ABSTRACT

Testlet Effects on Pass/Fail Decisions Under Competing Rasch Models

Kari J. Hodge, Ph.D.

Mentor: Grant B. Morgan, Ph.D.

The item response model chosen to estimate ability can influence proficiency classification, or pass/fail decisions, made about people based on test scores. This poses a potential problem for both the examinee and the decision makers because examinees may be misclassified based on the item response model used to estimate ability and not their actual proficiency in a domain of interest. The purpose of this study was to examine the use of an incorrect item response model and its impact on proficiency classification. A Monte Carlo simulation design was employed in order to directly compare competing models when the true structure of the data is known (i.e., testlet conditions). The conditions used in the design (e.g., number of items, testlet to item ratio, testlet variance, proportion of items that are testlet-based and sample size) reflect those found in the applied educational literature. An empirical example is also analyzed for pass/fail decisions with the competing models.

Overall, decision consistency (DC) was very high between the two models, ranging from 91.5% to 100%. The design factor that had the greatest effect on DC was the testlet effect or testlet variance. Other design factors that affected DC included

number of testlets, an interaction between testlet variance and the percent of total items in testlets, and an interaction between the number of testlets and the percent of total items in testlets. PISA is traditionally calibrated with a DRM, and contained 29 items in nine testlets. The classification agreement percent between the DRM and the TRM was 99.5%. When a testlet structure is present in applied data the testlet variance is unknown and as the testlet variance increases so does the misclassification of examinees. When measurement models are used that do not align with the structure of the data additional error is introduced into the parameter estimates. This directly impacts the decisions that are made about people.

Testlet Effects on Pass/Fail Decisions Under Competing Rasch Models

by

Kari J. Hodge, B.A., M.Ed.

A Dissertation

Approved by the Department of Educational Psychology

_____

Terrill F. Saxon, Ph.D., Chairperson

Submitted to the Graduate Faculty of
Baylor University in Partial Fulfillment of the
Requirements for the Degree
of

Doctor of Philosophy

Approved by the Dissertation Committee

_____

Grant B. Morgan, Ph.D., Chairperson

_____

A. Alexander Beaujean, Ph.D.

_____

A. Scott Cunnigham, Ph.D.

_____

Susan K. Johnsen, Ph.D.

_____

Terrill F. Saxon, Ph.D.

Accepted by the Graduate School

August 2015

_____

J. Larry Lyon, Ph.D., Dean

Page bearing signatures is kept on file in the Graduate School.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would first like to express my sincere gratitude to Dr. Grant Morgan for serving as my mentor and dissertation chair. Through our many conversations and the multitude of projects we collaborated on I have grown personally, professionally, and spiritually. Thank you for the many hours and thoughtfulness you invested in my doctoral journey. I look forward to future collaborations and conversations.

I would like to thank my committee for their time and thoughtful comments, suggestions and questions. All of which stretched and challenged my thinking, thus enabling me to develop stronger scholarly skills and a much richer dissertation.

Next, I would like to thank my advisor, Dr. Terrill Saxon for seeing something in me and encouraging me to pursue the path I really wanted. Without his encouragement my graduate experience would have been much different.

Last but certainly not least; I would like to express my abundant gratitude to my family. Without my mom, Dr. Lucinda Harman, who set an example by allowing me to participate in her doctoral journey I would have never know this was possible. She has been an avid supporter, dedicated proofreader, a sounding board, and unending source of spiritual encouragement. I would like to thank my husband, Tony Hodge, for his patience love, support, and absolute belief in me for which none of this would have been possible. Knowing he has been with me through this seemingly never ending journey made it all worthwhile. Seeing his smile and pride waiting at the end of the stage is the ultimate reward.

DEDICATION

To my husband and my mom.

Tony, this dissertation and my doctoral experience were possible because of your sacrifice, love and support. Thank you for never giving up on me, making me laugh through my tears, and cheering through my success.

and

Mom, you have always shown me I could do more than I thought I could. Thank you for teaching me I could accomplish anything I set my mind on. Thank you for the many hours of prayer, encouragement, and support.

CHAPTER ONE

Introduction

Assessment is a process of obtaining information for making decisions about people, curricula, or programs (Nitko & Brookhart, 2006). Educational decisions made based on results from these assessments may include placement into educational programs or interventions, retention or progression, scholarship or awards, or credentialing or certifying competency, to name a few. The way assessment results are interpreted and used has potentially serious consequences for those being assessed. Therefore, the trustworthiness of data provided by instruments is important for making decisions about people based on test results. Interpretation of data is described as the meaning assigned to the score and use is defined as the action or decision made based on the score (Nitko & Brookhart, 2006).

Decision-making in educational settings often requires the assessment of knowledge, skills, and abilities. Theoretically, knowledge, skills, and abilities are not directly observable and require the use of tests, or a collection of questions or items, to be measured (Crocker & Algina, 2008). The knowledge, skill or ability is referred to as a domain or construct and is inferred from a set of observations or test questions. Therefore, the purpose of a test is to provide information so that inferences can be made about a person's amount or type of a construct measured through these items. For instance, data collected from a K-12 assessment in reading or language arts might be used

to place students in remedial or advanced courses, to make decisions for retention, or to endorse grade completion.

The type and purpose of an assessment must align with the decisions that will be made about examinees (Miller, Linn, & Gronlund, 2009). For example, severe consequences may be imposed on people based on the decisions made in high stake assessment situations. Examples of high stakes decisions include awarding scholarships to some but not all students, admitting some but not all applicants into college, offering employment to some but not all applicants, or allowing some but not all professionals to practice through certification of some type. In such situations, assessment conditions are often standardized in an effort to increase objectivity. That is, standardized assessments usually consist of the same or very similar items, computerized scoring, and the same or very similar response formats. Mastery decisions reflect the degree to which an examinee has met or exceeded a specific standard, which could be minimal competence or high achievement standards. These tests often include basic skills, general knowledge, and professional/applied knowledge (Nitko & Brookhart, 2006).

Tools for collecting information may include observations of performances, paper-pencil tests, research projects, oral questioning, and essays, but tests tend to be used in large-scale assessments. A test is a type of assessment that is an instrument or systematic procedure for describing or observing an examinee based on well defined characteristics via a numerical scale or classification scheme (Nitko & Brookhart, 2006). Tests are usually scored for correctness and are used to measure ability, aptitude, or achievement (AERA/APA/NCME, 1999).

Tests are also generally designed to measure how much examinees have achieved or are able to do, and they are developed to cover breadth as well as the depth of the construct (AERA/APA/NCME, 1999). Item response formats (i.e. multiple choice, constructed response, or essay) are typically selected based on whether breadth or depth of content coverage is desired because it is difficult to simultaneously address both. For example, essays are intended to measure aspects of a construct in more depth by requiring the examinee to provide responses in his/her own words with more complex presentation. Essays require the examinee to provide more information and aim to measure complexity or higher order thinking. Yet, essays require a considerable amount of time to score and may be prone to subjectivity in the scoring. Thus, essays tend to be preferred when depth of coverage is desired. Multiple-choice items tend to be chosen when breadth of coverage is desired because multiple-choice items can cover a construct more broadly and efficiently (Haladyna, Downing, & Rodriguez, 2002). Multiple-choice items are also easier to score because the examinee selects the best answer out of response choices, one of which has been predetermined as correct. What multiple-choice items gain in efficiency is traded for content representativeness due to the lack of depth provided by a single item (Haladyna et al., 2002). More items are therefore needed to cover the domain adequately.

Multiple-choice items are frequently used in large-scale assessments and can be administered in multiple formats or delivery. Haladyna et al. (2002) reported a number of multiple-choice formats including true false, two and four option questions, matching, context-dependent items and context-dependent item sets. The authors described a context-dependent item as an item whose response is based on a given scenario, vignette,

reading passage, graph, chart, or other interpretive material or stimuli. The major benefit of context-dependent items is that they are able to measure higher-level thinking, such as problem solving and/or critical thinking. The major drawback is that they require considerable space on the test, and examinees may need more time to read and respond to the items. One way to solve the problem of inefficiency of single context-dependent item is to develop context-dependent item sets, which are described as several items paired with a common stimulus. These item sets are frequently called testlets, item bundles, super items, and context-dependent item sets. In this paper this type of item set is referred to as a testlet. One major benefit of the testlet is the ability to assess examinees' application of knowledge and skills of more complex constructs, such as reading comprehension, written communication skills, and/or problem solving. One potential drawback of using this type of item set is the violation of conditional, or local, independence, which is a statistical problem. This is discussed in more detail in the following section and is a major focus of this study.

Testlets

Multiple-choice items are able to measure a broader range of content more efficiently and are scored more easily than more complex constructed-response items, such as essays or performance assessments. The tradeoff for efficiency may be that individual items may not as deeply measure the domain of interest. One method of countering this problem was to introduce a more in-depth stimulus that is better able to represent the domain of interest and develop a group of items related to the stimulus (Wainer, Bradlow, & Wang, 2007). Examples of in-depth stimuli include reading

comprehension passages, science or mathematics graphs and tables, geography maps, cases, scenarios, or music compositions.

Wainer and Kiely (1987) proposed the term "testlet" to denote a group of items that are administered together (e.g., in computer adaptive situations where different sets of items might be administered). Similarly, Wainer et al. (2007) define a testlet as "a group of items that may be developed as a single unit that is meant to be administered together" (p 53). Wang and Wilson (2005a) described testlets as groups of items that have a common stimulus, such as a case-based scenario or reading passage. Although testlets have partially addressed the problem of the lack of depth in domain representation with multiple-choice items, the use of testlets may create additional relationships between items unaccounted for by the construct of interest.

Almond, Mulder, Hemet, and Yan (2009) described a situation in which familiarity or unfamiliarity of the context related to the reading passage provides additional context to the passage that may not provide information about the construct under investigation. They provided the example of a reading comprehension test where the reading passage was about dinosaurs. If an examinee was interested and familiar with dinosaurs, she or he may be more likely to recognize words like "pterodactyl" and "Paleolithic" whereas an examinee who was unfamiliar with dinosaurs might have to decode these words and infer their meaning from the context of the passage. Thus, it is necessary to account for the construct-irrelevant context from the response pattern in order to provide a more accurate estimate of the ability or proficiency in question, because the items may be unintentionally measuring more than one construct.

Purpose of Study

When testlets are present the model selected to estimate examinee ability may influence the pass/fail decisions made about examinees due to the additional relationship between items that is unaccounted for. This poses a potential problem for both the examinee and the decision makers. Therefore, this study seeks to compare the known proficiency status via the Rasch testlet model of each observation to the estimated pass/fail decision via the dichotomous Rasch model when there are testlets that are ignored. A Monte Carlo simulation design was employed in order to compare competing models when the true structure of the data is known (i.e., testlet conditions). The factors used in the design (e.g., number of items, testlet to item ratio, testlet variance, percent of total items that testlet-based and sample size) reflect those found in the applied education literature. An empirical data set is also analyzed in order to compare the pass/fail decisions between the competing models.

Overview of Procedures

First, the applied literature was reviewed in order to identify the empirical conditions under which educational assessments include testlets. These conditions were reported and formed the basis for generating conditions that mirror applied settings. The particular conditions included in this study were the number of items, item to testlet ratio, testlet variance, proportion of items that were testlet-based, and sample size. In order to reflect multiple choice large-scale educational assessments all items in this study were dichotomous (Fountas & Pinnell, 2012; Good, Wallin, Simmons, Kame'enui, & Kaminsji, 2002; Harcourt, 2003; NCES, 2014; SBAC, 2014).

All data simulated for this study were generated from a testlet population structure. Next, two competing models were fit to the simulated data and pass/fail decisions were recorded based on a predetermined cut score. Because the population model was known, the appropriate pass/fail decision for each simulated examinee is likewise known. This allowed for comparisons to be made between the true and estimated pass/fail decisions and thus evaluate the decision consistency between the competing models. The proportion of pass/fail decisions that differ between models was reported. The competing models were also applied to a publicly available secondary dataset (i.e., Program for International Student Assessment [PISA]) that contains item-level responses collected from a subset of students from the United States. The items were a part of the reading assessment, which included testlets. The pass/fail decision consistency produced by the models were examined and reported. Finally, the implications and considerations for the use of these models in high-stakes assessments are discussed.

Delimitations

A primary limitation of all studies that use Monte Carlo methods is that the results are generalizable only to the conditions simulated. Therefore, it is essential to generate conditions similar to those found in the applied literature. As is true with any simulation study, this study did not simulate all possible conditions reported in the literature. Instead the most commonly reported conditions were included, along with an empirical example. Finally, there are many models with varying degrees of complexity available to estimate ability about which decisions about examinees can be made. This study includes Rasch models, which are more parsimonious than other item response models. Rasch models were used based on the idea that if ignoring testlets affects pass/fail decisions in less

complex models, then the effect of ignoring testlets may be increased and possibly more difficult to detect in more complex models.

CHAPTER TWO

Literature Review

Introduction

Items linked via a reading passage, scenario, or a case are called testlets and are commonly used in educational assessments (Fountas & Pinnell, 2012; Good et al., 2002; Harcourt, 2003; NCES, 2014; SBAC, 2014). The use of testlets in test construction was developed out of the need to more efficiently test a person's ability to understand a stimulus such as a reading passage, map, graph, or music composition (Wainer, et al., 2007). One of the greatest advantages of using testlets, aside from efficiency, is the ability for test developers to construct tests in ways that may be more representative of the construct being measured. Yet, the use of testlets violates the assumption of local item independence (LID) and may introduce bias and instability into score interpretation (Zumbo & Rupp, 2004) because testlets are secondary dimensions that are unaccounted for in traditional models.

Unidimensional models that ignore the testlet effect are frequently used in large scale testing situations (Dickenson, 2005). Similarly, increases in the use of testlets by test developers created a need to investigate the potential differences in decisions made about people depending on the model used to calibrate tests containing testlets. In other words, the raw or sum score from a test is only an indication of a possible measure (Dickenson, 2005). The use of Rasch measurement models provide a means of constructing inferences from observations that are independent of who else is taking the

test or the difficulty of the items on the test in the following ways: a) constructing a linear measure, b) accommodating missing data, c) providing estimates of reliability, and d) providing a means of detecting misfit (Wright & Mok, 2004). This chapter will introduce the models used in this study and provide a review of the methodological studies related to testlets.

Testlets

Testlets are becoming increasingly prevalent in educational testing situations. Languages Arts and reading assessments are prime examples of tests that include testlets, as examinees are provided with a reading passage and then asked a group of items related to the passage. For example, the Smarter Balanced Assessment Consortium (SBAC, 2014) English language arts assessment includes testlets and was designed for alignment with the Common Core State Standards Initiative (e.g., reading critically, vocabulary, and comprehension), and results are used to make decisions about students proficiency, school accountability ratings, and implementation of programs or interventions for third through twelfth grade students. The Texas Primary Reading Inventory (TPRI), The Dynamic Indicators of Basic Early Literacy Skills (DIBELS) (Good et al., 2002), The Rigby PMs collection (Rigby) (Harcourt, 2003), and The Fountas and Pinnell Benchmark Assessment system (F&P BAS) (Fountas & Pinnell, 2012) are reading assessment programs that measure reading readiness or monitor reading progress for elementary students by asking the student to read a story or passage and then answer group of questions about the story. The National Assessment of Educational Progress (NAEP) is a large national assessment that measures academic achievement of U.S. students every two years in fourth, eighth, and twelfth grades (NCES, 2014). The NAEP reading

assessment measures reading comprehension and is used to compare the nation, states, and urban districts and is disaggregated by demographics (e.g., sex, socioeconomic status, race/ethnicity).

In all of these assessments each examinee is presented with a reading passage and then asked at least three questions about the passage. Due to the number of assessment programs using testlets (Fountas & Pinnell, 2012; Good et al., 2002; Harcourt, 2003; NCES, 2014; SBAC, 2014), it is necessary to investigate the extent to which the decisions made based on these tests are affected by the inclusion of testlets so that examinees are not misclassified.

Testlets are operationally defined in this study as groups of items that have a common stimulus. Almond et al. (2009) described a situation in which familiarity or unfamiliarity with the context or topic related to a reading passage provided additional context or differing prior knowledge about the passage that may not provide information about the latent construct under investigation (i.e., reading ability).

Testlets create patterns of local dependence (i.e., additional correlation between items) in item responses beyond the effect of the underlying latent trait. Imposing local independence, the probability of observing a correct response (i.e., coded as "1") on the *i*th item can be expressed:

$$\mathrm{P}(U_i = 1 \mid \Theta = \theta) = \mathrm{P}(U_i = 1 \mid \Theta = \theta, V = v), \text{ for all } V \qquad [2.1]$$

where P is the probability of the observed item response of item *i* given θ, U_i is the response to item *i,* Θ is the latent trait, θ is the specified value of person ability, V is any other variable that is unaccounted for. In equation [2.1], the latent trait, Θ, contains all necessary information for accounting for a set of item responses because conditioning on

any other variable, *V*, beyond the latent trait does not change the modeled probability of a correct response to item *i*. Local dependence means that the latent trait of interest does not adequately account for the covariance in item responses. Therefore, local dependence can be expressed:

$$\mathrm{P}(U_i = 1 \mid \Theta = \theta) \neq \mathrm{P}(U_i = 1 \mid \Theta = \theta, V = v), \text{ for some } V. \quad [2.2]$$

This additional dimension may influence parameter estimation when the model selected does not account for LID. The item response model selected can influence proficiency classification based on test scores. This may pose a problem for both the examinee and the decision makers because examinees may be misclassified based on the item response model and not their proficiency in a domain of interest. Therefore, this study seeks to compare the known proficiency status via the Rasch testlet model of each observation to the estimated pass/fail decision via the dichotomous Rasch model when there are testlets that are ignored.

Approaches for Modeling Testlets

Multiple approaches have been proposed for dealing with violations of local independence resulting from the use of testlets. A common approach is to ignore the dependency and proceed with a traditional measurement model (Wainer, et al., 2007). Most traditional measurement models assume local, or conditional, independence that includes the assumption that all item responses are uncorrelated after the latent trait of interest has been accounted for (Yen, 1993). Although it is rarely believed that this assumption holds completely, testlets directly violate the assumption of LID and may introduce bias and instability into score interpretation (Zumbo & Rupp, 2004). Wainer, et al. (2007) suggested that if the number of items included in each testlet (e.g., four testlets

with three items in each testlet) is relatively small, then ignoring the violation of local independence assumption might be acceptable. As the item-to-testlet ratio increases or the magnitude of the dependence in the testlet increases, then a lack of precision in the parameter estimates may also increase. In other words, as the number of items included in each testlet increases, the residual correlations (i.e., local item dependence) increase. Additionally, as the magnitude of the LID increases, there is more variance left unaccounted for by the primary dimension (i.e., latent trait of interest). Therefore, as more variance is left unaccounted for, the precision of the parameter estimates is negatively affected.

The second method is to aggregate these types of items into what have been called superitems, which are created by combining binary responses into a polytomous item (Wainer & Kiely, 1987). This alternative for modeling an assessment with testlets would treat testlets as polytomous items instead of individual dichotomous items (Thissen, Steinberg, & Mooney, 1989; Wainer & Kelly, 1987; Wainer & Lewis, 1990). For example, if 10 items relate to the same case or scenario, the scored responses would be added together to produce a possible range of scores from zero to ten. Then the ten items would be treated as a single polytomous item, and each testlet would be treated in this way. Items within testlets are considered locally dependent, but the separate testlets are considered locally independent. This approach solves the within-testlet dependency problem, but the item-level response pattern is lost. Combining items from a testlet into one superitem will provide information about the number of items in the testlet that a person answered correctly but not which specific items were answered correctly.

The third approach is to account for the additional dependence by including testlets as latent variables in a model specifically designed for testlet effects. This method accounts for the construct-irrelevant context by providing a more accurate and representative scoring pattern. Item responses are analyzed as individual items instead of polytomous items, where the additional relationship is accounted for by the latent testlet variable. Although testlet models are more complex to implement, they provide the most information about people while providing more flexibility in modeling item responses. Wang and Wilson (2005a) describe three advantages of the testlet model over the polytomous model with superitems. The first advantage is that the unit of analysis remains at the item level rather than the testlet level. Second, the item scoring rubrics remain the same, meaning each item is scored individually and not summed together. Third, item difficulty parameters are conceptualized the same way as the dichotomous model. This third approach is the approach taken in this study.

Rasch Models

The family of Rasch models are probabilistic models for estimating person ability and item difficulty parameters, where the measured proficiency or ability estimate does not depend on who else is taking the test or the difficulty of the test (Rasch, 1960). Rasch models separate the person parameter from the item parameters and provide information about a person's ability based on responses to a set of items. Similarly, item difficulty can be derived from the responses to an item from a set of people (Rasch, 1960). The probability of a person responding correctly depends on how much ability the person has and the difficulty of the item. The Rasch model also provides a method for ordering people based on their ability and items according to their difficulty. The traditional,

dichotomous Rasch model and the testlet Rasch model are described in more detail below.

Dichotomous Rasch Model

The traditional, dichotomous Rasch model (DRM) is appropriate for items that are scored as dichotomous outcomes of a probabilistic process that is a linear combination of ability and difficulty parameters (see Figure 2.1). This model can be expressed as (Linacre, 2004):

$$\log\left(\frac{P_{ni1}}{P_{ni0}}\right) = \theta_n - \beta_i \qquad [2.3]$$

where θ_n is the ability of person *n*, β_i is the difficulty of item *i*, P_{ni1} is the probability that person *n* will succeed on item *i*, and P_{ni0} is the probability that person *n* will fail on item *i.* Therefore P_{ni1} is expressed as

$$P_{ni1} = \frac{\exp(\theta_n - \beta_i)}{1 + \exp(\theta_n - \beta_i)}. \qquad [2.4]$$

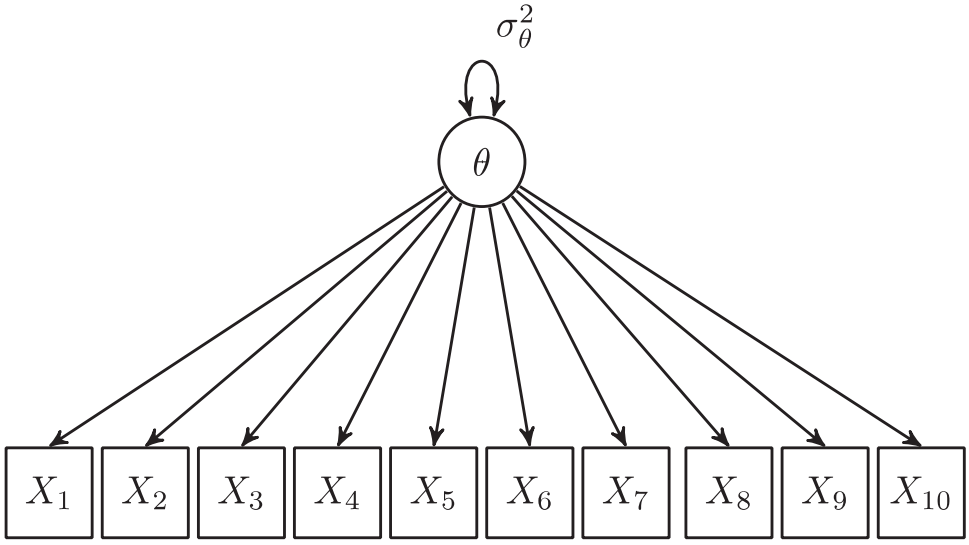

Figure 2.1: Rasch conceptual model for ten independent items with error terms excluded

The DRM sets item discriminations to be equal to one, meaning that items discriminate equally across items and people. The model assumes item response functions are monotonically increasing; meaning that as ability level increase the probability of correctly responding to an item does not decrease. The DRM further assumes that the collection of items is unidimensional, or that they measure one latent construct. The last assumption for the DRM is local independence, which means that after partialling out the correlation between items due to person ability, the items are no longer correlated (Linacre, 2004). The DRM is conceptually consistent to Spearman's model of intelligence (Spearman, 1927).

Rasch Testlet Model

The Rasch testlet model (RTM) is similar to the DRM except that the RTM includes one or more additional random effect parameters to account for the violation in the local independence assumption (see Figure 2.2). A testlet effect is an interaction between a person and the testlet, and each testlet may have a different effect. According to Wilson and Wang's (2005a) RTM the probability of a correct response on a dichotomously scored item can be defined as:

$$P_{ni1} = \frac{\exp(\theta_n - \beta_i + \gamma_{nd(i)})}{1+\exp(\theta_n - \beta_i + \gamma_{nd(i)})} \quad [2.5]$$

where $\gamma_{nd(i)}$ is the testlet effect and P_{ni1} is the probability of person *n* with ability θ_n will succeed on item *i* with difficulty β_i. One assumption of this model is that the testlet effect $\gamma_{nd(i)}$ is mean centered within testlets such that the sum of the testlet effect is equal to zero (i.e., $\Sigma_n \gamma_{nd(i)} = 0$), and the testlet effect is normally distributed with a mean of zero and some standard deviation (i.e., $\gamma_{nd(i)} \sim N(0, \sigma^2_{\gamma_{d(i)}})$). According to Wilson and Wang

(2005a) when the variance of $\gamma_{nd(i)}$ equals zero there is no testlet effect (reduces to dichotomous Rasch model), and $\sigma^2_{\gamma_{d(i)}}$ represents the amount of the testlet effect for testlet d(i). The larger the value of $\sigma^2_{\gamma_{d(i)}}$ the greater the proportion of total variance in test score that is attributable to the testlet. In the Rasch testlet model the variance is the same for all testlets ($\sigma^2_{\gamma_{d(i)}} = \sigma^2_{\gamma}$). An additional assumption for the RTM is that the ability θ_n parameter and the testlet $\gamma_{nd(i)}$ effect are independent, meaning that testlet effects are not different at varying levels of ability (Paek, Yon, Wilson, & Kang, 2009).

