ABSTRACT

Title of Document: CROSS-CLASSIFIED MODELING OF DUAL LOCAL ITEM DEPENDENCE

Chao Xie, Doctor of Philosophy, 2014

Directed By: Associate Professor, Hong Jiao
Measurement, Statistics and Evaluation
Department of Human Development and Quantitative Methodology

Previous studies have mainly focused on investigating one source of local item dependence (LID). However, in some cases, such as scenario-based science assessments, LID might be caused by two possible sources simultaneously. In this study, such kind of LID that is caused by two factors simultaneously is named as dual local item dependence (DLID).

This study proposed a cross-classified model to account for DLID. Two simulation studies were conducted with the primary purpose of evaluating the performance of the proposed cross-classified model. Data sets with DLID were simulated with both testlet effects and content clustering effects. The second purpose of this study was to investigate the potential factors affecting the need to use the more complex cross-classified modeling of DLID over the simplified multilevel modeling of LID by ignoring cross-classification structure. For both simulation studies, five

factors were manipulated, including sample size, number of testlets, testlet length, magnitude of the testlet effects represented by standard deviations (SDs), and magnitude of the content clustering effects represented by SDs. The difference between the two simulation studies was that, simulation study 1 constrained the SDs of the testlet effects and content clustering effects as the same across testlets and content areas, respectively; simulation study 2 released this constraint by having mixed SDs of the testlet effects and mixed SDs of the content clustering effects.

Results of both simulation studies indicated that the proposed cross-classified model yielded more accurate parameter recovery, including item difficulty, persons' ability, and random effects' SD parameters with smaller estimation errors than the two multilevel models and the Rasch model which ignored one or both item clustering effects. The two manipulated variables, the magnitude of the testlet effects and the magnitude of the content clustering effects, determined the necessity of using the more complex cross-classified model over the simplified multilevel models and the Rasch model: the larger the magnitude of the testlet effects and the content clustering effects, the more necessary to use the proposed cross-classified model. Limitations are discussed and suggestions for future research are presented at the end.

CROSS-CLASSIFIED MODELING OF DUAL LOCAL ITEM DEPENDENCE

By

Chao Xie

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2014

Advisory Committee:
Associate Professor Hong Jiao, Chair/Advisor
Associate Professor, Jeffrey R. Harring
Professor, Robert W. Lissitz
Professor, George B. Macready
Associate Professor, Laura M. Stapleton
Professor, Margaret J. McLaughlin, Dean's Representative

Dedication

This dissertation is dedicated to my husband Guangwen Zhu, who was as much a part of my doctoral experience as the research study itself. A special dedication to my parents, without their support I would never have achieved this academic milestone. Finally, I would like to especially thank my mother-in-law; I would not be able to finish my dissertation if she were not here to help me take care of my baby.

Acknowledgements

I would like first to acknowledge the help and support of my advisor, Dr. Hong Jiao. Her guidance during my years in the EDMS program was crucial for shaping my research. Thank her for encouraging me, pushing me, and challenging me no matter during the dissertation process but also during all my life in the EDMS program.

Thanks to the members of my dissertation committee: Dr. Lissitz, Dr. Macready, Dr. Stapleton, Dr. Harring, and Dr. McLaughlin. I believe that my dissertation has been made much better because of their helpful suggestions and insightful comments.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

This chapter introduces some background information about the formulation and development of the proposed model in this study. It describes the research purpose and research questions, and addresses the significance of the proposed study.

Background

Item response theory (IRT) models are broadly used in social sciences, especially the field of education, to measure persons' latent trait or ability based on item responses. The probability of answering an item correctly is modeled as a mathematical function of the person's ability and the item parameters. The main advantage of using IRT models over classical test theories (CTT) is the invariance property of IRT item/person parameters (Hambleton & Swaminathan, 1985; Embretson & Reise, 2000).

However, IRT requires stronger assumptions than CTT. One of the underlying fundamental assumptions is local item independence, which means that the probability of responding to one item correctly does not influence the probability of answering other items correctly controlling for ability (Hambleton & Swaminathan, 1985; Embretson & Reise, 2000). Previous studies indicated that ignoring the violation of local item independence assumption might have negative impacts, e.g. inaccurate estimation of both item and person parameters, over-estimation of test reliability, and equating errors (e.g., Ackerman, 1987; Chen & Thissen, 1997; Sireci, Thissen, & Wainer, 1991; Wainer, 1995; Wainer & Thissen, 1996; Tuerlinckx & De Boeck, 2001; Yen, 1984).

The violation of local item independence assumption, also called local item dependence (LID) problem, can be caused by a variety of factors. Thissen, Bender, Chen, Hayashi, and Wiesen (1992) classified LID into two categories based on the causes of LID: one is underlying local dependence (ULD) and the other is surface local dependence (SLD). The former category, ULD, mainly refers to LID where items share the same stimulus; for example, in reading tests, some items share the same reading passage. The second category, SLD, mainly refers to the similarity in item responses caused by speededness or content similarity; for example, examinees might omit some items at the end of a long test, then examinees' responses to these omitted items are similar.

Many current standardized educational tests contain items based on a common stimulus. Such a cluster of items that share a common stimulus is often referred to as a testlet (e.g., Thissen, Steinberg, & Mooney, 1989; Wainer & Kiely, 1987). Several response models have been proposed to account for LID within a testlet. Lu (2010) divided these models into two categories: one category is modeling the testlet effect as a second dimension (e.g., Reckase, 1997), and the other category is adding a random variable into a standard IRT model to account for the testlet effect, like the Rasch testlet model (Wang & Wilson, 2005), the two-parameter logistic testlet model (Bradlow, Wainer, & Wang, 1999), and the three-parameter logistic testlet model (Wainer, Bradlow, & Du, 2000).

Over the past two decades or so there has been an increasing interest in fitting multilevel/hierarchical models to large datasets in various fields, including education, social and behavioral sciences, psychology, and medical studies (Raudenbush &

Bryk, 2002). The multilevel modeling technique has attracted the interest of many educational and social researchers for handling clustered/nested data structures. Specifically, studies (e.g. Adams, Wilson, & Wu, 1997; Beretvas & Williams, 2004; Fox & Glas, 2001; Kamata, 1998, 2001) have shown that regular IRT models could be formulated as hierarchical generalized linear models (HGLM), in which item are treated as nested within people. Those reparameterized IRT models in HGLM framework are referred to as multilevel measurement models (MMMs) (Beretvas & Kamata, 2005).

From the multilevel modeling perspective, studies (Jiao, Wang, & Kamata, 2005; Beretvas & Walker, 2012) have extended the MMMs to handle testlet-based LID (hereafter termed as MMMT). The hierarchy of the MMMT model proposed by Jiao et al. (2005) is that items (Level 1) are nested within testlets (Level 2), which are then nested within persons (Level 3). Different from Jiao et al.'s three-level MMMT model, Beretvas and Walker (2012) suggested instead a two-level MMMT model, in which item are nested within persons, the scores are modeled as a function of both item and testlet difficulties, and the testlet-specific dependencies are modeled using dummy-coded testlet indicator variables at Level-1.

The advantages of using a multilevel parameterization of testlets are similar with the advantages of using multilevel models in statistics. The primary benefit is the capability to account for the dependences from higher levels of clustering. For example, Jiao, Kamata, Wang, and Jin (2012) proposed a four-level IRT model in which person clustering is also accounted for in addition to modeling the testlet-specific LID. The second benefit is that variables could be added to the appropriate

levels to model impact, differential item functioning (DIF), and differential testlet functioning (DTF) (Beretvas & Walker, 2012).

However, researchers have found that many data sets have more complex non-hierarchical structures. One such complexity involves cross-classified data structures that cannot be handled by the hierarchical linear modeling techniques. In cross-sectional studies, one illuminating example of cross-classified data structure is given by Goldstein (2003), i.e. students are cross-classified by the schools they attend and the neighborhoods they live in. In longitudinal studies, non-hierarchical structure occurs when, for example, students change schools overtime. In this case, occasions are cross-classified by students and schools. To model data with a cross-classified structure, cross-classified random effects modeling (CCREM) techniques have been developed to accommodate non-nested factors (Goldstein, 1986, 2003; Raudenbush, 1993; Rasbash & Goldstein, 1994).

