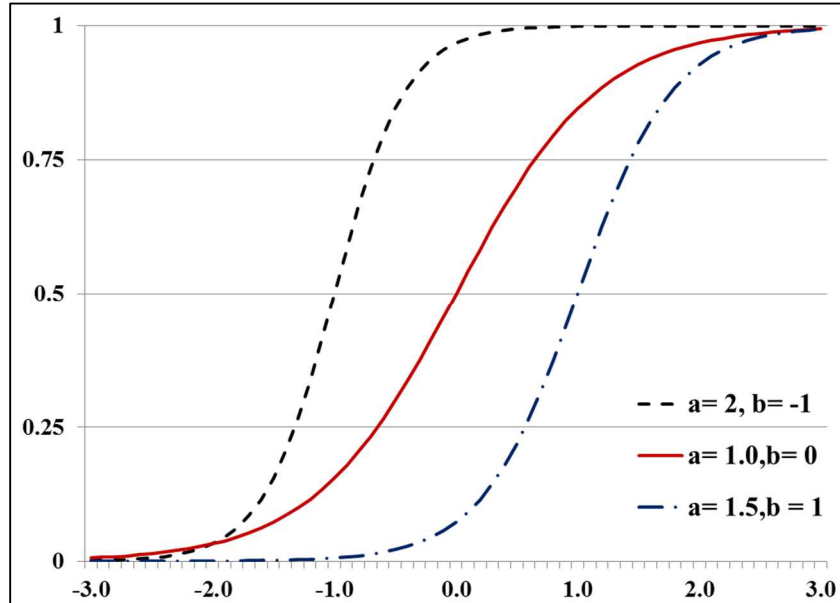


Figure 2. Item characteristic curve (ICC) for 2PLM

The 3PLM allows an ICC with non-zero pseudo-chance factor. Birnbaum (1968) modified the 2PLM to include a pseudo-chance parameter to the probability of correct response. The mathematical expression of the 3PLM can be expressed as,

$$P(Y_{ij} = 1) = c_j + (1 - c_j) \frac{1}{1 + e^{-Da_j(\theta_i - b_j)}} = c_i + \frac{(1 - c_i)}{1 + e^{-DL}}, \quad (10)$$

where c_j represents the probability that examinees with extremely lower ability answer correctly for item j . The logistic deviate (L) for 3PLM is the same as 2PLM ($a_j(\theta_i - b_j)$). In Figure 3, the ICC for the item with $a = 0.5$, $b = -1$, and $c = 0.5$ lost mathematical properties of the logistic function by a high value of the pseudo-chance parameter in some cases. Note that a represents the item discrimination value, b is the item difficulty, and c is the pseudo-chance parameter.

The ultimate duty of psychometricians is to estimate both item and person parameters utilizing various estimation methods including Maximum Likelihood Method and diverse Bayesian methods. The fundamental principle of MLE is to estimate the underlying proficiency of parameters with the likelihood function based on the pattern of item responses of a person. The likelihood function has two components which are the probability of correct responses and the probability of incorrect responses. The probability of correct response for item i was described in equations (8) through (10) depending on the number of parameters in the model.

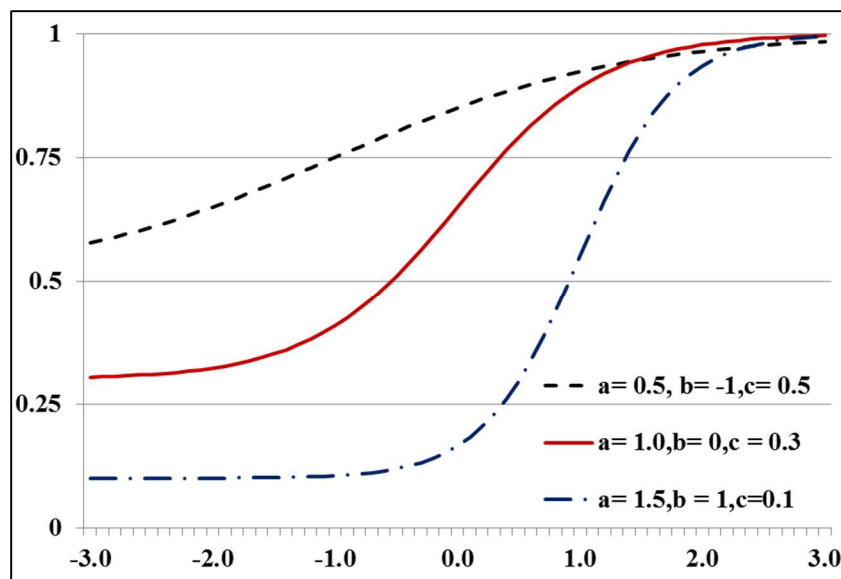


Figure 3. Item characteristic curve (ICC) for 3PLM

The probability of incorrect responses for the item can be addressed as $Q_j(\theta_i) = 1 - P_j(\theta_i)$. Then, the likelihood function of the item response is articulated as,

$$L \equiv L(\mathbf{u}|\theta) = \prod_{j=1}^n P_j^{u_j} Q_j^{1-u_j}, \quad (11)$$

where \mathbf{u} ' is the row vector of obtained item response, u_i is each item response, and $Q_j = 1 - P_j$. The Log-Likelihood function of $L (= l)$ and the first derivative of $l (= l')$ must be obtained for a given set of item responses (\mathbf{u}). Setting l' to zero and solving the equation for θ will result in the MLE. However, due to a few estimation issues, MLE may have severely biased estimates of parameters (Kim & Nicewander, 1993). The Bayesian methods use various prior distributions to compute the posterior probability based on the Bayes' principle in order to improve the accuracy of estimation through either maximum-a-posteriori (MAP, Samejima, 1969) or expected-a-posteriori (EAP, Bock & Mislevy, 1982). Some other Bayesian methods are also available with minor modifications from MAP and EAP. Research has shown that the Bayesian methods outperforms MLE with less biased and more accurate estimations (Kim & Nicewander).

Testlet Response Theory (TRT)

Two basic assumptions in the traditional IRT are unidimensionality and local independence of items. The unidimensionality assumption states that the items in a test should measure only one underlying dimension of the construct for ability or proficiency. Unidimensionality can be tested through factor analysis and other appropriate methods. Local independence is related to the correlation among items in a test. Given the ability level, the probability of answering an item correct should be independent of the probability of answering other items correctly. However, in the area of reading, a typical format of reading comprehension tests contains various passages followed by multiple

items stemmed from the same passage. In this case, items for the same passage are correlated to each other due to the fact that items ask questions about the same passage, which is a clear violation of the local independence assumption of the unidimensional IRT. Although bundled items may violate the local independence assumption of the unidimensional IRT, different IRT models have been used for the estimation of person ability and item parameters ignoring the violation of the assumption (de Ayala, 2009; Yen & Fitzpatrick, 2006).

Many researchers reported that ignoring local item dependency (LID) caused several problems: (1) overestimation of the person ability estimates; (2) underestimation of the standard error of estimates; and (3) biased item parameter estimates such as item difficulties or item discriminations (e.g., Sireci, Thissen, & Wainer, 1991; Tuerlinckx & De Boeck, 2001; Wainer & Wang, 2000; Yen, 1984). Thissen, Steinberg, and Mooney (1989) and Sireci, Thissen, and Wainer (1991) pointed out that LID would lead to an overestimate of reliability and test information, as well as an underestimate of the standard error of measurement. However, if appropriately modeled, item bundles (e.g., Wilson & Adams, 1995), context-dependent item sets (e.g., Keller, Swaminathan, & Sireci, 2003), or testlets (Wainer & Kiely, 1987) could allow for the measurement of interrelated tasks and skills.

To account for LID associated with items nested within a testlet, Bradlow, Wainer, and Wang (1999; Wainer, Bradlow, & Wang, 2007) proposed the 2PL testlet response theory (TRT) model. The 2PL TRT model includes a random-effect parameter representing the interaction of person i with testlet $d(j)$, which contains item j . In this

model, the probability of a correct response to item j nested in testlet $d(j)$ for a person ability θ_i is given by;

$$P(Y_{ij} = 1) = \Phi a_j(\theta_i - b_j - \gamma_{id(j)}), \quad (12)$$

where $a_j(\theta_i - b_j - \gamma_{id(j)})$ is equivalent to $\frac{1}{1 + e^{-1.7a_j(\theta_i - b_j - \gamma_{id(j)})}}$ with the testlet parameter of $\gamma_{id(j)}$. This equation is similar to Equation 2 except the c -parameter. One constraint of this model is that the item discrimination parameter (a -parameter) has a uniform effect on θ_i , b_j , and $\gamma_{id(j)}$. This testlet model is called the constrained testlet model. This constrained testlet model may not fully represent a situation where the item discrimination parameter has differential effects on θ_i and $\gamma_{id(j)}$. This model makes the interpretation of item discrimination parameter difficult.

The generalized testlet model was proposed to provide different effects of the item discrimination parameter on the testlet factor ($\gamma_{id(j)}$) and ability parameter (θ_i). The form of the generalized testlet model is (Li, Bolt, & Fu, 2006);

$$P(Y_{ij} = 1) = \Phi (a_{j1} \theta_i - b_j + a_{j2}\gamma_{id(j)}), \quad (13)$$

where $\gamma_{id(j)}$ and θ_i are uncorrelated and a_{j1} and a_{j2} indicate the item discrimination parameter with respect to $\gamma_{id(j)}$ and θ_i . For Equation 13, a_{j1} is the discriminating power for only the person ability and a_{j2} is for the testlet effects ($\gamma_{id(j)}$).

For a reading comprehension test with reading passages where items in a testlet are asking questions about the same passage, the traditional IRT models may raise issues related to an overestimation of item discrimination and item difficulty parameters along with an underestimation of standard error. Comparing CTT and IRT will benefit

researchers and educators by providing them with more item-related indices for the evaluation of strength and weakness of each item along with item and test information from IRT. An application of different TRT models to a reading comprehension test should offer valuable information about the utility of TRT models in comparison with the traditional IRT model.

In order to compare TRT and IRT models, the criteria of the χ^2 -test with -2LL (Log-Likelihood), AIC (Akaike, 1973), and BIC (Schwarz, 1978) are widely utilized. These criteria are the measures of the relative indices for statistical significance tests. The difference between two -2LL values from two comparing models is approximately distributed as a χ^2 distribution ($\Delta\chi^2 = \chi_1^2 - \chi_2^2$) with the degrees of freedom of $\Delta df (= df_1 - df_2)$. The -2 Log-Likelihood function (-2LL) value increases with added number of parameters in estimation when comparing traditional IRT or TRT models. Lower values of -2LL indicate a better model-fit.

$$\text{Likelihood function} \equiv L(\mathbf{u} | \theta, a, b, c) = \prod_{i=1}^n P_i^{u_i} Q_i^{1-u_i}, \quad (14)$$

where $L(\mathbf{u} | \theta, a, b, c)$ represents conditional probability of \mathbf{u} given $\theta, a, b,$ and c . The notation of u_i denotes item response, P_i is the probability of a correct answer, and Q_i is the probability of an incorrect response ($Q_i = 1 - P_i$).

Let $LL = \text{Log-Likelihood function}$,

$$LL = \ln L = \sum [u_i \ln P_i + (1 - u_i) \ln Q_i]. \quad (15)$$

Then, -2LL will be distributed approximately as a χ^2 distribution. One issue of the -2LL function is that it does not take into account the effect of sample size (N) and the number of parameters (p) in the model.

In order to accommodate the issue of the number of parameters in the model, Akaike (1973) proposed a new criterion for model fit index (AIC). The AIC is defined as;

$$AIC = -2LL + 2p, \quad (16)$$

where LL indicates the maximized log-likelihood estimate and p is the number of parameters in the model. Lower values of the AIC also indicate a better model-fit.

Unlike the AIC which does not take the sample size effect, BIC is the estimate from the Bayesian framework to compare models (Schwarz, 1978) as;

$$BIC = -2LL + p(\ln(N)), \quad (17)$$

where, N is the sample size. A higher BIC value indicates more complex model with larger sample sizes.

Research questions

The central goal of this study is to evaluate psychometric models such as CTT, traditional IRT, and TRT models using a fifth grade reading comprehension test with a large data set. These model comparisons and item analyses will help both researchers and practitioners in the area of reading comprehension with the guidelines for construction and selection of tests and items, and for the decision making process related to literacy research and education. In order to address best-fitting statistical model for a reading comprehension test and to provide precise items and test information which is constructed on the basis of CCSS, specific research questions of this present study are as follows:

1. What are the similarities and dissimilarities between CTT and IRT? The results from CTT would be compared to those of traditional IRT model.

2. Which IRT model shows the best-fit for a reading comprehension test? The best-fitting IRT model would be selected with three model comparison criteria (-2LL, AIC, and BIC).
3. Do the reading passages show testlet effects? IRT and TRT models would be compared with model comparison criteria (-2LL, AIC, and BIC).
4. What are the differences between the item parameter estimates obtained using TRT and IRT models? The estimated item parameters and the testlet parameters would be computed and compared.

CHAPTER THREE: METHODS

Participants

The archival data contained a total of 10,897 participants for a 5th grade reading comprehension test across 15 states. The data were collected during the 2012/2013 academic year as the results for a benchmark assessment. There were no missing data for each item and the total scores from the test. Some demographic variables have missing data including gender, race, English as a Second Language (ESL), Special Education (SpEd), and free lunches. These demographic variables will not be analyzed because the current study is mainly a psychometric validation project on test item levels. There was no information in the data set which may lead to the identification of participants. All data points were assigned by the subject ID numbers. In order to provide basic information on several demographic variables, some descriptive statistics were computed from the available data points. There were a total of 2,505 data points for the gender variable. There were 1,223 (48.8 %) females and 1,282 (51.2%) males. For the ethnicity variable, 1,858 students were responded. A total of 425 (14.9%) students reported to be African American, 48 (1.7%) American Indian, 69 (2.4%) Asian, 1,316 (46.0%) White students.

Measurement

The test items were developed by a for-profit testing company in the United States. The test items are based on CCSS and have 33 items, 11 passages, and 4 categories. In this study, only 22 items with 7 passages in the categories of reading

standards of literature (RL), reading standards of informational text (RI), reading standards of foundation skills (RF), and language standards (L) will be analyzed. The writing category (W) was excluded because it was not part of reading comprehension. CCSS (2014) addresses the RL code for 5th grade

(http://www.corestandards.org/assets/CCSSI_ELA%20Standards.pdf, p. 12) as:

RL.5.1. Quote accurately from a text when explaining what the text says explicitly and when drawing inferences from the text.

RL.5.2. Determine a theme of a story, drama, or poem from details in the text, including how characters in a story or drama respond to challenges or how the speaker in a poem reflects upon a topic; summarize the text.

RL.5.3. Compare and contrast two or more characters, settings, or events in a story or drama, drawing on specific details in the text (e.g., how characters interact).