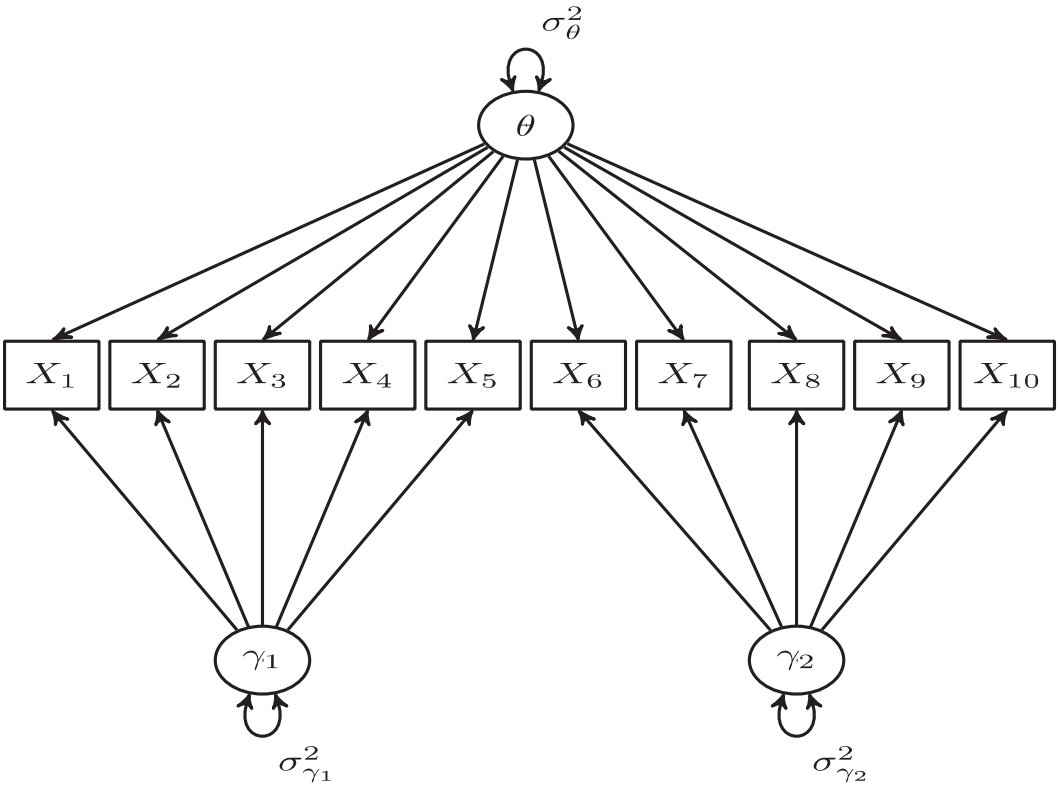

Figure 2.2: Testlet Rasch conceptual model for two testlets with five items in each testlet and error terms excluded

Applied and Methodological Studies

Applied Examples

Differential item functioning (DIF) studies examine the response patterns across groups, such as ethnic subgroups or linguistic groups compared with other subgroups. DIF studies are the most prevalent use of testlet models due to the fact that when testlets are not modeled, DIF is either more pronounced or less obvious (Wang & Wilson, 2005b). Two possible sources of DIF include the testlet effect and the item characteristics. For example, when testlet effects are large and difficulty is small for one group but not the other DIF is amplified at the item level. Similarly, when the testlet effects are small and difficulty is large for one group and not the other then DIF is amplified at the item level. This means that DIF estimates will be larger for those items, but may be influenced by the testlet and not necessarily a problem with the item.

Conversely, when both testlet effects and item difficulty are either very large or very small for one group but not the other then DIF cancellation occurs at the item level. This means that DIF estimates will be smaller for those items but may be influenced by the testlet and may be more problematic. However, when items that have small but systematic DIF are combined into testlets DIF is amplified at the testlet level. Items with large but un-systematic DIF would go undetected when combined into a testlet resulting in DIF cancellation at the testlet level. Fortunately, item response testlet models provide a means of investigating the conditional probabilities of correct responses for different groups by way of item and testlet characteristic curves (Wang & Wilson, 2005b).

Bao, Dayton, and Hendrickson (2009) investigated DIF between males (n=2,875) and females (n=3,078) and Caucasians (n= 3,171) and minorities (n=1,271) on the 1995

American College Testing (ACT) Reading exam. There were 40 items in four testlets. Model comparisons between the two-parameter logistic (2PL) model, 2PL testlet model, multi-group 2PL, and the multi-group 2PL testlet model were compared using deviance information criterion (DIC). They concluded that the multi-group 2PL testlet model was the best fitting model, and they found both DIF amplification and cancellation.

Bao et al. (2009) found that the first testlet variances between subgroups were similar and ranged from 0.2 and 0.3. However, they reported that males scored lower than females and minorities scored lower than Caucasians on the first testlet. The second testlet variance was smaller and ranged from 0.1 to 0.2. On average males scored higher than females and minorities scored lower than Caucasians. The third testlet variance was about the same (γ =0.46) for males and females but was different for minorities ($\gamma = 0.53$) and Caucasians ($\gamma = 0.40$). The fourth testlet variance was one. The authors noted that all participants scored especially low on this testlet but that minority group scored much lower than the Caucasians group. Item level DIF revealed that item seven ($\mu DIF = 0.8836$) and item twenty ($\mu DIF = -0.8027$) had the largest differences between subgroups. They concluded that the magnitude of DIF for item seven was attributed to the testlet effect where as the magnitude DIF of item twenty was due to item difficulty after controlling for the main latent trait and the testlet effect. Items one, four, and 34 were reported to have large magnitudes of DIF and favored Caucasians and items eight and nine had moderate magnitudes of DIF and favored minorities. All other item difficulties were reported as negligible. Bao, et al. (2009) concluded that there was a person-testlet interaction effect and that the magnitude of the effect varied between testlets and subgroups. In other words, when testlets are present a test may have DIF at

the item and or testlet level and when the testlets are not modeled, DIF at the testlet level may diminish or exacerbate DIF at the item level.

Testlets affect the accuracy of individual proficiency classification. Hooker and Finkelman (2010) investigated what they termed paradoxical results (when one person passes and the other fails even though the person who failed may have answered more items correctly) with testlet-structured data in a reading test. They assumed a single dimension of the ability θ for the tests that include testlets. They found that using multidimensional item response models, to capture the nuanced testlet effect, were more susceptible to the paradoxical results, which can have adverse consequences for examinees. However, when item bundles are treated as independent effects in the item response models this phenomenon is minimized. .

The reliability of proficiency classification is often over estimated due to the additional correlation in items when testlets are present. Zhang (2010) investigated decision consistency of proficiency classification for examinees taking the Examination for the Certification of Proficiency in English (ECPE) of the listening and grammar, cloze, vocabulary, and reading (GCVR) sections. The listening section had 35 independent items and three testlets with five items each. The cloze section had one testlet with 20 items and the reading section had four testlets with 5 items each. The grammar and vocabulary section each had 30 independent items. The sample reported in included 5,000 randomly selected examinees from the ECPE (Zhang, 2010). A 2PL and a 3PL testlet model were used in this study. The 3PL dichotomous and polytomous model were estimated using marginal maximum likelihood (MMLE) with the MULTILOG computer program whereas the 3PL testlet model was estimated using Markov chain

Monte Carlo (MCMC) procedures in the SCORIGHT computer program. Zhang (2010) reported testlet effects for all testlets were ≥ 0.30. The cut score for the listening section were similar across models but the GCVR and reading cut scores were more varied across models. Zang (2010) compared the accuracy of classification as the percentage of examinees classified correctly. The listening, (3PL testlet=84.8, 3PL dichotomous=85.0, 3PL polytomous=84.3) GCVR (3PL testlet=88.1, 3PL dichotomous=88.5, 3PL polytomous=87.9), and the reading (3PL testlet=81.8, 3PL dichotomous=84.8, 3PL polytomous=80.4) section were all compared across the three models and in all cases the dichotomous IRT model resulted in higher accuracy of classification. Zang (2010) concluded that the higher percentage in accuracy was due to inflated estimates due to the violation in LID. He noted that this inflation was more pronounced on the reading section and he called for the use of a testlet model.

Testlets create psychometric problem for test equating (Eckes, 2013). Eckes (2013) investigated the structure of Test of German as a Foreign Language Test (TestDaf) listening section for testlet-based dependency. The author compared the TestDaf structure using three item response models (i.e., traditional 2PL, graded response, and 2PL testlet IRT). This portion of the test included 25 items linked to three listening passages with eight, ten, and seven items, respectively. The participants consisted of two samples. The first sample had 2,859 examinees (1,855 females and 1,004 males) and the second sample had 2,214 examinees (1,429 females and 785 males) and were all foreign students applying for admission to a higher education institute in Germany. The test measures language proficiency ranging from intermediate to operational proficiency with three levels of classification. The test contains both short

answer and true/false items that are all scored correct/incorrect by expert raters. Eckes (2013) used the SCORIGHT computer program to estimate the testlet models using Bayesian estimation by drawing samples from the posterior distribution using MCMC. However, at the time of the study SCORIGHT was limited to nine categories for the polytomous model so the author collapsed categories from two of the testlets. Eckes (2013) reported using five chains, with 4,000 iterations and 3,000 as burn-in iterations. The first 3,000 iterations called the burn-in are not included in the posterior distribution as this burn-in phase allows the estimates to fluctuate in order to reduce the serial correlation in sampling distribution. Eckes (2013) found that in both samples only testlet two had a variance above 0.25, which was the *a priori* value below which was determined to have negligible effects. The person ability correlation estimates compared across models ranged from .982 to .999. The mean difference in person ability ranged from 0.00 to 0.01. Root mean square difference (RMSD) ranged from 0.004 to 0.17. The testlet model revealed lower precision estimates for person separation reliability (R) (*R*=0.71, 0.74) compared to the graded response (*R*= 0.78, 0.85) and the 2PL models (*R*=0.76, 0.82). The testlet model revealed higher RMSE estimates for the testlet model (RMSE= 0.48, 0.45) compared to the graded response (RMSE= 0.42, 0.36) and the 2PL models (R= 0.44, 0.39). The graded response model tended to underestimate the ability at lower levels of the ability scale. The 2PL model slightly overestimated difficulty at the lower end of the difficulty scale. Although, the testlet effect was reported as small, Eckes (2013) concluded that there were small but noticeable differences between the three models.

While the inclusion of testlets in test construction have many appealing features for reading assessment, they create psychometric problems for score interpretation and use, due to amplification or cancelation of DIF (Bao, et al., 2009; Wang & Wilson, 2005b), biased parameter estimates (Bradlow, et al.,1999), errors in examinee classification (Hooker & Finkelman, 2010; Zang, 2010), and test equating (Eckes, 2013).

Methodological Studies

Wang and Wilson (2005b) investigated traditional DIF detection techniques applied to testlet items calibrated with a testlet model. Specifically they used simulations to investigate how well the testlet models recovered the DIF, ability, and difficulty parameters for Rasch testlet models. The conditions varied in this study were test type, anchor item method, difference between the mean estimates and the generated values, and the root mean square error of the estimates. The dichotomous model had 40 items across four testlets with 10 items per testlet, and the difficulty of the items ranged from -2.00 to 2.00 with a mean of zero. The polytomous model contained 24 three-point items across four testlets with six items per testlet with step difficulties that ranged from -2.00 to 2.00 with a mean of zero. The mixed model has 20 dichotomous items in two testlets and 12 three point polytomous items across two testlets. Variances of the random testlet effect were set at 0.25, 0.5, 0.75, and 1.00. Members of the reference group were distributed as $N(0.5, 1)$ and the focal group were $N(-0.5, 1)$ such that the reference group was on average one logit higher on ability than the focal group. The differences in item difficulty between the reference and focal group for DIF items were 0.4 and 0.6, so that a negative value indicates a lower difficulty for the reference group than the focal group. This would suggest that the item favored the reference group. The authors set the

proportion of DIF items at 40% and noted this is considered a high proportion. Two anchor methods were used in the detection of DIF. The first approach is to anchor one item that is known to be DIF free and estimate the rest of the items with a DIF parameter. The other approach is to anchor all items as DIF free so that the model was the generating model. Wang and Wilson (2005b) reported that one of the 85 estimators of the dichotomous model was biased and that the magnitude of that bias was -0.030 to 0.021. The minimum, maximum and mean of the RMSE were 0.98, 1.97, and 1.13 and they concluded that the all item anchor method was more efficient than the one item anchor method. Similar results were noted for the polytomous model where five out of 100 estimates were biased. The magnitude of the bias ranged from -0.024 and 0.0029. The minimum, maximum and mean for RMSE were 0.95, 1.30, 1.05 indicating that the all anchor method was more efficient than the one item anchor method but that these differences were very small. Four of the 92-parameter estimates in the mixed model were biased with magnitudes that ranged from -0.029 to 0.030. The minimum, maximum and mean for RMSE were 0.70, 1.53, and 1.01 indicating that the two anchor methods were the same with respect to efficiency. The authors concluded that the traditional DIF techniques could be used to detect DIF in testlet models and that both anchor methods recovered the parameter well.

Bradlow, Wainer and Wang (1999) conducted a simulation study to investigate parameter estimates with data augmented Gibbs sampler models (DAGS) with and without the testlet effect. Seven conditions were studied with 1,000 observations, 60 items with 30 testlet items and 30 independent items, and the same set of population distributions, $\theta_n \sim N(0, 1^2)$, $\beta i \sim N(\mu_\beta = 0, \sigma_\beta^2 = 1)$, $ai \sim N(\mu_a = 0.8, \sigma_a^2 = 0.2^2)$. The

distributional values were modeled after the Scholastic Aptitude Test (SAT) and then compared to real data analysis. Two conditions were varied 1) items per testlet (5 and 10) and 2) testlet variance ($\frac{\sigma_y^2}{\sigma_\beta^2+\sigma_y^2} = \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, = \sigma_y^2 = \frac{1}{2}, 1, 2$). Mean absolute error (MAE) and correlation ranks between the estimated and true values (RRank= cor(ranked estimates, ranked true values)) for ability, difficulty and discrimination and the 95% coverage probability (95% CP=$\Sigma 1(\text{true value } \epsilon\ (\hat{F}_{0.025}, \hat{F}_{0.975}))$) for all parameter estimates summarized simulation conditions (where $\hat{F}$ is the empirical cumulative distribution function taken from the posterior draws). They found that MAE was always smaller for the DAGS testlet model than the non-testlet DAGS and that the magnitude of the improvement is monotonically related to the increase in σ_y^2. Posterior intervals (akin to a confidence interval) for all parameter estimates were narrower when the testlet was unaccounted for. When the testlet is unaccounted for this narrow interval is misleading.

Similarly, coverage properties were better for the testlet model. In other words the coverage properties for the ability or proficiency estimates are less biased. Difficulty was either too low or too high when the testlet was not accounted for. The purpose of the SAT is to rank order people and make decisions about people at the top of the ability distribution. The simulation study found that the DAGS testlet model had higher rank correlations than the other two models, meaning that when the testlet model was used a higher percentage of the top 100 examinees were identified. SAT scores are frequently used to award scholarships or acceptance into top colleges. This study demonstrated that when the testlet model is not used fewer examinees would receive benefits for their test scores than they should have had they been modeled with a testlet model. However, when

the models were applied to real data all models performed about the same due to the posterior mean testlet variance of $\sigma_y^2 = 0.11$, classified by the authors as a modest effect.

Implications of testlet effects are clearly impacted due to the magnitude of the testlet variance. DeMars (2012) investigated model based approaches to detecting testlet effects. The conditions varied in this study include test length and items per testlet (25 items with 5:5 item: testlet ratio, 50 items with 5:10 items: testlet ratio, and 50 items with 10:5 items: testlet ratio). Item parameters were randomly selected for each replication where $\beta i \sim N(0,1)$ with a range of -3 to 3. Guessing parameters were constrained to c_i = 0.2, discrimination parameters were $ai \sim N(0, 0.5)$ with a range of 0.5 to 2. The testlet discrimination to general discrimination were constrained to be constant with in each testlet thus making the bi-factor model equal to the random-effects testlet model, where the ratios were 0, 0.3, 0.6, 0.9, 1.2 (ratios of zero indicate no testlet effect). Simulated data included 2,000 observations with 1,000 replications for each condition. The models used to investigate testlet effects included the unidimensional model, a single testlet model, an all-but-one testlet model, and a complete bi-factor model. Model fit statistics included Testfact and -2 Log-Likelihood (-2LL), Dimtest's test of Essential Unidimensionality, Akaike's Information Criteria (AIC), sample-size adjusted Bayesian Information Criteria (SSA-BIC), and Bayesian Information Criteria (BIC). Over all SSA-BIC was most accurate in detecting the true model. However, all indices were increasingly more effective when the testlet effect was larger (DeMars, 2012).

Jiao, Wang, and He (2013) reported that the Rasch testlet models and the three-level one-parameter logistic (1PL) testlet model were equivalent. Where level one expresses the log-odds of a person *j* answering item *i* in testlet *d* using linear regression.

Level two models the testlet effect and level three models the person effect. The Rasch testlet model and the three level 1PL model are algebraically equivalent (interested readers should see Jiao, Wang, & He, 2013). If testlet effects are considered additional dimensions of the general dimension being measured by the test, then the testlet model is a special case of the multidimensional random coefficients multinomial logit model. Jiao, et al.(2013) described this model as a Rasch version of a bi-factor model, which is conceptualized similarly to the Bayesian random-effects testlet model. The 1PL testlet model was used to compare three estimation methods: MCMC in WINBUGS, marginalized maximum likelihood estimation (MMLE) with the expectation-maximization (EM) algorithm in ConQuest, and the sixth-order approximation Laplace (Laplace) method in HLM6. Simulation conditions included: 1,000 observations, 54 multiple-choice items with six testlets and nine items in each testlet. True values of person ability and item difficulty were randomly selected from a normal distribution, $\theta_n \sim N(0,1)$ and $\beta i \sim N(0,1)$. The testlet effects were simulated from a normal distribution $\gamma_{d(i)n} \sim N(0, \sigma^2_{\gamma_{d(i)n}})$ with $\sigma^2_{\gamma_{d(i)n}}$. = 0, 0.25, 0.5625, and 1. When $\sigma^2_{\gamma_{d(i)n}} = 0$, the testlet model reduces to the Rasch model. Twenty-five replications were generated for each condition.

The authors compared the three estimation methods by comparing the true values with the estimated ones for all three estimation methods as well as comparing bias, standard errors, and root mean square error. For the MCMC method 1,000 iterations were used to burn-in the conditions (for more information on MCMC estimation see Chapter 3 section on Bayesian estimation) that had a $\sigma^2_{\gamma_{d(i)n}} = 0$ and 2,000 for the test of the conditions. The results indicated that MCMC slightly overestimated true values when

testlet effects were small. However, MCMC estimation produced estimates with less average bias when the testlet effects were moderate to large. According to Jiao et al. (2013), the estimation method produced significantly different estimates of bias in the testlet and ability variance, random error of ability, and bias in item difficulty. The authors concluded that MCMC was more efficient than the Laplace or MMLE estimation methods.

Li, Bolt, and Fu (2006) compared four testlet models using pseudo-Bayes factor (PsBF), deviance information criteria (DIC), and posterior predictive checks (Bayesian chi-square). The first model is the two-parameter normal ogive model that includes a random effect that represents the interaction between people and testlets where the variance of the testlet is allowed to vary and is an indication of the amount of local dependence in each testlet. Items with in testlets are considered conditionally independent.

The second model is a general model that is a multidimensional model that treats each testlet as a separate ability dimension, where the latent trait underlying an examinee's response to a testlet contains an ability estimate and a random dimension for each testlet. This model does not make assumptions about the relationship between item discrimination in regards to the ability and the secondary dimension.

The third model is a general model with constrained slopes containing multidimensional discrimination parameter that are constant across items. This model may be useful for well-designed tests where the influences of the testlet effects are minimized and secondary dimensions are small. Model 2 is a special case of model 1 with a reduction in the number of parameters estimated.

The fourth model is a general model with constant discrimination parameters across all items but the slope is not the same for the ability estimate and the testlet or secondary dimension. However, overall influence of the testlet factor can still be varied. Model 3 is also a special case of model 1.

The small simulation study showed that the PsBF and the Bayes Chi-square based on the odds ratios were able to distinguish between the true model and the fitted models more effectively than the information criteria fit indices. Li et al. (2006) also found that the second model and the first model fit the real data better than the special case models (third and fourth). This is because the second model makes no assumptions about item discrimination parameters in regards to the intended ability and the secondary dimensions. This model also provides more information about items in relation to ability and each testlet, allowing the study of items or item types that may be more influenced by passage or scenario factors and which passages or passage types contribute to passage effects on items.

In summary, Wang and Wilson (2005b) found that traditional DIF detection techniques could be applied to testlet models and that the anchor method was able to recover the parameters efficiently. The detection of testlet effects is aided by fit indices that perform better for larger testlet effects (DeMars, 2012). Bradlow et al. (1999) found differences in examinee classification between testlet and non-testlet models, but when the testlet effect was small classification between the two models were more consistent.

The methodological studies presented here provide support for the use of testlet models for calibrating items and people as well as psychometric investigations including model fit (DeMars, 2012; Li, et al., 2006), DIF (Wang & Wilson, 2005b), and examinee

classification (Bradlow, et al., 1999). Estimation techniques were also compared and MCMC was found to be more efficient (Jiao et al., 2013).

Frame the Problem

Bradlow, Howard, Wainer and Wang (1999) reported overestimation of precision of ability and item parameters. DeMars (2012) investigated fit indices for detecting testlet effects and found that all indices were increasingly more effective when the testlet effect was larger. Jiaoet al. (2013) investigated testlet model estimation and concluded that MCMC was less biased and more efficient than MML or Laplace estimation procedures. Zhang (2010) found that when testlet effects were present and traditional item response models were used, decision consistency estimates were overestimated compared to those estimated with a testlet model. Although each of the studies presented thus far recommend using a testlet model when testlets are present, according to Dickenson (2005) non-testlet models are still being used in large-scale testing programs. The following research questions will be addressed:

1. Does the pass/fail decision made for each examinee change depending on whether a DRM or a RTM was used under conditions reflective of those found in applied settings?
    a. Number of testlets
    b. Number of items per testlets
    c. Testlet variance
    d. Sample size
    e. Proportion of items that are in testlets
2. Does the pass/fail decision made for each examinee change depending on whether a DRM or a RTM was used in the Program for International Student Assessment (PISA) dataset?

This study adds to the growing body of literature on testlets and violation of the LID assumption in two ways. First, in the model comparison studies described above ability estimates were compared based on bias estimates or rank order correlations. This study seeks to compare actual pass/fail decisions based on a predetermined cut score. Second, in all of the studies described above, traditional IRT/Rasch models were estimated in one analytic program and estimated with a likelihood algorithm, while the testlet models were typically estimated in SCORERIGHT or ConQuest with MCMC. However, the models used in this study will be generated and estimated in the same program or software with the same estimation procedure. This sets the model estimates up for direct comparison and eliminates alternative explanations for differences found.

CHAPTER THREE

Method

Introduction

As stated in Chapter Two a testlet is a group of items that are administered together and have a common stimulus (Wang & Wilson, 2005a). However, the use of testlets created additional relationships between items, which violates the assumption of LID and may introduce bias and instability into score interpretation (Zumbo & Rupp, 2004). This chapter provides the models used in this study are reintroduced, an overview of the empirical conditions on which the generated parameters were derived, the selected parameters, the statistical software, estimation techniques, and a summary of how the outcome variables were developed.

Competing Models

Rasch analysis involves the use of mathematical models to measure individual examinees' ability or latent trait, where the probability of a correct response is modeled through a function of the ability parameter and the item parameters. The family of Rasch models separates the person ability parameter from the item difficulty parameters to provide information about a person's ability based on the response to a specific item (Rasch, 1960). According to Linacre (2004) the probability of a correct response on a dichotomously scored item can be defined as:

$$P_{ni} = \frac{\exp(\theta_n - \beta_i)}{1+\exp(\theta_n - \beta_i)}, \qquad [3.1]$$

where P_{ni} is the probability of person *n* with ability θ_n, responding correctly to item *i* with difficulty β_i.

When testlets are present and traditional or unidimensionality models are used the LID assumption is violated. According to Wilson and Wang (2005a) when LID was ignored reliability was overestimated. Therefore, interpretations made based on ability scores are influenced by model specification.

According to Wilson and Wang's (2005a) Rasch testlet model, the probability of a correct response on a dichotomously scored item can be defined as:

$$P_{ni} = \frac{\exp(\theta_n - \beta_i + \gamma_{nd(i)})}{1+\exp(\theta_n - \beta_i + \gamma_{nd(i)})} \quad [3.2]$$

where P_{ni} is the probability of person *n* with ability θ_n, responding correctly to item *i* with difficulty β_i and the testlet effect $\gamma_{nd(i)}$, is the interaction between persons and items.

Many large-scale assessment programs still use unidimensional models even when testlets are present (Dickenson, 2005). As a result this study used Monte Carlo methods to investigate differences in the decisions made about people based on model specification, and the influences of various conditions (i.e. number of items, number of testlets, testlet effects, item to testlet ratio, and sample size).

Monte Carlo Methods

Monte Carlo (MC) simulation studies are appropriate techniques for evaluating statistical estimators and procedures under varying conditions that may include sample size, normality, data of varying metric, model complexity, and model misspecification (Boomsma, 2013; Paxton, Curran, Bollen, Kirby, & Chen, 2001). According to Harwell,

Stone, Hsu, and Kirisci (1996) MC simulation studies are designed as experiments where random numbers are generated stochastically to create sampling distribution intended to reflect conditions found in empirical studies. According to Boomsma (2013) the two main reasons for conducting a MC simulation study include: a) investigate model assumptions that are violated in applied research, and b) investigate model selection for varying empirical conditions. Many widely used statistical techniques have theoretical assumptions (e.g., Rasch models) and the validity of the results may be questionable when the assumptions are violated (e.g., LID). MC studies may also inform researchers about the seriousness of the consequences of violated assumptions (Fan, 2012).