Such non-strict hierarchical data structures could also exist in assessments. For example, in scenario-based science assessments, the test usually covers multiple subject areas, like physical science, life science, earth and space science, science and technology, science in personal and social perspectives, history and nature of science. In this case, LID could be caused by two sources simultaneously: one is the testlet effect from scenarios and the other is the content clustering effect due to coverage of multiple content areas. Appendix A shows an example of scenario-based science test with coverage of multiple content areas. In this study, such LID from two sources is referred to as dual local item dependence (DLID).

Since testlets (or scenarios) are not nested within content areas nor vice-versa, the two are said to be cross-classified. This study proposed a cross-classified measurement model to account for such DLID as existing in scenario-based science assessments. The structure of this proposed model is that items (Level 1) are cross-classified by testlets and content areas (Level 2), and both testlets and content areas are nested within persons (Level 3).

In addition to the scenario-based science assessment described above, another example of DLID is that LID might come from testlet and subskills, sub-content domains, content strands or clusters simultaneously. Take TOEFL Reading as an example, there are 3 to 5 passages (testlets) in the reading test, and each passage contains 12 to 14 multiple-choice questions, which generally belong to one of the following subskills: detail/fact, vocabulary, reference questions, and summary. Appendix B shows an example of a TOEFL Reading passage followed by several items assessing different skills. In this case, the structure of the item responses is that each item is cross-classified by testlets (passages) and skills, which are nested within persons. In short, it is common that LID might be caused by more than one single factor.

Research Purpose

The main objective of this study was to formulate a cross-classified model to deal with the DLID issue. It demonstrated that the proposed cross-classified modeling of DLID is algebraically equivalent with a constrained version of the testlet model accounting for two types of LID proposed by Jiao and her colleagues (Jiao, Wang, Wan, & Lu, 2009). Two simulation studies were conducted with the primary purpose

of evaluating the performance of the proposed cross-classified modeling of DLID. Data sets with DLID were simulated with both testlet effects and content clustering effects.

Previous research in the field of statistics (Luo, 2007; Meyers, 2004; Meyers & Beretvas, 2006; Ren, 2011) carried out simulation studies to investigate the impact of ignoring cross-classification on model fit and parameter estimates, including fixed and random. Thus, a secondary purpose of this study was to extend previous research to the field of measurement by investigating the potential factors affecting the need to use the more complex cross-classified modeling of DLID over the simplified multilevel modeling of LID by ignoring cross-classification structure.

Research Questions

This study was designed to address the following research questions:

1. How are the item and person parameter estimates affected when ignoring the effects of testlets and/or content areas versus correctly modeling the two effects via the proposed cross-classified models?
2. Which manipulated factors, including sample size, number of testlets, testlet length, magnitude of the testlet effects, and magnitude of the content clustering effects, influence the estimates of the model parameters? How is the significant effect represented?
3. Which model fit index performs well in correctly identifying the proposed cross-classified model as the best fitting model under different simulation conditions?

To answer these research questions, two simulation studies were conducted. For both simulation studies, five factors were manipulated, including sample size, number of testlets, testlet length, magnitude of the testlet effects, and magnitude of the content clustering effects. The difference between the two simulation studies was that, simulation study 1 constrained the testlet effects' SDs as well as the content clustering effects' SDs as the same across the testlets and content areas, respectively; simulation study 2 released this constraint by having mixed SDs of the testlet effects and mixed SDs of the content clustering effects.

Significance of the Study

The problem of DLID could exist in many contexts, like the scenario-based science assessment and the TOEFL Reading assessment. However, little research has been conducted to explore how to deal with the issue of DLID and investigate the impact of ignoring one source of LID. The only introductory investigation was conducted by Jiao et al. (2009). Therefore, this study contributed to the literature on LID and provided empirical evidence about the impact of DLID.

In addition, this methodological study tried to deal with the issue of DLID from the cross-classified modeling perspective, which was an extension based on the multilevel measurement models and multilevel testlet models. Raudenbush (1993) claimed that, in practice, there are almost no purely nested data structures. Therefore, in reality, the cross-classified modeling should be better reflective of the real data structure than the multilevel modeling. Thus, the proposed model is more generalized and flexible in dealing with complex LID issues than the current multilevel models.

Chapter 2: Literature Review

The first three sections in this chapter capture the formulation process of the proposed model in this study. Specifically, this chapter demonstrates how the proposed model is evolved from IRT to testlet response theory (TRT) models, then to multilevel measurement models for testlets (MMMT), and finally to the proposed cross-classified modeling of DLID. Section 1 presents a brief review of the three commonly used dichotomous IRT models and model assumptions, especially the assumption of local item independence; then it presents how the TRT models are formulated based on the three dichotomous IRT models. The second section demonstrates how IRT models and TRT models described in Section 1 are parameterized from a multilevel modeling perspective. The third section first provides the cross-classified parameterization of IRT models and then proposes the cross-classified modeling of DLID.

Item Response Theory and Testlet Response Theory Models

Item Response Theory Models and Assumptions

Models

Based on IRT models, the probability of answering an item correctly is modeled as a mathematical function of the person's ability and the item parameters. According to how the items are scored, IRT models are divided into two categories: dichotomous IRT models with two response categories, and polytomous IRT models with multiple score categories. In this study, since only dichotomous models are used as base models to formulate testlet response theory (TRT) models as well as the

corresponding multilevel and/or cross-classified parameterizations, a brief introduction to the three commonly used dichotomous IRT models, including the Rasch (Rasch; 1960), the two-parameter logistic (2PL; Birnbaum, 1968), and the three-parameter logistic (3PL; Birnbaum, 1968) models, are provided here.

The Rasch, 2PL and 3PL models employ 1, 2 and 3 item parameters respectively to characterize the item response functions. Among them, the most generalized 3PL model is represented as follows:

$$p_{ij}(X_{ij}=1|\theta_j,a_i,b_i,c_i)=c_i+\frac{1-c_i}{1+\exp[-a_i(\theta_j-b_i)]}, \quad (1)$$

where p_{ij} is the probability for person *j* answering item *i* correctly. θ_j represents person *j*'s ability. b_i represents item *i*'s difficulty, which corresponds to the point on the ability continuum at which a person has a $\frac{1+c_i}{2}$ probability of getting a correct response. a_i designates the item discrimination, where the larger the value of a_i, the more discriminating of the item in separating examinees at the difficulty level, b_i, of the item (Hambleton, Swaminathan & Rogers, 1991). c_i indicates guessing parameters, which corresponds to the lower asymptote of the item characteristic curves (ICCs) (Embretson & Reise, 2000).

The 2PL model is a constrained version of the 3PL model by setting the lower asymptote as 0:

$$p_{ij}(X_{ij}=1|\theta_j,a_i,b_i)=\frac{1}{1+\exp[-a_i(\theta_j-b_i)]}, \quad (2)$$

where the difficulty parameter b_i here corresponds to the point on the ability continuum where the probability of getting a correct response is 0.5.

The Rasch model assumes that all items share a common discrimination parameter 1 in addition to the assumption of zero probability of guessing:

$$p_{ij}(X_{ij}=1|\theta_j,b_i)=\frac{1}{1+\exp[-(\theta_j-b_i)]}, \quad (3)$$

where the only item parameter in the Rasch model is the item difficulty parameter b_i, which has the same interpretation as in the 2PL model.

Assumptions

There are three key assumptions underlying an IRT model, including dimensionality, monotonicity, and local independence (Hambleton, 1989). Dimensionality designates the number of latent trait that the test items intend to measure. Even though multidimensional IRT models have been developed and discussed (e.g., Reckase, 1997, 2009), unidimensional IRT models have generally been used by many testing programs (Kolen & Brennan, 2004). Also, since the focus of this study is unidimensional IRT models, unidimensionality should be assumed here, which means that a single latent trait is assumed to underlie item performance (Hambleton et al., 1991).

The second assumption, monotonicity, relates to that the mathematical function that describes the relationship between the probability of correctly responding to an item and the latent trait is monotonically increasing, that is, as the latent trait becomes higher, the probability of getting a correct response becomes higher.

The implication of the local independence assumption constitutes two parts: local item independence and local person independence. Local independence is obtained when the relationship among items (or persons) is fully characterized by the

IRT model (Embretson & Reise, 2000). Specifically, local item independence (Lord & Novick, 1968) means that a person's response to one item does not influence his/her responses to other items; mathematically, it can be represented as (Lord, 1980):

$$p(U = u \mid \theta) = \prod_{i=1}^{I} p(u_i \mid \theta) = p(u_1 \mid \theta)\, p(u_2 \mid \theta) \cdots p(u_I \mid \theta)\,. \tag{4}$$

It indicates that the probability of a response pattern, *u*, for a person with latent trait of θ, $p(U = u \mid \theta)$, is the product of the probabilities of the individual responses, u_i, to the *i*th item on a test, $p(u_i \mid \theta)$.