RL.5.4. Determine the meaning of words and phrases as they are used in a text, including figurative language such as metaphors and similes.

RL.5.7. Analyze how visual and multimedia elements contribute to the meaning, tone, or beauty of a text (e.g., graphic novel, multimedia presentation of fiction, folktale, myth, poem).

The RI skills for 5th grade are described as

(http://www.corestandards.org/assets/CCSSI_ELA%20Standards.pdf, p. 14);

RI.5.1. Quote accurately from a text when explaining what the text says explicitly and when drawing inferences from the text.

RI.5.2. Determine two or more main ideas of a text and explain how they are supported by key details; summarize the text.

RI.5.3. Explain the relationships or interactions between two or more individuals, events, ideas, or concepts in a historical, scientific, or technical text based on specific information in the text.

RI.5.5. Compare and contrast the overall structure (e.g., chronology, comparison, cause/effect, problem/solution) of events, ideas, concepts, or information in two or more texts.

RI.5.6. Analyze multiple accounts of the same event or topic, noting important similarities and differences in the point of view they represent.

RI.5.7. Draw on information from multiple print or digital sources, demonstrating the ability to locate an answer to a question quickly or to solve a problem efficiently.

RI.5.9. Integrate information from several texts on the same topic in order to write or speak about the subject knowledgeably.

The RF in CCSS is (http://www.corestandards.org/assets/CCSSI_ELA%20Standards.pdf, p. 17);

RF.5.4 Read with sufficient accuracy and fluency to support comprehension.

The Language standards (L) skills are

(http://www.corestandards.org/assets/CCSSI_ELA%20Standards.pdf, p. 28-29);

L.5.2. Demonstrate command of the conventions of standard English capitalization, punctuation, and spelling when writing.

L.5.3. Use knowledge of language and its conventions when writing, speaking, reading, or listening.

L.5.4. Determine or clarify the meaning of unknown and multiple-meaning words and phrases based on grade 5 reading and content, choosing flexibly from a range of strategies.

L.5.5. Demonstrate understanding of figurative language, word relationships, and nuances in word meanings.

Passage 1 contains four items: Item 1 (Standard Code = RL.5.2), Item 2 (RL.5.1), Item 3 (RL.5.3), and Item 4 (L.5.4). Passage 2 has Item 5 (RL.5.4), Item 6 (RL.5.4), and Item 7 (L.5.5). In Passage 3, Item 8 (RL.5.1), Item 9 (RL.5.7), Item 10 (RL.5.2), and Item 11 (L.5.3) are assigned. Passage 4 contains Item 12 (RL.5.3), Item 13 (RL.5.3), and Item 14 (L.5.7). In Passage 5, Item 15 (RL.5.4), Item 16 (RL.5.1), and Item 17 (L.5.2) are involved. Passage 6 has two items, (L.5.2) and (L.5.3). Passage 7 has Item 20 (RL.5.5),

Item 21 (RL.5.9), and Item 22 (L.5.6). All items were scored 1 and 0 where 1 represents a correctly answered item, and 0 represents a wrong item. The total scores were computed as the sum of all correctly answered item scores.

Procedure

The mean and standard deviation for each item and the total score were computed for CTT analyses. The mean of each item was the proportion of students who correctly answered the item. The mean value of each item is identical to the p -value in CTT. Cronbach's α along with item-test correlation was computed as part of CTT analyses. In order to apply any IRT analyses, an exploratory factor analysis (EFA) should be utilized to confirm the unidimensionality assumption. Once unidimensionality was confirmed, three IRT models (1, 2, and 3PLM) were tested with the data. The best-fitting model was selected, and the three comparison criteria (-2LL, AIC, and BIC) were recorded. The final step of validation was the application of the constrained and general TRT models. The traditional IRT, constrained TRT, and general TRT were compared with three comparison criteria. The estimated item parameters and the testlet parameters along with item and test information were also computed and compared. The advantages and disadvantages of IRT over CTT, and TRT over IRT were discussed in conjunction with previous findings.

The SAS (Statistical Analysis System) software was used to compute descriptive statistics, item-test correlation for each item, and EFA in CTT. The WinBUGs software (Spiegelhalter, Thomas, & Best, 2003) was used to analyze the traditional IRT and two

TRT models with the Markov chain Monte Carlo (MCMC) algorithm. The chain length was set to 10,000 with the burn-in of 5,000. The estimated item and testlet parameters were posterior means and variances which were obtained by the only last 1,000 draws of MCMC chain. The first 4,000 iterations were discarded. In order to estimate item difficulty parameters, b_i , and person ability, θ , of traditional IRTs and testlet effects, γ_{di} , of TRT models, normal priors were used: $b_i \sim N(0, 1)$, $\theta \sim N(0, 1)$, and $\gamma_{di} \sim N(0, 1)$. Truncated normal distribution priors were used for item discrimination parameters in the IRT and TRT models: $a_i \sim N(0, 1) I(0, \infty)$, where $I(0, \infty)$ indicated that observations of the item discrimination parameters occur above zero. For the pseudo-chance parameter, c_i , the beta distribution priors were used: $c_i \sim \text{beta}(1, 1)$ which indicated the uniform distribution over the interval between 0 and 1. Three comparison criteria (-2LL, AIC, and BIC) were applied to test significant differences among various IRT and TRT models through the WinBUGs program. The WinBUGs codes for three IRT and two TRT models were presented in Appendix 1.

Model comparison between any pair of models can be conducted by treating any of the comparison criteria (-2LL, AIC, and BIC) as an approximation of the χ^2 distribution with a corresponding df . The computational formula of $df = 2^p - np - 1$ for is commonly used with any approximation of the the χ^2 distribution for the IRT models (Cai, Maydeu-Olivares, Coffman, & Thissen, 2006). For this formula, n represents the degrees of freedom (df) for a given IRT model such as 1PLM ($n = 1$), 2PLM ($n = 2$), or 3PLM ($n = 3$), and p represents the number of items in the test. However, when test items (p), are greater than 20, the df formula by Cai, Maydeu-Olivares, Coffman, and

Thissen (2006) is not appropriate for any IRT models (Mair, Reise, & Bentler, 2008). In this study, following Guyer and Thompson (2011), the degree of freedom for the goodness-of-fit test was computed with $df = G - \#$ of parameters, where $G = 15$ in each model.

Once the comparison criterion and corresponding df s are identified, a model comparison can be performed with the following procedure. First, the $\Delta\chi^2$ must be computed from any pair of the comparing models. This $\Delta\chi^2$ can be obtained by computing the difference value of any chosen comparison criterion from -2LL, AIC, or BIC for the two comparing models, either for any two IRT models or two TRT models. Then, the $\Delta\chi^2$ would be divided by the Δdf which was the df difference between any comparing models. For example, if one wants to compare 1PLM and 2PLM IRT models, the $\Delta\chi^2$ can be computed by subtracting -2LL (or AIC, or BIC) of the 2PLM from that of the 1PLM. Then, the Δdf can be obtained by subtracting the df of the 2PLM from the df of the 1PLM. The final step is to divide the $\Delta\chi^2$ by the Δdf , and to follow through the regular χ^2 test.

CHAPTER FOUR: RESULTS

CTT and IRT analysis results

In this study, CTT, three different IRT and two TRT models were applied to a large data set ($N = 10,897$) for a 5th grade reading comprehension test based on CCSS. Of the total 33 items, 22 items from 7 reading passages were employed for analysis, excluding 11 items and 4 passages associated with writing items. For the item and test analyses in CTT, the p -value and item-test correlation along with the Cronbach's alpha were computed. The p -value is identical to the mean of each item. Table 1 showed the CTT analysis results. The mean values of the test items were ranged from .35 ($SD = .48$) to .82 ($SD = .38$). Easier items were Item 2 ($M = .82, SD = .38$) in Passage 1 and Item 7 ($M = .82, SD = .39$) in Passage 2. About 82% students (8,935) answered both items correctly. Item 19 ($M = .35, SD = .48$) in Passage 6 and Item 12 ($M = .38, SD = .49$) in Passage 4 were two of the hardest items on the test. Only 35% of 5th graders (3,813) answered Item 19 correctly. The item-test correlation is similar to the item discrimination index in CTT. It indicates the correlation between each item and the total test score excluding the comparing item. The higher the item-test correlation, the higher the relationship between the item and the total test score. The item-test correlation values ranged from .17 (Item 12) to .50 (Item 17). The correlations of Items 5 and 12 with the overall test were .21 and .17, while Item 7 and Item 17 correlated at .47 and .50, respectively. The Cronbach's alpha for the 22-item test was .79, which was a relatively good internal consistency reliability index.

Table 1

Descriptive statistics of test items from CTT and estimated item parameters of 3PLM

Testlets	Items	CTT			3PLM					
		<i>M</i>	<i>(SD)</i>	<i>Item-Total Correlation</i>	<i>a</i>	<i>(SD)</i>	<i>b</i>	<i>(SD)</i>	<i>c</i>	<i>(SD)</i>
1	1	0.57	(0.50)	.37	1.04	(0.04)	-0.27	(0.04)	0.02	(0.02)
	2	0.82	(0.38)	.43	1.71	(0.08)	-1.24	(0.07)	0.06	(0.04)
	3	0.63	(0.48)	.35	1.03	(0.05)	-0.49	(0.09)	0.06	(0.04)
	4	0.69	(0.46)	.35	1.00	(0.04)	-0.90	(0.07)	0.04	(0.03)
2	5	0.63	(0.48)	.21	0.52	(0.03)	-0.90	(0.17)	0.05	(0.04)
	6	0.73	(0.45)	.43	1.49	(0.08)	-0.79	(0.08)	0.09	(0.04)
	7	0.82	(0.39)	.47	2.12	(0.11)	-1.07	(0.08)	0.10	(0.05)
3	8	0.45	(0.50)	.24	0.60	(0.03)	0.47	(0.08)	0.02	(0.02)
	9	0.62	(0.49)	.24	0.65	(0.04)	-0.62	(0.18)	0.08	(0.05)
	10	0.76	(0.43)	.42	1.37	(0.04)	-1.07	(0.05)	0.03	(0.03)
	11	0.58	(0.49)	.42	1.77	(0.11)	0.06	(0.05)	0.18	(0.02)
4	12	0.38	(0.49)	.17	1.74	(0.17)	1.56	(0.05)	0.28	(0.01)
	13	0.75	(0.43)	.43	1.64	(0.08)	-0.75	(0.07)	0.16	(0.04)
	14	0.73	(0.45)	.36	1.36	(0.09)	-0.52	(0.10)	0.26	(0.04)
5	15	0.48	(0.50)	.25	0.68	(0.06)	0.38	(0.16)	0.07	(0.04)
	16	0.53	(0.50)	.36	1.27	(0.06)	0.19	(0.05)	0.13	(0.02)
	17	0.73	(0.45)	.50	2.55	(0.14)	-0.49	(0.04)	0.20	(0.02)
6	18	0.61	(0.49)	.30	0.80	(0.04)	-0.50	(0.09)	0.06	(0.03)
	19	0.35	(0.48)	.23	0.78	(0.08)	1.41	(0.07)	0.10	(0.02)
7	20	0.54	(0.50)	.30	1.13	(0.09)	0.37	(0.07)	0.20	(0.03)
	21	0.48	(0.50)	.27	0.68	(0.03)	0.25	(0.07)	0.03	(0.02)
	22	0.69	(0.46)	.36	1.08	(0.06)	-0.77	(0.10)	0.07	(0.04)

Note. $N = 10,897$

Before conducting IRT analyses, an exploratory factor analysis was utilized to confirm the unidimensionality assumption. As shown in Table 2, the results demonstrated a relatively clear one-factor solution, explaining around 20% of the data variance by the first factor (eigenvalue = 4.27). Figure 4 also showed that the slope of the curve was clearly reaching a steady rate from the second factor. Based on the variance of eigenvalue and scree plot, we could conclude that the data met the unidimensionality assumption. Therefore, the application of IRT models to the data was justified.

As shown in Table 3, the model-fit indices of three IRT models indicated that 3PLM fitted the given data best when 3PLM was compared with 1PLM and 2PLM. The -2LL difference (= 264000 – 261700) test with Δdf (= 308 – 286) between 1PLM and 2PLM was 2300 ($p < .01$). The AIC difference was 2200 ($p < .01$) and the BIC difference was 2100 ($p < .01$) between 1PLM and 2PLM. The comparison between 1PLM and 2PLM revealed that 2PLM was a better fit for the 5th grade reading comprehension test. The -2LL difference between 2PLM (= 261700) and 3PLM (= 261000) with Δdf (= 286 – 264) was 700 ($p < .01$). The AIC difference with Δdf (= 22) was 600 ($p < .01$) and the BIC difference was 500 ($p < .01$). All three comparison criteria showed that the 3PLM was the best-fitting model for the data.

Table 2

Eigenvalues of the correlation matrix

	<i>Eigenvalue</i>	<i>Difference</i>	<i>Proportion</i>	<i>Cumulative</i>
1	4.27	3.21	0.19	0.19
2	1.06	0.07	0.05	0.24
3	0.99	0.02	0.05	0.29
4	0.97	0.01	0.04	0.33
5	0.96	0.03	0.04	0.38
6	0.93	0.00	0.04	0.42
7	0.92	0.01	0.04	0.46
8	0.91	0.01	0.04	0.50
9	0.90	0.02	0.04	0.54
10	0.88	0.02	0.04	0.58
11	0.87	0.02	0.04	0.62
12	0.85	0.03	0.04	0.66
13	0.82	0.01	0.04	0.70
14	0.81	0.01	0.04	0.73
15	0.80	0.02	0.04	0.77
16	0.78	0.03	0.04	0.81
17	0.75	0.01	0.03	0.84
18	0.75	0.01	0.03	0.87
19	0.74	0.04	0.03	0.91
20	0.70	0.02	0.03	0.94
21	0.68	0.04	0.03	0.97
22	0.63		0.03	1.00

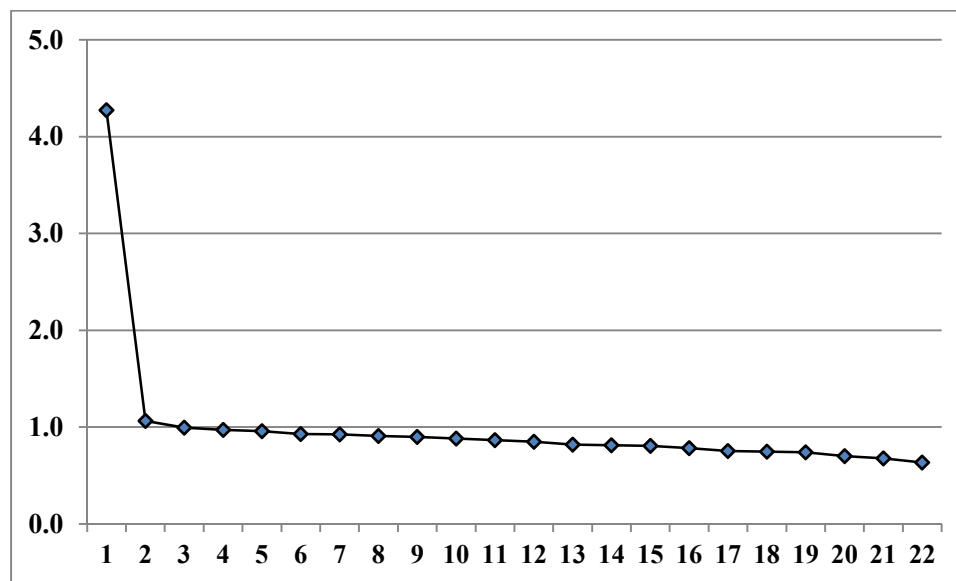


Figure 4. Scree plot of eigenvalues

Table 3

Model-fit indices of three traditional IRT models.