MC studies allow the researcher to artificially produce sampling distributions of parameter estimates in order to study the finite sampling performance of parameter estimates (Fan, 2012). This is done by first creating a model where the population parameters are defined by the researcher and are therefore known. Then, repeated samples are drawn from the population and parameters of interest are estimated for each sample. Next, all parameter estimates are collected into a sampling distribution and the properties of that distribution are reported.

In order to control for internal and external validity the selection of independent and dependent variables, the number of replications, and the measurement models selected should maximize generalizability and replicability. There are multiple steps in conducting a MC study including: a) developing the research question, b) selecting variables, c) selecting the appropriate experimental design, d) selecting the number of replications, e) generating item responses from random numbers, f) estimating model

parameters, and g) analyzing, summarizing, and reporting the results (Paxton, et al., 2001). These steps are followed for the presentation of the study methods.

Developing the Research Question

The research question to be answered must be informed by the literature. The hypothesis being tested is the operationalization of the problem or question, and the effects being measured must be sensitive to the conditions or variables being manipulated. Paxton et al (2001) and Boomsma (2013) recommend developing a research question from theory and previous research. MC research questions must be grounded in statistical theory and should be questions that cannot be answered by empirical studies (Harwell, et al., 1996). Some examples of such studies include: a) determination of sampling distributions, b) comparison of algorithms, and c) comparison of models.

Selecting Variables

Empirical conditions. A review of applied studies utilizing testlets in high stake assessments was conducted in order to inform the conditions created for this study by searching the Academic Search Complete, Elton B Stephens Company (EBSCO), Education Full Text, and Education Resources Information Center (ERIC) databases. The keywords used to find relevant studies included: testlets, item bundles, super items, linked items, locally dependent items, cases, scenarios, paragraphs, passages, reading, certification, multiple choice items based on paragraphs or cases, and Rasch. Of the studies identified those selected for review were from peer-reviewed education, measurement, and assessment journals between the years of 2003-2014.The search was limited to peer-reviewed, full text articles as well as dissertations. The large database

search does not guarantee all relevant studies were included, and unpublished technical reports and masters theses may have been missed or are underrepresented.

Based on the inclusion criteria of K-12 high stake testing situations, seventeen studies were selected that described a testlet structure. The number of items reported in the selected studies ranged from 15 to 136 items with a median of 28 and a mean of 39 (see Table 3.1). The number of testlets reported ranged from one to 14 with a median of four and mean of 5.4. The number of items per testlet reported ranged from 1 to 47 with a median of 8 and a mean of 12.5. The testlet variance (i.e., testlet effect) ranged from zero to 1.2 with a median of 0.30 and a mean of 0.44. The number of examinees in K-12 papers ranged from 100 to 28,593, with a median of 1,024, and a mean of 3,266. The five number summaries of the selected conditions from the selected applied studies are presented in Table 3.1.

Table 3.1

Summary of K-12 Exam Conditions in 17 Applied Papers from 2003- 2013

Statistic	No. of items	No. of testlets	Item-to-Testlet Ratio	Testlet variance	Sample size	Difficulty Range	Ability range
Minimum	15	1	1:1	0.0	100	-1.5 to 1	0 to 1
Q1	24.5	3	5:1	0.16	414	-2 to 1.5	-2 to 2
Median	28	4	8:1	0.30	1024	-2 to 2	-3 to 3
Mean	39	5.4	12.5:1	0.44	3266	-3 to 3	-3 to 3
Q3	42	7.5	14.5:1	0.7	2859	-3 to 3	-5 to 5
Maximum	136	14	47:1	1.2	28593	-4 to 4	-6 to 6

The five number summaries from Table 3.1 can be used to judge which conditions are commonly encountered in applied research where testlets are included in K-12 exams. In the current study, this information was used to create simulation conditions that

represent empirical research conditions in order to investigate the effects of model specification on high stake decisions. This strategy can provide useful information for applied researchers when selecting item response models for situations where items violate local independence.

*Population models***.** Independent and dependent variables should be selected based on the research question and the relationship between the variables and interpretability of the results (Boomsma, 2013; Paxton, et al., 2001). Independent variables included number of testlets, number of items in testlets, testlet effect, percent of total items that are in testlets, and sample size.

Dependent variables should be able to measure the effect of manipulated independent variables and in a form that simplifies the results of inferential analysis. Examples of dependent variables include root mean square error (RMSE) to measure successful parameter recovery, correlations between estimated and true parameters to measure estimation procedures, or mastery classification estimates to measure differences between measurement models (Harwell, et al., 1996; Zhang, 2010). The dependent variable in this study was decision consistency (DC) defined as the proportion of pass/fail decisions that were the same for the DRM and the TRM. The two most important variables in a simulation are sample size and number of replications (Boomsma, 2013; Fan, 2012). Other considerations in selecting variables and conditions in an MC study include: distribution of variables, estimation methods, level of each variable, and model misspecification as a condition.

MC studies are especially appropriate methods for investigating model misspecification because they can answer questions about the effect of misspecification

on parameter estimates. This is possible because the true population model is known, samples are drawn from that population and models are estimated from those samples, then models that are the same and different from the populations are estimated. Model comparisons are specified by adding or deleting paths between variables and misspecification can range from trivial to severe (Boomsma, 2013, Paxton et al., 2001).

The population structure (variables) used in the current study is a testlet model with known number of items, testlets, items within testlets, testlet effects, and the proportion of total items in testlets. Then the DRM and the RTM were fit to the data to compare the pass/fail decision based on ability estimates of the two models.

Selecting the Appropriate Experimental Design

For this Monte Carlo study, the design factors and conditions that were varied are: number of testlets, number of items in testlets, testlet effect, percent of total items that are in testlets, and sample size, yielding 72 possible conditions (3 numbers of items × 3 number of items in testlets × 2 testlet variances × 2 sample sizes × 2 proportion of items in testlet = 72). For each condition, 100 replications were generated, which is consistent with Wilson and Wang (2005a).

Number of testlets. In order to represent conditions most frequently found in applied research, the quartiles were used to inform the levels of each condition (see Table 3.1). Therefore, the number of items per testlet included four, eight, and 12.

Number of items per testlet. The number of items per testlet was selected based on the applied papers that described testlets as items related to the same stimuli (e.g.,

paragraph or passage). Therefore, the number of items per testlet included three, four, and seven.

Testlet effect. Testlet effects are the amount of local dependence between items that is unaccounted for by the dominant latent trait. The Rasch testlet model described by Wilson and Wang (2005) includes a testlet parameter ($\gamma_{d(i)j}$) which is a random effect capturing the interaction between person j with testlet d(i) when the dominant latent trait is held constant. Thus, the testlet effects are simulated from a normal distribution with a mean of zero and a standard deviation of the square root of the testlet variance, where $\gamma_{d(i)j} \sim N\left(0, \sigma^2_{\gamma d(i)}\right)$. Testlet effects selected for the current study were based on the first and third quartiles of the reviewed applied studies (see Table 3.1). The testlet effects included 0.2 and 0.7.

Proportion of items in testlets. The number of items in testlets varied. Therefore, the proportion of testlets was directly related to the purpose of the exam. In the K-12 studies the items were distributed equally over the testlets (e.g., reading comprehension exams), or approximately 75% of the items were clustered in testlets while the remaining items were independent (i.e. math exams that may have a few testlets related to a graph or table and the rest are independent). The proportion of items in testlets selected for this study were 75% and 100%.

Sample size. The sample sizes varied in this study included 1,400 and 6,000. The sample sizes selected were informed by the five number summaries of the reviewed empirical studies reported in Table 3.1.

Selecting the number of replications. The number of replications is akin to sample size and is influenced by the purpose of the study, minimization of sampling variance of estimated parameters, and power to detect the effects of interest. A small number of replication (e.g. 10) is acceptable for studies comparing item response methodologies where empirical distributions are not of primary consideration. In these types of studies, Harwell, et al. (1996) recommends a minimum of 25 replications.

The number of replications selected should minimize sampling variance (Bandalos & Leite, 2013). Testlet variance is the parameter of interest for this study (0.2 and 0.7) so the needed number of replications would minimize the sampling variance for the testlet variance. The mean and standard deviation for the testlet variance should be less than .05. Relative bias of the testlet variance was investigated where relative bias is defined as (Foreroet al., 2009):

$$(\bar{\gamma} - \gamma)/\gamma \qquad [3.3]$$

where $\bar{\gamma}$ is the estimated testlet variance and γ is the true testlet variance.

$$(\overline{SE_\gamma} - sd_\gamma)/sd_\gamma \qquad [3.4]$$

where $\overline{SE_\gamma}$ is the estimated standard error of the testlet and sd_γ is the true standard deviation of the testlet. Consistent with recommendations from Forero et al., (2009) RB below .1 was considered acceptable, .1 to .2 indicated substantial bias, and values above .2 were considered unacceptable.

Generation Of Item Response From Random Numbers

Ability and difficulty parameters. In order to generate item responses similar to those in applied K-12 studies, ability (θ) population parameters were generated from a normal distribution with a mean of zero and a standard deviation of one ($\theta_j \sim N(0,1)$. Item difficulties ($\boldsymbol{b_i}$) population parameters were generated from a normal distribution with a mean of zero and a standard deviation of one ($b_i \sim N(0,1)$. The standard normal distribution was selected so that items were appropriately targeted to the ability distribution and is consistent with previous research (Glas, 2012; Jiao, Wang, & He, 2013).

Statistical software. Data generation was conducted in R (version 3.1.2 R Development Core Team, 2014) and estimation was conducted using the **sirt** package (Robitzsch, 2014).

Estimating Model Parameters

Bayesian estimation. In IRT applications, the item difficulty distribution and ability are continuous values so Bayes theorem will be described in terms of a continuous density function. The continuous density function represents the likelihood of all possible outcomes and is generally a normal density function that resembles a bell-shaped curve. Bayes theorem can be expressed as (Kim & Bolt, 2007):

$$\mathrm{f}(\Omega|X) = \mathrm{f}(\mathrm{X}|\Omega) * \mathrm{f}(\Omega) / \left[\int_{\Omega} f(X|\Omega) f(\Omega) d\Omega\right] \qquad [3.5]$$

where X denotes all of the item response data and Ω represents all of the unknown item and person parameters and represents multiple events. The left side of the equation is the

joint posterior density given the data and is used in the determination of model parameter estimates. The function, $f(X|\Omega)$ express the likelihood of the item response data given the model parameters and is defined by the specific item response model selected and the assumption regarding local independence.

On the right hand side of the equation $f(\Omega)$ is the prior density of the model parameters and indicates the relative likelihood of specific parameter values prior to data collection. The denominator on the right is a normalizing constant where the joint posterior density is proportional to the product of the quantities. This is the basis for the sampling procedures in Markov chain Monte Carlo (MCMC) estimation (Kim & Bolt, 2007). The purpose of MCMC estimation is to reproduce $f(\Omega|X)$ distribution by sampling observations. Sampling enough observations provides information about the specific characteristics of the distribution that includes the mean and standard deviation that is the basis for model parameter estimates.

MCMC estimation is based on the Bayesian framework where each parameter of interest is assumed to vary by a specified probability distribution called a prior distribution. The first step in MCMC estimation is to specify priors, where the prior distribution is a pre-determined distribution based on prior knowledge of the parameter in question. This procedure allows known information about the characteristics of items and examinees to be incorporated into the estimation process (Dickenson, 2005). The second consideration in selecting a prior is the strength of the prior. This is done by increasing or decreasing the variance of the population mean. The prior distributions for the parameters in this study are

$$\theta_n \sim N(0, 10{,}000) \qquad [3.6]$$

$$\beta_i \sim N(0, 10{,}000) \quad [3.7]$$

where ability and difficulty are drawn from a normal distribution with a mean of zero and a variance of 10,000 , where the large variance is specified such that the priors do not influence the posterior distribution (Robitzsch, 2014**)**. The priors selected in this study are intended to be weak or less influential.

The second step in MCMC estimation is the selection of sampling procedures (Kim & Bolt, 2007). In IRT models Gibbs sampling has been used for normal ogive models but requires a data augmentation process. Gibbs sampling is a straightforward sampling procedure because of the use of known conditional distributions for sampling. Thus, the estimated model is

$$P(X_{ni} = 1) = \Phi\left(\theta_n + \gamma_{n,d(i)} + \beta_i\right) \quad [3.8]$$

where Φ is the cumulative normal distribution (ogive), and the distributional assumptions are

$$\theta_n \sim N(0,1) \quad [3.9]$$

$$\beta_i \sim N(\mu_\beta, \sigma_\beta^2) \quad [3.10]$$

$$\gamma_{n,d(i)} \sim N(0, \sigma_t^2) \quad [3.11]$$

The estimated model is a normal ogive model instead of a logistic model thus a transformation was applied such that the traditional DRM is approximated in the normal ogive model as:

$$\mathrm{P}(P_{ni} = 1) = (\theta_n - \beta_i) \quad [3.12]$$

and the RTM as:

$$\mathrm{P}(P_{ni} = 1) = \theta_n - \beta_i + \gamma_{n,d(i)} \quad [3.13]$$

The third step in MCMC estimation is to monitor the Markov chain for convergence. When the Markov chain has converged priors are updated or combined with the sample data to create a new distribution called the posterior distribution. This is done iteratively in order to produce a chain that is stationary but retains random fluctuation (Gill, 2014). Through the selected sampling procedure, random observations are repeatedly sampled iteratively in order to produce a set of observations that represent states in the Markov chain. The first set of values, the starting state, must be specified either as random numbers or specific point estimates (Gill, 2014). The sequence of values in the chain are not independent because new states are partially defined by previous states. This creates positive correlations between sampled states in the chain, making it necessary to discard a large number of initial states called the burn-in states, and estimate the posterior distribution from the remaining observations. According to Kim and Bolt (2007) this requires a large number of iterations for model parameters to be reliably estimated because it is necessary to discard at least the first 500 states as burn-in iterations.

There is no standard for the number of iterations as each study is different. The default values in the sirt package (Robitzsch, 2014) for burn-in are 500, and 1,000 for the posterior distribution. However, Patz and Junker (1999) used 1,000 burn-in and 7,400 for the posterior distribution. Wainer, Bradlow, and Wang (1999) reported convergence after the 1,000th iterations. Therefore they used 1,000 iterations for the burn-in and the remaining 500 or 1,000 for the posterior distribution. Jiao et al. (2013) found convergence of non-testlet models after 1,000 iterations and with a testlet model after 2,000 iterations. Glas (2012) used 4,000 iterations, with the first 1,000 as burn-in

iterations. Similar to the studies reported here the burn-in phase will start with 1,000 iterations and be adjusted as need for the chain to converge. The posterior distribution iterations ranged greatly in the studies presented. The burn-in phase for this study was 1,000 iterations, and the posterior distribution was 2,000 iterations beyond convergence.

Analyzing, Summarizing, and Reporting Results

Numerous studies have compared unidimensional and testlet model parameters ordering via correlation, RMSE and bias and found that ability ordering is quite similar between models (Eckes, 2013; Jiao et al., 2013; Paek, Yon, Wilson, & Kang, 2008; Wang & Wilson, 2005a; Wang & Wilson, 2005b). However, proficiency classification decisions have not yet been studied. One might reasonably expect that people on either extreme of the ability distribution to be fairly easily classified. That is, those who score low on the assessment will have a lower ability estimate regardless of the model used. Conversely, those who score high in the assessment will have a higher ability estimate regardless of the model fit to the data. However, those who are close to the cut score may have more variability depending on what model is used to estimate ability. In other words, someone may pass if a TRM model is used and fail if a DRM model is used.

The decision made about people around the cut score is of interest. In order to select a cut score that is representative of applied settings the cut score was set based on the average passage rate of an operational reading assessment in a large southern state. Review of the state's technical manual revealed that on average 75% of students passed the previous years' reading assessment, across tested grade levels. The ability distribution from the reading assessment is assumed to follow a standard normal

distribution; therefore, the cut score was set to the normal quantile (i.e., *z*-score) above which approximately 75% of the distribution falls (z = -0.67, see equation 3.14 below).

$$z = \Phi^{-1}(.75) \approx -0.67, \quad [3.14]$$

where Φ^{-1} is the inverse of the cumulative standard normal distribution function. An ability estimate from the testlet model above -0.67 is coded as "pass" and below -0.67 is coded as "fail". DC was established with the following procedures:

1. Data was generated from a testlet model
2. A DRM was fit to the data and the pass/fail status for each examinee was recorded by comparing the estimated ability to the predetermined cut score
3. A RTM was fit to the data and the pass/fail status for each examinee was recorded by comparing the estimated ability to the predetermined cut score
4. Compared the pass/fail decisions from step two and three. When the decisions from both estimated models were the same the DC was coded as one or consistent. When the DC was discrepant DC was coded as 0.

The estimated DRM and the estimated TRM were compared to each other in order to reflect applied situations where the truth is not known. When the unidimensional model produces the same decision as the testlet model that is a correct decision. When the unidimensional model contradicts the testlet model an incorrect decision is produced.

A false positive would indicate an incorrect decision of passing a student who may benefit from additional instruction or does not possess adequate knowledge. A false negative would indicate an incorrect decision of failing a student who should be classified as proficient. A true positive indicates that the correct decision to pass an individual and a true negative would indicate the correct decision to fail an individual. The true decision is indicated by the ability estimates in relation to the cut score for the testlet model, because the data were generated from testlet models.

The models investigated in this study will be evaluated based on the consistency of decisions (proportion of decisions that are the same for the testlet and Rasch model)

made using the ability estimates produced from the true testlet model and the unidimensional model. The frequency or proportion of estimated and true pass/fail decisions will be presented to determine how many incorrect decisions would have been made based on model selection and not person ability (see Table 3.2).

Table 3.2

*Proportion of Pass/Fail Decision Between the True Model and the Estimated Mode*l

		DRM	
		Pass=1	Fail=0
RTM	Pass=1	π_{pp}	π_{pf}
	Fail=0	π_{fp}	π_{ff}

Estimating design factor effects. Harwell, et al. (1996) stated that the characteristics of the independent variables indicate the proper experimental design and must be explicitly stated. The researcher should state whether the study is balanced or unbalanced, fully crossed, and describe between and within subject factors. Designing specific experimental conditions must include a reasonable rational for why each design factor is included in the study (Fan, 2012). The rational should be based on previous findings, a belief that the factor could affect the outcome and a theoretical consideration based on relevant literature (Fan, 2012). For making causal inferences or investigating the effect of explanatory variables the experimental design factors must have multiple levels. According to Boomsma (2013) the response variable, and the explanatory variables create a full factorial design where the contents of each cell is determined by the number of replications and the response variable. The product of the number of levels of each design factor times the number of replications equals the total number of simulations to be analyzed where the factors are fully crossed.

The use of descriptive statistics and graphs are the predominant means of analyzing and reporting MC studies. However, in order to detect the effects of interest and the magnitude of those effects inferential analysis is needed. Most of inferential analyses of MC studies are analyses of variance (ANOVA) or regression based analyses. Summarizing results of a MC study begins with a statistical technique that is applied to the data sample for analysis, and the statistic of interest is computed. Second, each statistic for each sample is obtained and accumulated across samples under all design factors (Fan, 2012). Third, descriptive statistics (e.g., mean, variance, mean relative bias, mean square error, and correlation or covariance) are included to summarize the findings (Fan, 2012). Fourth, graphical displays (e.g., figures, boxplots, scatterplots, or power curves) of observable trends are presented (Fan, 2012). Inferential statistics are then provided to evaluate design factors main effects and interactions, estimate effect sizes, and to investigate explanatory variables x that account for variability in Y using ANOVA (Boomsma, 2013). A performance criterion is often reported as an effect size, where Bandalos (2002, p. 88) used partial eta squared, and Gerbing (1984, p.163) used omega squared.

In order to investigate the effects of each condition on decision consistency, factorial ANOVA was conducted using each design factor as an effect in the ANOVA model and the correct classification proportion of each replication as the outcome measure. All two-way interactions as well as main effects will be considered in determining the effect each design factor has on decision consistency.

Eta-squared (η^2) and Partial eta- squared (η_p^2) were computed for all examined effects and presented in descending ordered to determine which design factor(s) most strongly

influence decision consistency. Eta-squared is defined as the proportion of variance accounted for an effect and can be expressed as:

$$\eta^2 = \frac{SS_{effect}}{SS_{Total}} \quad [3.15]$$

where SS_{effect} is the sum of squares attributed to an effect and SS_{Total} is the total sum of squares for in the model (Tabachnick & Fidell, 2001). Partial- eta squared is defined as the proportion of total variation attributed to the factor after partilalling out the other factors from the total variation:

$$\eta_p^2 = \frac{SS_{effect}}{(SS_{effect} + SS_{Error})} \quad [3.16]$$

where SS_{effect} is the sum of squares attributed to an effect and SS_{Error} is the error sum of squares for in the model (Pierce, Block, & Aguinis, 2004).

An Empirical Example

The 2012 PISA measures the performance of fifteen-year-old students in mathematics, science, and reading from 65 countries (Kastberg, Roey, Lemanski, Chan, Murray, 2014). For each country the population was described as fifteen-year-old students in the 7th grade or higher from education systems that had a minimum of 4,500 students in 150 schools. A stratified two-stage sampling procedure was used to sample schools and 35 students within each school.

The PISA assessment included 222 items with 85 mathematics, 44 reading, 53 science, and 40 financial literacy items grouped into 17 test booklets. Each booklet made up of four clusters: seven mathematics item clusters (M1-M7), three reading item clusters (R1-R3), three science item clusters (S1-S3), and two financial literacy item clusters. Items within a cluster were in a fixed order. A rotated block design was used to

administer the test and each student was randomly assigned to one of the thirteen forms. The subset of reading items within clusters included anchor items for equating purposes. For more information regarding the design, measures, and procedures interested readers should consult the 2012 Technical Report (Kastberget al., 2014).

The PISA (2012) U.S. public-use data files from Institute of Education Sciences (IES) contains a subset of students' mathematics, reading, and science responses (http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2014028). For the purpose of this study only the reading portion of booklet two was used to study the testlet effect on decision consistency between the DRM model and the RTM models (see Appendix A for a description of the items). Booklet two contains 29 reading items with nine testlets. The items per testlet range from two to four (see Figure 3.2). This sub sample consists of 422 fifteen-years-old students in the U.S.

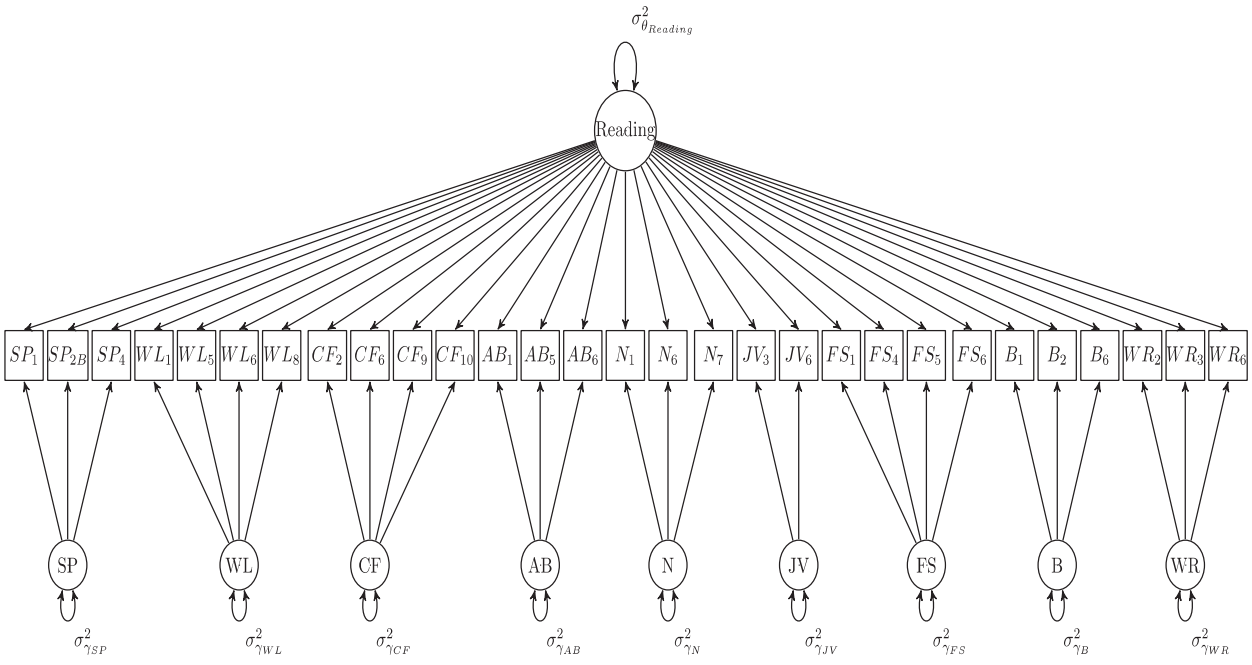

Figure 3.1: Conceptual model for PISA with error terms excluded

PISA is traditionally calibrated with a Rasch model, thus making this data set ideal for this study. Two models will be estimated: a) the DRM and b) the RTM. For each model the pass/fail decisions of each examinee will be recorded and then the proportion of examinees whose pass/fail decisions differed between the two models will be reported. The cut score for the PISA data was set based on the percentage of examinees that scored at a level two or higher on the 2012 exam (OECD, 2013) because scores below a level two indicates a struggling reader. According to the OECD (2013), 81.2 % of students scored at a level two or higher. The ability distribution is assumed to follow a standard normal distribution, therefore, the cut score is set to the value above which approximately 81.2% of the distribution falls (z = -0.89, see equation 3.17 below).

$$z = \Phi^{-1}(.812) \approx -0.89, \qquad [3.17]$$

where Φ^{-1} is the inverse of the cumulative standard normal distribution function.