The local item independence assumption is related to the dimensionality assumption: local item independence can be achieved for both unidimensional data and multidimensional data as long as the IRT model contains person parameters for each dimension of latent traits underlying item performance (Embretson & Reise, 2000). In addition, if the IRT model contains person parameters on only one dimension, the unidimensionality assumption holds when the local item independence assumption is achieved.

Local person independence can be mathematically represented by:

$$p(U_i = u_i \mid \boldsymbol{\theta}) = \prod_{j=1}^{n} p(u_{ij} \mid \theta_j) = p(u_{i1} \mid \theta_1)\, p(u_{i2} \mid \theta_2) \cdots p(u_{in} \mid \theta_n)\,. \tag{5}$$

It indicates that the probability of the response to a single item, *i*, by *n* persons with abilities θ_j in the vector $\boldsymbol{\theta}$, is the product of the probabilities of each person *j*'s (*j* =1, 2,…, *n*) response to the *i*th item, $p(u_{ij} \mid \theta_j)$. When local person independence holds,

one person's response to the item will not be associated with another person's response to the same item.

The violation of local item independence assumption is referred to as local item dependence (LID), which is the focus of the current study. Therefore, the following section presents a more detailed description of LID, including the causes and impacts.

Local Item Dependence

Previous studies indicated that ignoring the violation of local item independence assumption might have negative impacts such as, inaccurate estimation of both item and person parameters, over-estimation of test reliability, and equating errors (e.g., Ackerman, 1987; Chen & Thissen, 1997; Sireci et al., 1991; Wainer, 1995; Wainer & Thissen, 1996; Tuerlinckx & De Boeck, 2001; Yen, 1984). Ackerman (1987) found that when LID exists, item discrimination parameters tend to be overestimated, item difficulty parameter tend to become homogenous, and ability estimates are affected as the degree of dependency increased. Sireci et al. (1991) showed that the estimates of reliabilities are substantially overestimated when not accounting for the testlet structure on two reading comprehension tests. Further, Yen (1984) demonstrated substantial unsystematic errors of equating tests with LID.

Yen (1993) stated that "the basic principle involved in producing LID is that there is an additional factor that consistently affects the performance of some students on some items to a greater extent than others" (p. 188). She listed and described 10 possible causes of LID, including external assistance or interference, speededness, fatigue, practice, item or response format, passage dependence, item chaining,

explanation of previous answer, scoring rubrics or raters, and content knowledge or abilities.

Some researchers have classified LID based on the types of causes. Hoskens and De Boeck (1997) divided LID into two categories: order dependency and combination dependency. Order dependency means the response to early items affects the responses to subsequent items; thus, some of the causes in Yen's list, like item chaining and explanation of previous answer, are consistent with this category. Combination dependence refers to items that share the same stimulus content, and in Yen's list, passage dependence is a good example of this.

Thissen et al. (1992) classified LID into underlying local dependence and surface local dependence. The former category assumes that each set of locally dependent items share a common trait that is not shared by the rest of the items; this is similar with passage dependence listed by Yen. The latter category means that examinees tend to give identical answers to similar items like in speeded tests.

Among the causes listed above, passage dependence, or a cluster of items by common stimuli, is a common source of LID and have been explored broadly (e.g., Ferrara, Huynh, & Baghi, 1997; Ferrara, Huynh, & Michaels, 1999; Lee, 2004; Sireci et al., 1991; Thissen et al., 1989). Such a cluster of items that share a common stimulus is often referred to as a testlet (e.g., Thissen et al., 1989; Wainer & Kiely, 1987). To account for LID within a testlet, various models have been proposed, which will be reviewed in details in the following section.

Testlet Response Theory Models

Testlets are broadly used in educational tests. For example, in reading tests, examinees might be presented a passage and a bundle of items related to that passage; in math tests, several items may depend on the same data table; and science tests commonly use a graph as the central stimulus for a set of items. The section above illustrates that, since testlet is one source of LID, fitting standard IRT models to testlet responses will result in negative impacts. Therefore, more complex models have been formulated to account for the effect of testlets based on the standard IRT models. In literature, these models have been generally referred to as testlet response theory (TRT) models.

2PL-TRT Model

Bradlow et al. (1999) formulated the two-parameter testlet response theory (2PL-TRT) model by adding a random effect to the standard 2PL IRT model (Equation 2):

$$p_{ij}(X_{ij}=1\,|\,\theta_j, a_i, b_i, \gamma_{jd(i)}) = \frac{1}{1+\exp[-a_i(\theta_j - b_i + \gamma_{jd(i)})]}, \tag{6}$$

where p_{ij} is the probability of correctly responding to item *i* nested within testlet *d* for person *j* with ability θ_j. The parameter $\gamma_{jd(i)}$ is interpreted as a person-specific testlet effect, which is the same for all items within a testlet for a particular examinee *j* and is modeled as independent of ability and item parameters. The variance of the testlet effect is constrained to be the same across all testlets within a test, that is, $\gamma_{jd(i)} \sim N(0, \sigma^2)$. All the other parameters have the same interpretations as in Equation 2.

3PL-TRT Model

Wainer et al. (2000) further extended the 2PL-TRT model into the 3PL-TRT model by including a pseudo-guessing parameter c_i:

$$p_{ij}(X_{ij}=1|\theta_j,a_i,b_i,c_i,\gamma_{jd(i)}) = c_i + \frac{1-c_i}{1+\exp[-a_i(\theta_j-b_i+\gamma_{jd(i)})]}. \quad (7)$$

By including an additional guessing parameter c_i, the other difference between the 3PL-TRT and the 2PL-TRT model is that, the 3PL-TRT model allows variation in the random effects across testlets, that is, $\gamma_{jd(i)} \sim N(0,\sigma_d^2)$.

Rasch-TRT Model

The Rasch testlet (Rasch-TRT) model proposed by Wang and Wilson (2005) is a special case of the 3PL-TRT model by having $a_i = 1$ and $c_i = 0$:

$$p_{ij}(X_{ij}=1|\theta_j,b_i,\gamma_{jd(i)}) = \frac{1}{1+\exp[-(\theta_j-b_i+\gamma_{jd(i)})]}, \quad (8)$$

which is the same as in the 3PL-TRT model, the variances of the testlet effects $\gamma_{jd(i)}$ are also allowed to vary across testlets with $\gamma_{jd(i)} \sim N(0,\sigma_d^2)$, where σ_d^2 is the variance of the testlet *d*.

Alternative Models for Testlets

In addition to the above three TRT models that are extended from standard unidimensional IRT models by including a random testlet effect, alternative models that account for LID within testlets have also been proposed. Li, Bolt, and Fu (2006) interpreted testlet effect from a confirmatory multidimensional modeling perspective, and treated $\gamma_{jd(i)}$ as another ability dimension in a multidimensional IRT model. Under the multidimensional framework, Li et al. first proposed a general model:

$$p_{ij}(X_{ij}=1\,|\,\theta_j, a_{i1}, a_{i2}, b_i, \gamma_{jd(i)}) = \frac{1}{1+\exp[-(a_{i1}\theta_j - b_i + a_{i2}\gamma_{jd(i)})]}, \quad (9)$$

where the distributions for both θ_j and $\gamma_{jd(i)}$ are fixed as $N(0,1)$ for model identification. Like in the 2PL-TRT model, θ_j and $\gamma_{jd(i)}$ are assumed to be independent. a_{i1} and a_{i2} are item discrimination parameters for general ability θ_j and second ability dimension $\gamma_{jd(i)}$ respectively. Relative to the 2PL-TRT model, this model provides more information about each item within testlets and is helpful in identifying which items within a testlet are most influenced by γ_d. Li et al. have also shown that the 2PL-TRT model is a special case of this multidimensional model.

Li et al. (2006) assumed that, "if an item has high discriminating power on θ, the ability intended to be measured, this item's discriminating power on the secondary dimensional might be expected to be low" (p. 5). Therefore, they imposed constrains on slope parameters and proposed a second model:

$$p_{ij}(X_{ij}=1\,|\,\theta_j, a_{i1}, b_i, \gamma_{jd(i)}) = \frac{1}{1+\exp[-(a_{i1}\theta_j - b_i + \sqrt{MDISC^2 - a_{i1}^2}\gamma_{jd(i)})]}, \quad (10)$$

where *MDISC* is a multidimensional discrimination parameter that is constant across items. The discrimination parameter for the secondary dimension, $\sqrt{MDISC^2 - a_{i1}^2}$, implies that the two discrimination parameters are inversely related.