IRT models	<i>NP</i>	<i>df</i>	<i>-2LL</i>	<i>AIC</i>	<i>BIC</i>	<i>-2LL_{difference}</i>
1PLM	22	308	264000	264000	264200	
2PLM	44	286	261700	261800	262100	$\chi^2 (22) = 2300$
3PLM	66	264	261000	261200	261600	$\chi^2 (22) = 700$

In accordance with the results of the item-test correlation in CTT, item discrimination parameters in 3PLM showed that Item 7 ($a_7 = 2.12$) and Item 17 ($a_{17} = 2.55$) were the best items in terms of the item discrimination parameter estimates on the location of the person ability level of -1.07 and -0.49, respectively. The item-test

correlation for Item 7 was .47, and for Item 17 was .50 in CTT. Both CTT and IRT analyses indicated that these two items were the best discriminating items. There were several additional items with high a -parameter values. In general, there was agreement between CTT and IRT results in terms of item discrimination indices. For the middle of ability distribution ($\theta \approx 0$), Item 11 ($a_{11} = 1.77$) in Passage 3 yielded a better discrimination value than Item 21 ($a_{21} = .68$) in Passage 5.

Item 12 in Passage 4 showed a higher item discrimination function ($a_{12} = 1.74$) for a relatively higher ability level ($\theta = 1.56$) of 5th graders. On the contrary, item-total correlation in CTT revealed that Item 12 ($r = .17$) was worst discriminating item in the test. This item demonstrated an evident difference between CTT and IRT. The IRT results showed that the following items were listed in an ascending order from the lowest to highest values in item discriminating parameter estimates: Item 5 ($a_5 = 0.52$), Item 8 ($a_8 = 0.60$), Item 9 ($a_9 = 0.65$), Item 15 ($a_{15} = 0.68$), and Item 21 ($a_{21} = 0.68$). They functioned poorly to distinguish between examinees who had knowledge of the item and those who did not. The results in CTT showed that Items 12 ($r = 0.17$), 5 ($r = 0.21$), 19 ($r = 0.23$), 8 ($r = 0.24$), and 9 ($r = 0.24$) were discriminating poorly. There were minor discrepancies between CTT and IRT results. On the other hand, The IRT results showed that the following items showed high item discriminating indices in a descending order: Items 17 ($a_{17} = 2.58$), 7 ($a_7 = 2.09$), 11 ($a_{11} = 1.78$), and 12 ($a_{12} = 1.76$). They provided high item discrimination parameter estimates. In CTT, Items 17 ($r = 0.50$), 7 ($r = 0.47$), 13 ($r = 0.43$), 6 ($r = 0.43$), and 2 ($r = 0.43$) were all relatively highly correlated with the overall test. The ICCs for the test items which had good item discrimination values

(Items 7, 11, 12, and 17) and poor item discrimination values (Items 5, 8, 9, and 15) were presented in Figure 5. The ICCs of Items 7, 11, 12, and 17 had sharper slopes on their ability ranges than Items 5, 8, 9, and 15.

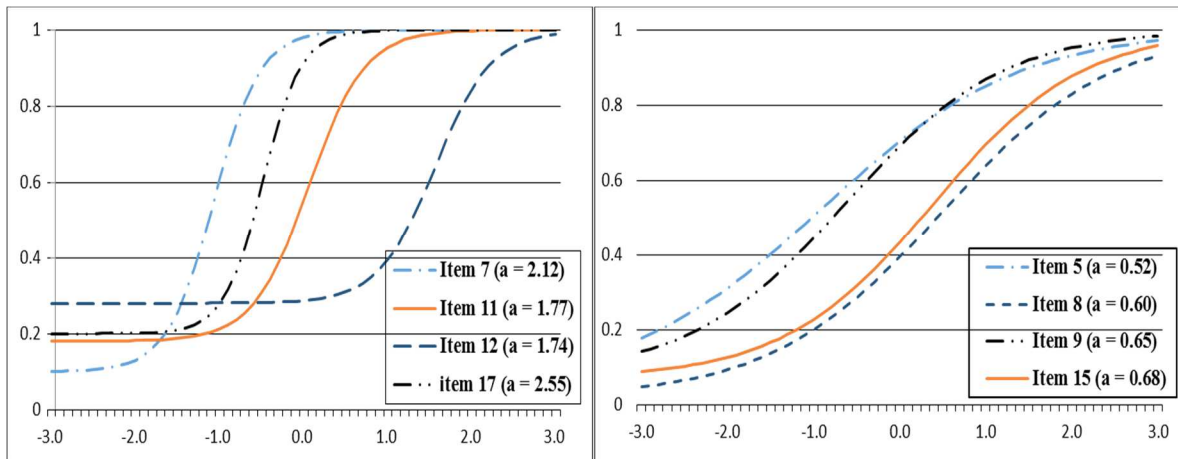


Figure 5. Each four items for good and poor item discrimination parameters

In order to specify positive values for item discrimination parameters, normal priors to item discrimination parameters with a mean of 0 and standard deviation of 1 were given with a truncated threshold value of 0. The MCMC algorithm in WinBUGs estimated the positive posterior means for item discrimination parameters. Figure 6 provided an illustration of the posterior probability density functions (PDFs) for four items with good item discrimination parameter estimates and four items with poor item discrimination parameters.

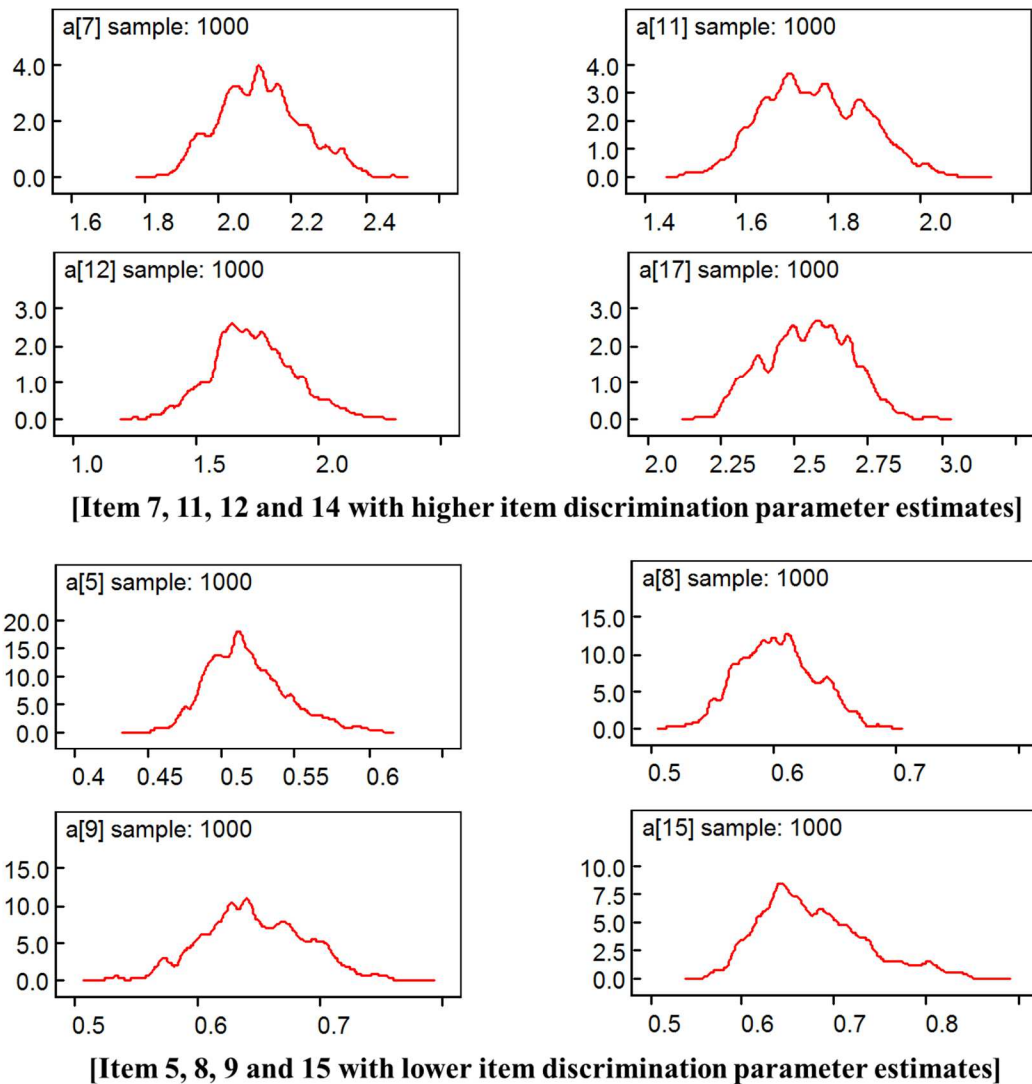


Figure 6. Illustration of posterior PDFs of item discrimination parameters

For the posterior PDF distributions, the a -parameter estimates are placed along the horizontal abscissa, while the frequencies by the last 1,000 samples of MCMC chain are located along the vertical ordinate. As shown in Figure 6, the mean of the estimated a -parameter values of the high discriminating items (Items 7, 11, 12, and 17) ranged between 1.74 ($SD = .17$) and 2.55 ($SD = .14$). The mean of low item discriminating items

(Items 5, 8, 9, and 15) ranged between .52 ($SD = .03$) and .68 ($SD = .06$). If the a -parameter values were above 1.00, the items were considered as high discriminating items.

According to item difficulty estimates in IRT, Items 12 and 19 were the most difficult items in the reading comprehension test, $b_{12} = 1.56$ and $b_{19} = 1.41$, respectively. The p -values in CTT also indicated that Item 12 ($M = .38$) and Item 19 ($M = .35$) were the hardest items to answer correctly. Interestingly, as commented above, Item 12 ($a_{12} = 1.74$) discriminated well for examinees who were higher level of ability ($\theta = 1.56$) while Item 19 ($a_{19} = 0.78$) distinguished poorly for examinees who were at a similar level of ability ($\theta = 1.41$). Item 2 ($b_2 = -1.24$) in Passage 1, Item 7 ($b_7 = -1.07$) in Passage 2, and Item 10 ($b_{10} = -1.07$) in Passage 3 were easy items to answer. Among these items, Item 7 and Item 10 had the same level of item difficulty parameter values while, in CTT, Item 7 ($M = 0.82$) was easier than Item 10 ($M = 0.76$). The PDFs of item difficulty for two difficult items and two easy items were displayed in Figure 7.

The c -parameter in 3PLM is the probability of answering an item correctly by only guessing. Item 12 ($c_{12} = .28$) and Item 14 ($c_{14} = .26$) in Passage 4 had the highest pseudo-chance parameter values. Item 1 ($c_1 = .02$) in Passage 1, Item 8 ($c_8 = .02$) in Passage 3, and Item 21 ($c_{21} = .03$) in Passage 7 showed the lowest probabilities for getting the correct answers by guessing. The PDFs of item pseudo-chance parameter for these five items (12, 14, 1, 8, and 21) were displayed in Figure 8. The probability density graphs of Items 1, 8, and 21 were positively skewed with mean values of 0.02 ($SD = 0.02$), 0.02 ($SD = 0.02$), and 0.03 ($SD = 0.02$), respectively.