Summary

The purpose of this simulation study was to compare the consistency of pass/fail decision made based on competing models. The factors included in this study were the number of items, number of testlets, proportion of items in testlets, testlet variance, and sample size, yielding 72 conditions. Each condition was replicated 100 times. The proportion of estimated and true pass/fail decisions was presented to determine how many incorrect decisions were made based on model selection and not person ability. An empirical example was also included in this study.

The contribution this study made to the literature on testlets and violation of the LID assumption is two fold. First, in previous model comparison studies ability estimates were compared based on bias estimates or rank order correlations. This study sought to

compare actual decisions made about examinees based on a predetermined cut score. Second, in previous studies, traditional measurement models were estimated in one analytic program and estimated with a likelihood algorithm, while the testlet models were typically estimated in another analytic program with MCMC. The Rasch models in this study were generated and estimated with the sirt package (Robitzsch, 2014) with Markov Chain Monte Carlo methods (MCMC) (Glas, 2012) estimation procedure. In doing so the models were directly compared and alternative explanations were eliminated. Based on the findings, some recommendations were provided in chapter 5 regarding the use of testlet models for estimating testlet items under conditions that are likely to be encountered in empirical research.

CHAPTER FOUR

Results

Evaluation of Simulation Data

Data were generated using R (version 3.1.2 R Development Core Team, 2014) and models were estimated using the **sirt** package using Markov Chain Monte Carlo methods (MCMC) (Glas, 2012; Robitzsch, 2014). For each condition, 100 replications were generated, which is consistent with Wilson and Wang (2005a). The number of replications was selected in order to minimize sampling variance (Bandalos & Leite, 2013). The testlet variance was the parameter of interest for this study so number of replications needed should adequately minimize the sampling error for the testlet variance. The specified testlet variances are 0.2 and 0.7. The mean and empirical standard error for the testlet variance in the smallest condition (four testlets, three items per testlet, sample size of 1,400, and all items in a testlet) and the largest condition (12 testlets, seven items per testlet, sample size of 6,000, and all items in a testlet) were M= 0.19, SE= 0.006 and M= 0.697, SE=0.0002, respectively. These values indicate 100 replications were sufficient because the standard errors were very small.

In Monte Carlo studies, data generation begins with population parameters as starting values where user-defined distributions or transformations are specified. Next, it is necessary to check the validity of data generation by generating sample data and comparing characteristics to target population parameters. Next, one must examine the empirical sampling distribution characteristics of the simulated data sample relative to the

population parameters (Paxton, et al. 2001). All simulated datasets were examined using R (version 3.1.2 R Development Core Team, 2014) to evaluate whether the simulated data sets were structured according to population specifications. The structure of the sample matched the structure specified by the population model within three-hundredths or less of the true parameter, on average, for the vast majority of simulated conditions. When the testlet variance was set at 0.2, the largest difference between the true and estimated testlet variance was 0.0236 (0.2- 0.1764= 0.0236). When the testlet variance was set at 0.7, the largest difference between the true and estimated testlet variance was 0.0041 (0.7- 0.6959= 0.0041).

Similar to Forero et al (2009), relative bias (RB) was computed as:

$$RB = {(\bar{\gamma} - \gamma)}/{\gamma} \quad [4.1]$$

where $\bar{\gamma}$ is the estimated testlet variance and γ is the true testlet variance.

$$RB = {(\overline{SE_\gamma} - sd_\gamma)}/{sd_\gamma} \quad [4.2]$$

where $\overline{SE_\gamma}$ is the estimated standard error of the testlet and sd_γ is the true standard deviation of the testlet. RB values below .10 were considered acceptable, .1 to .2 indicated substantial bias, and values above .2 were considered unacceptable. Flora and Curran (2004) recommended slightly more conservative values, where values less than .05 indicated trivial bias, .05 to .1 indicated moderate bias, and values above .1 indicated substantial bias. For the smallest condition, in which there were three items per testlet, a total of twelve items and testlet variance was 0.2 the largest RB was -.118. The RB for this testlet variance was above the upper bound for acceptable estimation. The RB for

another testlet in the same condition was -.07. that was slightly above the upper bound of trivial bias (Flora & Curran, 2004). The condition discussed here was selected because all other RB values were well below the .05 cut off. These RB estimates were also an indication that 100 replications was sufficient overall. A summary of the specified parameters, mean parameter estimates, and RB estimates for the smallest and the largest condition are provided in Table 4.1.

Table 4.1

Sample Testlet Variance for Data Generated from the Testlet Population with the Smallest and Largest Number of Testlets and Testlet Variance of 0.2 and 0.7

	Testlet Variance of 0.2				Testlet Variance of 0.7			
Testlet	True	Estimated	SD	Relative Bias	True	Estimated	SD	Relative Bias
1	0.2	.1869	.0054	-.0655	0.7	.6964	.0002	-.0051
2	0.2	.1764	.0065	-.1180	0.7	.7008	.0002	.0011
3	0.2	.1903	.0068	-.0485	0.7	.6978	.0002	-.0031
4	0.0	.0336	.0049	.0336	0.7	.7037	.0003	.0053
5					0.7	.7019	.0003	.0027
6					0.7	.6994	.0003	-.0009
7					0.7	.6982	.0003	-.0026
8					0.7	.6959	.0003	-.0059
9					0.7	.7008	.0003	.0011
10					0.7	.7006	.0002	.0009
11					0.7	.7021	.0003	.0030
12					0.7	.7032	.0003	.0050

Note. The smallest testlet condition includes 3 testlets with a variance of 0.2 and one testlet with variance of 0.0 (75% of total items in testlets).

Model Estimation and Diagnostics

Model estimation methods may not necessarily be dictated by a research question and should be evaluated for appropriateness. Furthermore, Harwell, et al., (1996) recommended using a well-established commercial program. Two considerations for selecting an estimation method include the handling of starting values and convergence.

The default starting values in most commercial programs are generally adequate (Harwell et al., 1996) and were used in this study. When models fail to converge, another estimation method may be considered, provided it is appropriate for distributions of the observed data. Harwell, et al. (1996) stated that studies with small samples or complex models might exhibit increased problems with default starting values and/or non-convergence. In such cases, the authors recommend the use of Bayesian methods because correct specification of prior distribution and accompanying parameter estimates may mitigate these problems. Given the model complexity of models examined in this study, Bayesian estimation via Markov Chain Monte Carlo (MCMC) methods were used to estimate the model parameters.

MCMC estimation comes with its own considerations for use. One consideration is the number of burn in states to discard based on the autocorrelations in the chain. If autocorrelations remain present following the burn-in phase, then a second consideration is whether or not to thin the chain. Gill (2015) recommended extending rather than thinning the chain because thinning may cause variance estimates to be higher and does not improve the mixing properties of the chain. Following the recommendation of Gill (2015), the chains were extended rather than thinned in this study. Examples of autocorrelation plots for a Markov chain for condition 36 (seven items in 12 testlets and testlet variance was 0.7, and sample size was 6,000 and all items were in testlets) are presented in Figure 4.1 (a) and (b). The autocorrelations using the default values (burn in= 500 and iterations=1,000) in the sirt package (Glas, 2012; Robitzsch, 2014) are presented in Figure 4.1a. The autocorrelations after the chain was extended (burn in=

1,000 and iterations=2,000) are shown in Figure 4.1b. The autocorrelations in Figure 4.1a are higher, but the autocorrelations decrease after extending the chain (see Figure 4.1b).

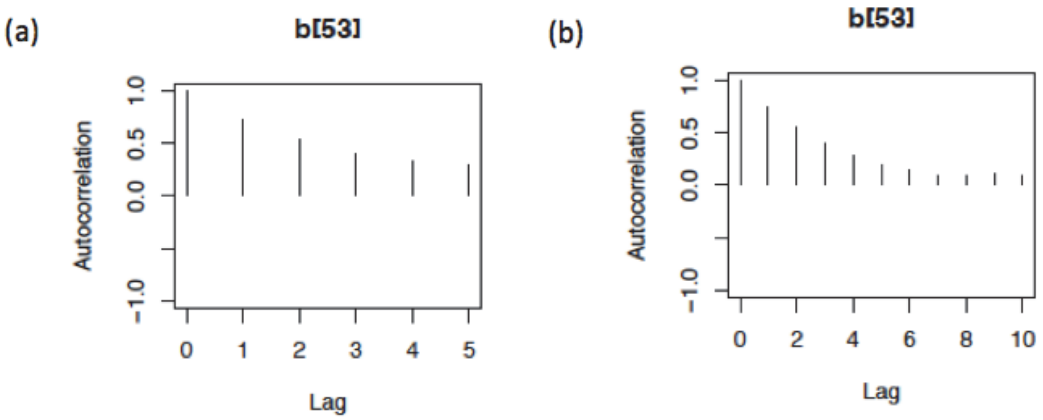

Figure 4.1: Autocorrelation plots of the difficulty of item 53

The Markov chain must also be evaluated for chain convergence. This can be done graphically by plotting the sampling history of the chain (see Figure 4.2). Two different trace plots of Markov chains are provided in Figure 4.2 (a) and (b). Again, the traces presented in Figure 4.2 differ based on the number of iterations. The trace presented in Figure 4.2a is an example a Markov chain for condition 36, which is the condition discussed above, using the default values (burn in= 500 and iterations=1,000). There are sections of the trace in Figure 4.2a where the chain appears to have consecutive high or low values, which may indicate a problem with convergence (Kim & Bolt, 2009). After extending the chain, the trace shown in Figure 4.2b appears to bounce up and down from one iteration to the next. That trace is more indicative of a chain that has converged. In addition, the marginal densities shown in the right-hand column of Figure 4.2 are indicative of improved chain mixing with an extended chain. That is, the marginal

density with an increased number of iterations is smoother compared to the marginal density with fewer iterations.

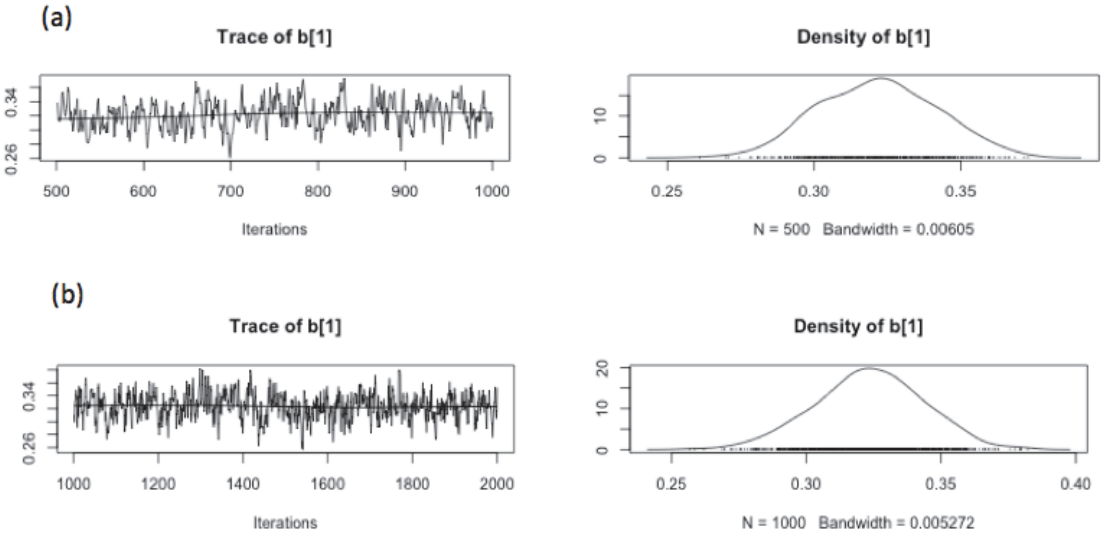

Figure 4.2: Example of sampling histories trace plots and marginal density curves associated with the Markov chains displaying evidence of convergence (a) and nonconvergence (b).

Another approach for examining convergence is Geweke's criterion, which involves computing a Z-score for each parameter (Gill, 2015). The Z-scores were computed by taking the difference between the mean of the first 10% of the states and the mean of the last 50% of the states, and then dividing by their pooled standard deviation (Kim & Bolt, 2009). Z-scores within -1.96 to 1.96 are not statistically different from zero and indicate chain convergence. It can also be plotted to provide a visual investigation for Z-scores outside this range. Examples of two different Geweke criterion plots of z-scores of Markov chains are presented Figure 4.3 (a) and (b). In Figure 4.3a, the Z-scores occasionally fell outside of the ± 1.96 boundaries, which indicates that convergence may be problematic. After extending the chain (see Figure 4.3b), there is evidence that the

chain converged because all Z-scores are within -1.96 to 1.96 standard deviations (Kim & Bolt, 2009).

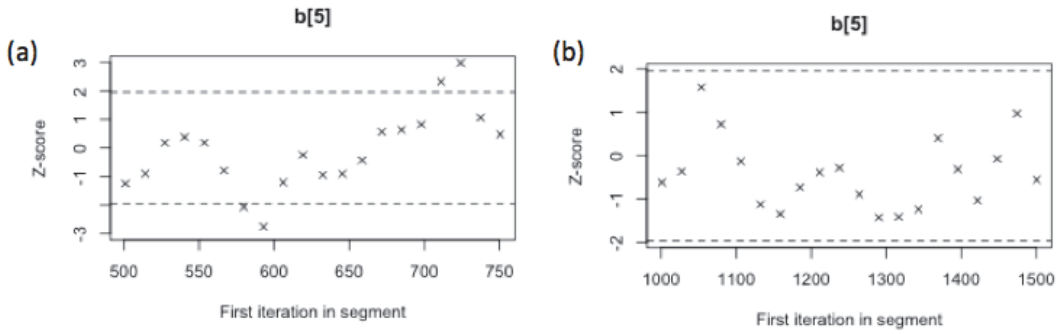

Figure 4.3: Geweke plots for the difficulty of item five

The Raferty and Lewis criterion index can also be used to help researcher diagnose a possible problem with convergence. Raferty and Lewis diagnostics provide conservative recommendations for the number of sampled states needed to reach convergence accounting for the autocorrelation in the chain (Gill, 2014; Kim & Bolt, 2009). The Raferty Lewis criterion index provided evidence that a burn-in of 1,000 and iteration of 2,000 were well beyond the minimum requirements for convergence, with an *a priori* posterior quantile of 0.025, acceptable tolerance (r) is 0.025, and probability of being within that tolerance (s) is 0.95 (Kim & Bolt, 2009).

The dependence factor, *I*, is the proportional increases in the number of iterations attributed to serial dependence. Values of *I* greater than five may indicate influential starting values, high correlations between coefficients, or poor mixing (Gill, 2014). In condition 36 when the default iterations were used, *I* ranged from one to 5.49. However, the dependence factor for condition 36 with increased iterations ranged from 1.13 to 4.69, indicating that the iterations used in this study were acceptable.

A fourth chain convergence check is the Heidelberg and Welch diagnostic that was based on the Cramer-von Mises test statistic that tests that null hypothesis that the Markov chain is different from a stationary distribution (Gill, 2014). This is a two-step procedure in which the first step involves calculating the test statistic on the entire chain. Second, the test statistic is repeated after discarding 10% of the chain after each rejection of the null hypothesis up to 50% of the chain. If the null hypothesis is still rejected then the chain needs to be extended. The second step tests the remainder of the chain after 50% was discarded, and this test is called the halfwidth test. The null hypothesis of the Heidelberg and Welch test is that the chain has reached a stationary chain where the test starts with the complete set of iterations produced from the running sampler from the first iteration. The number of iterations (N), accuracy ($\in$), and alpha (α) are established *a priori*. If the mean divided by the halfwidth is lower than $\in$, then the halfwidth test is passed, where accuracy was set at 0.1 ($\in$), and the Type I error rate (α) was set to .05. Condition 36 with default iterations and those used in this study passed the first step and only three out of 98 estimates did not pass the halfwidth test. This indicates that the burn-in and iterations selected for this study produced stationary chains with 95% confidence.

The last consideration is how many sampled states are necessary to reduce the Monte Carlo standard errors due to the posterior distribution being constructed from samples. This means that when moments, such as a mean, are computed from the posterior distribution error in the estimate can be attributed to standard error of the point estimate and sampling. This type of error is referred to as Monte Carlo error. According to Kim and Bolt (2007), Monte Carlo error for each parameter of interest should be less than 5% of the sample standard deviation. The burn-in phase was 1,000 iterations and the

posterior distribution was an additional 2,000 iterations beyond convergence. According to Robitzsch (2014), the percent of the standard error ratio (PSER) indicates the proportion of the Monte Carlo standard error in relation to the total standard deviation of the posterior distribution for each parameter estimate (see Table 4.2). The PSER ranged from 4.3 to 10.5 for the DRM and 4.3 to 13 for the RTM. The majority of the PSER were within acceptable limits recommended by Kim and Bolt (2007).

Table 4.2

MCMC parameter Estimates with Standard Deviation and Monte Carlo Standard Error

Parameters	Mean	SD	Minimum	1st Quartile	Median	3rd Quartile	Maximum
DRM	-0.09	0.75	-2.40	-0.54	-0.14	0.52	1.48
SD	0.02	0.00	0.02	0.02	0.02	0.02	0.04
PSER	5.79	0.89	4.30	5.30	5.60	5.90	10.50
TRM	0.02	0.95	-3.08	-0.60	0.09	0.83	1.88
SD	0.20	0.00	0.01	0.02	0.02	0.02	0.04
PSER	5.92	2.00	4.30	4.70	5.10	6.00	13.00

Note- DRM. Dbar = 482428.5, Dhat = 476600, pD = 5828.51 where pD = Dbar - Dhat, DIC = 488257 where DIC = Dhat + pD, EAP Reliablity = 0.96
Note- TRM. Dbar = 378369.6, Dhat = 332442.8, pD = 45926381 where pD = Dbar - Dhat, DIC = 424296.4 where DIC = Dhat + pD, EAP Reliablity = 0.914

All of the diagnostic tests presented were performed for all conditions. Chain mixing and convergence was slightly more problematic in condition with fewer items. Overall increasing the burn-in to 1,000 and total iterations to 2,000 improved diagnostic indicators of convergence in every condition.

Analyzing Results of Monte Carlo Studies

Monte Carlo studies are considered experiments and are randomized factorial designs and the use of descriptive statistics and graphs are the predominant means of

analyzing and reporting MC studies (Harwell, et al., 1996). The outcome of interest in this MC study was the proportion of pass/fail decisions that were the same between the DRM and the TRM. Then, DC was computed for each replication and was aggregated across replications under all design factors (Fan, 2012). Analysis of variance (ANOVA) was then conducted to evaluate main and two-way interaction effects of design factors and also to estimate effect sizes (Boomsma, 2013). The use of ANOVA implies null hypothesis significance testing but was not the primary objective of this MC study due to the large number of replications. Instead, performance criteria were established using eta-squared and partial eta-squared effect size estimates (Bandalos, 2002, p. 88). ANOVA and effect size estimates for all main and two-way interaction effects are presented in Table 4.3.

Next, descriptive statistics (i.e. mean, and standard deviation) were included to summarize each condition (Fan, 2012). Then graphical displays (i.e. boxplots) were presented of observable trends for the main and two-way interaction effects (Fan, 2012).

Table 4.3

ANOVA for All Simple and Two-Way Interactions for Each Design Factor

Conditions	Df	Sum Sq	Mean Sq	F value	Pr(>F)	η^2	η_p^2
Testlet variance	1	1.1179	1.1179	17041.109	<.001	0.6181	0.6975
Number of testlet	2	0.0560	0.0280	427.185	<.001	0.0309	0.1034
Testlet variance:Percent of total items in testlets	2	0.0531	0.0266	404.918	<.001	0.0302	0.1012
Number of testlet:Percent of total items in testlets	2	0.0391	0.0195	297.995	<.001	0.0244	0.0835
Number of testlet: Testlet variance	2	0.0089	0.0044	67.651	<.001	0.0069	0.0252
Percent of total items in testlets	1	0.0030	0.0030	45.194	<.001	0.0022	0.0083
Number of testlet:Sample size	1	0.0027	0.0027	41.211	.106	0.0015	0.0056
Number of items	2	0.0008	0.0004	6.115	.002	0.0007	0.0027
Percent of total items in testlets:Sample size	1	0.0004	0.0004	5.920	.048	0.0002	0.0008
Number of items:Number of testlet	4	0.0015	0.0004	5.833	<.001	0.0020	0.0074
Testlet variance:Sample size	1	0.0003	0.0003	3.921	.009	0.0001	0.0005
Number of items:Testlet variance	2	0.0005	0.0002	3.579	.021	0.0001	0.0004
Number of items:Sample size	2	0.0004	0.0002	2.677	.014	0.0001	0.0005
Number of items:Percent of total items in testlets	2	0.0003	0.0001	2.085	<.001	0.0003	0.0011
Sample size	1	0.0000	0.0000	0.060	.800	0.0001	0.0002
Residuals	7173	0.4705	0.0001				

Testlet Variance

Testlet variance had the strongest effect on DC ($\widehat{\eta^2} = .618, \widehat{\eta_p^2} = .698$). When testlet variance was 0.7 DC was lower, on average, with more variability across conditions than those with testlet variance of 0.2 (see Figure 4.4). The average DC between the TRM and DRM was 98.53% (SD=0.01), and 96.04% (SD=0.01) for testlet variance of 0.2 and 0.7, respectively. When testlet variance was 0.2 the number of people misclassified on average was 21 and 88 for sample sizes of 1,400 and 6,000, respectively. When testlet variance was 0.7 the number of people misclassified on average was 56 and 238 for sample sizes of 1,400 and 6,000, respectively (see Table 4.4 and 4.5).

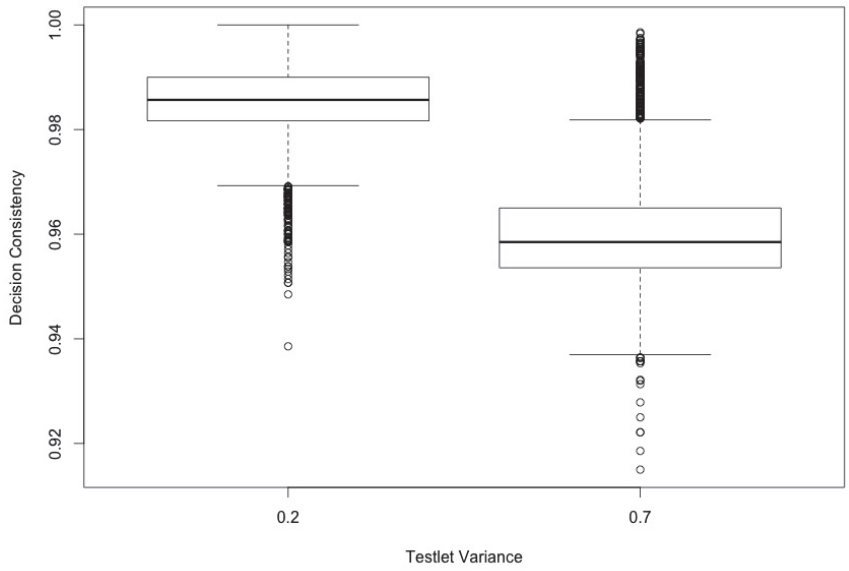

Figure: 4.4: Boxplot of testlet variance

Number of Testlets

The second largest effect on DC was the number of testlets ($\widehat{\eta^2} = .031, \widehat{\eta_p^2} = .103$). The average DC between the DRM and the TRM was 97.67% (SD=0.02), 97.2% (SD=0.01), and 97% (SD=0.02) for number of testlets of four, eight, and 12, respectively

(see Figure 4.5). When there were four testlets the number of people misclassified on average was 33 and 140 for sample sizes of 1,400 and 6,000, respectively. When there were eight testlets the number of people misclassified on average was 40 and 168 for sample sizes of 1,400 and 6,000, respectively. When there were 12 testlets the number of people misclassified on average was 42 and 180 for sample sizes of 1,400 and 6,000, respectively (see Table 4.4 and 4.5).