The third model that Li et al. (2006) proposed assumes a constant item discrimination parameter with respect to γ_d:

$$p_{ij}(X_{ij}=1\,|\,\theta_j, a_{i1}, b_i, \gamma_{jd(i)}) = \frac{1}{1+\exp[-(a_{i1}\theta_j - b_i + \gamma_{jd(i)})]}. \quad (11)$$

Thus, all the items have the same discriminating power for the random dimension γ_d.

This section first reviews the three commonly used unidimensional dichotomous IRT models as well as the associated assumptions, and then summarizes the TRT models that have been formulated to account for LID. The next section will review how both the IRT models and the TRT models can be parameterized from the multilevel modeling perspective, and the advantages of using multilevel parameterizations.

Multilevel IRT Models and Multilevel Testlet Models

Over the past two decades or so there has been an increasing interest in fitting multilevel/hierarchical models to large datasets in various fields, including education, social and behavioral sciences, psychology, and medical studies (Raudenbush & Bryk, 2002). The multilevel modeling technique has attracted the interest of many educational and social researchers for handling clustered/nested data structures. Such data sets are either cross-sectional (e.g., students nested within schools) or longitudinal (e.g., occasions nested within individuals). A typical example of a multilevel data structure in educational research is that, students are nested within classrooms and schools, where students (Level 1), classrooms (Level 2), and schools (Level 3) form a three-level hierarchical structure (Raudenbush & Bryk, 2002).

The primary reason for the multilevel modeling technique applied in situations with nested data is its capability of dealing with the issue of within-cluster dependencies. In nested data, for example, students within classrooms, the assumption of independence might be violated, because units (e.g. students) within the same cluster (e.g., classrooms) might share some inherent similarities. In this case, the multilevel modeling technique could appropriately deal with the issue of

dependencies by allowing the intercept and the effect of explanatory variables to vary across higher-level units (Snijders & Bosker, 1999).

This main objective of this section is to review how the standard unidimensional dichotomous IRT models and the TRT models have been parameterized from a multilevel modeling perspective. Then, the advantages of using multilevel parameterization will be summarized. Moreover, to illustrate the applicability of such parameterization, some published applications that are drawn directly from the field of measurement will be presented.

Multilevel Measurement Models

Earlier studies (e.g. Adams et al., 1997; Beretvas & Williams, 2004; Fox & Glas, 1998; Kamata, 1998, 2001) have shown that regular IRT models could be formulated as hierarchical generalized linear models (HGLM), in which items are treated as nested within people. Those reparameterized IRT models from the HGLM framework are referred to as multilevel measurement models (MMMs) (Beretvas & Kamata, 2005). Figure 1 depicts a graphical representation of how items are clustered within persons.

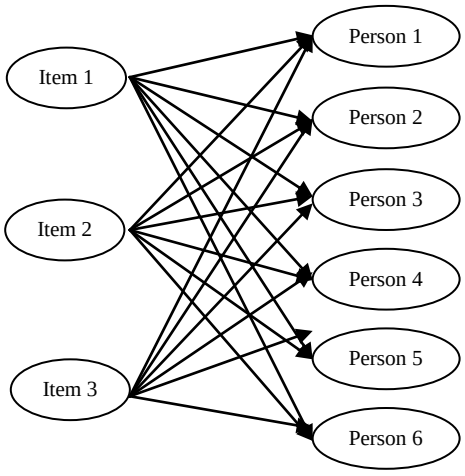


Figure 1: Graphical Representation of the Clustering of Responses within Persons

Take the Rasch model as an example, the corresponding multilevel parameterization (Kamata, 2001) is represented as follows:

$$\text{Level 1}: \log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \eta_{ij} = \pi_{0j} + \pi_{1j}X_{1ij} + \pi_{2j}X_{2ij} + \cdots + \pi_{kj}X_{kij}$$
$$= \pi_{0j} + \sum_{q=1}^{k-1}\pi_{qj}X_{qij} \quad , \qquad (12)$$

$$\text{Level 2}: \begin{cases} \pi_{0j} = \beta_{00} + u_{0j} \\ \pi_{1j} = \beta_{10} \\ \vdots \\ \pi_{kj} = \beta_{k0} \end{cases}$$

where p_{ij} is the probability that person *j* responds to item *i* correctly; X_{qij} represents the *q*th dummy coded variable for person *j*, with values -1 when $q = i$ and 0 when $q \neq i$ for item *i*. Coding with a value of negative one instead of positive one results in a more straightforward correspondence between the standard Rasch model and the corresponding multilevel parameterization (Chen, 2010), which will be illustrated later. It should be noted that the dummy variable for the last item is dropped in order to achieve full rank for the design matrix of the model. π_{0j} is the intercept term at Level 1, which is modeled to vary across persons at Level 2 with $u_{0j} \sim N(0, \tau_{00})$. π_{qj} is the coefficient associated with X_{qij} ($q = 1,...,k-1$), and represents the fixed item effect. Since the item effects are modeled as fixed across persons, no error term is associated with the Level-2 equations for each item parameter.

The log-odds of the probability of a correct response to item *i* for person *j* is obtained by combining the above Level 1 and Level 2 equations as:

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right)=\eta_{ij}=u_{0j}-\beta_{i0}, \tag{13}$$

thus, the probability of a correct response can be expressed as

$$p_{ij}=\frac{1}{1+\exp[-(u_{0j}-\beta_{i0})]}. \tag{14}$$

Comparing between Equation 14 and Equation 3 (the standard Rasch model equation), it can be concluded that the multilevel formulation is algebraically equivalent to the Rasch model by having the ability parameter θ_j in the Rasch model corresponding to the error term u_{0j} in the multilevel formulation and the item difficulty parameter b_i corresponding directly with the fixed item effect β_{i0} in the multilevel formulation. However, if coding item indicators with positive ones rather than negative ones, the item difficulty parameter b_i would correspond with $-\beta_{i0}$. Therefore, negative coding for item indicators were used in this study in order to have more straightforward correspondence.

By relaxing the equal discrimination assumption in the multilevel Rasch model above, the multilevel parameterization of the 2PL model has been formulated and studied (Fox, 2003; Fox & Glas, 2001; Skrondal & Rabe-Hesketh, 2004). In addition, Skrondal and Rabe-Hesketh (2004) provided the multilevel parameterization of the 3PL model by allowing the persons to respond to items with guessing.

Why do the researchers develop the multilevel parameterizations of IRT models? Consider a research scenario of investigating the effects of student characteristics on student abilities. In the traditional two-step analysis, a standard IRT model would be used first to estimate student abilities, which are then used as an

outcome variable in a linear model with student characteristics as predictors. However, such two-step analysis may not provide accurate results because of biased parameter estimates and measurement errors associated with the ability estimates. According to Kamata (2001), a one-step analysis could be performed instead by including student characteristics into the multilevel IRT model, which would facilitate the modeling of measurement error. Similarly, from the multilevel modeling framework, group characteristics can also be evaluated by considering a three-level model with persons (Level 2) nested within groups (Level 3), and thereby avoid the need to perform separate analysis (Adams et al., 1997; Kamata, 1998, 2001). Therefore, a primary benefit of supporting the use of multilevel IRT models is its capability in including person-level or group-level predictors and modeling the clustering effects commonly found in data.

Built upon the theoretical development, application studies of multilevel IRT models are also flourishing in educational measurement literature, such as detection of differential item functioning (DIF) (e.g., Cheong, 2001; Kamata, 1998, 2001; Luppescu, 2002), test equating (e.g., Chu & Kamata, 2005), and dimensionality assessment (e.g., Beretvas & Williams, 2004). Take Chu and Kamata's (2005) study as example, the authors used the multilevel Rasch model in test equating by controlling for differential item functioning (DIF) effects, and their results demonstrated that the multilevel IRT model performed better than the multiple-group concurrent equating designs in terms of the accuracy and stability of item and ability parameter estimates.