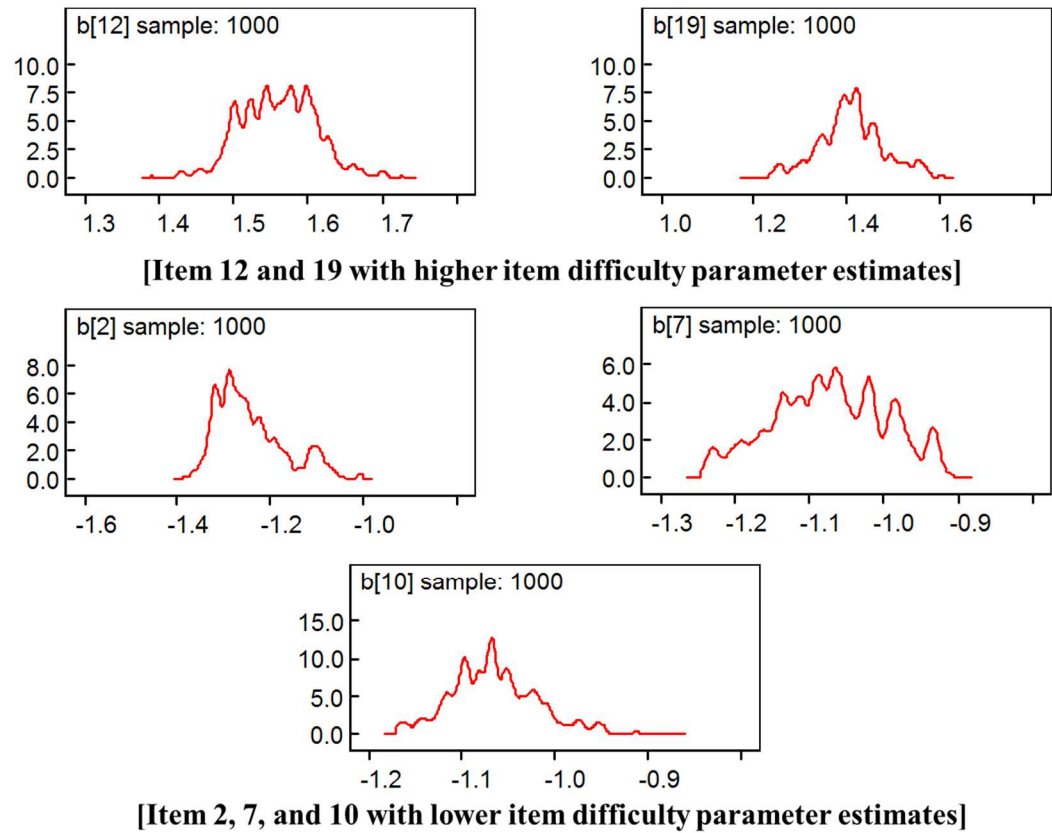


Figure 7. Probability density functions of item difficulty parameters

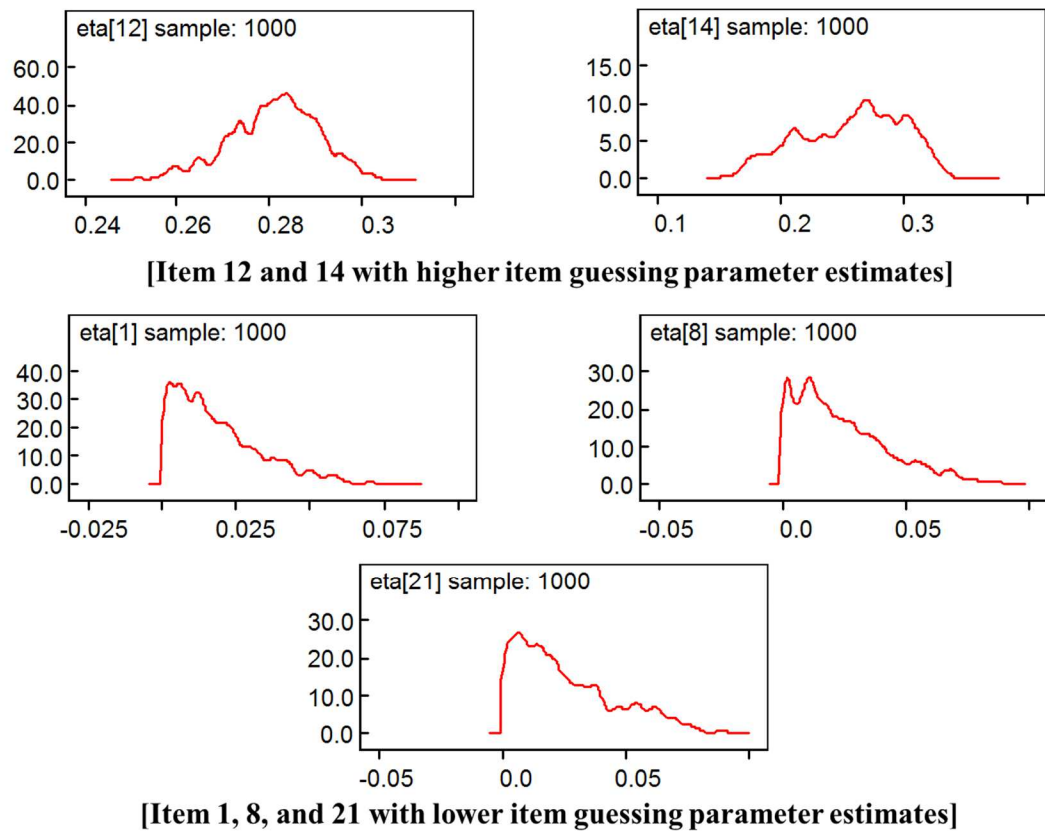


Figure 8. Probability density functions of the pseudo-chance parameters

The convergence processes for the pseudo-chance parameters which were associated with MCMC algorithm experienced more computational complications than those with item discrimination and difficulty parameters. Figure 9 illustrated examples of the item parameter estimation histories of Item 9. Although 3PLM fitted best for a reading comprehension test (see Table 3), iteration histories with Markov chains provided the evidence of convergence problem when estimating pseudo-chance parameters.

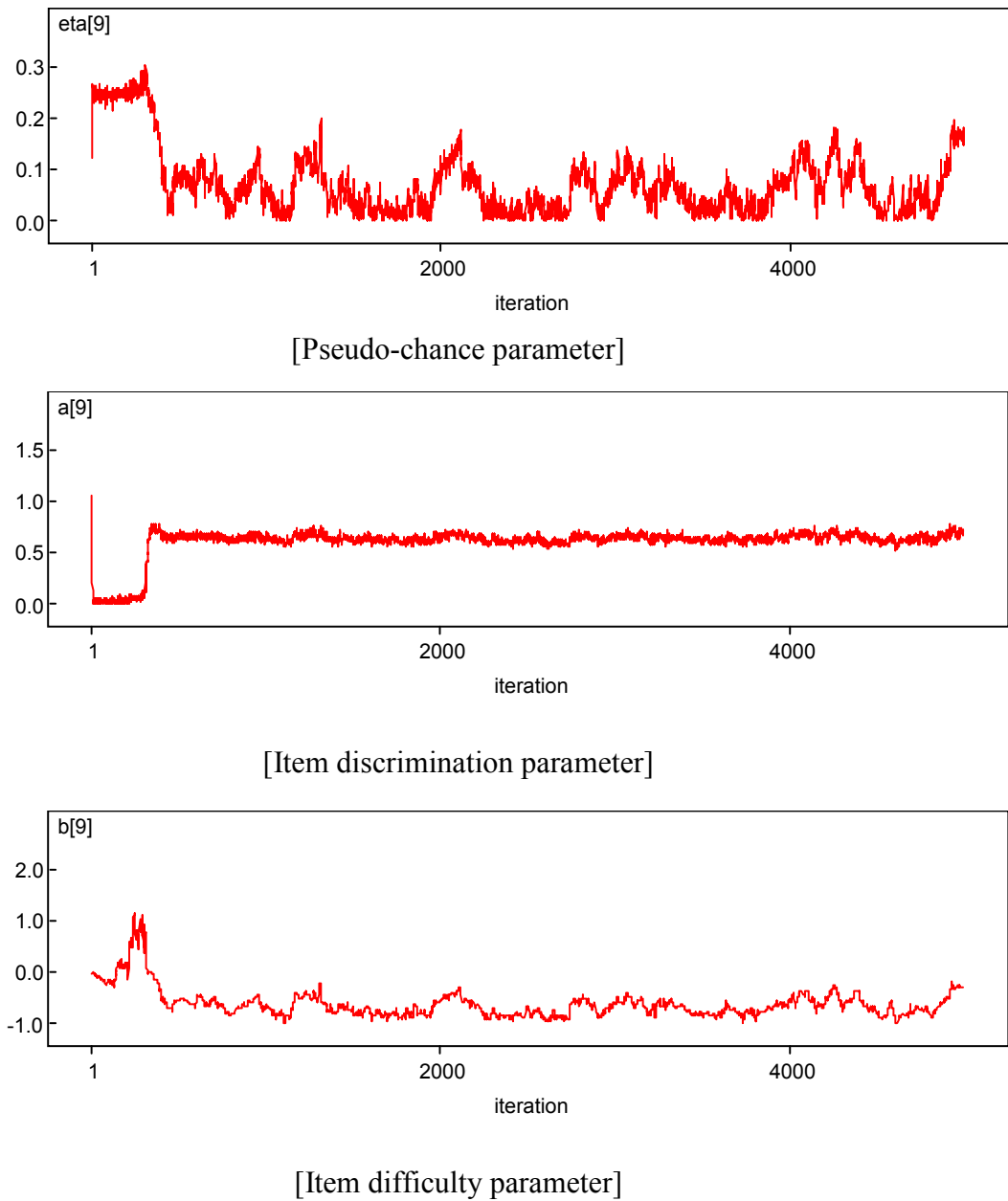


Figure 9. Iteration histories of item parameter estimates of 3PLM

In order to solve this convergence problem and provide stable estimates, the number of iterations in MCMC algorithm was set to 5,000 in this study. The first 4,000 iterations were discarded and only the last 1,000 draws of iteration was recorded for the

IRT estimates. Figure 10 depicted the acceptance rates of 3PLM with 5,000 iterations. The rate was stabilized after approximately 1,500 iterations.

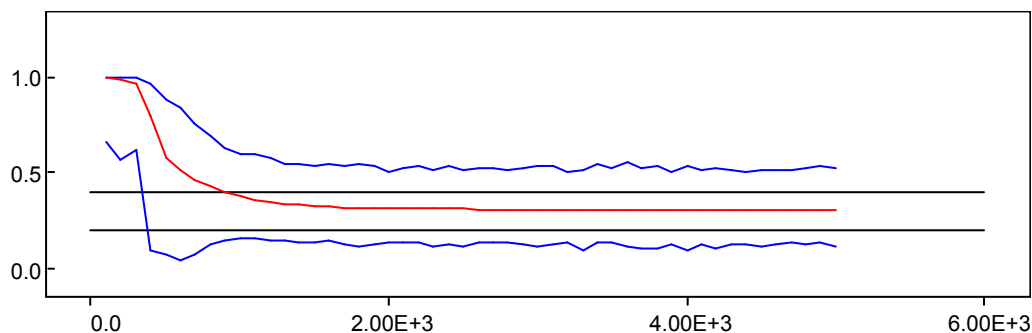


Figure 10. Acceptance rates of MCMC iterations for 3PLM

Model comparisons of 3PLM and TRT models

Because the 3PLM unidimensional IRT model was selected as the best-fitting model for the data, two 3PL TRT models (constrained and generalized) were applied and compared to the 3PLM unidimensional IRT model in order to discern the effect of testlet parameters. As shown in Table 4, no significant model fit difference was found with the -2LL statistics among the unidimensional 3PLM, constrained TRT, and generalized TRT model. One issue with the -2LL statistics has been known as a test statistic which does not consider the sample size and the model complexity associated with the number of parameters in the model. The AIC index also showed an identical value of 261,200 for all three models. As one can see from the formula of AIC, it does not consider the complexity of the model associated with the number of parameters. The BIC takes into

account both sample size and complexity of the model. It is somewhat puzzling that the BIC indices demonstrate a favorable result towards the unidimensional 3PLM IRT model. The testlet effect was not evidently revealed for this data set.

Table 4

Model-fit indices of 3PLM and two TRT models.

IRT/ TRT models	<i>NP</i>	<i>df</i>	<i>-2LL</i>	<i>AIC</i>	<i>BIC</i>
Unidimensional 3PLM	66	264	261000	261200	261600
Constrained 3PL TRT	73	257	261000	261200	261700
Generalized 3PL TRT	95	264	261000	261200	261900

Investigating other estimates including ability, item parameters, and testlet parameters may shed light to this phenomenon. In order to compare person ability and item parameter estimates among unidimensional 3PLM, constrained TRT, and generalized TRT, correlation coefficients were computed for different parameter estimates and scatter plots were constructed. Figure 11 presented the ability estimate scatter plots between unidimensional 3PLM and constrained TRT, and between unidimensional 3PLM and generalized TRT along with correlation coefficients. As shown in Figure 11, the estimated person abilities from three different models were highly correlated to each other ($r = .99, p < .01$). It was clearly shown that different IRT models, either the unidimensional IRT model or two different TRT models, yielded

consistent estimates for the person ability. The invariance of parameter estimates has been known for IRT models, which is an obvious advantage of IRT over CTT.

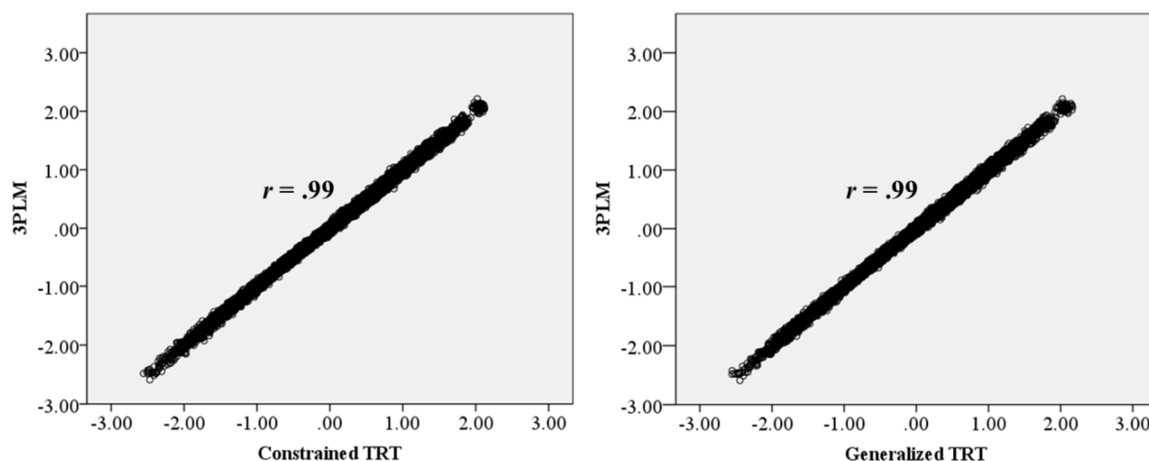


Figure 11. Comparison of the person ability estimates between 3PLM and TRT models

The estimated ability from 3PLM ranged from -2.57 to 2.16 with a mean of .001 and standard deviation of .895. The constrained TRT model yielded the ability range of -2.56 through 2.12 with a mean of .0002 and standard deviation of .895. The range of the person ability from the generalized TRT model was from -2.55 to 2.17 with a mean of -.0005 and standard deviation of .892. The histograms of the estimated ability parameters from the unidimensional IRT and two TRT models were displayed in Figure 12. As it was expected, all three IRT models produced consistent ability estimations with a mean value around 0.00 and standard deviation of 1.00. Along with a correlation among the three IRT-TRT models in Figure 11 and Figure 12, they demonstrated the invariant parameter estimates of IRT-TRT models.

The difference of estimates of a person's abilities between 3PLM and constrained TRT model, between 3PLM and generalized TRT, and between constrained TRT and generalized TRT was normally distributed with a mean of 0.00 and standardized deviation of .05. The histogram for the ability parameter estimate differences among the IRT-TRT models was presented in Figure 13. These graphs illustrated a minimal difference among estimated ability values from unidimensional IRT and two TRT models. All three graphical displays of estimated ability distributions from Figures 11 through 13 revealed an almost identical result for invariant estimates of students' ability.