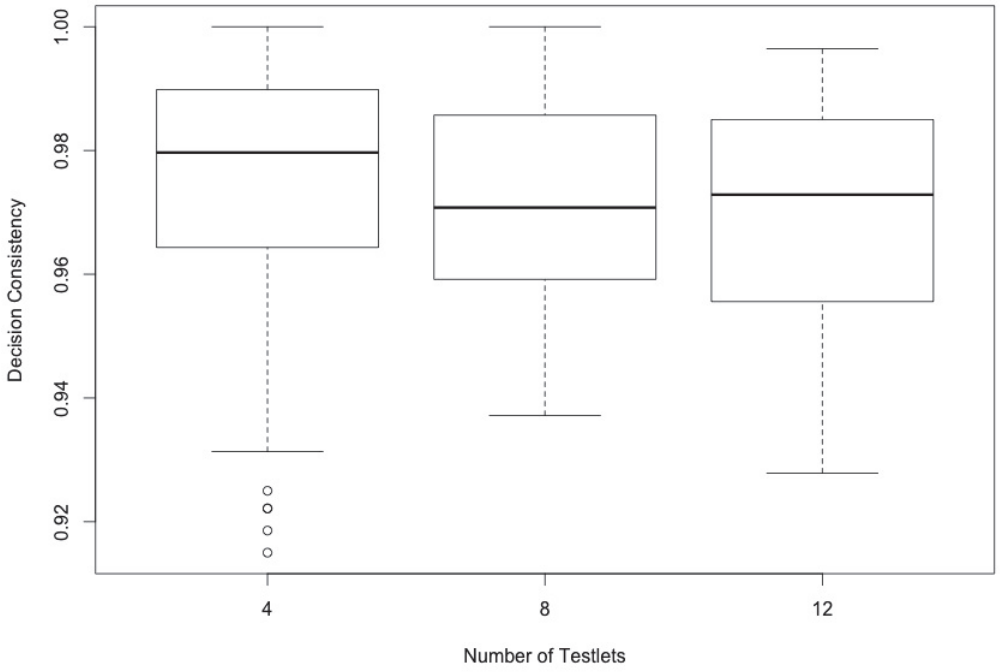

Figure: 4.5: Boxplot of number of testlets

Testlet Variance and Percent Total Items in Testlets

The third largest effect on DC was the percent of total items in testlets interaction ($\widehat{\eta^2} = .030, \widehat{\eta_p^2} = .101$). When testlet variance was 0.2, DC was 98.53% (SD=0.01) and 98.54% (SD=0.01) for percent of items in testlets of 75% and 100%, respectively. When testlet variance was 0.2 and 75% of items were in testlets the number of people

misclassified on average was 21and 88 for sample sizes of 1,400 and 6,000, respectively. When testlet variance was 0.2 and 100% of items were in testlets the number of people misclassified was 21 and 88 for sample sizes of 1,400 and 6,000, respectively (see Table 4.4 and 4.5).

When testlet variance was 0.7, DC was 95.9% (SD=0.01) and 96.19% (SD=0.01) for percent of items in testlets of 75% and 100%, respectively (see Figure 4.6). When testlet variance was 0.7 and 75% of items were in testlets, the number of people misclassified on average was 58 and 246 for sample sizes of 1,400 and 6,000, respectively. When testlet variance was 0.7 and 100% of items were in testlets the number of people misclassified on average was 54 and 229 for sample sizes of 1,400 and 6,000, respectively (see Table 4.4 and 4.5).

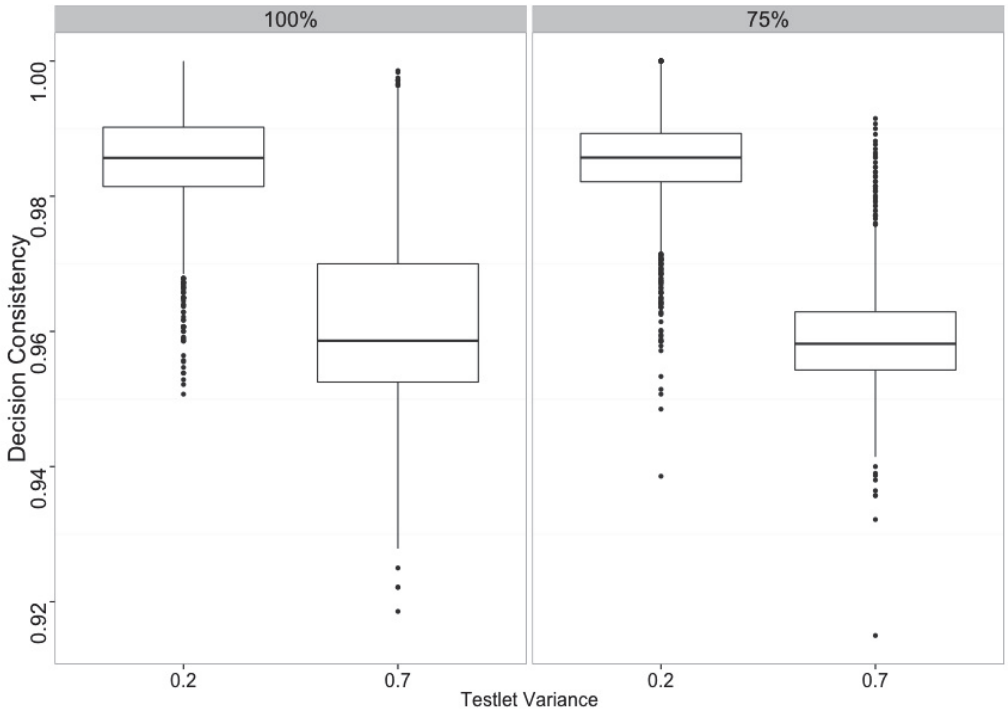

Figure: 4.6: Boxplot of testlet variance by percent of total items in testlets

Number of Testlets and Percent Total Items in Testlets

Then, the design factor with the next largest effect was the number of testlets and percent of total items in testlets interaction ($\widehat{\eta^2} = .024, \widehat{\eta_p^2} = .084$) (see Table 4.3). The average DC between the DRM and the TRM was 97.18% (SD=0.02), and 97.4% (SD=0.02) for percent of items in testlets of 75% and 100%, respectively. When there were four testlets DC was 97.12% (SD=0.02) and 98.14% (SD=0.01) when 75% and 100% of items were in a testlet. When there were four testlets and 75% of items were in testlets the number of people misclassified on average was 41 and173 for sample sizes of 1,400 and 6,000, respectively. When there were four testlets and 100% of items were in testlets the number of people misclassified on average was 26, and 112 for sample sizes of 1,400 and 6,000, respectively (see Table 4.4 and 4.5).

When there were eight testlets DC was 97.21% (SD=0.01) and 97.19% (SD=0.01) when 75% and 100% of items were in a testlet, respectively. When there were eight testlets and 75% of items were in testlets the number of people misclassified on average was 39 and 168 for sample sizes of 1,400 and 6,000, respectively. When there were eight testlets and 100% of items were in testlets the number of people misclassified on average was 40 and169 for sample sizes of 1,400 and 6,000, respectively (see Table 4.4 and 4.5).

When there were 12 testlets DC was 97.2% (SD=0.01) and 96.81% (SD=0.02) when 75% and 100% of items were in a testlet, respectively. When there were 12 testlets and 75% of items were in testlets the number of people misclassified on average was 40 and 168 for sample sizes of 1,400 and 6,000, respectively. When there were 12 testlets and 100% of items were in testlets the number of people misclassified on average was 45 and192 for sample sizes of 1,400 and 6,000, respectively (see Table 4.4 and 4.5). There

was more variability in DC when all items were in testlets compared to conditions where 75% of items were in testlets (see Figure 4.7).

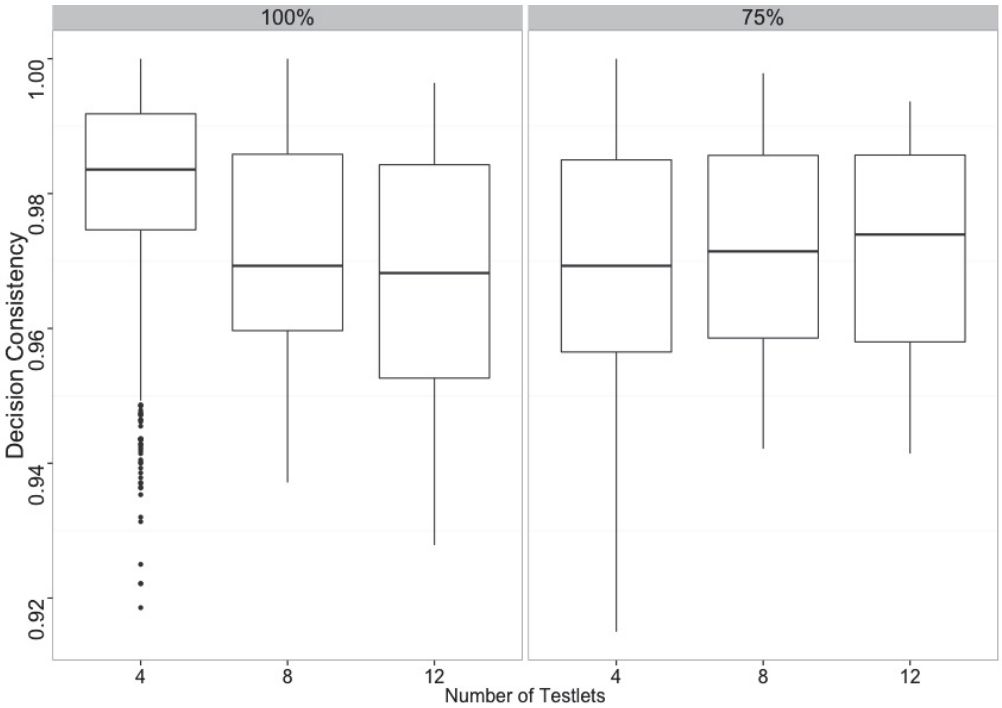

Figure: 4.7: Boxplot of Total number of items in testlets by number of testlets

Number of Testlets and Testlet Variance

The next strongest effect was the number of testlets and testlet variance interaction ($\widehat{\eta^2} = .0069, \widehat{\eta_p^2} = .025$) (see Table 4.5). As the number of testlets increased and testlet variance increased DC decreased and the number of people misclassified increased (see Figure 4.8). When there were four testlets and variance is 0.2 and 0.7 respectively, DC was 98.61% (SD=0.01) and 96.73% (SD=0.01). When there were four testlets and testlet variance is 0.2 the number of people misclassified on average was 20 and 84 for sample sizes of 1,400 and 6,000, respectively. When there were four testlets and testlet variance is 0.7 the number of people misclassified on average was 46 and 197 for sample sizes of 1,400 and 6,000, respectively (see Table 4.4 and 4.5).

When there were 8 testlets and variance was 0.2 and 0.7, respectively DC was 98.52% (SD=0.01) and 95.88% (SD=0.01), respectively. When there were eight testlets and testlet variance is 0.2 the number of people misclassified on average was 21 and 89 for sample sizes of 1,400 and 6,000, respectively. When there were eight testlets and testlet variance is 0.7 the number of people misclassified on average was 58 and 247 for sample sizes of 1,400 and 6,000, respectively (see Table 4.4 and 4.5).

When there were 12 testlets and variance was 0.2 and 0.7 DC was 98.48% (SD=0.01) and 95.55% (SD=0.01), respectively (see Table 4.4 and 4.5). When there were 12 testlets and testlet variance is 0.2 the number of people misclassified on average was 22 and 92 for sample sizes of 1,400 and 6,000, respectively. When there were 12 testlets and testlet variance is 0.7 the number of people misclassified on average was 63 and 267 for sample sizes of 1,400 and 6,000, respectively (see Table 4.4 and 4.5).

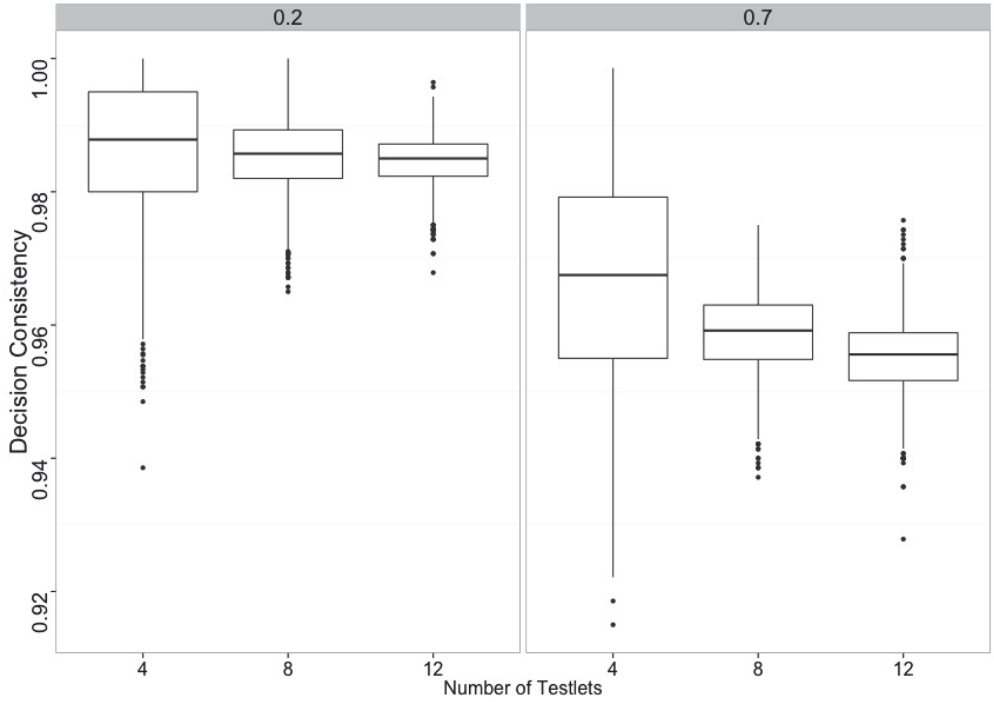

Figure: 4.8: Boxplot of number of testlets by testlet variance

Percent Total Items in Testlets

Then the percent of total items in testlets effect was next largest ($\widehat{\eta^2} = .002, \widehat{\eta_p^2} = .008$) (see Table 4.5). When the percent of total items in testlets was 100%, DC was higher, on average, with more variability across conditions than those with 75% of total items in testlets (see Figure 4.9). The average DC between the DRM and the TRM was 97.18% (SD=0.02), and 97.4% (SD=0.02) for percent of items in testlets of 75% and 100%, respectively (see figure 4.9). When 75% of items were in testlets the number of people misclassified on average was 40 and 170 for sample sizes of 1,400 and 6,000, respectively. When 100% of items were in testlets the number of people misclassified on average was 37 and156 for sample sizes of 1,400 and 6,000, respectively (see Table 4.4 and 4.5).

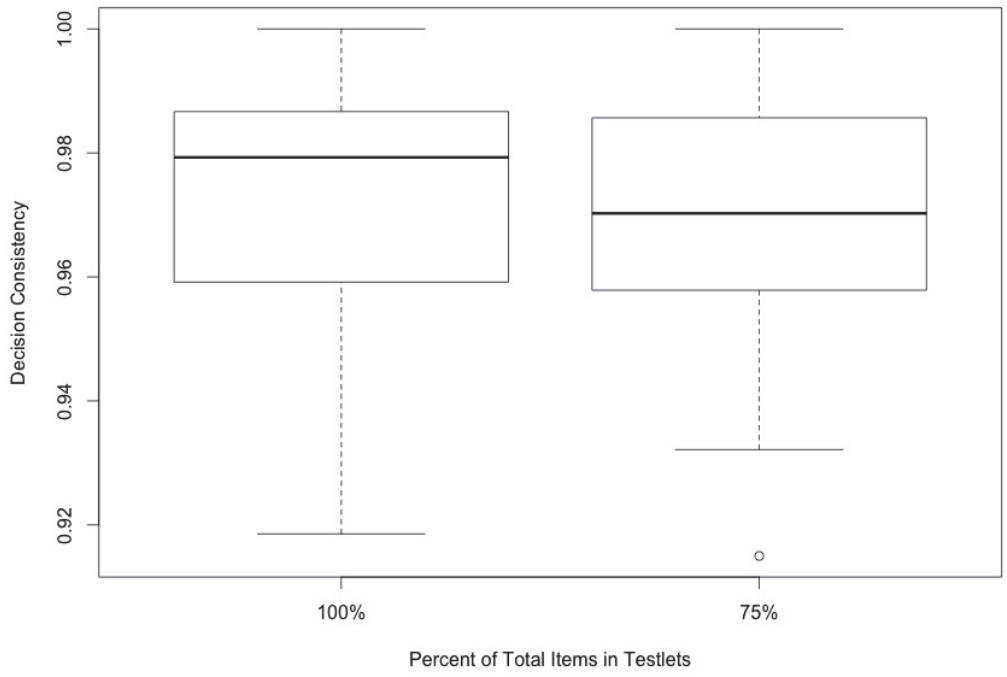

Figure: 4.9: Boxplot of the percent of total items in testlets

Number of Items and Number of Testlets

The next strongest effect was for the interaction between the number of items and number of testlets ($\widehat{\eta^2} = .002, \widehat{\eta_p^2} = .007$) (see Table 4.5). The average DC between the DRM and the TRM was 97.27% (SD=0.02), 97.34% (SD=0.02), and 97.27% (SD=0.01) for number of items in each testlet of three, four, and seven respectively.

When there were three items per testlet, DC was 97.70% (SD=0.02), 97.12% (SD=0.02), and 97% (SD=0.02), when the number of testlets were four, eight, and 12, respectively. When there were three items in each testlet and four testlets the number of people misclassified on average was 33 and 138 for sample sizes of 1,400 and 6,000, respectively. When there were three items in each testlet and eight testlets the number of people misclassified on average was 41 and173 for sample sizes of 1,400 and 6,000, respectively. When there were three items in each testlet and 12 testlets the number of people misclassified was 42 and 180 for sample sizes of 1,400 and 6,000, respectively. (see Table 4.4 and 4.5).

When there were four items per testlet DC was 97.74% (SD=0.01), 97.27% (SD=0.01) and 97.01% (SD=0.02) when the number of testlets were four, eight, and 12, respectively. When there were four items in each testlet and four testlets the number of people misclassified on average was 32 and136 for sample sizes of 1,400 and 6,000, respectively. When there were four items in each testlet and eight testlets the number of people misclassified on average was 39 and 164 for sample sizes of 1,400 and 6,000, respectively. When there were four items in each testlet and 12 testlets the number of people misclassified was 42 and180 for sample sizes of 1,400 and 6,000, respectively (see Table 4.4 and 4.5).

When there were seven items per testlet DC was 97.58% (SD=0.01), 97.21% (SD=0.02), and 97.01% (SD=0.02), when the number of testlets were four, eight, and 12, respectively (see Figure 4.10). When there were seven items in each testlet and four testlets the number of people misclassified on average was 34 and 145 for sample sizes of 1,400 and 6,000, respectively. When there were seven items in each testlet and eight testlets the number of people misclassified on average was 39 and 173, for sample sizes of 1,400 and 6,000, respectively. When there were seven items in each testlet and 12 testlets the number of people misclassified on average was 42 and 180 for sample sizes of 1,400 and 6,000, respectively (see Table 4.4 and 4.5).

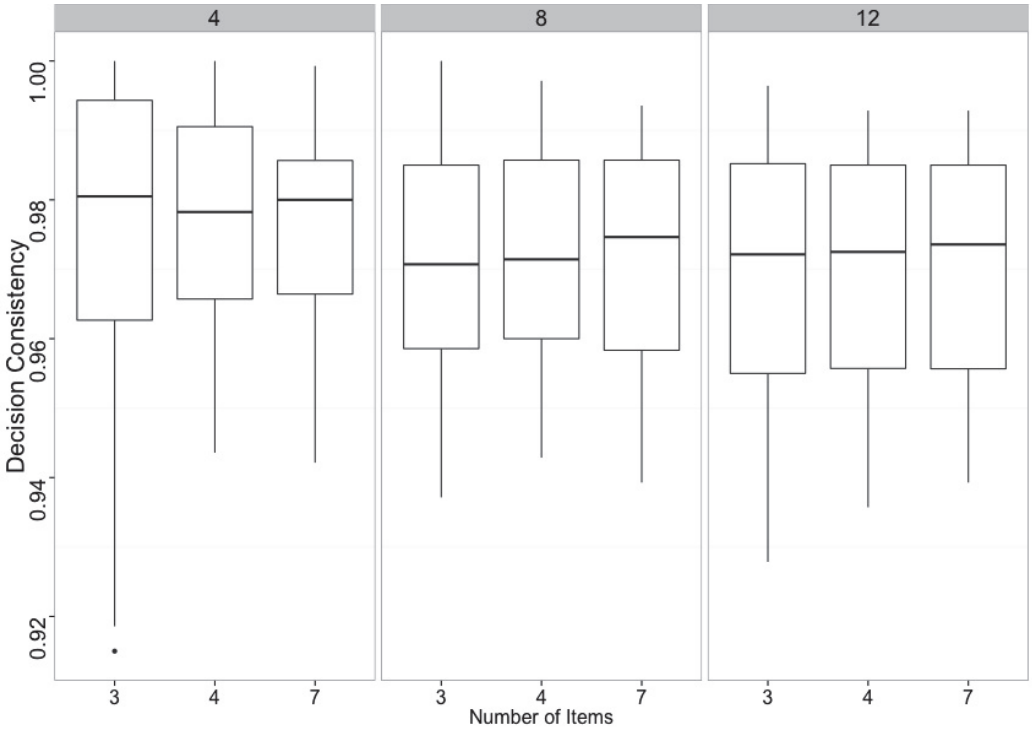

Figure: 4.10: Boxplot of the number of items by the number of testlets

Number of Testlets and Sample Size

The next strongest effect was for the interaction between the number of testlets and sample size ($\widehat{\eta^2} = .002, \widehat{\eta_p^2} = .006$) (see Table 4.5). The average DC between the DRM and the TRM was 97.3% (SD=0.02), and 97.28% (SD=0.02) for sample sizes of 1,400 and 6,000, respectively (see Table 4.4 and 4.5). When there were four testlets, DC was 97.7% (SD=0.02) and 97.66% (SD=0.02) for sample sizes of 1,400 and 6,000, respectively. When there were four testlets the number of people misclassified were 33 and 140 for sample sizes of 1,400 and 6,000, respectively. When there were eight testlets DC was 97.2% (SD=0.01) and 97.2% (SD=0.01) for sample sizes of 1,400 and 6,000, respectively. When there were eight testlets the number of people misclassified were 40 and 169 for sample sizes of 1,400 and 6,000, respectively. When there were 12 testlets DC was 97.01% (SD=0.02) and 97% (SD=0.02) for sample sizes of 1,400 and 6,000, respectively (see Figure 4.11). When there were 12 testlets the number of people misclassified were 84 and 180 for sample size of 1,400 and 6,000, respectively.

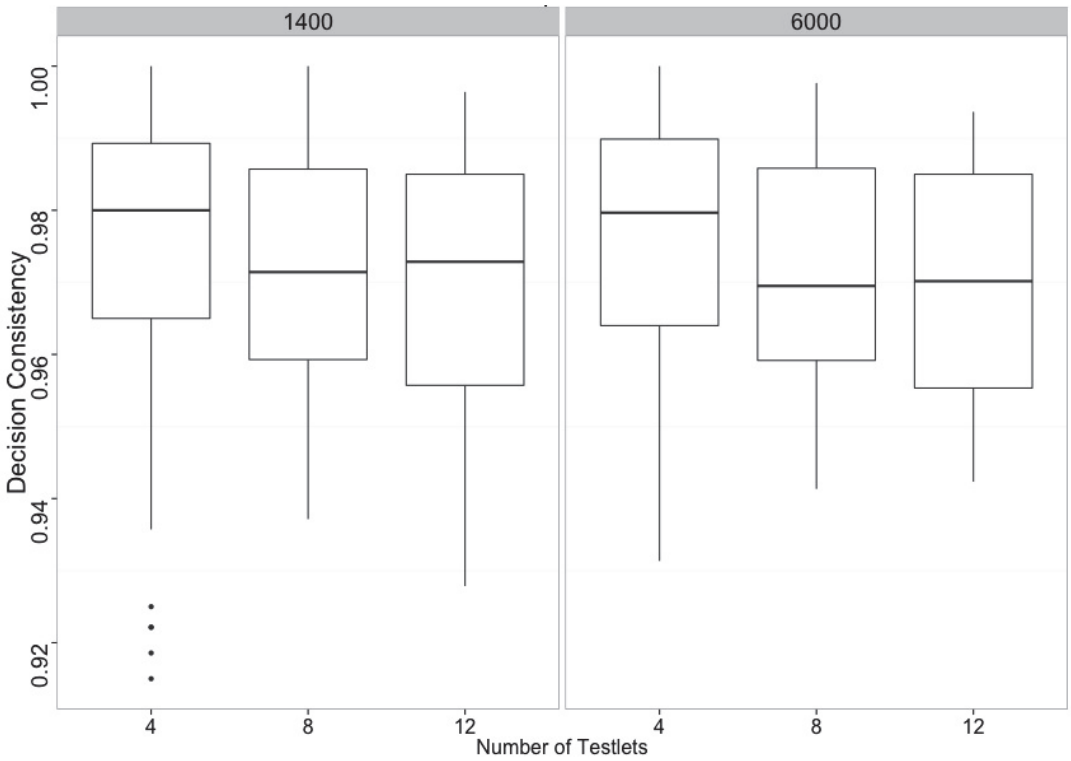

Figure: 4.11: Boxplot of number of testlets by sample size

Number of Items

The last effect greater than .001 for either eta squared or partial eta squared was the simple effect number of items ($\widehat{\eta^2} = .001, \widehat{\eta_p^2} = .003$) (see Table 4.5). When there were three items per testlet, DC had more variability (see Figure 4.8). The average DC between the DRM and the TRM was 97.27% (SD=0.02), 97.34% (SD=0.02), and 97.27 (SD=0.01) for three, four, and seven items per testlet, respectively (see figure 4.12). When the number of items was three the number of people misclassified on average was 39 and 164, for sample sizes of 1,400 and 6,000, respectively. When the number of items was four the number of people misclassified on average was 38 and 160, for sample sizes of 1,400 and 6,000, respectively. When the number of items was seven the number of

people misclassified was 39 and 164, for sample sizes of 1,400 and 6,000, respectively (see Table 4.4 and 4.5).