Multilevel Testlet Models

The last section illustrates the multilevel parameterizations of the standard IRT models, including the model representations, advantages, and some examples of applications. This section focuses on how this multilevel parameterization could incorporate the clustering of items, e.g. testlets. Based on the multilevel parameterization of the Rasch model, Jiao et al. (2005) proposed a multilevel modeling of local item dependence due to testlet effects. The hierarchy (Figure 2, Jiao et al., 2005, p. 5) of their proposed model is that items (Level 1) are nested within testlets (Level 2) which are modeled as nested within persons (Level 3). Mathematically, this three-level model is represented as

$$\begin{aligned}
\text{Level 1}: \log\left(\frac{p_{itj}}{1-p_{itj}}\right) &= \eta_{itj} = \pi_{0tj} + \pi_{1tj}X_{1itj} + \pi_{2tj}X_{2itj} + \cdots + \pi_{ktj}X_{kitj} \\
&= \pi_{0tj} + \sum_{q=1}^{k-1}\pi_{qtj}X_{qitj} \\
\text{Level 2}: &\begin{cases} \pi_{0tj} = \beta_{00j} + w_{0tj} \\ \pi_{1tj} = \beta_{10j} \\ \vdots \\ \pi_{ktj} = \beta_{k0j} \end{cases} \\
\text{Level 3}: &\begin{cases} \beta_{00j} = \gamma_{000} + u_{00j} \\ \beta_{10j} = \gamma_{100} \\ \vdots \\ \beta_{k0j} = \gamma_{k00} \end{cases}
\end{aligned} \quad , \qquad (15)$$

where p_{itj} is the probability that person *j* responds to item *i* in testlet *t* correctly. X_{qitj} represents the *q*th dummy coded variable for person *j*, with values -1 when $q = i$ and 0 when $q \neq i$ for item *i* in testlet *t*. Same as the multilevel parameterization of the Rasch model, the dummy variable for the last item is dropped

to achieve full rank for the design matrix of the model. π_{0tj} is the intercept term at Level 1, which is modeled to vary across persons and testlets at Level 2. π_{qtj} is the coefficient associated with X_{qitj} $(q = 1,...,k-1)$, and represents the item effect. Since the item effects are modeled as fixed across persons, no error term is associated with the Level-3 equations for each item parameter. The random effect w_{0tj} at Level 2 is interpreted as an interaction effect between testlets and persons and is assumed that $w_{0tj} \sim N(0, \tau_t)$. u_{00j} at Level-3 is the person specific random effect and is assumed that $u_{00j} \sim N(0, \tau)$. Two assumptions were made when the authors proposed this model: first, no interdependence occurs between testlets; second, independence is present between any two items that come from different testlets.

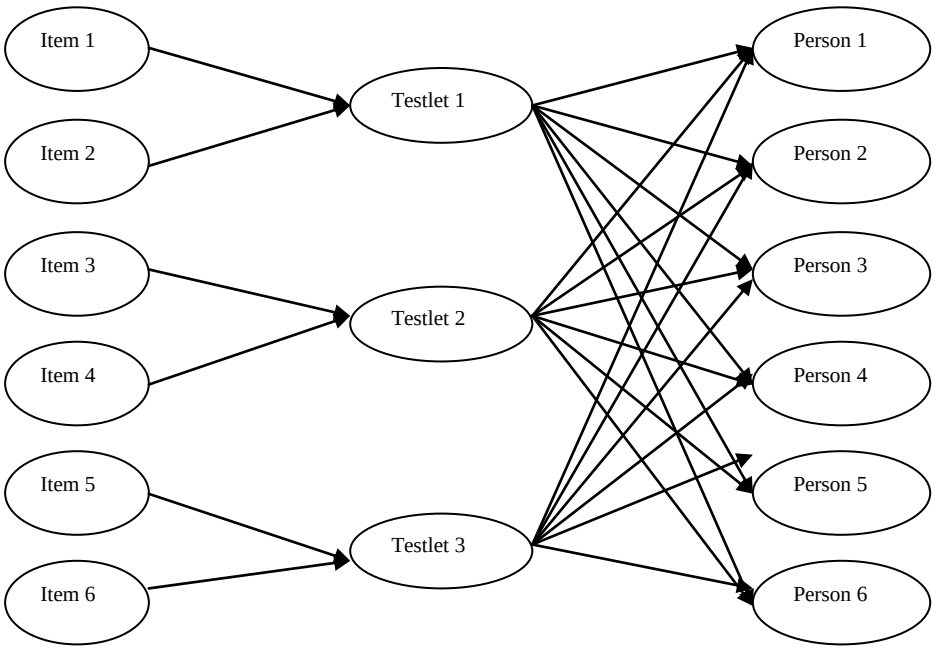


Figure 2: Hierarchy of Multilevel Modeling of LID Caused by Testlet Effects (Adapted from Jiao et al., 2005, p.5)

The probability of a correct response to item *i* in testlet *t* for person *j* is obtained by combining the three levels from Equation 15:

$$p_{itj} = \frac{1}{1+\exp[-(u_{00j} - \gamma_{i00} + w_{0tj})]}. \quad (16)$$

Comparing Equations 16 and 8 (the Rasch-TRT model equation), it can be seen that this three-level testlet model formulated by Jiao et al. (2005) is algebraically equivalent with the Rasch-TRT model (Wang & Wilson, 2005) by having the ability parameter θ_j in the Rasch-TRT model corresponding with the error term u_{00j}, the item difficulty parameter b_i corresponding directly with the fixed item effect γ_{i00}, and the person specific testlet effect $\gamma_{jd(i)}$ corresponding with the Level-2 random effect w_{0tj} in the multilevel formulation (Jiao, Wang, & He, 2013).

Since the multilevel testlet model proposed by Jiao et al. (2005) consists of three levels, it has been referred to as MMMT-3 (Chen, 2010; Beretvas & Walker, 2012). An alternative parameterization of the multilevel measurement model for testlets, which consists of two levels (hence referred to as MMMT-2), is proposed by Beretvas and Walker (2012). Considering a test consisting of *mq* items with *m* testlets each consisting of *q* items, this two level model is represented as:

$$\text{Level 1}: \log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \eta_{ij} = \pi_{0j} + \pi_{1j}X_{1ij} + \pi_{2j}X_{2ij} + \cdots + \pi_{(q-1)j}X_{(q-1)ij} + \pi_{(q+1)j}X_{(q+1)ij} + \cdots + \pi_{(mq-1)j}X_{(mq-1)ij} + \pi_{T1j}T_{T1ij} + \cdots + \pi_{Tmj}T_{Tmij}$$

$$\text{Level 2}: \begin{cases} \pi_{0j} = u_{0j} \\ \pi_{1j} = \beta_{10} \\ \vdots \\ \pi_{(q-1)j} = \beta_{(q-1)0} \\ \pi_{(q+1)j} = \beta_{(q+1)0} \\ \vdots \\ \pi_{(mq-1)j} = \beta_{(mq-1)0} \\ \pi_{T1j} = \beta_{T10} + u_{T1j} \\ \vdots \\ \pi_{Tmj} = \beta_{Tm0} + u_{Tmj} \end{cases}, \qquad (17)$$

where X_{ij} and T_{ij} are dummy-coded item indicator and testlet indicator, and both are coded with "-1" for the relevant item and testlet, respectively. For each testlet of *q* items, there are (*q*-1) dummy-coded item indicators; and for the *m* testlets, there are *m* testlet indicators. The level 2 random residuals u_{0j} and $u_{Tdj}(d = 1 \cdots m)$ correspond to the person abilities and the testlet abilities, respectively. Same as the conventional TRT models (e.g., Wainer et al., 2000; Wang & Wilson, 2005), the residuals for this model are assumed independently normally distributed with means of zero and the following covariance structure:

$$\text{cov}\begin{bmatrix} u_{0j} \\ u_{T1j} \\ \vdots \\ u_{Tmj} \end{bmatrix} = \begin{bmatrix} \tau_0 & 0 & \cdots & 0 \\ 0 & \tau_{T1} & \cdots & \vdots \\ \vdots & \cdots & \vdots & 0 \\ 0 & \cdots & 0 & \tau_{Tm} \end{bmatrix}.$$

By constraining off diagonals as zero in the covariance structure, Beretvas and Walker (2012) assumed that the general ability and testlet ability factors are uncorrelated, which is consistent with the same assumption made by the conventional TRT models. However, Beretvas and Walker also indicated that it is possible to model nonzero covariances among any of these effects as extensions.