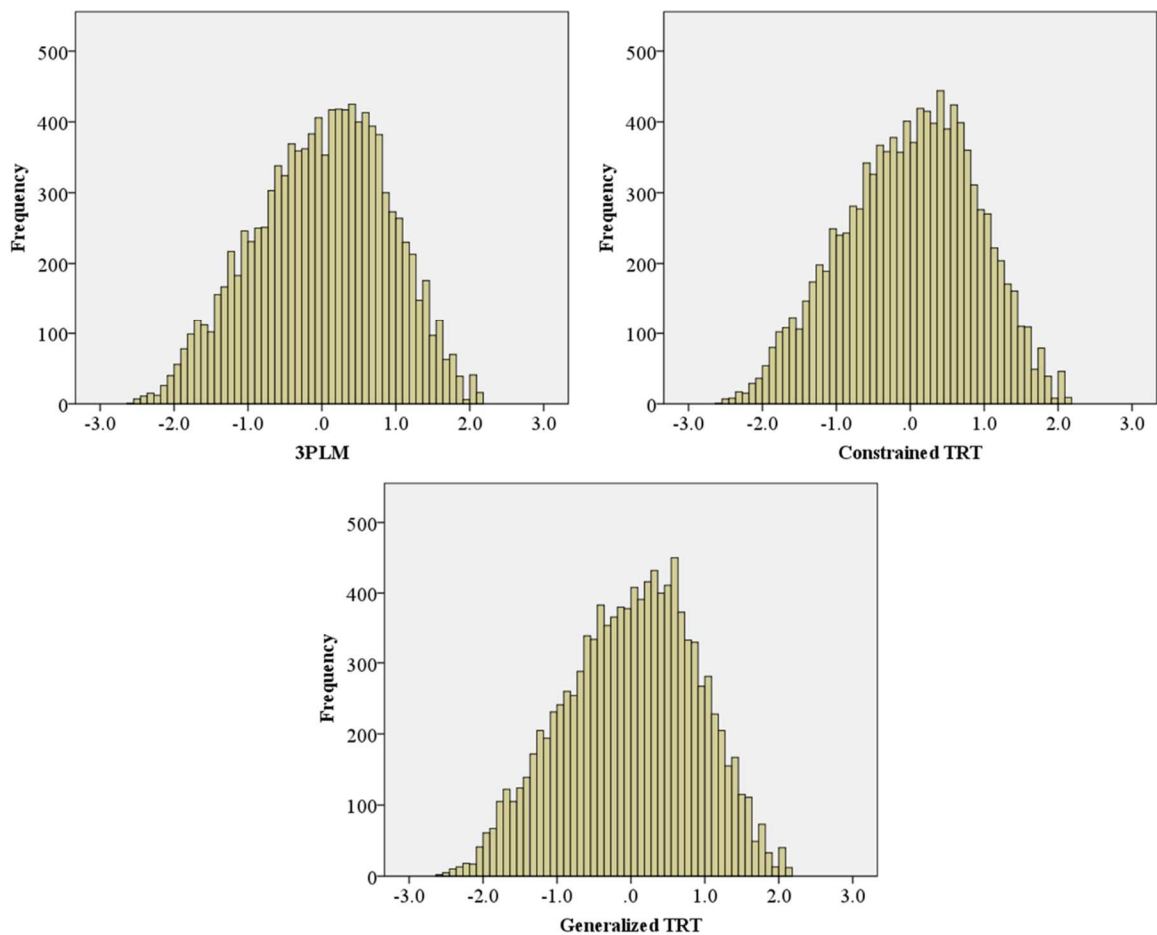


Figure 12. Distribution of the person ability estimates from the IRT and TRT models

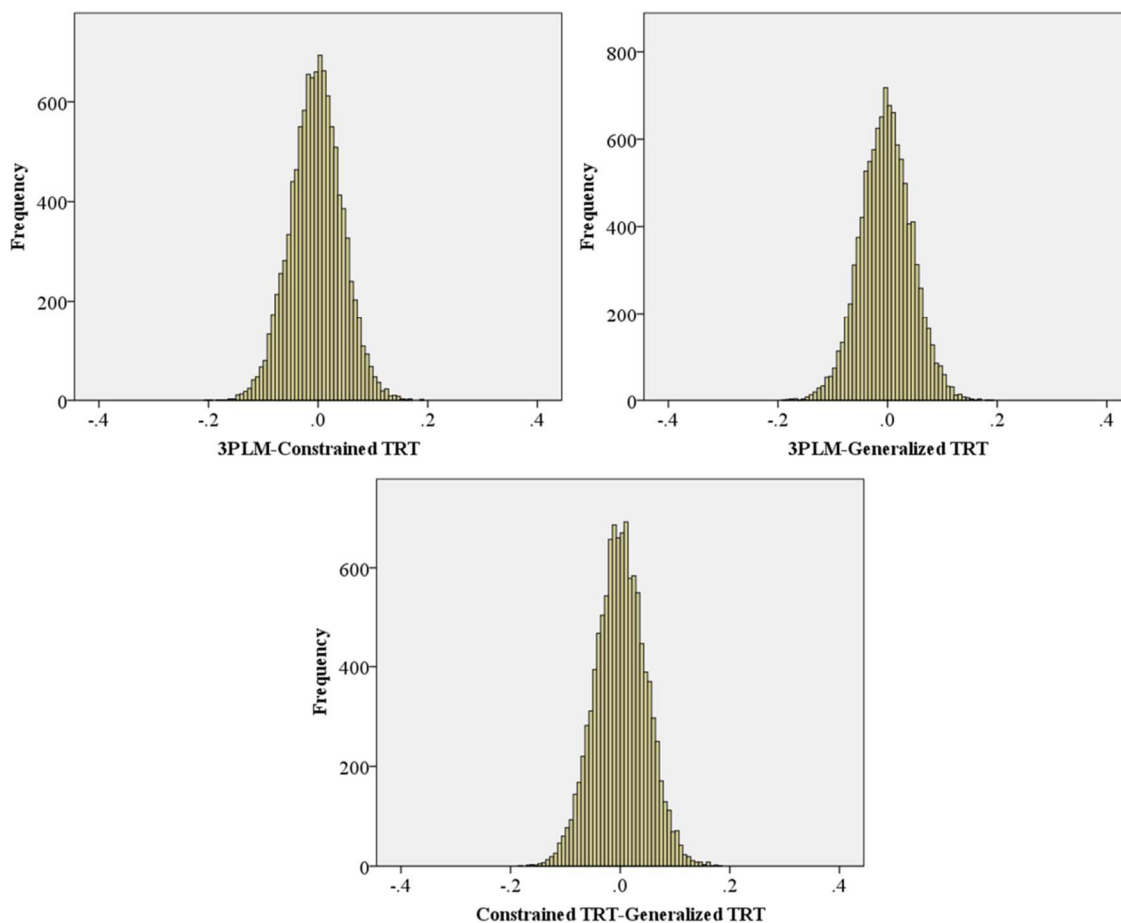


Figure 13. Difference score distributions of ability estimates from IRT and TRT models

Testlet effects

The item parameter estimates and their standard errors along with the testlet mean and variance are presented in Table 5. As shown in Table 5 for the constrained TRT model and in Table 1 for the traditional 3PLM, item discrimination parameter estimates were not dramatically different between the two models. The estimates of item discrimination parameters were ranged from 0.55 to 2.58 in constrained TRT model. These item discrimination parameter estimates of 22-item by the constrained TRT model

were associated with both θ and γ_{ai} . However, item difficulty parameter estimates in Table 5 were quite different to those of 3PLM in Table 1. For instance, the constrained TRT model yielded that item 17 ($b_{17} = -1.57$) and 18 ($b_{18} = -1.34$) were easiest in the test while, in 3PLM, these two items were placed on the range of moderately easy with values of $b_{17} = -0.48$ and $b_{18} = -0.54$, respectively. Item 3 ($b_3 = 0.48$) was a moderately hard item in Table 5 (TRT), but was classified as a moderately easy item ($b_3 = -0.59$) in Table 1 (3PLM).

As described in Chapters 1 and 2, the testlet effects should be considered in order to properly interpret the difference between item difficulty parameter estimates of the constrained TRT model and that of the 3PLM. The variance of testlet (σ^2_{γ}) indicated the degree of LID as the strength of relationship among items in each passage (Wainer, Bradlow, & Wang, 2007). Wang, Bradlow, and Wainer (2002) addressed that there was no indication of a testlet effect among items when the testlet variance was less than 0.04. In the current study, the variances of testlet effects were negligibly small ($\widehat{\sigma}_{\gamma_1}^2 = 0.01$, $\widehat{\sigma}_{\gamma_2}^2 = 0.00$, $\widehat{\sigma}_{\gamma_3}^2 = 0.01$, $\widehat{\sigma}_{\gamma_4}^2 = 0.00$, $\widehat{\sigma}_{\gamma_5}^2 = 0.01$, $\widehat{\sigma}_{\gamma_6}^2 = 0.01$, and $\widehat{\sigma}_{\gamma_7}^2 = 0.01$).

However, in Table 5, the mean of the testlet parameter estimates were non-negligible and were related to item difficulty estimates. For example, the item difficulty estimates of Items 1, 2, 3, and 4 in the first testlet ($\widehat{M}_{\gamma_1} = -1.01$) were relatively higher ($b = 0.73, -0.24, 0.48, \text{ and } 0.10$) than the item difficulty estimates in 3PLM ($b = -0.27, -1.24, -0.49, \text{ and } -0.90$), respectively. Similar to the testlet 1, the items in the Testlet 2 ($\widehat{M}_{\gamma_2} = -0.06$) and Testlet 4 ($\widehat{M}_{\gamma_4} = -0.14$) from the constrained TRT model provided higher values

of item difficulties than those of 3PLM. The lower the testlet mean, the higher the item difficulty parameter estimates in the constrained TRT model. With high testlet means, the estimates of item difficulties were lower. The item difficulty parameter estimates of Items 15, 16, and 17 ($b = -0.72, -0.92, \text{ and } -1.57$) in Testlet 5 ($\widehat{M}_{\gamma_5} = 1.09$) were relatively lower than the items in 3PLM ($b = 0.38, 0.19, \text{ and } -0.49$), respectively.

The generalized TRT model provided quite different results from the constrained TRT model as presented in Table 6. The generalized TRT model provided two types of item discrimination parameter (a_{1i} and a_{2i}). The first a -parameter (a_{1i}) estimates of the generalized TRT model ranged from 0.53 to 2.60 which were associated with only θ . The second a -parameter (a_{2i}) estimates for γ_{di} ranged from 0.12 to 2.19 which indicated small to large testlet effects. The first a -parameter (a_{1i}) estimates in the generalized TRT model were very similar to the estimates of 3PLM and constrained TRT model while the second a -parameters did not have any similarity to the 3PLM and the constrained TRT models.

In order to compare the b -parameters between 3PLM and generalized TRT model, the testlet mean should be considered as a comparison criterion between 3PLM and constrained TRT. The b -parameter estimates of Items 15, 16, and 17 in Passage 5 ($\widehat{M}_{\gamma_1} = -1.84$) and Items 20, 21, and 22 in Passage 6 ($\widehat{M}_{\gamma_1} = -1.45$) in Table 6 were relatively higher ($b_{15} = 0.77, b_{16} = 0.70, b_{17} = 0.37, b_{20} = 0.99, b_{21} = 0.88, \text{ and } b_{22} = 0.05$) than the estimates of 3PLM in Table 1 ($b_{15} = 0.38, b_{16} = 0.19, b_{17} = -0.49, b_{20} = 0.37, b_{21} = 0.25, \text{ and } b_{22} = -0.77$), respectively. The mean value of the testlet effects and item

difficulty parameter estimates were negatively related. A lower mean of the testlet effect yielded a higher item difficulty parameter estimates in generalized TRT.

Table 5

Estimated item parameters and testlet effects of constrained TRT model

Testlets	Items	Item parameter estimates						Testlet effects	
		<i>a</i>	(<i>SD</i>)	<i>b</i>	(<i>SD</i>)	<i>c</i>	(<i>SD</i>)	M_{γ}	σ_{γ}^2
1	1	1.04	(0.04)	0.73	(0.11)	0.02	(0.02)	-1.01	(0.01)
	2	1.71	(0.06)	-0.24	(0.09)	0.05	(0.03)		
	3	1.01	(0.04)	0.48	(0.12)	0.05	(0.03)		
	4	1.00	(0.04)	0.10	(0.09)	0.04	(0.03)		
2	5	0.55	(0.04)	-0.57	(0.22)	0.11	(0.06)	-0.06	(0.00)
	6	1.46	(0.05)	-0.77	(0.06)	0.07	(0.03)		
	7	2.09	(0.09)	-1.04	(0.06)	0.09	(0.04)		
3	8	0.61	(0.03)	-0.04	(0.15)	0.03	(0.02)	0.53	(0.01)
	9	0.67	(0.05)	-1.01	(0.18)	0.11	(0.06)		
	10	1.41	(0.06)	-1.53	(0.10)	0.06	(0.04)		
	11	1.78	(0.10)	-0.46	(0.11)	0.19	(0.02)		
4	12	1.76	(0.19)	1.70	(0.07)	0.28	(0.01)	-0.14	(0.00)
	13	1.64	(0.10)	-0.62	(0.09)	0.16	(0.05)		
	14	1.39	(0.08)	-0.35	(0.09)	0.27	(0.04)		
5	15	0.68	(0.06)	-0.72	(0.16)	0.07	(0.05)	1.09	(0.01)
	16	1.25	(0.09)	-0.92	(0.08)	0.13	(0.03)		
	17	2.58	(0.14)	-1.57	(0.07)	0.21	(0.02)		
6	18	0.81	(0.05)	-1.34	(0.15)	0.08	(0.04)	0.90	(0.01)
	19	0.76	(0.11)	0.47	(0.11)	0.09	(0.04)		
7	20	1.14	(0.09)	-0.21	(0.11)	0.20	(0.03)	0.58	(0.01)
	21	0.68	(0.03)	-0.34	(0.11)	0.02	(0.02)		
	22	1.10	(0.07)	-1.33	(0.16)	0.08	(0.05)		

Table 6

Estimated item parameters and testlet effects of generalized TRT model

Testlets	Items	Item parameter estimates								Testlet effects	
		<i>a1</i>	(<i>SD</i>)	<i>a2</i>	(<i>SD</i>)	<i>b</i>	(<i>SD</i>)	<i>c</i>	(<i>SD</i>)	M_γ	σ_γ^2
1	1	1.05	(0.04)	0.55	(0.17)	0.18	(0.13)	0.02	(0.02)	-0.81	(0.00)
	2	1.71	(0.07)	0.36	(0.27)	-1.85	(0.21)	0.05	(0.04)		
	3	1.00	(0.04)	0.23	(0.16)	-0.36	(0.14)	0.04	(0.03)		
	4	1.01	(0.04)	2.19	(0.20)	0.90	(0.17)	0.05	(0.03)		
2	5	0.53	(0.04)	0.76	(0.58)	-0.46	(0.13)	0.07	(0.05)	0.06	(0.02)
	6	1.50	(0.08)	0.67	(0.55)	-1.19	(0.11)	0.09	(0.04)		
	7	2.11	(0.10)	0.57	(0.49)	-2.27	(0.09)	0.11	(0.04)		
3	8	0.60	(0.03)	0.12	(0.11)	0.35	(0.08)	0.02	(0.02)	-0.72	(0.04)
	9	0.63	(0.03)	0.44	(0.32)	-0.12	(0.26)	0.05	(0.04)		
	10	1.40	(0.06)	0.61	(0.32)	-1.04	(0.17)	0.04	(0.03)		
	11	1.79	(0.10)	0.71	(0.33)	0.62	(0.26)	0.19	(0.02)		
4	12	1.72	(0.16)	0.56	(0.48)	2.71	(0.27)	0.28	(0.01)	-0.17	(0.18)
	13	1.67	(0.09)	0.79	(0.46)	-1.06	(0.33)	0.17	(0.04)		
	14	1.35	(0.09)	0.52	(0.38)	-0.59	(0.24)	0.25	(0.04)		
5	15	0.67	(0.05)	0.29	(0.13)	0.77	(0.24)	0.06	(0.04)	-1.84	(0.19)
	16	1.25	(0.09)	0.30	(0.20)	0.70	(0.20)	0.12	(0.03)		
	17	2.60	(0.14)	0.91	(0.26)	0.37	(0.33)	0.21	(0.02)		
6	18	0.81	(0.05)	0.81	(0.61)	-0.30	(0.23)	0.07	(0.04)	-0.05	(0.06)
	19	0.81	(0.08)	1.01	(0.65)	1.26	(0.29)	0.11	(0.03)		
7	20	1.14	(0.08)	0.38	(0.16)	0.99	(0.25)	0.20	(0.03)	-1.45	(0.04)
	21	0.68	(0.04)	0.49	(0.12)	0.88	(0.12)	0.03	(0.02)		
	22	1.10	(0.06)	0.59	(0.20)	0.05	(0.28)	0.09	(0.04)		

The testlet variances for Passages 1 and 2 were small, $\widehat{\sigma}_{\gamma_1}^2 = 0.00$ and $\widehat{\sigma}_{\gamma_2}^2 = 0.02$, respectively. However, the variances of Testlets 3 through 7 were ranged from 0.04 to 0.19. The Testlets 4 and 5 caused moderate local dependence among items, $\widehat{\sigma}_{\gamma_4}^2 = 0.18$ and $\widehat{\sigma}_{\gamma_5}^2 = 0.19$, respectively. While all variances of the testlets in constrained TRT were

negligibly small, the variances of Testlets 4, 5, and 6 in the generalized TRT model were considerably high ($\hat{\sigma}^2_{\gamma_4} = 0.18$, $\hat{\sigma}^2_{\gamma_5} = 0.19$, and $\hat{\sigma}^2_{\gamma_6} = 0.06$).