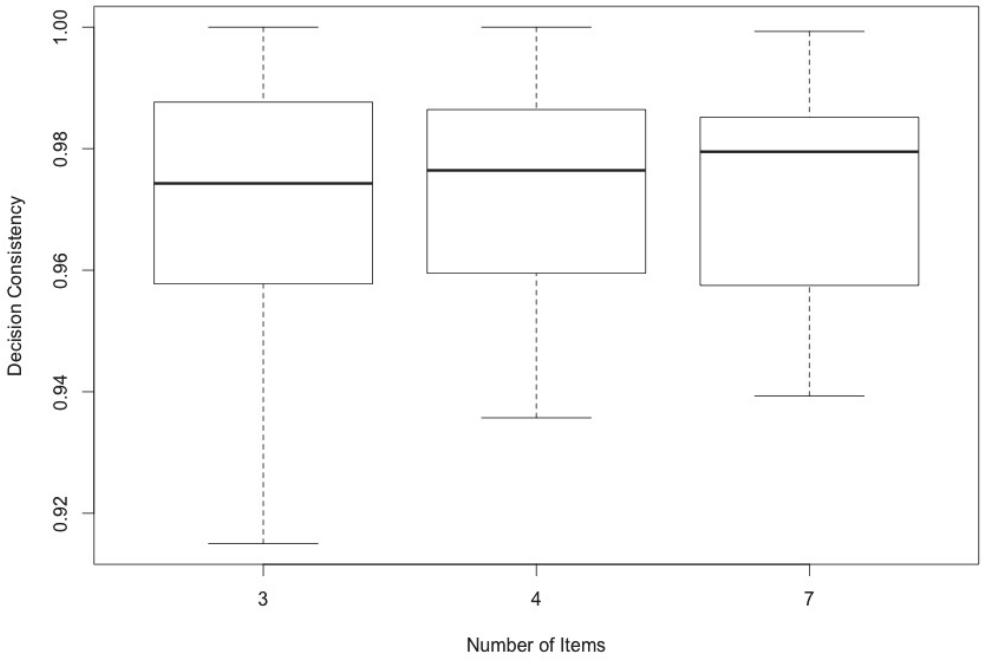

Figure: 4.12: Boxplot of number of items

Number of Items and Percent of Total Items in Testlets

The next strongest effect was for the interaction between the number of items and the percent of total items in testlets ($\widehat{\eta^2} = .0003, \widehat{\eta_p^2} = .001$) (see Table 4.5). When the percent of total items in testlets was 75% DC is lower on average (see Figure 4.13). For three item testlets DC is 97.32% (SD=0.01), and 97.21% (SD=0.02), when the percent of total items in testlets were 75% and 100%, respectively. When there were three items in each testlet and 75% of items were in testlets the number of people misclassified on average was 38 and 161, for sample sizes of 1,400 and 6,000, respectively. When there were three items in each testlet testlets and 100% of items were in testlets the number of

people misclassified on average was 39 and 168, for sample sizes of 1,400 and 6,000, respectively (see Table 4.4 and 4.5).

For four item testlets DC is 97.3% (SD=0.01), and 97.38% (SD=0.02), when the percent of total items in testlets were 75% and 100%, respectively. When there were four items in each testlet and 75% of items were in testlets the number of people misclassified was 38 and 162, for sample sizes of 1,400 and 6,000, respectively. When there were four items in each testlet testlets and 100% of items were in testlets the number of people misclassified on average was 37, and 157, for sample sizes of 1,400 and 6,000, respectively (see Table 4.4 and 4.5).

For seven item testlets DC is 96.89% (SD=0.01), and 97.58% (SD=0.02), when the percent of total items in testlets were 75% and 100%, respectively. When there were seven items in each testlet and 75% of items were in testlets, the number of people misclassified on average was 44 and 187 for sample sizes of 1,400 and 6,000, respectively. When there were seven items in each testlet testlets and 100% of items were in testlets the number of people misclassified on average was 34 145, for sample sizes of 1,400 and 6,000, respectively (see Table 4.4 and 4.5).

When all items were in testlets as the number of items increased DC also increased. However, when 75% of the items are in testlets and the number of items increased DC decreased (see Figure 4.12).

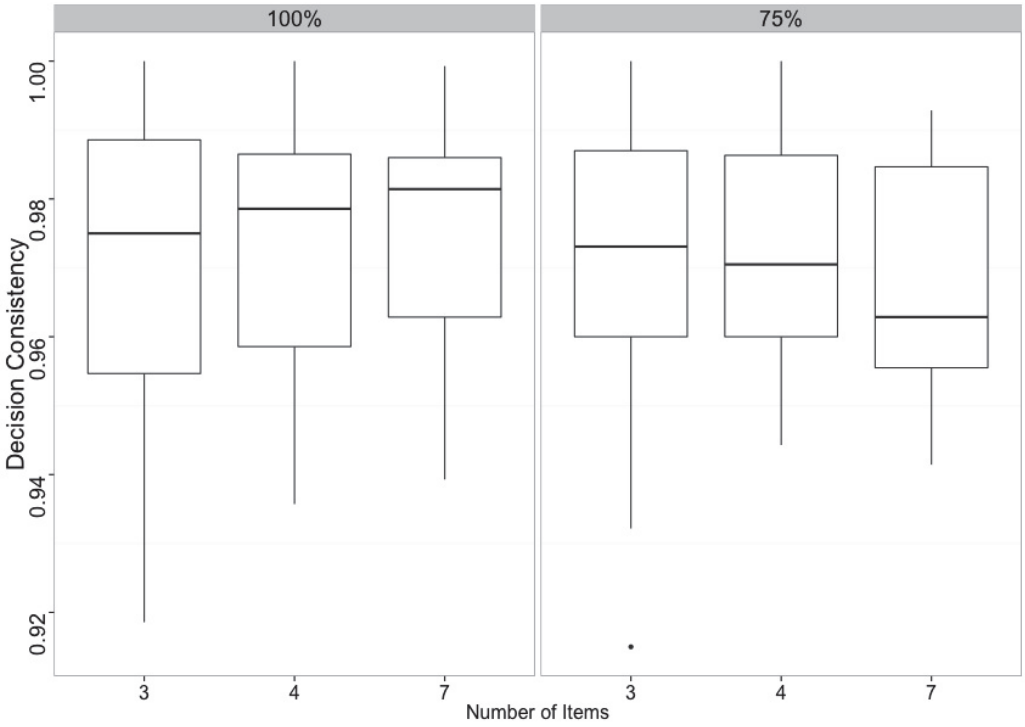

Figure: 4.13: Boxplot of number of items by percent of total items in testlets

Percent Total Items in Testlets and Sample Size

The next strongest effect was for the interaction between the percent of total items in testlets and sample size ($\widehat{\eta^2} = .0002, \widehat{\eta_p^2} = .001$) (see Table 4.5). When sample size was 1,400, DC was 97.17% (SD=0.02) and 98.41% (SD=0.02) when 75% and 100% of items were in a testlet. When sample size was 1,400, the number of people misclassified when 75% and 100% of items were in a testlet is 40 and 23, respectively. When sample size was 6,000, DC was 97.19% (SD=0.01) and 97.39% (SD=0.02) when 75% and 100% of items were in a testlet. When sample size was 6,000, the number of people misclassified when 75% and 100% of items were in a testlet is 169 and 157, respectively (see Table 4.4 and 4.5). When the percent of total items in testlets was 75% DC was lower on average (see Figure 4.14).

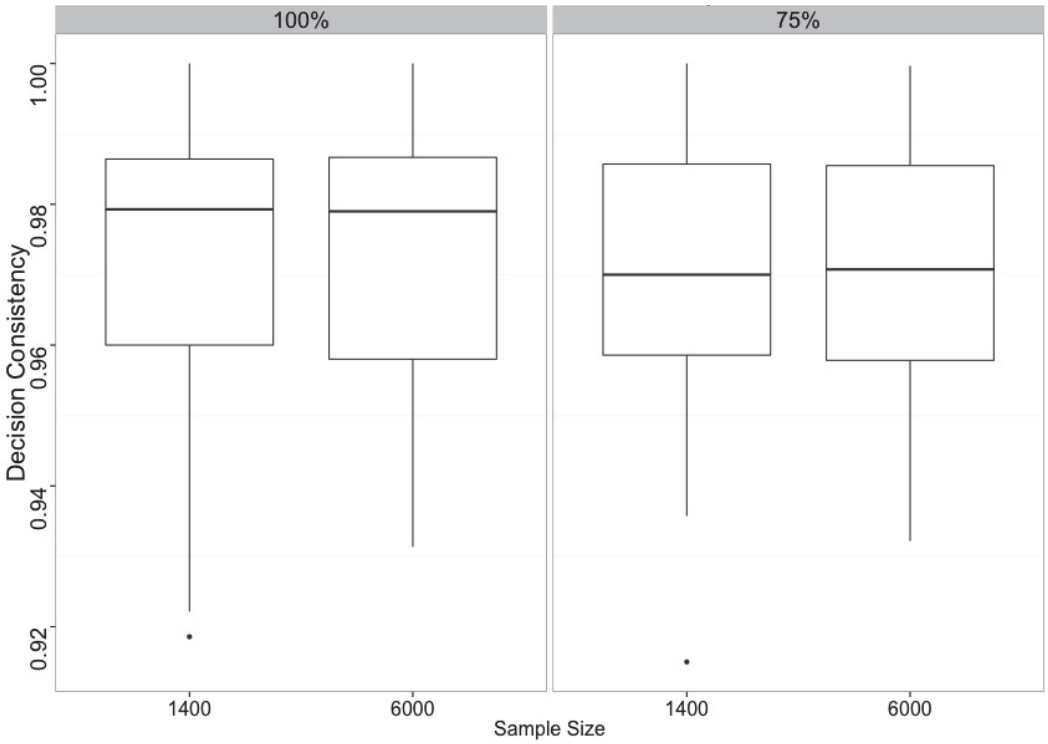

Figure: 4.14: Boxplot of percent of total items in testlets by sample size

Number of Items and Sample Size

Eta-squared was .0001 and partial eta- squared was .001 for the interaction between the number of items and sample size was (see Table 4.5). When there are three items DC was 97.32% (SD=0.01), and 97.21% (SD=0.02) when the percent of total items in testlets were 75% and 100%, respectively. When there were three items, the number of people misclassified on average was 100 and 104 when sample size was 1,400 and 6,000, respectively. When there were four items DC is 97.3% (SD=0.01), and 97.38% (SD=0.02) when sample size was 1,400 and 6,000, respectively. When there were four items, the number of people misclassified on average was 100 and when sample size was 1,400 and 6,000, respectively. When there were seven items, DC is 96.89% (SD=0.01), and 97.58% (SD=0.02), when sample size is 1,400 and 6,000, respectively. When there were seven items the number of people misclassified on average was 100 and 104 when

sample size is 1,400 and 6,000, respectively (see Table 4.4 and 4.5). When there were three items DC is lower on average and there was little variability between sample sizes (see Figure 4.15).

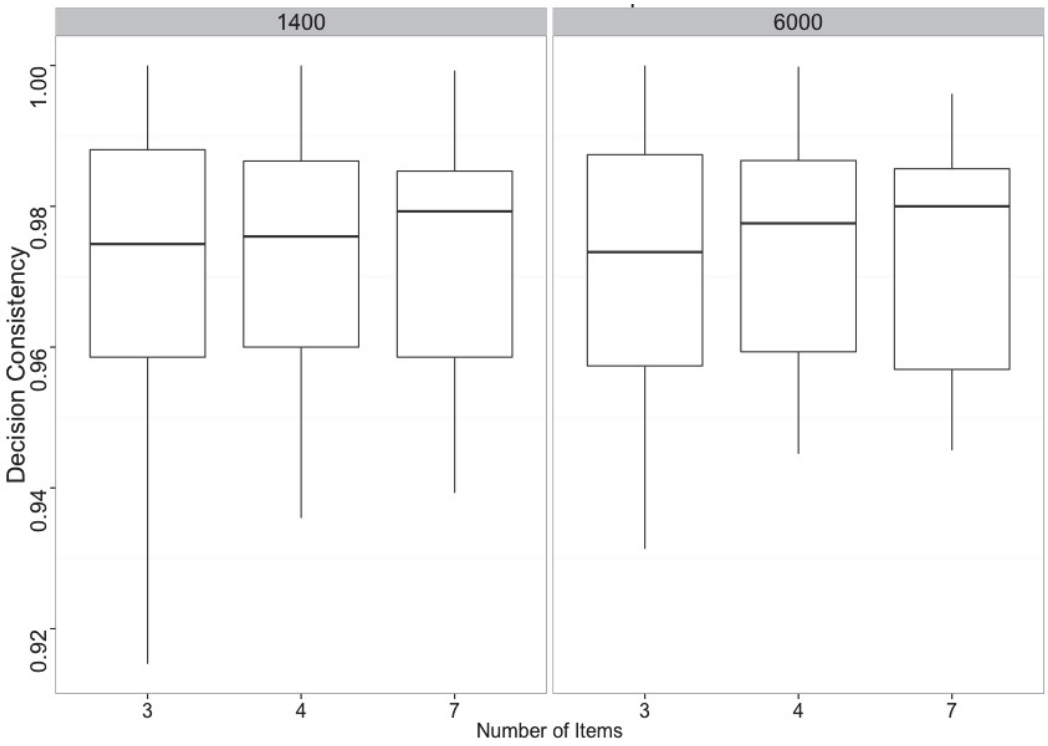

Figure: 4.15: Boxplot of the number of items by sample size

Testlet Variance and Sample Size

Eta-squared was .0001 and partial eta-squared was .001 for the interaction between testlet variance and sample size. When testlet variance was 0.2, DC was 98.53% (SD=0.01) and 98.54% (SD=0.01) when sample size was 1,400 and 6,000, respectively. When testlet variance was 0.2, the number of people misclassified were 21 and 88 when sample size was 1,400 and 6,000, respectively. When testlet variance was 0.7, DC was 96.06% (SD=0.01) and 96.03% (SD=0.01) when sample size was 1,400 and 6,000, respectively (see Figure 4.16). When testlet variance was 0.7, the number of people

misclassified were 55 and 1237 when sample size was 1,400 and 6,000, respectively (see Table 4.4 and 4.5). While there were minimal differences in DC as sample size increased more examinees were misclassified.

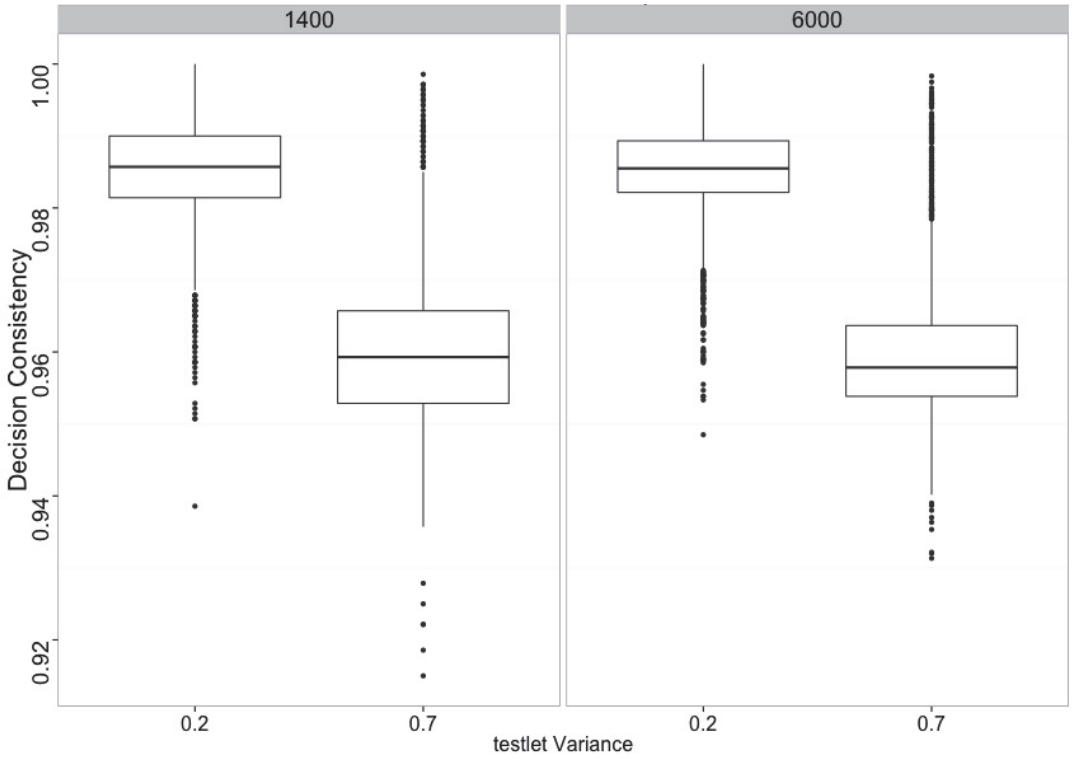

Figure: 4.16: Boxplot of testlet variance by sample size

Number of Items and Testlet Variance

Eta-squared was .0001 and partial eta-squared was .0004 for the interaction between the number of items and testlet variance (see Table 4.3). As the number of items increased and testlet variance increased, DC decreased and the number of people misclassified increased (see Figure 4.17). As the number of items increases variability in DC decreased regardless of the testlet variance. DC was similar across the number of items but decreased as testlet variance increased.

When there were three items and variance is 0.2 and 0.7, DC was 98.51% (SD=0.01) and 96.02% (SD=0.01), respectively. When there were three item testlets and testlet variance was 0.2, the number of people misclassified on average was 21, and 90 for sample sizes of 1,400 and 6,000, respectively. When there were three item testlets and testlet variance is 0.7, the number of people misclassified on average was 56 and 239 for sample sizes of 1,400 and 6,000, respectively (see Table 4.4 and 4.5).

When there were four items and variance was 0.2 and 0.7, respectively DC and number of people misclassified were 98.55% (SD=0.01) and 96.13% (SD=0.01), respectively. When there were four item testlets and testlet variance is 0.2, the number of people misclassified on average was 21 and 87 for sample sizes of 1,400 and 6,000, respectively. When there were four item testlets and testlet variance is 0.7, the number of people misclassified on average was 55 and 233 for sample sizes of 1,400 and 6,000, respectively (see Table 4.4 and 4.5).

When there were seven items and variance was 0.2 and 0.7, respectively DC and number of people misclassified on average were 98.55% (SD=0.01) and 95.99% (SD=0.01), respectively. When there were seven item testlets and testlet variance is 0.2, the number of people misclassified on average was 21 and 87, for sample sizes of 1,400 and 6,000, respectively. When there were seven item testlets and testlet variance is 0.7, the number of people misclassified on average was 57 and 241, for sample sizes of 1,400 and 6,000, respectively (see Table 4.4 and 4.5).

Table 4.4

Proportion Tables for Conditions with 100% of Items in Testlets

Testlet Variance	No. of Testlets	No. of Items	Sample Size	Testlet and Rasch					
				FF	*FP*	*PF*	*PP*	*PCC*	*NMP*
0.2	4	3	1400	0.223	0.002	0.010	0.765	0.988	17
			6000	0.225	0.002	0.012	0.761	0.986	84
		4	1400	0.226	0.003	0.012	0.759	0.985	21
			6000	0.228	0.002	0.010	0.761	0.989	66
		7	1400	0.233	0.003	0.010	0.755	0.988	17
			6000	0.233	0.002	0.008	0.756	0.989	66
	8	3	1400	0.225	0.000	0.015	0.759	0.984	23
			6000	0.226	0.000	0.016	0.757	0.983	102
		4	1400	0.229	0.000	0.015	0.755	0.984	23
			6000	0.230	0.000	0.014	0.755	0.985	90
		7	1400	0.231	0.000	0.014	0.754	0.985	21
			6000	0.233	0.000	0.013	0.754	0.987	78
	12	3	1400	0.229	0.000	0.016	0.755	0.984	23
			6000	0.229	0.000	0.016	0.754	0.983	102
		4	1400	0.230	0.000	0.016	0.754	0.984	23
			6000	0.230	0.000	0.016	0.754	0.984	96
		7	1400	0.232	0.000	0.016	0.752	0.984	23
			6000	0.233	0.000	0.016	0.751	0.984	96
0.7	4	3	1400	0.196	0.001	0.027	0.776	0.972	40
			6000	0.194	0.000	0.029	0.776	0.970	180
		4	1400	0.204	0.002	0.024	0.770	0.974	37
			6000	0.205	0.001	0.020	0.774	0.979	126
		7	1400	0.210	0.003	0.020	0.767	0.977	33
			6000	0.210	0.003	0.020	0.767	0.977	138
	8	3	1400	0.191	0.000	0.044	0.765	0.956	62
			6000	0.191	0.000	0.044	0.765	0.956	264
		4	1400	0.197	0.000	0.041	0.762	0.959	58
			6000	0.196	0.000	0.042	0.762	0.958	252
		7	1400	0.201	0.000	0.039	0.760	0.961	55
			6000	0.202	0.000	0.039	0.759	0.961	234
	12	3	1400	0.192	0.000	0.049	0.759	0.951	69
			6000	0.192	0.000	0.049	0.760	0.952	288
		4	1400	0.194	0.000	0.048	0.758	0.952	68
			6000	0.193	0.000	0.048	0.759	0.952	288
		7	1400	0.198	0.000	0.047	0.756	0.954	65
			6000	0.197	0.000	0.046	0.757	0.954	276

Note. FF= Fail Fail, FP=Fail Pass, PF=Pass Fail, PP= Pass Pass, PCC=Proportion Correctly Classified, NPM= Number of People Misclassified. The number of people misclassified was rounded up to the nearest whole number.

Table 4.5

Proportion Tables for Conditions with 75% of Items are in Testlets

Testlet Variance	No. of Testlets	No. of Items	Sample Size	Testlet and Rasch					
				FF	*FP*	*PF*	*PP*	*PCC*	*NMP*
0.2	4	3	1400	0.226	0.003	0.100	0.760	0.986	20
			6000	0.226	0.003	0.011	0.760	0.986	84
		4	1400	0.230	0.004	0.011	0.756	0.986	20
			6000	0.229	0.005	0.011	0.756	0.985	90
		7	1400	0.233	0.005	0.011	0.751	0.984	23
			6000	0.231	0.005	0.012	0.751	0.982	108
	8	3	1400	0.232	0.002	0.014	0.753	0.985	21
			6000	0.231	0.001	0.014	0.754	0.985	90
		4	1400	0.233	0.001	0.012	0.075	0.986	20
			6000	0.234	0.002	0.012	0.753	0.987	78
		7	1400	0.236	0.002	0.013	0.750	0.986	20
			6000	0.235	0.002	0.013	0.750	0.985	90
	12	3	1400	0.234	0.000	0.014	0.751	0.985	21
			6000	0.232	0.001	0.014	0.753	0.985	90
		4	1400	0.235	0.001	0.013	0.750	0.985	21
			6000	0.234	0.000	0.013	0.752	0.986	84
		7	1400	0.236	0.001	0.014	0.749	0.985	21
			6000	0.236	0.001	0.014	0.749	0.985	90
0.7	4	3	1400	0.199	0.006	0.030	0.765	0.964	51
			6000	0.200	0.006	0.030	0.764	0.964	216
		4	1400	0.204	0.010	0.028	0.759	0.963	52
			6000	0.203	0.010	0.031	0.757	0.960	240
		7	1400	0.207	0.014	0.031	0.749	0.956	62
			6000	0.205	0.014	0.033	0.748	0.953	282
	8	3	1400	0.204	0.002	0.038	0.756	0.960	56
			6000	0.203	0.002	0.028	0.757	0.960	240
		4	1400	0.207	0.003	0.037	0.753	0.960	56
			6000	0.204	0.004	0.035	0.757	0.961	234
		7	1400	0.207	0.005	0.038	0.749	0.956	62
			6000	0.207	0.006	0.039	0.748	0.955	270
	12	3	1400	0.206	0.001	0.040	0.754	0.960	56
			6000	0.205	0.001	0.040	0.754	0.959	246
		4	1400	0.207	0.001	0.040	0.752	0.959	58
			6000	0.207	0.001	0.040	0.752	0.959	246
		7	1400	0.209	0.002	0.040	0.749	0.958	59
			6000	0.208	0.003	0.040	0.749	0.957	258

Note. FF= Fail Fail, FP=Fail Pass, PF=Pass Fail, PP= Pass Pass, PCC=Proportion Correctly Classified, NPM= Number of People Misclassified. The number of people misclassified was rounded up to the nearest whole number.

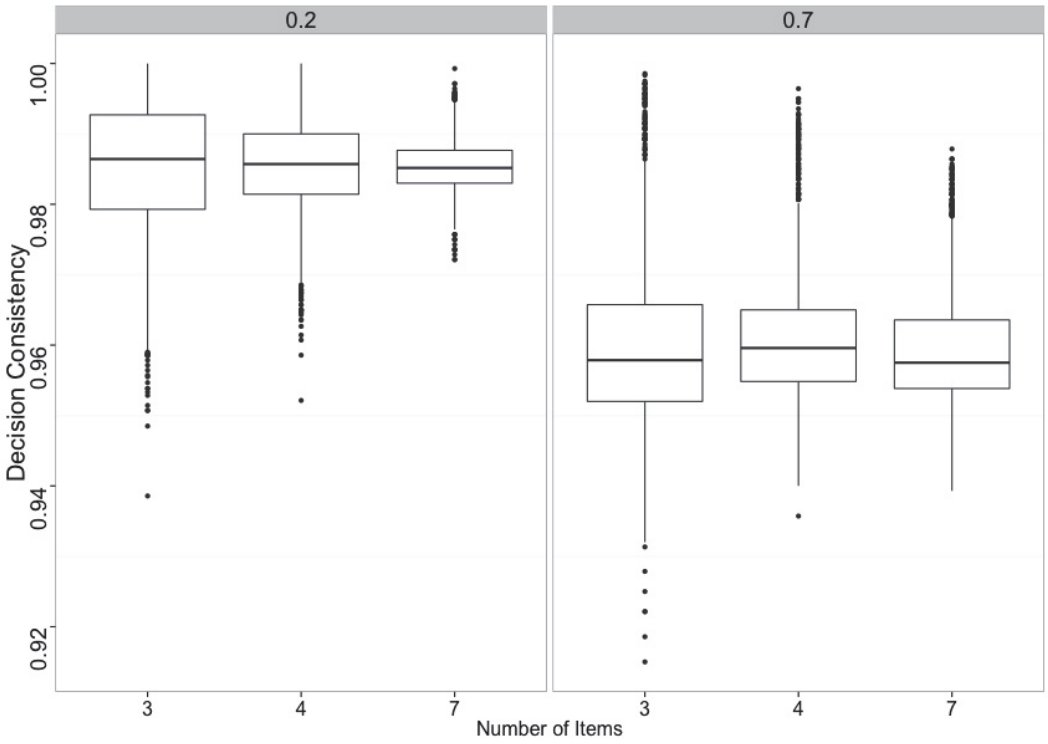

Figure: 4.17: Boxplot of the number of items by testlet variance

Sample Size

Eta- squared was .0001 and partial eta-squared was .0002 for the design factor sample size (see Table 4.5). DC was almost the same regardless of the sample size (see Figure 4.18). The average DC between the DRM and the TRM was 97.62% (SD=0.02), and 97.68% (SD=0.02) when sample size was 1,400 and 6,000, respectively (see figure 4.9). The number of people misclassified, on average, when sample size was 1,400 and 6,000 was 33 and 140, respectively (see Table 4.4 and 4.5).