Combining Level 1 and Level 2 equations, the probability of a correct response to non-reference indicator item *i* in testlet *d* for person *j* is:

$$p_{ij} = \frac{1}{1+\exp[-((u_{0j}+u_{Tdj})-(\beta_{0j}+\beta_{Td0}))]}, \tag{18}$$

where β_{0j} corresponds with the item specific difficulty and β_{Td0} represents with testlet specific difficulty. Different from conventional TRT models and the multilevel testlet model formulated by Jiao et al. (2005), Beretvas and Walker (2012) decomposed the testlet effect for examinee *j* on testlet *d* into the person-specific random effect, u_{Tdj}, and the fixed (across examinees) testlet effect, β_{Td0}. When there is no person-specific random effect, u_{Tdj}, but only fixed testlet effect, β_{Td0}, the model is referred to as MMMT-2f; and when both fixed and random effects are included, the model is referred to as MMMT-2r.

Chen (2010) did a simulation study to compare the performance of the three multilevel testlet models, MMMT-3, MMMT-2r, and MMMT-2f. She found that, no matter whether the MMMT-2r model was the generating model or not, the MMMT-2r model yielded the best parameter bias in estimation on fixed item effects, fixed testlet effects, and random testlet effects under conditions with nonzero equal pattern of random testlet effects' variance. She concluded that model differences were of little

practical significance, and MMMT-2r had the greatest flexibility from a modeling perspective.

The benefits of using the multilevel measurement model for testlets are similar with that of using multilevel IRT models. First, higher levels of clustering could be modeled. As an extension of the MMMT-3 model, Jiao et al. (2012) proposed a four-level IRT model to simultaneously account for local item dependence due to item clustering and local person dependence due to person clustering. The authors fitted their proposed model to real data from a reading comprehension test, and concluded that the proposed four-level IRT model was the best fitting one compared with the three-level Rasch model for person clustering, the Rasch-TRT model, and the Rasch model in terms of the Deviance Information Criterion (DIC).

The second benefit is that, person-level predictors could be added to model DIF as well as differential testlet functioning (DTF) with the presence of testlet effects. Beretvas and Walker (2012) applied their proposed MMMT-2 model to measure impact, DIF, and DTF for tests that include testlet-based dichotomous items, and found that the MMMT-2 parameterization of DTF was not affected by differential functioning cancellation nor amplification that occur in differential bundle functioning (DBF).

This section reviews how both the IRT models and the TRT models can be parameterized from the multilevel modeling perspective, the advantages, and some applications of using multilevel parameterizations. However, researchers have found that many data sets are not strictly hierarchical but cross-classified. Thus, the next section will first review how the standard IRT models can be parameterized from the

cross-classified modeling perspective, and then the proposed cross-classified modeling for dual local item dependence (DLID) will be presented.

Cross-Classified IRT Models and Cross-Classified Modeling of DLID

Researchers have found that many data sets have more complex non-hierarchical structures. One such complexity involves cross-classified data structures that cannot be handled by the multilevel modeling techniques. In cross-sectional studies, one illuminating example of cross-classified data structure is given by Goldstein (2003), i.e. students are cross-classified by the schools they attend and the neighborhoods they live in. In longitudinal studies, non-hierarchical structure occurs when, for example, students change schools overtime. In this case, occasions are cross-classified by students and schools. To model data with a cross-classified structure, cross-classified random effects modeling (CCREM) techniques have been developed to accommodate non-nested factors (Goldstein, 1986, 2003; Rasbash & Goldstein, 1994; Raudenbush, 1993).

Raudenbush (1993) claimed that, in practice, there are almost no pure nested data structures. Therefore, in reality the CCREMs is expected to be more reflective of the real data structure than the multilevel hierarchical models. However, because of the complexity of this technique, many researchers are still inclined to use multilevel models to fit the data. For example, in the case of cross-sectional studies, they choose to ignore the cross-classified structure of their data sets by treating one of the cross-classified factors hierarchically and disregarding information on the second cross-classified factor (e.g., Ainsworth, 2002; Ma & Wilkins, 2002). On the other hand, in longitudinal studies, if students change schools, researchers have chosen to delete

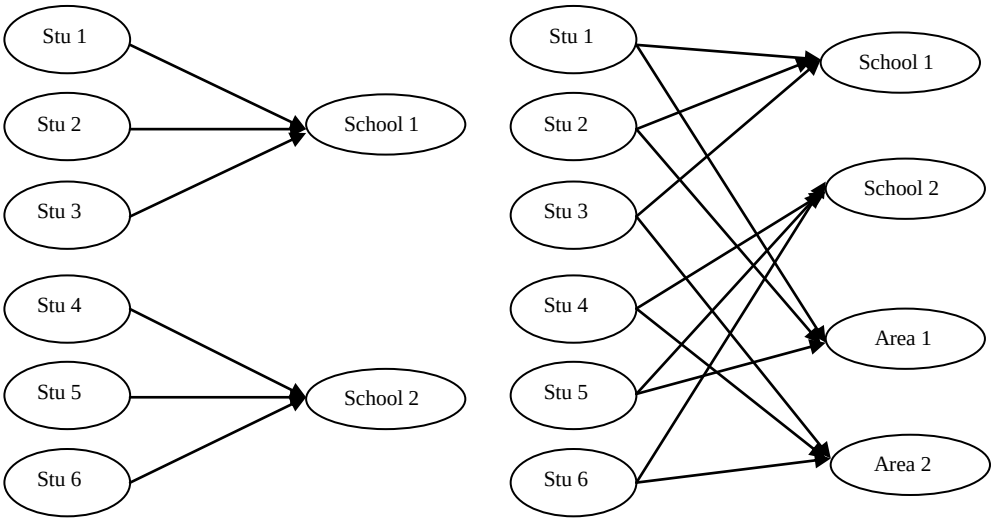
data for mobile students or using only information from one of the schools that those mobile students have attended (e.g., Lee, 2000; McCoach, O'Connell, Reis, & Levitt, 2006; Noble & Schnelker, 2007).

In order to illuminate the consequences of misspecifying CCREMs, simulation studies have been conducted to enhance our understanding of the functioning of this class of models (Meyers, 2004; Meyers & Beretvas, 2006; Luo, 2007; Ren, 2011). Those simulation studies demonstrate that, inappropriate modeling of cross-classified data structures would cause inaccurate estimates of parameters and their associated standard errors.

Recently some researchers start to put CCREMs into real applications because they have realized the inappropriateness of using multilevel models to analyze cross-classified data. Further, the availability of computer programs such as HLM 7.0, MLwiN 2.0, SAS PROC MIXED, R package lme and lme4 to estimate CCREMs have increased the likelihood of applying these models in real applications. For example, Fielding and Goldstein (2006) found in their review of real applications that CCREMs have been applied in such areas like health, survey, social networks, veterinary epidemiology, missing identification of units, generalizability theory, psychometrics, and education.

Multiple authors (e.g., Beretvas, 2010; Browne, Goldstein, & Rasbash, 2001; Fielding & Goldstein, 2006) have formulated cross-classified models. Graphically, researchers have either used diagrams or tables to explain cross-classified data structures. Figure 3 presents both a traditional two-level hierarchical model (left

diagram) and a two-level cross-classified model (right diagram). The classifications are represented by arrows from the lowest level units to the classification unit.



(I) (II)

Figure 3: (I) Two-level Hierarchical Linear Model (II) Two-level Cross-Classified Model

Tables are also used to describe cross-classified data structures. Table 1 depicts purely nested data in which individuals are nested within Factor 1 (e.g., schools) and Factor 1 is nested within Factor 2 (e.g., neighborhoods), while Table 2 depicts a cross-classification structure where students are cross-classified by Factor 1 (e.g., middle school attended) and Factor 2 (e.g., high school attended).

Table 1

Individuals (X) nested within Factor 1 nested within Factor 2

	Factor 2		
Factor 1	A	B	C
1	XXX		
2	XXX		
3		XXX	
4		XXX	
5			XXXX
6			XXXX

Table 2

Individuals (X) cross-classified by Factor 1 and Factor 2

	Factor 2		
Factor 1	A	B	C
1	XXX		XXX
2	XXX		
3		XXX	
4		XX	
5	XXX		XXXX
6	XX		XXXX

Note: Table 1 and Table 2 are adapted from Meyers & Beretvas (2006, p. 474, p. 475)

Mathematically, an unconditional two-level cross-classified random effect model with two cross-classified factors can be expressed as:

$$\begin{aligned} &\text{Level 1: } Y_{i(j_1,j_2)} = \beta_{0(j_1,j_2)} + e_{i(j_1,j_2)} \\ &\text{Level 2: } \beta_{0(j_1,j_2)} = \gamma_{000} + u_{00j_1} + u_{00j_2} \end{aligned}, \tag{19}$$

where $Y_{i(j_1,j_2)}$ is the outcome for individual *i* belonging to Factor 1, j_1, and Factor 2, j_2; γ_{000} is the grand mean outcome on *Y* across individuals, Factor 1, and Factor 2; $e_{i(j_1,j_2)}$ is the Level 1 residual term; u_{00j_1} is the Level 2 residual for Factor 1, j_1; and u_{00j_2} is the Level 2 residual for Factor 2, j_2. In terms of the residuals' distributions,

the following is assumed: $e_{i(j_1,j_2)} \sim N(0,\sigma^2)$, $u_{00j_1} \sim N(0,\tau_{uj_100})$, and $u_{00j_2} \sim N(0,\tau_{uj_200})$. The Level 1 and Level 2 equations can be combined into a single equation as:

$$Y_{i(j_1,j_2)} = \gamma_{000} + u_{00j_1} + u_{00j_2} + e_{i(j_1,j_2)}. \quad (20)$$

As with the conventional multilevel models, explanatory variables can be added to unconditional cross-classified models at Level 1 and Level 2 to explain variability in the outcome at those levels.