Comparisons of item parameters

Item discrimination parameter estimates were highly correlated between 3PLM and constrained TRT model ($r = .99$, $p < .01$), and between constrained TRT and generalized TRT model ($r = .99$, $p < .01$). However, the second slope parameter (a_{2i}) of the generalized TRT model, which was related only to the testlet effect estimates, was not significantly related with the estimates of both 3PLM ($r = .08$, $p > .05$) and constrained TRT ($r = .08$, $p > .05$).

The scatter plots of correlation coefficients among the IRT-TRT item difficulty parameters in Figure 15 indicated that item difficulty estimates were moderately related between 3PLM and constrained TRT ($r = .59$, $p < .05$), between 3PLM and generalized TRT ($r = .84$, $p < .05$), and between constrained TRT and generalized TRT ($r = .52$, $p < .05$). The correlation between 3PLM and constrained TRT was relatively low, which might indicate that the item difficulty parameter estimates were influenced by the testlet effects.

As one can see Equations 10 and 12, 3PLM ($P = \Phi a_j(\theta_i - b_j)$) is embedded in constrained TRT ($P = \Phi a_j(\theta_i - b_j - \gamma_{id(j)})$). Item difficulty parameter (b_j) in 3PLM is separated into two parts of item difficulty (b_j) and testlet effect ($\gamma_{id(j)}$) in constrained TRT. The testlet effects caused irregularity on the estimates of 3PLM and constrained TRT. The generalized TRT model in Equation 13 ($P = \Phi (a_{j1} \theta_i - b_j + a_{j2}\gamma_{id(j)})$)

estimated the b -parameter differently with the constrained model ($P = \Phi (a_j \theta_i - a_j b_j - a_j \gamma_{id(j)})$). The constrained TRT model ($P = \Phi (a_j \theta_i - a_j b_j - a_j \gamma_{id(j)})$) and 3PLM ($P = \Phi (a_j \theta_i - a_j b_j)$) estimated the b -parameter with the portion of $a_j * b_j$ in the formula while the generalized TRT model ($P = \Phi (a_{j1} \theta_i - b_j + a_{j2} \gamma_{id(j)})$) in Equation 13 estimated the b_j without the a -parameter (a_j). In the current study, these estimation processes caused the discrepancy among the b -parameter estimates of three different models.

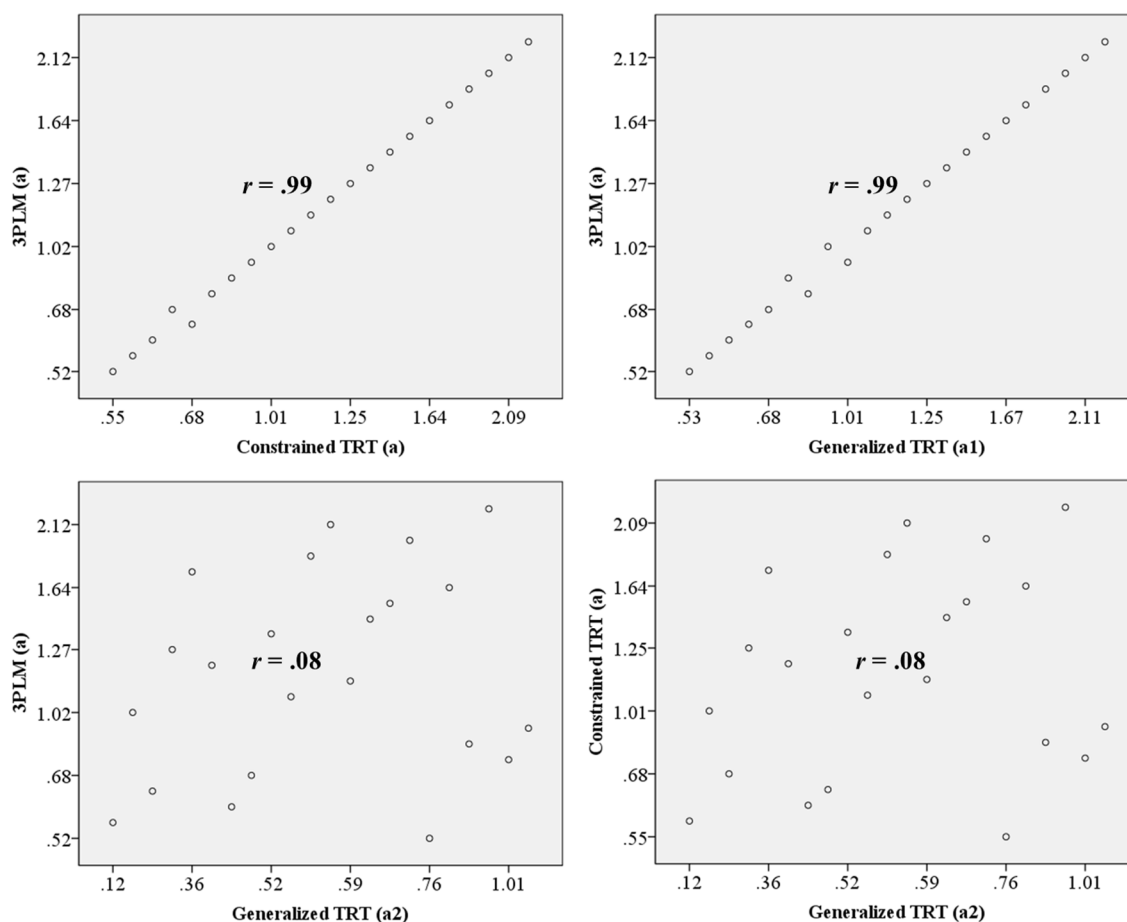


Figure 14. Comparisons of the item discrimination estimates (a_1 and a_2) between 3PLM and TRT models

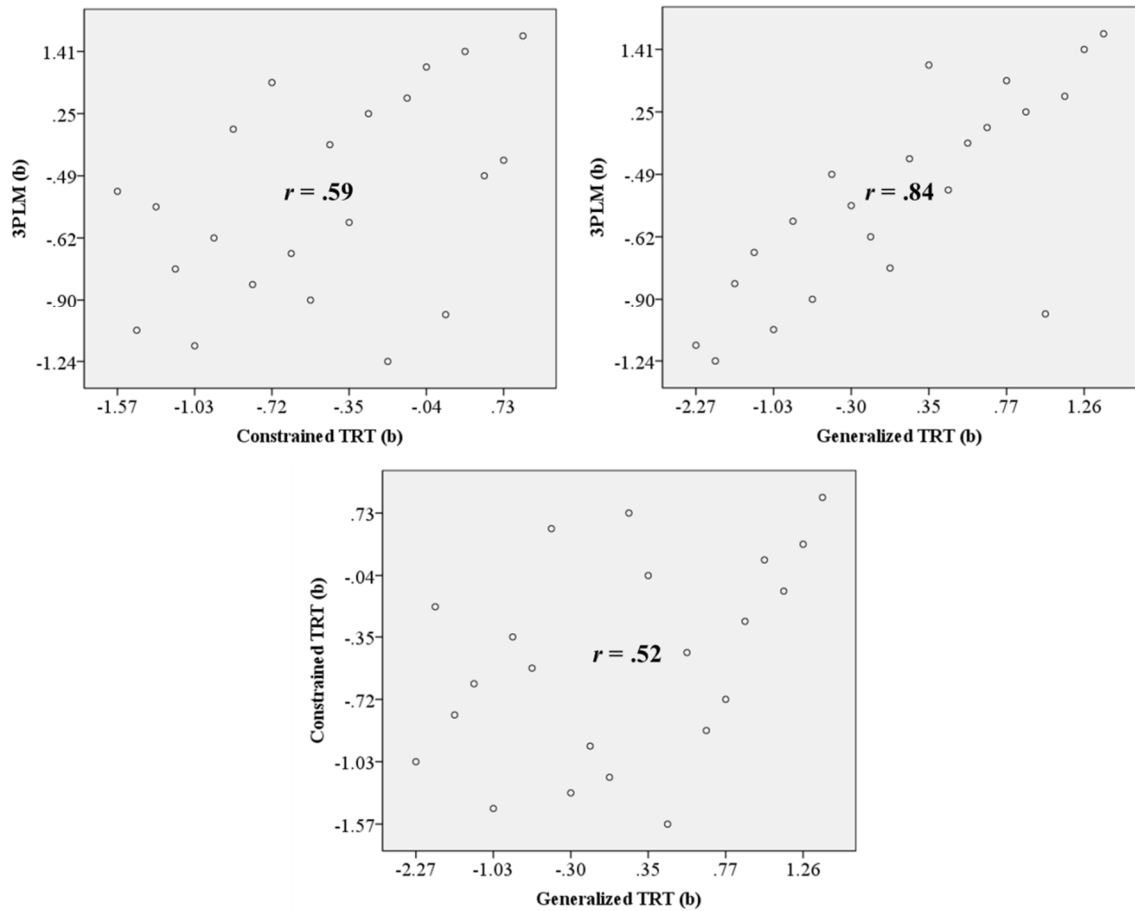


Figure 15. Comparisons of the item difficulty estimates between 3PLM and TRT models

In Figure 16, the estimated pseudo-chance parameters of 3PLM were highly correlated with those of the constrained TRT ($r = .97, p < .01$) and generalized TRT ($r = .99, p < .01$). The pseudo-chance parameter between constrained TRT and generalized TRT was also highly related, $r = .97, p < .01$. These results indicated that the pseudo-chance parameters were not affected by the testlet effect as one could infer from Equations 2 and 3.

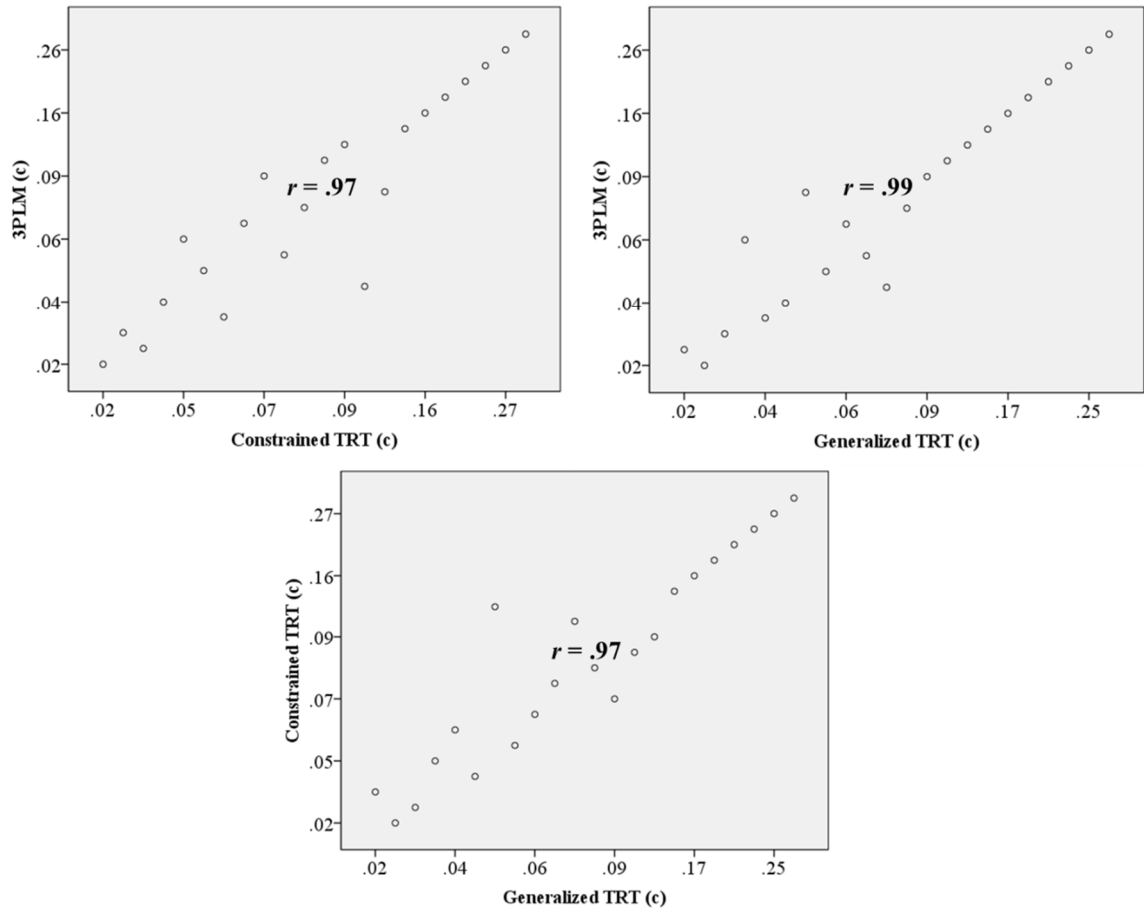


Figure 16. Comparisons of the pseudo-chance parameter estimates between 3PLM and TRT models

CHAPTER FIVE: DISCUSSION

This study was an endeavor to compare various psychometric models (CTT, three IRTs, and two TRTs) which have been commonly used to evaluate test or item information on a reading comprehension test constructed for 5th grade students based on CCSS. Computation of item and test indices from CTT were done, and these indices were used to compare the item and testlet characteristics from 3PLM, constrained TRT, and generalized TRT. Before the IRT analysis, an exploratory factor analysis was performed to assure the unidimensional assumption for the IRT models. Then, the best-fitting IRT model was selected from the 1PLM, 2PLM, and 3PLM models utilizing three comparison statistics of -2LL, AIC, and BIC. The 3PLM was proved as the best-fitting IRT model for the data. A large sample size ($N = 10,897$) of the employed data set enabled to compare three different psychometric models, CTT, IRT, and TRT models because CTT required a large sample size in order to develop stabilized items and test indices. The reading comprehension test (22 items with 7 passages) was composed with 4 areas of RL, RI, RF, and L based on CCSS. This chapter delivers an overview of the results, discussion along with previous findings, limitations, and recommendations for future research.