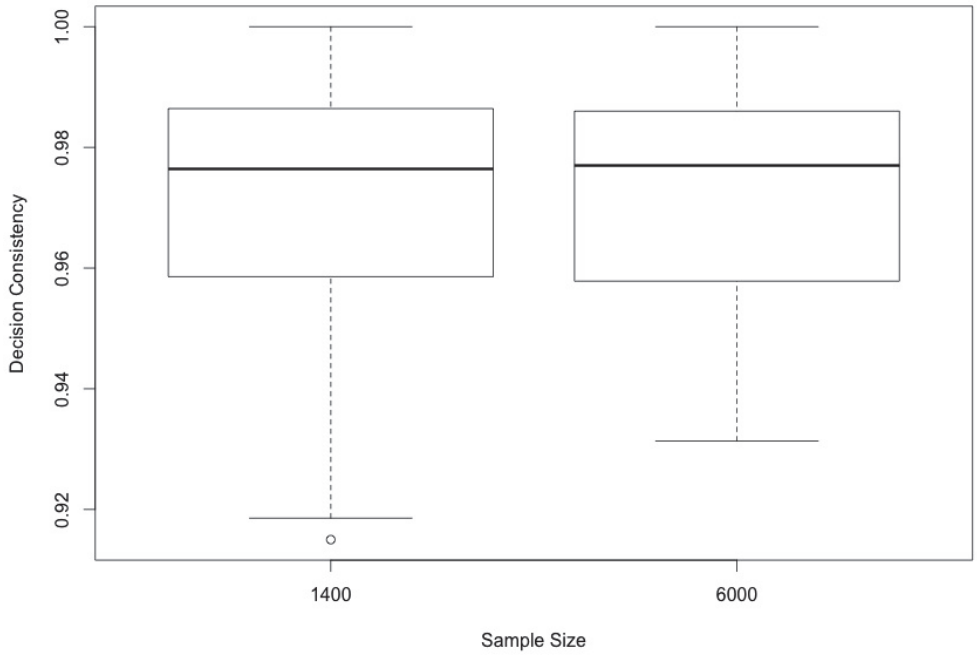

Figure: 4.18: Boxplot of sample size

An Empirical Example

The PISA (2012) U.S. public-use data files from Institute of Education Sciences (IES) contains a subset of students mathematics, reading, and science responses (http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2014028). For the purpose of this study only the reading portion of Booklet Two was used to study the testlet effect on DC between the DRM and the RTM (see Appendix A for a description of the items). Booklet Two contains 29 reading items with nine testlets. The items per testlet range from two to four (see Figure 3.1). This subsample consists of 422 U.S. students fifteen-years-old. The cut score for the PISA data was set based on the percentage of examinees that scored at a level two or higher on the 2012 exam (OECD, 2013). According to the OECD (2013) 81.2 % scored at a level two or higher. The ability distribution is assumed to follow a standard normal distribution; therefore, the cut score is set to the value above which approximately 81.2% of the distribution falls (z = -0.89, see equation 4.2).

$$z = \Phi^{-1}(.812) \approx -0.89, \quad [4.2]$$

where Φ^{-1} is the inverse of the cumulative standard normal distribution function.

PISA is traditionally calibrated with a DRM, thus making this data set ideal for this study. Two models will be estimated: a) the DRM and b) the RTM. For each model the pass/fail decisions of each examinee were recorded and the proportion of examinees whose pass/fail decisions differed between the two models is reported.

The classification agreement percent between the DRM and TRM was 99.5%, only two students were classified differently. The classification comparison is presented in Table 4.6 and the testlet variances and standard deviations are presented in Table 4.7.

Table 4.6

Proportion of Pass/Fail Decision Between the DRM and the RTM

		Rasch	
		Pass=1	Fail=0
Testlet	Pass=1	0. 879	0.005
	Fail=0	0.000	0.116

Table 4.7

Testlet Variance and Standard Deviation

Testlet	Testlet Variance	Standard Deviation
1	0.118	0.014
2	0.055	0.003
3	0.105	0.011
4	0.084	0.007
5	0.084	0.007
6	0.114	0.013
7	0.084	0.007
8	0.118	0.014
9	0.077	0.006

The testlet variance was relatively small and there were four or fewer items per testlet so the difference between classifications was also small. This is consistent with the Monte Carlo simulation findings.

Summary

In summary, diagnostics were conducted that provided evidence that 100 replications, population parameters, and 1,000 burning and 2,000 iterations were sufficient for this study. The true model was the RTM so when the DRM differed from the RTM the disagreement was considered misclassification. The percent of pass/fail decision agreement or DC between the DRM and the RTM was very high. The percent of DC ranged from 91% to 100%. However, the design factor that influenced DC the most was testlet variance. In other words, as testlet variance increased DC decreased. Misclassification was greater when the testlet variance was larger and when more testlets are present. The number of items alone did not affect DC. There was no difference in DC for differing sample sizes. Those misclassified were more often false negative rather than false positive. As with the empirical example when testlet variance is small the percent agreement is very high. . The empirical example with the PISA data followed the same trends found in the MC simulation study. Implications and recommendations are provided in the next chapter.

CHAPTER FIVE

Discussion

Testlets are becoming increasingly more prevalent in educational and cognitive testing situations. Languages Arts and reading assessments are prime examples of tests that include testlets, as examinees are provided with a reading passage and are then asked a group of items related to the passage (Fountas & Pinnell, 2012; Good, Wallin, Simmons, Kame'enui, & Kaminsji, 2002; Harcourt, 2003; SBAC, 2014; TEA, 2014; Texas Institute for Measurement, Evaluation, and Statistics, 2010). The theoretical structure of reading assessments that include reading passages and multiple items related to those passages is a testlet structure. Testlets create additional correlation between items that is unexplained by the latent trait. In applied settings, researchers are faced with the decision of selecting an appropriate model for use in large-scale assessments. Among the multitude of models available unidimensional models are more frequently used even when testlets are included in the structure and design of the assessment (Dickenson, 2005).

This study investigated classification consistency between competing Rasch models, a DRM and a TRM, where classification consistency was defined as the proportion of agreement between the two models based on a predetermined cut score.

The cut score was *a priori* set to the average passage rate of a large southern state reading assessment (~75%) or $z = \Phi^{-1}(.75) \approx -0.67,$ where Φ^{-1} is the inverse of the cumulative standard normal distribution function.

The population structure used in the current study is a TRM with known number of testlets, items within testlets, and testlet effects. For this Monte Carlo simulation study, the design factors and conditions that varied were: (a) number of testlets, (b) number of items in testlets, (c) testlet effect, (d) percent of total items that are in testlets, and (e) sample size, yielding 72 possible conditions (3 numbers of items × 3 number of items in testlets × 2 testlet variances × 2 sample sizes × 2 proportion of items in testlet = 72). For each condition, 100 replications were generated.

Discussion

Overall, DC was very high between the two models, ranging from 91.5% to 100%. These findings are consistent with those reported by Lu (2010). He compared a 3PL IRT and a testlet IRT model, and found DC ranged from 91% to 96%. Similar to Lu (2010), the design factor that had the greatest effect on DC was the testlet effect or testlet variance. Other design factor that affected DC included number of testlets, an interaction between testlet variance and the percent of total items in testlets, and an interaction between the number of testlets and the percent of total items in testlets.

Testlet Variance

In condition where testlet variance was 0.7 the number of examinees misclassified was more than double the number of examinees misclassified when testlet variance was 0.2. It was expected that there would be a decrease in DC as testlet variance increased because there are stronger relationships between items that was unaccounted for by the DRM. In applied settings the testlet variance is unknown and have been reported to be as high as 2.8 (Wanget al., 2002), meaning that the effects of testlet variance may be more

pronounced in applied settings increasing the number of people misclassified. Lu (2010) found that ability estimates were negatively affected by larger testlet variance and this in turn decreased DC.

Number of Testlets

In conditions where there were more testlets DC was lower. These findings are consistent with those found by Bradlow, Wainer and Wang (1999), where they concluded that an increase in the number of testlets affected the quality of the estimates and in turn significantly impacted the final inferences made about people. This was an expected outcome because when there are more testlets there is an increase in the number of additional unmolded dimensions and increasing the difference between the classification consistency between the DRM and TRM. Similarly, Yen and Fitzpatrick (2006) reported that when there are fewer testlets the effects of the testlet effect can be minimized.

Interaction Between Percent of Total Items in Testlets by the Number of Testlets

In conditions where 75% of the total number of items was in testlets, the number of testlets did not change the number of people misclassified. However, when 100% of the total number of items was in testlets, an increase in the number of testlets was associated with an increase in the proportion of people misclassified. There was more variability in DC when all items were in testlets compared to conditions where 75% of items were in testlets. This implies that the design structure of the assessment affects DC, meaning that when more testlets are present the proportion of DC may become less stable. However, when the proportion of items in testlets decreases this may minimize the impact of the testlet effect. This is consistent with Hembry's (2014) findings that when

there are more testlets and more of the percent of total items in testlets there is more of a testlet effect that impacts all outcomes (e.g. DC, precision in parameter estimates).

Interaction of Number of Testlets by Testlet Variance

As the number of testlets increased and testlet variance increased DC decreases and the proportion of people misclassified increased. This is consistent with findings reported by Lu (2010) who also found that as the number of testlets and testlet variance increased ability estimation precision decreased. The decrease in ability estimation precision, due to the unaccounted for LID, caused classification errors to spread across the ability distribution to more ability levels. This caused a decrease in DC (Lu, 2010).

As the number of testlets increased, thereby increasing the overall length of the test, variability in DC decreased. In general, when more information was present, the consistency of estimates increased. In conditions where there are smaller numbers of testlets the total amount of information is limited and the effect of testlet variance on the consistency of DC is also limited.

Other Design Factors

All other main effects and two-way interaction effects were trivial. The number of items in testlets and related interactions did not influence DC. However, this may be an artifact of directly specifying testlet variance in this study. The small effect was unexpected as previous research indicated that as the number of items in testlets increased DC decreased (Chen, 2014; Hembry, 2014; Lu, 2010). Chen (2014) reported an increase in the testlet variance as testlet length, more items in testlets, increases. Applied

researchers should proceed with caution when there are a large number of items in testlets.

Likewise, sample size and related interactions did not have a strong effect on DC, as it was the same for both sample sizes. Estimates were more precise when sample size was larger.

PISA

PISA is traditionally calibrated with a DRM, and contained 29 items in nine testlets. The cut score for the PISA data was set based on the percentage (81.2%) of examinees that scored at a level two or higher on the 2012 exam (OECD, 2013), therefore, the cut score is set to the value above which approximately 81.2% of the distribution falls $z = \Phi^{-1}(.812) \approx -0.89$, where Φ^{-1} is the inverse of the cumulative standard normal distribution function.

The classification agreement percent between the DRM and the TRM was 99.5%. That means that 2 students were classified differently. The classification comparison is presented in Table 4.8 and the testlet variances and standard deviations are presented in Table 4.9. The testlet variance is relatively small and there are four or fewer items per testlet so the difference between classifications is also small. This is consistent with the Monte Carlo simulation findings and Bradlow et al.'s (1999), findings that when the testlet and non-testlet models were applied to real data all models performed about the same due to the posterior mean testlet variance of 0.11 (σ_y^2), classified by the authors as a modest effect. Similarly, Wainer, et al. (2007) suggested that when the number of items included in each testlet (i.e., four testlets with three items in each testlet) is relatively small ignoring the violation of local independence assumption may be tolerable, but

cautioned that as the item-to-testlet ratio increased or the magnitude of the dependence in the testlet increased a lack of precision in the parameter estimates would also increase.

Conclusion

These findings are consistent with previous research. Numerous studies have compared unidimensional and testlet model parameter ordering via correlation, RMSE and bias and found that ability ordering is quite similar between models (Eckes, 2013; Jiaoet al., 2013; Paeket al., 2008; Wang & Wilson, 2005a; Wang & Wilson, 2005b). However, in high-stakes testing when DC is as high as 0.986 in a sample of 6,000, 84 examinees were misclassified. Those who were close to the cut score may have more variability depending on what model was used to estimate ability. In other words someone may pass if a testlet model was used and fail if a unidimensional model was used.

The misclassification was more often a false negative meaning that an examinee failed when they should have passed. Bradlow et al. (1999) reported a similar phenomenon in their study in the use of SAT scores for scholarship awards and acceptance into college. While the selection was norm referenced instead of criterion referenced and the top of the ability distribution was of interest the authors found that when the testlet model was used a higher percentage of the top 100 examinees were identified. This study demonstrated that when the testlet model was not used fewer examinees received benefits for their test scores that they should have had they been modeled with a testlet model.

The development of multidimensional item response models parallels the development of factor analytic models in the early nineteen hundreds. Spearman (1904)

developed a theory of intelligence that was thought to consist of one general factor, which is conceptually equivalent to the one parameter item response model. Later, Holzinger and Swineford (1937) found that one general factor did not adequately account for intercorrelations and thus developed the bi-factor theory. The bi-factor model accounts for a general factor and group factors that were described as mutually exclusive (Holzinger & Harman, 1939) similar to the testlet model described by Bradlow et al. (1999). More recently, Li et al. (2006) found that the testlet model is a constrained version of the bi-factor model. Rijmen (2010) concluded that the testlet model and the bi-factor model were formally equivalent. Therefore, the findings of this study may also inform model selection decisions in research involving the bi-factor model. For details and characteristics of bi-factor models, interested readers should see Holzinger and Harman (1939), Holzinger and Swineford (1937), Reise (2012), Reise Widaman, and Pugh (1993), and Yung, Thissen, and McLeod (1999).

Recommendations for use of Testlet Models

When a testlet structure is present in applied data the testlet variance is unknown and as the testlet variance increases so does the misclassification of examinees. When measurement models are used that do not align with the structure of the data additional error is introduced into the parameter estimates. This has the potential to directly impact the decisions that are made about people.

Therefore, when testlets are present practitioners should investigate the testlet effect by using a testlet model. An initial investigation of the assessment should include a testlet model in order to investigate the magnitude of the testlet effect before other statistical properties such as DIF, equating, or reliability are investigated. If the testlet

effect is found to be small .3 or less then a traditional model may be acceptable. On the other hand if testlet effects are larger than .3 the testlet model is recommended.

Limitations

One advantage of a MC study is the ability to study the effects of multiple factors simultaneously (i.e. sample size, number of items, number of dimensions, variety of models). Another advantage is the ability to generate random samples using a specified model in order to compare estimated results when the "truth" is known.

However, A major limitation of MC studies is that the studies are only as useful as the generalizability of the conditions modeled. Another limitation is the quality of the random number generator, and results may vary depending on the number of replications (Harwell, et al., 1996). Therefore, it is essential to generate conditions similar to those found in the applied literature. However, this study did not simulate all possible conditions reported in the literature. Instead the most commonly reported conditions were included, along with an empirical example.

There are many measurement models with varying degrees of complexity for calibrating and making decisions about examinee ability. This study includes Rasch models, which are more parsimonious than other item response models. If testlet variance effects pass/fail decisions in less complex models then the effect of testlet variance may be increased for more complex models.

Another limitation of this study was the use of Bayesian or MCMC estimation because the researcher is required to specify prior distributions as well as chain length. MCMC estimation includes the use of priors that influence the estimation process. One of the limitations of using priors is that results may be different when different priors are

used. Similarly, the selected number of burn-in iterations and total number of total iterations in conditions with fewer items per testlet and fewer testlets were more problematic in achieving the stationary chain.

A limitation of this study is that the distribution used to generate the item difficulties was different than the cut score. Using a different cut score than the item difficulty mean was motivated by the effort to approximate standard setting procedures in applied settings Another limitation is that DC was compared from the two estimated model parameters. The consistency with the true classification was not examined in this study.

Future Research

As noted above, the findings of this study are only generalizable to study conditions that are similar to those conditions in this study. Additional research is needed to continue to develop our understanding of pass/fail decisions under competing models when testlets are present. This study focused on those conditions most often found in K-12 settings and future research may include other testing situations such as licensure and credentialing exams. In order to expand on this study future research should examine DC when testlet variance is different for each testlet (i.e. four testlets with testlet variance of .2, .5, .75, & 1). This study investigated only conditions where the testlet variance was the same for each testlet with in a given condition. Similarly, alternate cut scores may be of interest to investigate decisions at various points on the ability distribution. The cut score in this study was -0.67, but alternative cut scores may be of interest. Given that there were several main effects and two-way interaction effects that included the design

factor, total percent of items in testlets, this is another area in need of further investigation.

Future research should include a comparison between the true pass/fail classification and the classification from each estimated model. This would provide additional information of the performance of both models. Similarly, replication studies that include alternate cut scores and item difficulty distributions targeted around the cut score are needed in order to provide additional more information about the DC between the DRM and the RTM.

Additional research is also recommended with regard to estimation. This study employed MCMC estimation techniques that require the researcher to specify prior distributions for parameter estimation as well as chain lengths in order to provide estimates from the sample distribution. Future research may include both informative and uninformative priors and different distributions (i.e. uniform instead of normal, or skewed distributions). Applied researchers may benefit from an investigation of the effects on DC based on varying chain lengths (e.g. burn-in iterations and total iterations). Similarly, a comparison of estimation techniques, such as maximum likelihood estimation, on DC may also be of interest.

Future research in the area of testlets and DC includes the use of receiver operator curve (ROC) investigations to determine if the DRM and the TRM select the same cut score that maximizes correct classification and minimizes incorrect classification. Considering most of the misclassification was false negatives it would also be of interest to use ROC to investigate the sensitivity and specificity for minimizing false negatives.

Similarly, an extension of the current study should be to investigate item difficulty and person ability in relation to the false negative misclassification.

DRM is more parsimonious than models that include discrimination and guessing parameters. Another area of future research might include replicating this study with less parsimonious models in order to investigate DC in more complex situations, such as, multidimensional TRM (i.e. the addition of problem solving to a reading assessment).

In summary, this study provides evidence that there are small but important differences between the DRM and TRM for making decisions about people when assessments include testlets. While this study adds to the growing body of literature on the effects of violating the assumption of LID by directly comparing pass/fail decisions between the DRM and the TRM, there is much more to learn about the use of testlet models in item analysis for large scale testing programs. The results of this study add to the growing support for the use of testlet models when testlets are present and may inform practitioners of testing situations that may be more sensitive to the violation of LID.

APPENDICES

APPENDIX A

Item Indicator and Description

Table A.1

PISA Item Descriptions

Testlet Indicator	Item Indicator	Description
SP	PR220Q01	READ - P2000 South Pole Q1
	PR220Q02B	READ - P2000 South Pole Q2B
	PR220Q04	READ - P2000 South Pole Q4
WL	PR412Q01	READ - P2009 World Languages Q1
	PR412Q05	READ - P2009 World Languages Q5
	PR412Q06T	READ - P2009 World Languages Q6
	PR412Q08	READ - P2009 World Languages Q8
CF	PR420Q02	READ - P2009 Children's Futures Q2
	PR420Q06	READ - P2009 Children's Futures Q6
	PR420Q09	READ - P2009 Children's Futures Q9
	PR420Q10	READ - P2009 Children's Futures Q10
AB	PR432Q01	READ - P2009 About a book Q1
	PR432Q05	READ - P2009 About a book Q5
	PR432Q06T	READ - P2009 About a book Q6
N	PR437Q01	READ - P2009 Narcissus Q1
	PR437Q06	READ - P2009 Narcissus Q6
	PR437Q07	READ - P2009 Narcissus Q7
JV	PR446Q03	READ - P2009 Job Vacancy Q3
	PR446Q06	READ - P2009 Job Vacancy Q6
FS	PR453Q01	READ - P2009 Find Summer Job Q1
	PR453Q04	READ - P2009 Find Summer Job Q4
	PR453Q05T	READ - P2009 Find Summer Job Q5
	PR453Q06	READ - P2009 Find Summer Job Q6
B	PR456Q01	READ - P2009 Biscuits Q1
	PR456Q02	READ - P2009 Biscuits Q2
	PR456Q06	READ - P2009 Biscuits Q6
WR	PR466Q02	READ - P2009 Work Right Q2
	PR466Q03T	READ - P2009 Work Right Q3
	PR466Q06	READ - P2009 Work Right Q6

Note. SP = South Pole; WL = World Languages; CF = Children's Futures; AB = About a book; N = Narcissus; JV = Job Vacancy; FS = Find Summer Job; B = Biscuits; WR = Work Right.