Cross-Classified IRT Model

The conventional multilevel formulation of IRT models treats items as fixed effects and persons as random effects. This way of parameterization generally regards persons as a random sample from a population and the purpose of the analysis is to evaluate the difficulties of some specific items. However, if the purpose of the analysis is to evaluate the abilities of some specific persons, rather than to evaluate the difficulties of some specific items, it is more reasonable to consider items as random and persons as fixed (Van den Noortgate, De Boeck, & Meulders, 2003). In psychometrics, it is uncommon to treat items as random. In order to demonstrate that random items are reasonable, De Boeck (2008) illustrated this concept from both theoretical and practical perspectives.

De Boeck (2008) summarized three theoretical reasons for being interested in random item models by reviewing the literature. The first is "the clearly random nature of the items, such as randomly drawn words from a vocabulary" (p. 534). The second is "the study of ability change in a longitudinal design with randomly drawn

item samples" (p. 534). The third is "modeling item families", which are defined as "sets of items with sufficient communalities within the set and sufficient differentiation from other sets"; in this case, the research focus is on the family parameters, like mean and variance, instead of item-specific parameters (p. 534). Based on his review of literature, De Boeck provides two more reasons for considering items as random. The first is that, items can be treated as drawn from a population, e.g., an item bank in computer adaptive testing can be considered as an item population; and in the context of criterion-referenced measurement, the concept of "universe" and "domain" has been used in the process of item generation (Hively, Patterson, & Page, 1968; Popham, 1978). The second is the uncertainty about the parameters, and therefore prior distribution is used as in the fully Bayesian approach. De Boeck argued that the uncertainty embedded in the prior distribution is equivalent with a population distribution where the elements are random.

De Boeck also demonstrated that the random item approach is promising to handle several issues from a practical point of view. The first issue he mentioned is the measurement of people's ability, where the generalization over items is wanted; and therefore, a model with fixed person effects and random item effects is ideal. The second issue is the explanation of item difficulties. De Boeck argued that it is unrealistic to use the linear logistic test model (LLTM; Ficher, 1973) in which the item difficulty is perfectly explained by several item properties; instead, treating the items as random by adding an error term to the LLTM model is more realistic. The third issue is related to DIF. He pointed out that both of the two global strategies for investigating the presence of DIF, anchoring strategy and free parameter strategy,

have drawbacks, which could be effectively handled through the use of random item models.

Therefore, both item and person effects can be simultaneously treated as random. Under this line of research, researchers have proposed an alternative parameterization of the IRT models, namely, the cross-classified IRT model, where item responses are considered cross-classified with items and persons (Meulders & Xie, 2004; Van den Noortgate et al., 2003). Figure 4 presents a graphical representation of how responses maybe cross-classified with persons and items, where the four responses, R_{11}, R_{12}, R_{21}, and R_{22}, are cross-classified by two persons, person 1 and person 2, and two items, item 1 and item 2.

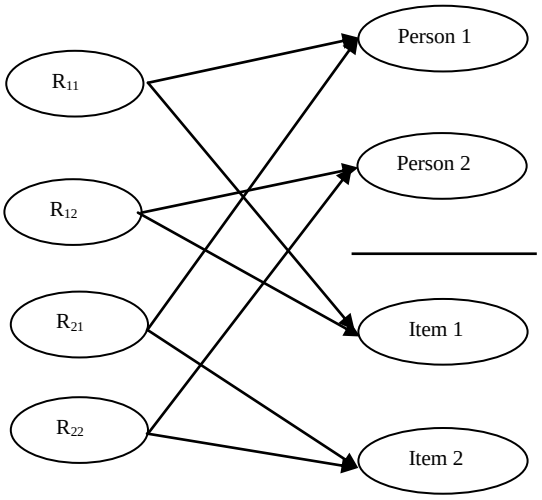


Figure 4: Graphical Representation of Reponses Cross-Classified by Items and Persons

The fully unconditional cross-classified IRT model is represented as:

$$\begin{aligned} &\text{Level 1}: \log\left(\frac{p_{i(j_1,j_2)}}{1-p_{i(j_1,j_2)}}\right) = \eta_{i(j_1,j_2)} = \pi_{0(j_1,j_2)}, \\ &\text{Level 2}: \pi_{0(j_1,j_2)} = u_{0j_10} + u_{00j_2} \end{aligned} \tag{21}$$

where j_1 and j_2 represents persons and items respectively; u_{0j_10} and u_{00j_2} represent the person and items residuals respectively. u_{0j_10} and u_{00j_2} are assumed independent with means of zero and constant variances of τ_1 and τ_2, respectively. Combining Level 1 and Level 2 equations, we can specify the probability of a correct response to item q for person j as:

$$p_{i(j,q)} = \frac{1}{1+\exp[-(u_{0j0}+u_{00q})]}. \quad (22)$$

By comparing Equations 22 and 3 (the standard Rasch model equation), it can be noted that the cross-classified parameterization is algebraically equivalent to the Rasch model where the ability parameter θ_j in the Rasch model corresponds to the person residual u_{0j0} in the cross-classified model and the item difficulty parameter b_q corresponds directly to the item residual $-u_{00q}$ in the cross-classified model.

Studies have been conducted to explain and compare the similarities and differences between the two parameterizations of IRT models, multilevel model and cross-classified model (Beretvas, Cawthon, Lockhart, & Kaye, 2012; Van den Noortgate & De Boeck, 2005). Beretvas et al. (2012) concluded that both parameterizations could estimate person abilities and item difficulties. The difference between these two parameterizations is that, while the multilevel parameterization may be used to estimate DIF, the cross-classified parameterization allows for the estimation of differential facet functioning (DFF) as well as the interaction between item and person descriptors.

Proposed Cross-Classified Modeling of DLID

Previous sections demonstrated how to model testlet effects from a multilevel modeling perspective. However, both the TRT models and the corresponding multilevel parameterizations are formulated to deal with the issue of LID that is caused by a single factor. In practice, LID could be caused by two or more factors simultaneously, which is referred to in this study as dual local item dependence (DLID).

One typical example of assessments where the issue of DLID may arise is scenario-based science assessment. Such science tests have at least two sources of LID, one resulting from item clustering, which is caused by scenarios; the other resulting from the coverage of multiple content areas. Appendix A provides an example of scenario-based science test with coverage of multiple content areas. Several other scenarios, like the one in Appendix A, could be included in one test. Each scenario is followed by several items that are created to test students' capability in different content areas. In Appendix A, there are three items following the same scenario; the three items are created to assess students' capability in population dynamics, population dynamics, and classified organisms, respectively. Therefore, two content areas, population dynamics and classified organisms, are assessed in one scenario. In addition, scenarios are not nested within content areas nor vice-versa, therefore, the two are said to be cross-classified.

Another example of assessments that have the issue of DLID is TOEFL Reading. In each TOEFL Reading section, there are 3 to 5 passages and each passage contains 12 to 14 multiple-choice questions, which generally belong to one of the

following subskills: detail/fact, vocabulary, reference questions, and summary. Appendix B provides an example of a TOEFL Reading passage followed by several subskills. In this case, LID is caused by testlet effects (passage) and subskills simultaneously and the structure of the item responses is that each item is cross-classified by testlets (passages) and subskills, which are nested within persons.