In order to address the strength and weakness of test items of the benchmark reading comprehension test as well as benefits of IRT over CTT, and TRT over IRT, four specific research questions were set:

1. What are the similarities and dissimilarities between CTT and IRT?

2. Which IRT model shows the best-fit for a reading comprehension test?
3. Do the reading passages show testlet effects?
4. What are the differences between the item parameter estimates obtained using TRT and IRT model?

The results from CTT were compared to those of traditional IRT model for the first research question, “What are the similarities and dissimilarities between CTT and IRT?” Due to a large sample size, one of shortcomings of CTT, the issue of “sample-dependent item indices” could be partially overcome. In general, item discriminations and item difficulties in CTT demonstrated similar characteristics to the estimates in IRT. In addition to a high Cronbach’s alpha of .79, difficult items from the CTT result were also considered difficult items from the IRT analysis and easy items from the CTT analysis were reported as easy items from the IRT analysis. Specifically, Items 8, 12, and 19 were regarded as hard items from both CTT and IRT analyses. Items 2, 7, and 10 were reported as easy items from both CTT and IRT.

Item 17 was an exception. According to CCSS, Item 17 is a question about “compare and contrast two or more characters, settings, or events in a story or drama, drawing on specific details in the text (http://www.corestandards.org/assets/CCSSI_ELA%20Standards.pdf, p. 12).” In CTT, this item was very hard to answer and not good for 5th graders because it poorly discriminated students. However, in IRT, Item 17 was hard but a good item for students who had high ability. One weakness of this item was that the probability was relatively

high to answer correctly by guessing. One of the advantages of IRT over CTT is that item discrimination is interpreted with the specific ability level as well as the probability of guessing. Unlike CTT, IRT provides invariant item parameter estimates as well as information on specific ability levels where item discrimination parameter estimates reach their maximum values. It may be recommended that Items 5, 8, 9, and 15 might be reconsidered for 5th graders in this bench marked reading comprehension test because of low item discrimination parameters. Thus, the first research question is fully answered by both CTT and IRT analyses, demonstrating clear advantages of IRT over CTT with more in depth information for items (*a*, *b*, and *c*-parameter estimates) along with detailed interpretations of item parameters for specific ability levels.

Exploratory factor analysis (EFA) is a prerequisite for any IRT analyses in order to assure that the unidimension assumption is satisfied. The EFA results showed that the first factor explained 20% of variability of the data along with the eigenvalue of 4.27. The other factors explained trivial proportions of data variance which was less than 5%.

For the second research question, “Which IRT model shows the best-fit for a reading comprehension test?”, the overall model-fit indices were employed to select best-fit model for a reading comprehension test. According to the three model comparison criteria (-2LL, AIC, and BIC), 3PLM fitted best for 5th grade reading comprehension test. Although 3PLM is good for a reading comprehension test, 2PLM is commonly used due to the computational efficacy (Min & He, 2014). In their study, 6 items out of 30 items had poorly estimated in 3PLM even though 3PLM provided best model-data fit for a testlet based reading comprehension test. In the current study, severe fluctuations in

estimation for the pseudo-chance parameters lasted during the iterations process as shown in Figure 9. The a -parameter and b -parameter estimation demonstrated a relatively stable process during the estimation. We experienced a difficulty in estimation of the pseudo-chance parameters with 1,000 or 2,000 iterations with the Markov Chains Monte Carlo (MCMC) method. In order to utilize 3PLM with MCMC algorithm on a reading comprehension test, the Markov chain length was set to 10,000 with the iterations of 5,000. The second research questions is also fully answered with the 3PLM as an evident winner for the model-data fit analysis although some cautions should be exercised when the 3PLM is applied due to computational issues in estimation iterations.

For the third question, “Do the reading passages show testlet effects?”, 3PLM and two 3PL TRT models (constrained and generalized) were compared to determine the effects of testlet parameters. As shown in Table 4, no significant model-fit differences among these three models were found with the -2LL and AIC statistics. Only the BIC indices showed small difference among the models because the BIC considered the effects of both sample size and the estimated number of parameters. The model comparison criteria demonstrated that unidimensional 3PLM was a better fit model than the TRT models for the data. This result indicated that the data of 5th grade reading comprehension test did not contain significant testlet effects even though several testlet effect variances in the generalized TRT model were significantly higher ($\hat{\sigma}^2_{\gamma_4} = 0.18$, $\hat{\sigma}^2_{\gamma_5} = 0.19$, and $\hat{\sigma}^2_{\gamma_6} = 0.06$) than the reference variance of .04 which was suggested by Wang, Bradlow, and Wainer (2002). It is speculated that this phenomenon occurred due to the computational problems in the 3PL TRT models. As one can see in Equation 12

and 13, the testlet parameters ($\gamma_{id(j)}$) are associated between examinee and testlets. With the testlet parameter ($\gamma_{id(j)}$), the WinBUGs program was stopped during iterations. In order to solve this problem, the testlet parameter ($\gamma_{d(j)}$) should be utilized instead of the $\gamma_{j \cdot di}$. The third research question is partially answered. Although the analysis results did not confirm the testlet effects from this particular data, some possible causes may be speculated. The estimation issue for 3PLM could be a possible cause of this puzzling result of showing no testlet effect from the data. Another possible interpretation of the result may be that the LID is not strong in the test. Since it is impossible to discern the exact level of LID in this reading comprehension test, other alternative methods (e.g., simulation study) may be needed, stipulating all levels of each parameter.

For the last question, “What are the differences between the item parameter estimates obtained using TRT and IRT model?”, we compared the estimates of 3PLM and two TRT models. In both the constrained TRT and generalized TRT model, item difficulty parameters were associated with the testlet effect means. Lower testlet means were related to higher item difficulty parameters. The higher the testlet means, the lower the item difficulty parameters. The graphs with correlation coefficients in Figure 11 revealed that IRT and two TRT models yielded invariant estimates for the person ability ($r = .99, p < .01$). The a -parameter estimates were highly correlated between 3PLM and constrained TRT model ($r = .99, p < .01$), and between constrained TRT and generalized TRT model ($r = .99, p < .01$). The invariance of person and item parameter estimates proved the advantage of IRT over CTT. However, item difficulty estimates did not show perfect linear relationship between 3PLM and constrained TRT ($r = .59, p < .05$),

between 3PLM and generalized TRT ($r = .84, p < .05$), and between constrained TRT and generalized TRT ($r = .52, p < .05$). This phenomenon could be explained by the estimation formula in Equations 10, 12, and 13. As interpreted in the result section, item difficulty parameters which were estimated by different psychometric models such as 3PLM ($P = \Phi (a_j \theta_i - a_j b_j)$), constrained TRT ($P = \Phi (a_j \theta_i - a_j b_j - a_j \gamma_{id(j)})$), and generalized TRT ($P = \Phi (a_{j1} \theta_i - b_j - a_{j2} \gamma_{id(j)})$) were not invariant. The last research questions is also fully answered with the association of item difficulty parameters and the testlet mean as well as with the invariant estimates of item discrimination and pseudo-chance parameters from 3PLM and two TRT models.

Limitation and recommendations for future research

No research project is without limitations, and this project is not an exception. The first limitation of this study is computational issues in WinBUGs which is associated with the testlet parameter ($\gamma_{id(j)}$). Using the $\gamma_{id(j)}$, the WinBUGs program failed to estimate item and testlet parameters. The performance of MCMC algorithm is affected by the number of items per testlet (Wainer, Bradlow, & Wang, 2007). In our study, passages contained from 2-item to 4-item. The suggestion for future study is that the model comparisons may be conducted with various numbers of items per testlet.

In order to confirm the association of a testlet mean and item difficulty parameters as well as to investigate testlet effects on a reading comprehension test, one may employ various reading data because, in the current study, only one reading comprehension test was used although it had a large sample size. In addition, the factor structure of the data

which we used was not supporting both the 4 areas (RL, RI, RF, and L) of the CCSS criteria and the multiple underlying constructs which defined by various reading comprehension theories. In our literature review, the unidimensionality may not fully represent a reading comprehension test due to various reading comprehension activities and process with the multidimensional constructs (Sweet, 2005). In order to endorse the unidimensionality or multidimensionality on a reading comprehension test, various reading comprehension test data should be used with various psychometric techniques.

In conclusion, the current study made a significant contribution to the field of reading comprehension by utilizing both the traditional CTT and more advanced and falsifiable IRT models in analyzing item level data for a reading comprehension test. Also, the finding of the association between testlet means and item difficulty parameters was meaningful to researchers and educators. However, without true values of testlet effect and item parameters, it is impossible to determine which models provide precise testlet mean, testlet variance, and item parameters. In future studies, simulation methods may give more information about appropriate testlet sizes along with computational issues in order to select best-fit model for a reading comprehension test and in order to interpret the relation between testlet effect and item parameters. The results of the current study shed light to both researchers and practitioners that we are all in need of research which utilizes both actual data and simulated data in order to achieve the closest approximation to the truth.

REFERENCES

- Adlof, S. M., Catts, H. W., & Little, T. D. (2006). Should the simple view of reading include a fluency component? *Reading and Writing, 19*, 933-958.
- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second international symposium on information theory* (p. 267-281). Budapest: Akademiai Kiado.
- Allen, M. J., & Yen, W. M. (2002). *Introduction to measurement theory*. Waveland Press.
- Binet, A., & Simon, Th. (1916). *The development of intelligence in children: The Binet-Simon Scale*. Publications of the Training School at Vineland New Jersey Department of Research No. 11. E. S. Kite (Trans.). Baltimore: Williams & Wilkins.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores* (pp. 395-479). Reading MA: Addison-Wesley.
- Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*, 443-459.
- Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6*, 431-444.
- Bollen, K. A. (1989). *Structural equations with latent variables* (p. 179-225). New York: Wiley.
- Bradley, L., & Bryant, P. E. (1983). Categorizing sounds and learning to read—a causal connection. *Nature, 301*, 419-421.
- Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64*(2), 153-168.
- Cai, L., Maydeu-Olivares, A., Coffman, D. L., & Thissen, D. (2006). Limited-information goodness-of-fit testing of item response theory models for sparse 2^p tables. *British Journal of Mathematical and Statistical Psychology, 59*, 173-194.
- Cain, K., Oakhill, J., & Bryant, P. E. (2004). Children's reading comprehension ability: Concurrent prediction by working memory, verbal ability, and component skills. *Journal of Educational Psychology, 96*, 31-42.
- Campbell, D.T., & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105.
- Carpenter, R. D., & Paris, S. G. (2005). Issues of validity and reliability in early reading assessments. In S. G. Paris and S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (p. 279-304). Mahwah, NJ: Lawrence Erlbaum Associates.
- Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. New York: Cambridge University Press.
- Catts, H. W., Hogan, T., & Fey, M. E. (2003). Subgrouping poor readers on the basis of individual differences in reading-related abilities. *Journal of Learning Disabilities, 36*(2), 151-164.

- 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.
- Cunningham, A., Perry, K. E., & Stanovich, K. E. (2001). Converging evidence for the concept of orthographic processing. *Reading and Writing: An Interdisciplinary Journal*, 14, 549-568.
- Curtis, M. E. (1980). Development of components of reading skills. *Journal of Educational Psychology*, 72, 656-669.
- 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.
- de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York: Guilford Press.
- De Corte, E., Verschaffel, L., & Van De Ven, A. (2001). Improving text comprehension strategies in upper primary school children: A design experiment. *British Journal of Educational Psychology*, 71(4), 531-559.
- DeMars, D. E. (2006). Application of the bi-factor multidimensional item response theory model to testlet-based tests. *Journal of Educational Measurement*, 43(2), 145-168.
- DeMars, D. E. (2012). Confirming testlet effects. *Applied Psychological Measurement*, 36(2), 104-121.
- Duke, N. K. (2005). Comprehension of what for what: Comprehension as a nonunitary construct. In Paris, S. G. & Stahl, S. A. (Eds.), *Children's reading comprehension and assessment* (p. 93-104). Lawrence Erlbaum Associates: Mahwah, New Jersey.
- Durkin, D. (1993). *Teaching them to read* (6th Ed.). Boston, MA: Allyn and Bacon.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates: Mahwah, New Jersey.
- Fisher, D., & Frey, N. (2009). *Background knowledge: The missing piece of the comprehension puzzle*. Portsmouth, NH: Heinemann.
- Gessaroli, M. E., & Folske, J. C. (2002). Generalizing the reliability of tests comprised of testlets. *International Journal of Testing*, 2, 277-295.
- Gough, P. B., & Tunmer, W. E. (1986). Decoding, reading, and reading disability. *Remedial and Special Education*, 7, 6-10.
- Guyer, R., & Thompson, N.A. (2011). *User's Manual for Xcalibre item response theory calibration software, version 4.1.3*. St. Paul MN: Assessment Systems Corporation.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage Publications, Inc.
- Hambleton, R. K., & van der Linden, W. J. (1982). Advances in item response theory and applications: An introduction. *Applied Psychological Measurement*, 6(4), 373-378.