APPENDIX B

Example R Syntax

```
# load package sirt
library(sirt)
# number of replications
reps<-1:100
#conditions to be varied
cond<-1
#testlet variance
tvar<-.2
#sample size
N<-1400
#samplesize
TT<-4
#total number of items
TnI<-12
testlet.sim<-function(reps){
  # generate dataset
  N <- N   # number of persons
  I <- 3      # number of items per testlet
  TT <- TT    # number of testlets

  ITT <- I*TT
  # b is difficulty with a mean of 0 and sd of 1

  b <- round( rnorm( ITT , mean=0 , sd = 1 ) , 2 )
  sd0 <- 1 # sd trait
  sdt <- rep(sqrt(tvar), TT ) # sd testlets
  sdt <- sdt

  # simulate theta
  theta <- rnorm( N , sd = sd0 )
  # simulate testlets
  ut <- matrix(0,nrow=N , ncol=TT )
  for (tt in 1:TT){ ut[,tt] <- rnorm( N , sd = sdt[tt] ) }
  ut <- ut[ , rep(1:TT,each=I) ]
  # calculate response probability
  prob <- matrix( pnorm(theta + ut + matrix(b , nrow=N , ncol=ITT , byrow=TRUE ) ) ,
N, ITT)
  Y <-matrix(0,nrow=N , ncol= TnI )
```

```
 Y <- (matrix(runif(N*ITT), N , ITT) < prob )*1
  # true pass/fail from generated data set
 true<- theta
 true<-as.data.frame(true)
 true$pass<- ifelse (true$true >= -0.67, 1, 0)
# estimate Rasch model in sirt package
 rasch<- mcmc.3pno.testlet( dat=Y ,  est.slope=FALSE , est.guess=FALSE ,
                      burnin=1000, iter=2000 )
  # Rasch estimated pass/fail
 person<-rasch$person
 person$pass<- ifelse (person$EAP >= -0.67, 1, 0)

 # define testlets
 testlets <- rep(1:TT , each=I)
# estimate Rasch testlet model in sirt package
 rasch.testlet<-mcmc.3pno.testlet(Y, testlets=testlets, est.slope=FALSE,
est.guess=FALSE, burnin=1000, iter=2000 )
  # estimated testlet pass/fail
 person.testlet<-rasch.testlet$person
 person.testlet$pass<- ifelse (person.testlet$EAP >= -0.67, 1, 0)
  # combine pass fail decisions for generated data , Rasch and testlet
 ability<-as.data.frame(cbind(true$true, person$EAP, person.testlet$EAP))
 colnames(ability)<-c("true", "rasch", "testlet")

 pass<-as.data.frame(cbind(true$pass, person$pass, person.testlet$pass))
 colnames(pass)<-c("true", "rasch", "testlet")
  # create table to write out of pass/fail decisions
 rasch.table<-table(pass$true, pass$rasch)
 rasch.d<-matrix(rep(0,4), nrow=1, ncol=4)
 rasch.d[1,1]<-rasch.table[1]
 rasch.d[1,2]<-rasch.table[2]
 rasch.d[1,3]<-rasch.table[3]
 rasch.d[1,4]<-rasch.table[4]
 colnames(rasch.d)<-c("rasch.Tfail", "rasch.Fpass", "rasch.Ffail", "rasch.Tpass")

 testlet.table<-table(pass$true, pass$testlet)
 testlet.d<-matrix(rep(0,4), nrow=1, ncol=4)
 testlet.d[1,1]<-testlet.table[1]
 testlet.d[1,2]<-testlet.table[2]
 testlet.d[1,3]<-testlet.table[3]
 testlet.d[1,4]<-testlet.table[4]
 colnames(testlet.d)<-c("testlet.Tfail", "testlet.Fpass", "testlet.Ffail", "testlet.Tpass")

 rasch.testlet.table<-table(pass$testlet, pass$rasch)
 rasch.testlet.d<-matrix(rep(0,4), nrow=1, ncol=4)
 rasch.testlet.d[1,1]<-rasch.testlet.table[1]
```

```
  rasch.testlet.d[1,2]<-rasch.testlet.table[2]
  rasch.testlet.d[1,3]<-rasch.testlet.table[3]
  rasch.testlet.d[1,4]<-rasch.testlet.table[4]
  colnames(rasch.testlet.d)<-c("rasch.testlet.Tfail", "rasch.testlet.Fpass",
"rasch.testlet.Ffail", "rasch.testlet.Tpass")

  c1.all.d<-cbind(rasch.d, testlet.d, rasch.testlet.d)
c1.all.d<-as.data.frame(c1.all.d)

  write.csv(c1.all.d,
paste0("~/Desktop/Dropbox/Hodge_Dissertation/Analysis/cond1/c1_rep",reps,".csv"),
          row.names=FALSE)
}

  # run simulation above for each replication
sapply(reps, testlet.sim)
  # read back in propotion tables for each replication
setwd("~/Desktop/Dropbox/Hodge_Dissertation/Analysis/cond2")
fileList <- list.files(path= "~/Desktop/Dropbox/Hodge_Dissertation/Analysis/cond1/",
pattern=".csv")

c1temp<-as.data.frame(matrix(unlist(t(sapply(fileList, read.csv))), nrow=100))

colnames(c1temp)<-c("rasch.Tfail", "rasch.Fpass", "rasch.Ffail", "rasch.Tpass",
"testlet.Tfail", "testlet.Fpass", "testlet.Ffail", "testlet.Tpass", "rasch.testlet.Tfail",
"rasch.testlet.Fpass", "rasch.testlet.Ffail", "rasch.testlet.Tpass")
c1temp$rep = 1:nrow(c1temp)
c1temp$c<-cond
#proportion
propc1<-c1temp[,1:12]/ 1400
#mean of the proportion for rasch compared to testlet
Tf<-mean(propc2$rasch.testlet.Tfail)
Fp<-mean(propc2$rasch.testlet.Fpass)
Ff<-mean(propc2$rasch.testlet.Ffail)
Tp<-mean(propc2$rasch.testlet.Tpass)
rasch.testlet.mean<-cbind(Tf, Fp, Ff, Tp)
rasch.testlet.mean
# decision consitency of rasch compared to testlet
dc<-Tf + Tp
dc
# missclassifcation consitency
mc<-Fp + Ff
mc
# number of people misclassified
nFp<-mean(c1temp$rasch.testlet.Fpass)
 nFf<-mean(c1temp$rasch.testlet.Ffail)
```

```
 nFp + nFf

# ntestlet isnumber of testlets 1=4, 2=8, 3=12
# nitems is number of items in testlets 1= 3, 2=4, 3=7
#testletv is the tesltet variance where 0.2 is group 1 and 0.7 is group 2
# nper is percent of items in testlet 1= 100 2=.75
#nsample is sample size 1= 1600 and 2=6000
#c1temp is the data object that contains all of the reps for condition 1
#c2temp condition 2
#c9temp condition 9
#c10temp condition 10 ect

# this was done for all conditions
#group labels for c1
c1temp$ntestlet<- 1
c1temp$nitems<-1
c1temp$testletv<-1
c1temp$nper<-1
c1temp$nsample<-1
#write in file for each condition with added columns for condition
# this was done for all conditions
write.csv(c1temp,
paste0("~/Desktop/Dropbox/Hodge_Dissertation/AnalysisDissertation/Anova/c01.csv"),r
ow.names=FALSE)
#Rasch testlet by Rasch decision consistency
dc<-(c.all.anova$rasch.testlet.Tfail +
c.all.anova$rasch.testlet.Tpass)/(c.all.anova$rasch.testlet.Tfail +
c.all.anova$rasch.testlet.Fpass + c.all.anova$rasch.testlet.Ffail +
c.all.anova$rasch.testlet.Tpass)
#take a look at data before running ANOVA
# ntestlet
# nitems
# testletv
# nper
# nsample
# I manually added the dc to this csv file.
c.all.anova <-
read.csv("~/Desktop/Dropbox/Hodge_Dissertation/AnalysisDissertation/Anova/c.all.ano
va.csv")

#descriptives
describe(dc)

describeBy(dc, list(c.all.anova$nitems,c.all.anova$ntestlet))
describeBy(dc, list(c.all.anova$nitems,c.all.anova$testletv))
describeBy(dc, list(c.all.anova$nitems,c.all.anova$nper))
```

```
describeBy(dc, list(c.all.anova$nitems,c.all.anova$nsample))
describeBy(dc, list(c.all.anova$ntestlet,c.all.anova$testletv))
describeBy(dc, list(c.all.anova$ntestlet,c.all.anova$nper))
describeBy(dc, list(c.all.anova$ntestlet,c.all.anova$nsample))
describeBy(dc, list(c.all.anova$nper,c.all.anova$testletv))
describeBy(dc, list(c.all.anova$nper,c.all.anova$nsample))
describeBy(dc, list(c.all.anova$nsample,c.all.anova$testletv))

#interactions
interaction.plot(x.factor = c.all.anova$nitems, trace.factor = c.all.anova$ntestlet,
            response = dc, fun = mean, type = "b", legend = T, ylab = "dc",
            xlab = "nitems", trace.label = "ntestlet", main = "Interaction Plot",
            pch = c(1, 19))
interaction.plot(x.factor = c.all.anova$nitems, trace.factor = c.all.anova$testletv,
            response = dc, fun = mean, type = "b", legend = T, ylab = "dc",
            xlab = "nitems", trace.label = "testletv", main = "Interaction Plot",
            pch = c(1, 19))
interaction.plot(x.factor = c.all.anova$nitems, trace.factor = c.all.anova$nper,
            response = dc, fun = mean, type = "b", legend = T, ylab = "dc",
            xlab = "nitems", trace.label = "nper", main = "Interaction Plot",
            pch = c(1, 19))
interaction.plot(x.factor = c.all.anova$nitems, trace.factor = c.all.anova$nsample,
            response = dc, fun = mean, type = "b", legend = T, ylab = "dc",
            xlab = "nitems", trace.label = "nsample", main = "Interaction Plot",
            pch = c(1, 19))
interaction.plot(x.factor = c.all.anova$ntestlet, trace.factor = c.all.anova$testletv,
            response = dc, fun = mean, type = "b", legend = T, ylab = "dc",
            xlab = "ntestlet", trace.label = "testletv", main = "Interaction Plot",
            pch = c(1, 19))
interaction.plot(x.factor = c.all.anova$ntestlet, trace.factor = c.all.anova$nper,
            response = dc, fun = mean, type = "b", legend = T, ylab = "dc",
            xlab = "ntestlet", trace.label = "nper", main = "Interaction Plot",
            pch = c(1, 19))
interaction.plot(x.factor = c.all.anova$ntestlet, trace.factor = c.all.anova$nsample,
            response = dc, fun = mean, type = "b", legend = T, ylab = "dc",
            xlab = "ntestlet", trace.label = "nsample", main = "Interaction Plot",
            pch = c(1, 19))
interaction.plot(x.factor = c.all.anova$nper, trace.factor = c.all.anova$testletv,
            response = dc, fun = mean, type = "b", legend = T, ylab = "dc",
            xlab = "nper", trace.label = "testletv", main = "Interaction Plot",
            pch = c(1, 19))
interaction.plot(x.factor = c.all.anova$nper, trace.factor = c.all.anova$nsample,
            response = dc, fun = mean, type = "b", legend = T, ylab = "dc",
            xlab = "nper", trace.label = "nsample", main = "Interaction Plot",
            pch = c(1, 19))
interaction.plot(x.factor = c.all.anova$nsample, trace.factor = c.all.anova$testletv,
```

```
            response = dc, fun = mean, type = "b", legend = T, ylab = "dc",
            xlab = "nsample", trace.label = "testletv", main = "Interaction Plot",
            pch = c(1, 19))
hist(c.all.anova$dc, xlab="Decisison Consistency")
ntestlet<-factor(c.all.anova$ntestlet)
nitems<-factor(c.all.anova$nitems)
testletv<-factor(c.all.anova$testletv)
nper<-factor(c.all.anova$nper)
nsample<-factor(c.all.anova$nsample)

c.all.anova$testletv2<-factor(c.all.anova$testletv, levels= c("1", "2"), labels=c("0.2",
"0.7"))
c.all.anova$ntestlet2<-factor(c.all.anova$ntestlet, levels= c("1", "2", "3"), labels=c("4",
"8", "12"))
c.all.anova$nper2<-factor(c.all.anova$nper, levels= c("1", "2"), labels=c("100%",
"75%"))
c.all.anova$nitems2<-factor(c.all.anova$nitems, levels= c("1", "2", "3"), labels=c("3",
"4", "7"))
c.all.anova$nsample2<-factor(c.all.anova$nsample, levels= c("1", "2"), labels=c("1400",
"6000"))
describeBy(c.all.anova$dc,list(c.all.anova$nsample2,c.all.anova$ntestlet2))
data.out = aov(c.all.anova$dc ~ nitems+ntestlet+testletv+
            nper+nsample+
            nitems*ntestlet*testletv*nper*nsample )

#check residuals
qqnorm(resid(data.out), xlab = "Expected Normal Values", ylab = "Observed Values")
qqline(resid(data.out), col = "red", lwd = 3)
# Identify any observations that depart from the straight line on the
# plot
identify(qqnorm(resid(data.out)))
library(car)
qqPlot(resid(data.out), ylab = "Residuals", xlab = "Normal Quantiles",
col.lines = "red4", grid = FALSE)

hist(resid(data.out), probability=TRUE, xlab="Residuals")
 curve(dnorm(x, mean=mean(resid(data.out)), sd=sd(resid(data.out))), col="darkblue",
lwd=2, add=TRUE)

qqnorm(data.out$res)
# ANOVA estimates
summary(data.out)
# eta squared
library(lsr)
etaSquared(data.out)
#re-label factors by name
```

```
c.all.anova$testletv2<-factor(c.all.anova$testletv, levels= c("1", "2"), labels=c("0.2",
"0.7"))
c.all.anova$ntestlet2<-factor(c.all.anova$ntestlet, levels= c("1", "2", "3"), labels=c("4",
"8", "12"))
c.all.anova$nper2<-factor(c.all.anova$nper, levels= c("1", "2"), labels=c("100%",
"75%"))
c.all.anova$nitems2<-factor(c.all.anova$nitems, levels= c("1", "2", "3"), labels=c("3",
"4", "7"))
c.all.anova$nsample2<-factor(c.all.anova$nsample, levels= c("1", "2"), labels=c("1400",
"6000"))

#boxplots for conditions with effect size
boxplot(dc ~ testletv2, data = c.all.anova, ylab = "DC",
  main = "Testlet Variance")

boxplot(dc ~ ntestlet2*nper2, data = c.all.anova, ylab = "DC",
  main = " Number of Tesltets by Percent Total Items")

boxplot(dc ~ ntestlet2*testletv2, data = c.all.anova, ylab = "DC",
  main = "  Number of testlets by Testlet Variance")

boxplot(dc ~ nitems2* nper2, data = c.all.anova, ylab = "DC",
  main = "Number of Items by Percent Total Items ")

boxplot(dc ~ ntestlet2*testletv2*nper2, data = c.all.anova, ylab = "DC",
  main = "Testlets by Testlet Variance by Percent Total Items ")

boxplot(dc ~ nper2, data = c.all.anova, ylab = "DC",
  main = "Percent Total Items in Testlets")


#Mean and SD for all main and two-way interactions
aggregate(c.all.anova$dc~c.all.anova$testletv2, FUN="mean", digits=4)
describeBy(c.all.anova$dc,c.all.anova$testletv2)

aggregate(c.all.anova$dc~c.all.anova$ntestlet2, FUN="mean", digits=4)
describeBy(c.all.anova$dc,c.all.anova$ntestlet2)

aggregate(c.all.anova$dc~c.all.anova$testletv2+c.all.anova$nper2, FUN="mean",
digits=4)
describeBy(c.all.anova$dc,list(c.all.anova$testletv2,c.all.anova$nper2))

aggregate(c.all.anova$dc~c.all.anova$nper2, FUN="mean", digits=4)
describeBy(c.all.anova$dc,c.all.anova$nper2)
aggregate(c.all.anova$dc~c.all.anova$nper2+c.all.anova$ntestlet2, FUN="mean",
digits=4)
```

```
describeBy(c.all.anova$dc,list(c.all.anova$nper2, c.all.anova$ntestlet2))

aggregate(c.all.anova$dc~c.all.anova$testletv2+c.all.anova$ntestlet2, FUN="mean",
digits=4)
describeBy(c.all.anova$dc,list(c.all.anova$testletv2,c.all.anova$ntestlet2))

aggregate(c.all.anova$dc~c.all.anova$nitems2, FUN="mean", digits=4)
describeBy(c.all.anova$dc,c.all.anova$nitems2)

aggregate(c.all.anova$dc~c.all.anova$nitems2+ c.all.anova$ntestlet2, FUN="mean",
digits=4)
describeBy(c.all.anova$dc,list(c.all.anova$nitems2,c.all.anova$ntestlet2))

aggregate(c.all.anova$dc~c.all.anova$nsample2, FUN="mean", digits=4)
describeBy(c.all.anova$dc,c.all.anova$nsample2)

aggregate(c.all.anova$dc~c.all.anova$nsample2+ c.all.anova$ntestlet2, FUN="mean",
digits=4)


#mcmc objects for diagnostics
#this was done for all conditions

# MCMC summary this summary for posterior
summary(rasch)
summary(rasch.testlet)
# trace plots
#pdf("~/Desktop/Dropbox/Hodge_Dissertation/AnalysisDissertation/
#RaschMCMCplotCondition1rep1.pdf")
 plot(rasch)
 #dev.off()

#pdf("~/Desktop/Dropbox/Hodge_Dissertation/AnalysisDissertation/
#TestletMCMCplotCondition1rep1.pdf")
plot(rasch.testlet)
#dev.off()

# autocorralation plots
autocorr.plot(rasch$mcmcobj)
autocorr.plot(rasch.testlet$mcmcobj)

# geweke z score and plots
geweke.diag(rasch$mcmcobj,frac1=0.1, frac2=0.5 )
geweke.plot(rasch$mcmcobj,frac1=0.1, frac2=0.5 )
geweke.diag(rasch.testlet$mcmcobj,frac1=0.1, frac2=0.5 )
geweke.plot(rasch.testlet$mcmcobj,frac1=0.1, frac2=0.5 )
```

```
# Raferty and Lewis
raftery.diag(rasch$mcmcobj, q = 0.025, r = 0.025, s = 0.95)
raftery.diag(rasch.testlet$mcmcobj, q = 0.025, r = 0.025, s = 0.95)
#M: number of burn-ins necessary
# N: number of iterations necessary in the Markov chain
# Nmin: minimum number of iterations for the “pilot” sampler
# I: dependence factor, interpreted as the proportional increase
#in the number of iterations attributable to serial dependence.
#High dependence factors (> 5) are worrisome and may be due
#to influential starting values, high correlations between
#coefficients, or poor mixing.

# Heidelberg and Welch Diagnostic
heidel.diag(rasch$mcmcobj)
heidel.diag(rasch.testlet$mcmcobj)
#If the chain passes the first part of the diagnostic, then it takes the
#part of the chain not discarded from the first part to test the
#second part.
#The halfwidth test calculates half the width of the (1 − )%
#credible interval around the mean.
#If the ratio of the halfwidth and the mean is lower than some ,
#then the chain passes the test. Otherwise, the chain must be run
#out longer.
```

REFERENCES

Adams, R. J., Wilson, M., & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*(1), 1-23. doi: 10.1177/0146621697211001

AERA, APA, & NCME (1999). *Standards for educational and psychological testing*. Washington, D.C.: Author.

Bandalos, D.L., & Leite, W., (2013). Use of Monte Carlo studies in structural equation modeling research. In G.R. Hancock & R.O. Mueller (Eds.), *Structural equation modeling a second course (2nd ed.*, pp. 625-666). Information Age Publishing, Inc. Charlotte, NC.

Bao, H., Dayton, C. M., & Hendrickson, A. B. (2009). Differential item functioning amplification and cancellation in a reading test. *Practical Assessment, Research & Evaluation*, *14*(19), Available online: http://pareonline.net/getvn.asp?v=14&n=19.

Bond, T. G & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences (2nd ed.)*. Hillsdale, N.J: Lawrence Erlbaum Associates.

Boomsma, A. (2013). Reporting Monte Carlo studies in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 20*(3), 518-540. doi: 10.1080/10705511.2013.797839

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64*(2), 153-168. doi: 10.1007/BF02294533

Chen J. (2014) *Model selection for irt equating of testlet-based tests in the random groups design* (dissertation). ProQuest, UMI Dissertations Publishing (3680050).

Crocker, L., & Algina, J. (2006, 2008). *Introduction to classical and modern test theory*. Mason, Ohio: Cengage Learning.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford Press.

De Champlain, A. F. (2010). A primer on classical test theory and item response theory for assessments in medical education. *Medical Education, 44(1),* 109-117. doi: 10.1111/j.1365-2923.2009.03425.x.

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

Dickenson, T. S. (2005). *Comparison of various ability estimates to the composite ability best measured by the total test score.* (Order No. 3181941, University of South Carolina). *ProQuest Dissertations and Theses,* 161-161 p. Retrieved from http://ezproxy.baylor.edu/login?url=http://search.proquest.com/docview/305414375?accountid=7014. (305414375).

Eckes, T. (2013). Examining testlet effects in the TestDaF listening section: A testlet response theory modeling approach. *Language Testing*, *31*(1), 39-61. doi: 10.1177/0265532213492969.

Fan, X. (2012). Designing simulation studies. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf & K. J. Sher (Eds.), *APA handbook of research methods in psychology (Vol. 2.): Data analysis and research publication* (pp. 427-444). Washington, DC: American Psychological Association.

Fountas I.C., Pinnell, G.S., (2012) *Fountas and Pinnell Benchmark Assessment System. Heinemann*. Retrieved from http://www.heinemann.com/-fountasandpinnell/reading-assessment.aspx

Gill, J. (2014). *Bayesian methods: A social and behavioral sciences approach* (3rd Ed.). CRC press. New York, NY.

Glas, C. A. W. (2012). *Estimating and testing the extended testlet model: LSAC Research Report Series.* Law School Admission Council. Retrieved from http://www.lsac.org/docs/default-source/research-(lsac-resources)/rr-12-03.pdf

Good, R. H., Wallin, J., Simmons, D. C., Kame'enui, E. J., & Kaminski, R. A. (2002). *System-wide percentilerRanks for DIBELS benchmark assessment (Technical Report 9).* Eugene, OR: University of Oregon. Retrieved from https://dibels.uoregon.edu/docs/techreports/DIBELS_Percentiles.pdf

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied measurement in education*, *15*(3), 309-333.

Harcourt (2003). *A summary report of the instructional effectiveness of the PMs collection*. Harcourt Education Company. Retrieved from http://rigby.hmhco.com/en/rigbyPM_home.htm

Harwell, M., Stone, C. A., Hsu, T.-C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement, 20*(2), 101-125. doi: 10.1177/014662169602000201

Harwell, M. R. (1997). Analyzing the results of Monte Carlo studies in item response theory. *Educational and Psychological Measurement, 57*(2), 266-279. doi: 10.1177/0013164497057002006

Hembry, I.F. (2014) *Operational characteristics of mixed-format multistage tests using the 3PL testlet response theory model* (dissertation). ProQuest,UMI Dissertations Publishing (3691396).

Holzinger, K. J., & Harman, H. H. (1939). Chapter XIII: Factor Analysis. *Review of Educational Research*, *9*(5), 528-531. http://www.jstore.org/stable/11677754

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika*, *2*(1), 41-54.

Hooker, G., & Finkelman, M. D. (2010). Paradoxical results of item bundles. *Psychometrika*, *75*(2), 249-271. doi: 10.1007/s11336-009-9143-Y.

Jiao, H., Wang, S., & He, W. (2013). Estimation methods for one-parameter testlet models. *Journal of Educational Measurement*, *50*(2), 186-203. doi: 10.1111/jedm.12010

Kastberg, D., Roey, S., Lemanski, N., Chan, J.Y., & Murray, G. (2014). *Technical Report and User Guide for the Program for International Student Assessment (PISA).* (NCES 2014-025). U.S. Department of Education. Washington, DC: National Center for Education Statistics. Retrieved from http://nces.ed.gov/pubsearch.

Kim, J.S., & Bolt, D. M. (2007). Estimating item response theory models using Markov Chain Monte Carlo methods. *Educational Measurement: Issues and Practice, 26*(4), 38-51. doi: 10.1111/j.1745-3992.2007.00107.x

Linacre, J. M. (2004). Estimation methods for Rasch measures. In Smith, E.V., & Smith, R.M. (Eds.) *Introduction to Rasch measurement* (pp.25-47). Maple Grove, MN: JAM Press.

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

Lu, R. (2010). *Impacts of local item dependence of testlet items with the multistage tests for pass-fail decisions academic* (dissertation). ProQuest, UMI Dissertations Publishing (3443478).

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). Washington, DC: American Council on Education / Macmillan.

Miller, M. D., Linn, R. L., & Gronlund, N. E. (2009). *Measurement and assessment in teaching (10th ed.)*. New York, NY: Merrill.

National Center for Education Statistics (2014). *An Introduction to National Assessment of Educational Progress* (NAEP). Retrieved from http://nces.ed.gov/nationsreportcard/pdf/parents/2010468.pdf

Nitko, A. J., & Brookhart, S. M. (2006). Educational Assessment of Students (5th Edition). Prentice-Hall, Inc. Upper Saddle River, NJ.

OECD (2013), *PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy*, PISA, OECD Publishing. DOI: 10.1787/9789264190511-en

Paek, I., Yon, H., Wilson, M., & Kang, T. (2008). Random parameter structure and the testlet model: extension of the Rasch testlet model. *Journal of Applied Measurement*, *10*(4), 394-407.

Patz, R. J., & Junker, B. W. (1999). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, *24*(2), 146-178.

Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., & Chen, F. (2001). Monte Carlo experiments: Design and implementation. *Structural Equation Modeling: A Multidisciplinary Journal, 8*(2), 287 - 312. doi: 10.1207/S15328007SEM0802_7

Pearson (2009). *DRA K-8 technical manual*. *Developmental Reading Assessment* 2nd ed. Pearson Education, Upper Saddle River, NJ. Retrieved from http://www.pearsonschool.com/index.cfm?locator=-PSZ4Z4&PMDBSUBCATEGORYID=&PMDBSITEID=2781&PMDBSUBSOLUTIONID=&PMDBSOLUTIONID=&PMDBSUBJECTAREAID=&PMDBCATEGORYID=&PMDbProgramID=23662

Pierce, C. A., Block, R. A., & Aguinis, H. (2004). Cautionary note on reporting eta-squared values from multifactor ANOVA designs. *Educational and Psychological Measurement*, *64*(6), 916-924. doi: 10.1177/0013164404264848

R Development Core Team (2014). R: A language and environment for statistical computing 3.2.1. *R Foundation for Statistical Computing*, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, *47*(5), 667-696.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. *Psychological bulletin*, *114*(3), 552.

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

Robitzsch, A., (2014). *Supplementary Item Response Theory Models: SIRT V1.1 User's Manual.* http://cran.r-project.org/web/packages/sirt/

Smarter Balanced Assessment Consortium (2014). *Smarter Balanced Assessments*. Retrieved from http://www.smarterbalanced.org/smarter-balanced-assessments/

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, *28*(3), 237-247.

Spearman, C. (1904). " General Intelligence," objectively determined and measured. *The American Journal of Psychology*, *15*(2), 201-292.

Spearman, C. (1927). *The abilities of man*. New York, NY: Cambridge

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (5th ed.). Upper Saddle River, NJ: Pearson Allyn & Bacon.

Texas Education Agency (2014). *STAAR Released Test Questions*. Retrieved on November 14, 2014 from http://tea.texas.gov/-Student_Testing_and_Accountability/Testing/State_of_Texas_Assessments_of_Academic_Readiness_(STAAR)/STAAR_Released_Test_Questions/

Texas Institute for Measurement, Evaluation, and Statistics (2010). *Technical report Texas Primary Reading Inventory*. Children's Learning Institute University of Texas- Houston Health Science Center. Retrieved on December 19, 2012 from http://www.tpri.org/resources/documents/20102014TechnicalReport.pdf

Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: A use of multiple categorical response models. *Journal of Educational Measurement*, *26*(3), 247-260. doi: 10.1111/j.1745-3984.1989.tb00331.x.

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

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, *24*(3), 185-201.

Wainer, H., & Lewis, C. (1990). Toward a psychometrics for testlets. *Journal of Educational Measurement*, *27*(1), 1-14.

Wang, X., Bradlow, E. T., & Wainer, H. (2002). A general Bayesian model for testlets: Theory and applications. *Applied Psychological Measurement, 26(1),* 109-128. doi: 10.1177/0146621602026001007.

Wang, W. C., & Wilson, M. (2005a). The Rasch testlet model. *Applied Psychological Measurement*, 29(2), 126-149. doi: 10.1177/0146621604271053.

Wang, W. C., & Wilson, M. (2005b). Assessment of differential item functioning in testlet-based items using the Rasch testlet model. *Educational and Psychological Measurement*, *65*(4), 549-576. doi: 10.1177/0013164404268677.

Wright, B. D., & Mok, M. M. (2004). An overview of the family of Rasch measurement models. In E.V. Smith & R.M. Smith (Eds.) *Introduction to Rasch measurement* (pp. 1-24). Maple Grove, MN: JAM Press.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*(3), 187-213.

Yung, Y. F., Thissen, D., & McLeod, L. D. (1999). On the relationship between the higher-order factor model and the hierarchical factor model. *Psychometrika*, *64*(2), 113-128.

Zhang, B. (2010). Assessing the accuracy and consistency of language proficiency classification under competing measurement models. *Language Testing*, *27*(1), 119-140. doi: 10.1177/0265532209347363.

Zumbo, B. D., & Rupp, A. A. (2004). Responsible Modeling of Measurement Data for Appropriate Inferences. In D. Kaplan (Ed). *The SAGE handbook of quantitative methodology for the social sciences* (pp.73-92). Thousand Oaks, CA: Sage Publications, Inc.