- Heller, R., & Greenleaf, C. L. (2007). *Literacy instruction in the content areas: Getting to the core of middle & high school improvement*. Washington, DC: Alliance for Excellent Education. Retrieved from http://carnegie.org/fileadmin/Media/Publications/PDF/Content_Areas_report.pdf
- Henderson, L. (1982). *Orthography and word recognition in reading*. London: Academic press.
- Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing: An Interdisciplinary Journal*, 2, 127-160.
- Johnson, D. D., & Pearson, P. D. (1975). Skills management systems: A critique. *The Reading Teacher*, 28, 757-764.
- Johnston, P. H. (1984). Assessment in reading. In P. D. Pearson, R. Barr, M. Kamil, & P. Mosenthal (Eds.), *Reading comprehension assessment* (p.147-182). New York: Longman.
- Juel, C., Griffith, P. L., & Gough, P.B. (1986). Acquisition of literacy: A longitudinal study of children in first and second grade. *Journal of Educational Psychology*, 78, 243-255.
- 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.
- Keller, L. A., Swaminathan, H., & Sireci, S. G. (2003). Evaluating scoring procedures for context dependent item sets. *Applied Measurement in Education*, 16, 207-222.
- Kendeou, P., van den Broek, P., White, M. J., & Lynch, J. (2007). Comprehension in preschool and early elementary children: Skill development and strategy interventions. In MaNamara (Eds.), *Reading Comprehension Strategies: Theories, interventions, and Technologies* (p. 27-45). Lawrence Erlbaum Associates: New York.
- Kershaw, S., & Schatschneider, C. (2012). A latent variable approach to the simple view of reading. *Reading and Writing*, 25, 433-464.
- Kim, J. K., & Nicewander, W. A. (1993). Ability estimation for conventional tests. *Psychometrika*, 58(4), 587-599.
- Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge, England: Cambridge University Press.
- Kintsch, W., & Kintsch, E. (2005). Comprehension. In S. G. Paris and S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (p. 71-92). Mahwah, NJ: Lawrence Erlbaum Associates.
- Li, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternative models of testlets. *Applied Psychological Measurement*, 30(1), 3-21.
- Lord, F. M. (1952). A theory of test scores (Psychometric Monograph No. 7). Iowa City, IA: Psychometric Society.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale NJ: Erlbaum.
- Lord, F. M. (1984). Standard errors of measurement at different score levels. *Journal of Educational Measurement*, 21, 239-243.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. MA: Addison-Wesley.

- MacGinitie, W. H., MacGinitie, R. K., Maria, K., & Dreyer, L. G. (2000). *Gates–MacGinitie Reading Tests* (4th Ed.). Itasca, IL: Riverside.
- Mair, P., Reise, S. P., & Bentler, P. M. (2008). *IRT goodness-fit using approaches from logistic regression*. Department of Statistics Papers, UCLA. Retrieved from <http://escholarship.org/uc/item/2tc0s6k9>.
- McGregor, K. K. (2004). Developmental dependencies between lexical semantics and reading. In C. A. Stone, E. R. Silverman, B. J. Ehren, & K. Apel (Eds.), *Handbook of language literacy and disorders* (p. 302-317). New York: Guilford Press.
- Min, S., & He, L. (2014). Applying unidimensional and multidimensional item response theory models in testlet-based reading assessment. *Language Testing*, 1-25. DOI: 10.1177/0265532214527277
- Mutter, V., Hulme, C., Snowling, M. J., & Stevenson, J. (2004). Phonemes, rimes, vocabulary and grammatical skill as foundations of early reading development: Evidence from a longitudinal study. *Developmental Psychology*, 40, 665-681.
- National Center for Education Statistics (2013). *The Nation's Report Card: 2013 mathematics and reading (NCES 2014-451)*. Washington, DC: Institute of Education Sciences, U.S. Department of Education.
- Neufeld, P. (2005). Comprehension instruction in content area classrooms. *The reading Teacher*, 59(4), 302-312.
- No Child Left Behind Act of 2001, Pub. L. No. 107–110.
- Oakhill, J. V., & Cain, K. (2011). The precursors of reading ability in young readers: Evidence from a four-year longitudinal study. *Scientific Studies of Reading*, 16(2), 91-121.
- Paris, S. G., & Stahl, S. A. (2005). *Children's reading comprehension and assessment*. Lawrence Erlbaum Associates: Mahwah, New Jersey.
- Pearson, P. D., & Hamm, D. N. (2005). The assessment of reading comprehension: A review of practices-past, present, and future. In S. G. Paris and S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (p. 13-70). Mahwah, NJ: Lawrence Erlbaum Associates.
- Plaut, D. C. (2005). Connectionist approaches to reading. In M.J. Snowling & C. Hulme (Eds.), *The science of reading: A handbook*. (p. 24-38). Oxford UK: Blackwell publishing.
- Pritchard, M. E., Wilson, G. S., & Yamnitz, B. (2007). What predicts adjustment among college students? *Journal of American College Health*, 56, 15-21.
- RAND Reading Study Group (2002). *Reading for understanding: Toward an R&D program in reading comprehension*. Santa Monica, CA: Rand.
- Samejima, F. (1969). *Estimation of a latent ability using a response pattern of graded scores* (Psychometrika Monograph No. 17). Iowa City, IA: Psychometric Society.
- Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2), 461–464.
- Seigneuric, A., & Ehrlich, M-F. (2005). Contribution of working memory capacity to children's reading comprehension: A longitudinal investigation. *Reading and Writing*, 18, 617-656.

- Shankweiler, D. (1989). How problems of comprehension are related to difficulties in word reading. In D. Shankweiler & I. Y. Liberman (Eds.), *Phonology and reading disability: Solving the reading Puzzle* (p.35-68). Ann Arbor: University of Michigan Press.
- Siegel, L. S. (1993). The development of reading. In H. W. Reese (Ed.), *Advances in child development and behavior* (Vol. 24, p. 63-97). San Diego, CA: Academic Press Inc.
- Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, 28, 237-247.
- Snow, C. E. (2003). *Assessment of reading comprehension*. In A. P. Sweet & C. E. Snow (Eds.), *Rethinking reading comprehension* (p. 192-206). New York: Guilford.
- Snow, C. E., & Sweet, A. P. (2003). Reading for comprehension. In A. P. Sweet & C. E. Snow (Eds.), *Rethinking reading comprehension* (p. 1-11). New York: Guilford.
- Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1), 72–101.
- Spiegelhalter, D., Thomas, A., & Best, N. (2003). WinBUGS version 1.4 [Computer program]. Cambridge, UK: MRC Biostatistics Unit, Institute of Public Health.
- Stahl, S. A., Hare, V. C., Sinatra, R., & Gregory, J. F. (1991). Defining the role of prior knowledge and vocabulary in reading comprehension: The retiring of number 41. *Journal of Reading Behavior*, 23(4), 487–508.
- Sweet, A. P. (2005). Assessment of reading comprehension: The RAND reading study group vision. In S. G. Paris and S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (p. 3-12). Mahwah, NJ: Lawrence Erlbaum Associates.
- Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement*, 26, 247-260.
- Thorndike, R. L. (1973). *Reading comprehension education in fifteen countries: an empirical study*. New York: John Wiley & Sons.
- Tuerlinckx, F., & De Boeck, P. (2001). The effect of ignoring item interactions on the estimated discrimination parameters in item response theory. *Psychological Methods*, 6(2), 181-195.
- Tunmer, W. E., & Hoover, W. A. (1992). Cognitive and linguistic factors in learning to read. In P. E. Gough, L. C. Ehri, & R. Treiman (Eds.), *Reading acquisition* (p. 175–214). Hillsdale, NJ: Erlbaum.
- Vygotsky, L. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.
- Wainer, H., Bradlow, E. T., & Du, Z. (2000). Testlet response theory: An analog for the 3-PL useful in testlet-based adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing, theory and practice* (p. 245-270). Boston, MA: Kluwer-Nijhoff.
- Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. Cambridge University Press, New York.

- Wainer, H., & Kiely, G. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*, 185-202.
- Wagner, R. K., & Torgesen, J. K. (1987). The nature of phonological processing and its causal role in the acquisition of reading skills. *Psychological Bulletin, 101*, 192-212.
- Wainer, H., & Wang, X. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement, 37*, 203–220.
- Wang, X., Bradlow, E. T., & Wainer, H. (2002). A general Bayesian model for testlets: Theory and applications. *Applied Psychological Measurement, 26*, 190–128.
- 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.
- Wilson, M., and Adams, R. (1995). Rasch models for item bundles. *Psychometrika, 60*, 181-198.
- Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*, 125-145.
- Yen, W., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (4th Ed., p. 111-153). Westport, CT: Praeger Publishers.

APPENDICES

APPENDIX A

1PLM

```

# n: the number of students
# p: the number of items
# a: item discrimination
# b: item difficulty
model {

# Read in individual item responses
  for ( i in 1:n ) {
    for ( j in 1:p ) {
      x[i , j ] <- response[i, j ]
    }
  }

# Identify one-parameter logistic (1PL) model
  for (i in 1:n) {
    for (j in 1 : p) {
      x[i , j ] ~ dbern(prob[i,j])
      logit(prob[i,j]) <- a[j] *(theta[i] - b[j])
    }
  }

# Specify prior for examinee parameters
  for (i in 1:n) {
    theta[i] ~ dnorm(0,1)
  }

#Specify priors for item parameters
  for (j in 1:p) {
    b[j] ~ dnorm(0, 1)
    a[j] <- 1.0
  }

#Log Likelihood
  for ( i in 1:n ) {
    for ( j in 1:p ) {
      L[i, j ] <- log(prob[i , j ]) * x[i, j ] + log(1-prob[i , j ]) * (1- x[i , j ])
    }
  }

  loglik <- sum(L[ 1: n, 1: p])
  LL <- -2*loglik
  AIC <- -2*(loglik - np)
  BIC <- -2*loglik + np*log(n)
}

```


APPENDIX B

2PLM

```

model {
# Read in individual item responses
  for ( i in 1:n ) {
    for ( j in 1:p ) {
      x[i , j ] <- response[i, j ]
    }
  }
# Identify two-parameter logistic (2PL) model
  for (i in 1:n) {
    for (j in 1 : p) {
      x[i , j ] ~ dbern(prob[i,j])
      logit(prob[i,j]) <- a[j] *(theta[i] - b[j])
    }
  }
# Specify prior for examinee parameters
  for (i in 1:n) {
    theta[i] ~ dnorm(0,1)
  }
#Specify priors for item parameters
  for (j in 1:p) {
    b[j] ~ dnorm(0, 1)
    a[j] ~ dnorm(0, 1) I(0, )
  }
}

```

APPENDIX C

3PLM

```

model {
# Read in individual item responses
  for ( i in 1:n ) {
    for ( j in 1:p ) {
      x[i , j ] <- response[i, j ]
    }
  }
# Identify three-parameter logistic (3PL) model
  for (i in 1:n) {
    for (j in 1 : p) {
      x[i , j ] ~ dbern(prob[i,j])
      logit(prob.star[i,j]) <- a[j] *(theta[i] - b[j])
      prob[i, j] <- eta[j] + (1-eta[j]) * prob.star[i, j]
    }
  }
# Specify prior for examinee parameters
  for (i in 1:n) {
    theta[i] ~ dnorm(0,1)
  }
#Specify priors for item parameters
  for (j in 1:p) {
    b[j] ~ dnorm(0, 1)
    a[j] ~ dnorm(0,1) I(0, )
    eta[j] ~ dbeta (1, 1)
  }
}

```

APPENDIX D

3PL CONSTRAINED TRT

```

model {
# Read in individual item responses
  for ( i in 1:n ) {
    for ( j in 1:p ) {
      x[i , j ] <- response[i, j ]
    }
  }
# Identify constraint three-parameter logistic (3PL) TRT model
  for (i in 1:n) {
    for (j in 1 : p) {
      x[i , j ] ~ dbern(prob[i,j])
      logit(prob.star[i,j]) <- a[j] *(theta[i] - b[j] - test[d[j]])
      prob[i, j ] <- eta[j] + (1-eta[j]) * prob.star[i, j ]
    }
  }
# Specify prior for examinee parameters
  for (i in 1:n) {
    theta[i] ~ dnorm(0,1)
  }
# Specify prior for testlet parameter
  for (k in 1:T) {
    test[k] ~ dnorm(0, 1)
  }
#Specify priors for item parameters
  for (j in 1:p) {
    b[j] ~ dnorm(0, 1)
    a[j] ~ dnorm(0, 1) I(0, )
    eta[j] ~ dbeta (1, 1)
  }
#Log Likelihood
  for ( i in 1:n ) {
    for ( j in 1:p ) {
      L[i, j ] <- log(prob[i, j ]) * x[i, j ] + log(1-prob[i, j ]) * (1-x[i, j ])
    }
  }
  loglik <- sum(L[1: n, 1: p])
  LL <- -2*loglik
  AIC <- -2*(loglik - np)
  BIC <- -2*loglik + np*log(n) }

```

APPENDIX E

3PL GENERALIZED TRT

```

model {

# Read in individual item responses
  for ( i in 1:n ) {
    for ( j in 1:p ) {
      x[i , j ] <- response[i, j ]
    }
  }

# Identify generalized three-parameter logistic (3PL) TRT model
  for (i in 1:n) {
    for (j in 1 : p) {
      x[i , j ] ~ dbern(prob[i,j])
      logit(prob.star[i,j]) <- a1[j] *theta[i] - b[j] - a2[j] *test[d[j]]
      prob[i, j] <- eta[j] + (1-eta[j]) * prob.star[i, j]
    }
  }

# Specify prior for examinee parameters
  for (i in 1:n) {
    theta[i] ~ dnorm(0,1)
  }

# Specify prior for testlet parameter
  for (k in 1:T) {
    test[k] ~ dnorm(0, 1)
  }

#Specify priors for item parameters
  for (j in 1:p) {
    b[j] ~ dnorm(0, 1)
    a1[j] ~ dnorm(0, 1) I(0, )
    a2[j] ~ dnorm(0, 1) I(0, )
    eta[j] ~ dbeta(1, 1)
  }
}

```

APPENDIX F
IRB APPROVAL



1/28/2015

Investigator(s): Kyungtae Kim, Dr. Jwa K. Kim
Department: Literacy Studies
Investigator(s) Email Address: kk2w@mtmail.mtsu.edu; Jwa.Kim@mtsu.edu

Protocol Title: Model Comparisons among Testlet Response Theories (TRT) on a Reading Comprehension Test

Protocol Number: #15-168

Dear Investigator(s),

Your study has been designated to be exempt. The exemption is pursuant to 45 CFR 46.101(b)(4) Collection or Study of Existing Data.

We will contact you annually on the status of your project. If it is completed, we will close it out of our system. You do not need to complete a progress report and you will not need to complete a final report. It is important to note that your study is approved for the life of the project and does not have an expiration date.

The following changes must be reported to the Office of Compliance before they are initiated:

- Adding new subject population
- Adding a new investigator
- Adding new procedures (e.g., new survey; new questions to your survey)
- A change in funding source
- Any change that makes the study no longer eligible for exemption.

The following changes do not need to be reported to the Office of Compliance:

- Editorial or administrative revisions to the consent or other study documents
- Increasing or decreasing the number of subjects from your proposed population

If you encounter any serious unanticipated problems to participants, or if you have any questions as you conduct your research, please do not hesitate to contact us.

Sincerely,

Lauren K. Qualls, Graduate Assistant
Office of Compliance
615-494-8918