To account for such DLID in scenario-based science assessments or TOEFL Reading, this study proposed a cross-classified model. Take the scenario-based science assessment as an example, the structure of this proposed model is that items (Level 1) are cross-classified with testlets (scenarios) and content areas (Level 2), and both testlets and content areas are nested within persons (Level 3) (see Figure 5). Mathematically, it may be represented as follows:

$$\begin{aligned}
\text{Level 1}: \log\left(\frac{p_{i(t,c)j}}{1-p_{i(t,c)j}}\right) &= \eta_{i(t,c)j} = \pi_{0(t,c)j} + \pi_{1(t,c)j}X_{1i(t,c)j} + \\
&\quad \pi_{2(t,c)j}X_{2i(t,c)j} + \cdots + \pi_{k(t,c)j}X_{ki(t,c)j} \\
&= \pi_{0(t,c)j} + \sum_{q=1}^{k}\pi_{q(t,c)j}X_{qi(t,c)j} \\
\text{Level 2}: &\begin{cases} \pi_{0(t,c)j} = \beta_{00j} + w_{0tj} + w_{0cj} \\ \pi_{1(t,c)j} = \beta_{10j} \\ \vdots \\ \pi_{k(t,c)j} = \beta_{k0j} \end{cases} \\
\text{Level 3}: &\begin{cases} \beta_{00j} = u_{00j} \\ \beta_{10j} = \gamma_{100} \\ \vdots \\ \beta_{k0j} = \gamma_{k00} \end{cases}
\end{aligned} \quad , \tag{23}$$

where $p_{i(t,c)j}$ is the probability that person *j* responds to item *i* in testlet *t* and content area *c* correctly. $X_{qi(t,c)j}$ represents the *q*th dummy coded variable for person *j*, with

values -1 when $q = i$ and 0 when $q \neq i$ for item *i* in testlet *t* and content area *c*. $\pi_{0(t,c)j}$ is the intercept term at Level 1, which is modeled to vary across persons, testlets, and content areas at Level 2. $\pi_{q(t,c)j}$ is the coefficient associated with $X_{qi(t,c)j}$ ($q = 1,...,k$), and represents the item effect. Since the item effects are modeled as fixed across persons, no error term is associated with the Level-3 equations for each item parameter. The random effect w_{0tj} at Level 2 is interpreted as an interaction effect between testlets and persons and is assumed that $w_{0tj} \sim N(0, \tau_t)$. The random effect w_{0cj} at Level 2 is interpreted as an interaction effect between content areas and persons and is assumed as $w_{0cj} \sim N(0, \tau_c)$. The random effect u_{00j} at Level-3 is the person specific random effect and is assumed as $u_{00j} \sim N(0, \tau)$. It should be noted that no fixed effect is included in the level-3 equation for the intercept term, and thus no reference indicator is required (Beretvas et al., 2012; Chen, 2010).

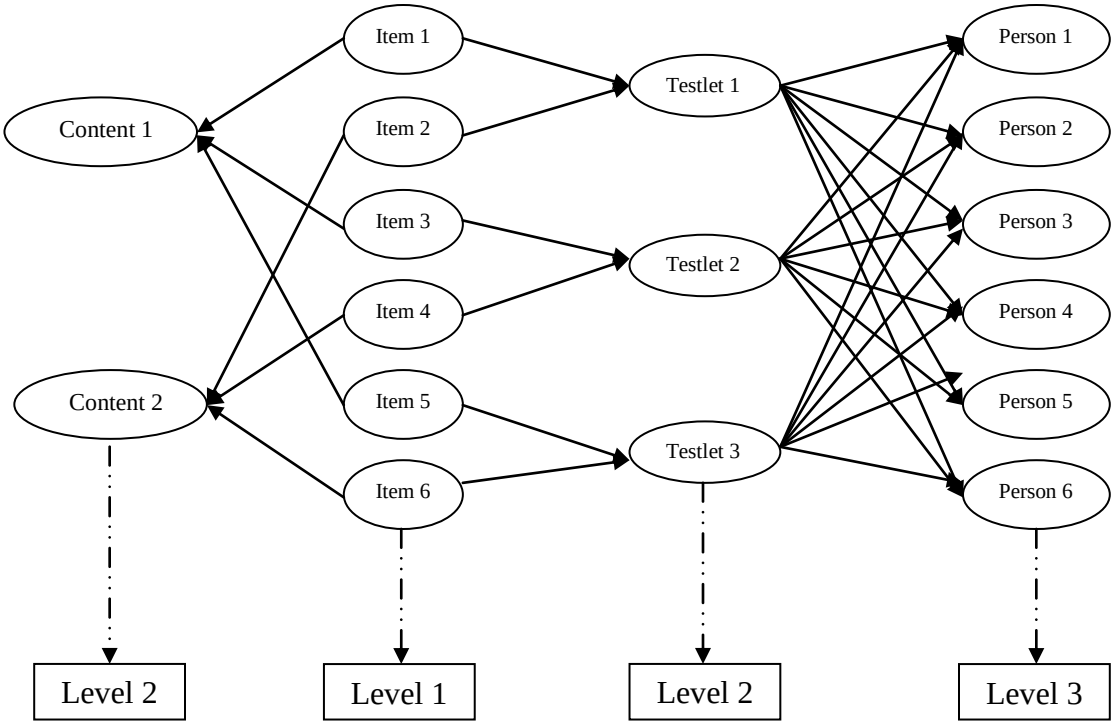


Figure 5: Graphical Representation of Cross-Classified Modeling of DLID

Combining Level 1, Level 2, and Level 3 equations above, the probability of a correct response to item *i* in testlet *t* and content area *c* for person *j* is:

$$p_{i(t,c)j} = \frac{1}{1+\exp[-(u_{00j} + w_{0tj} + w_{0cj} - \gamma_{i00})]}. \tag{24}$$

This is a constrained form of the 3PL testlet model accounting for two types of LID proposed by Jiao et al. (2009) by assuming a zero pseudo-guessing parameter ($c_i = 0$) and a constant discrimination parameter ($a_i = 1$) across items:

$$p_{jdi} = c_i + \frac{1-c_i}{1+\exp[-a_i(\theta_j + \gamma_{j1d(i)} + \gamma_{j2d(i)} - b_i)]}, \tag{25}$$

where $\gamma_{j1d(i)}$ is interpreted by Jiao et al. (2009) as the random-effects testlet-effect parameter for scenario-type of LID (Type 1 LID), which is equivalent to the residual, w_{0tj}, in the newly proposed cross-classified modeling of DLID, where $\gamma_{j2d(i)}$ is

interpreted as the random-effects testlet-effect parameter for content clustering-type of LID (Type 2 LID), and it is equivalent to the residual, w_{0cj}, in the newly proposed cross-classified modeling of DLID. In addition, the ability parameter θ_j corresponds to the person residual u_{00j} in the cross-classified formulation. The item difficulty parameter b_i corresponds to the fixed effect γ_{i00} in the cross-classified formulation.

Summary of the Theoretical Framework

This chapter described the formulation process of the proposed cross-classified modeling of DLID. One of the standard IRT models, the Rasch model, was utilized as the base model. Extending from the standard Rasch model, the Rasch testlet model was formulated by accounting for LID caused by testlet effects. Both the Rasch model and the Rasch testlet model were reparameterized from a multilevel modeling perspective. The primary benefit of this reparameterization is the capability of dealing with the issue of within-cluster dependence. Based on the multilevel parameterization of the Rasch model, Jiao et al. (2005) proposed a multilevel modeling of LID due to testlet effects. The hierarchy of their proposed model is that items (Level-1) are nested within testlets (Level 2) which are modeled as nested within persons (Level 3). However, both the Rasch testlet model and the corresponding multilevel parameterization are formulated to deal with the issue of LID that is caused by a single factor. In practice, LID could be caused by two or more factors simultaneously, which is referred to in this study as DLID. Therefore, a cross-classified model was proposed to account for DLID. The structure of this proposed model is that items (Level 1) are cross-classified with testlets (scenarios) and content

areas (Level 2), and both testlets and content areas are nested within persons (Level 3). It demonstrated that the proposed cross-classified model is algebraically equivalent with a constrained version of the testlet model accounting for two types of LID (Jiao et al., 2009).

In the next chapter (Chapter 3), two simulation studies were designed with the primary purpose of evaluating the performance of the proposed cross-classified model. Data sets with DLID were simulated with both testlet effects and content clustering effects. The second purpose was to investigate the potential factors affecting the need to use the more complex cross-classified modeling of DLID over the simplified multilevel modeling of LID by ignoring cross-classification structure.

Chapter 3: Methodology

This study conducted two simulation studies to investigate the performance of the proposed cross-classified model for DLID and the impact of ignoring the cross-classified structure under a variety of simulated study conditions.

Simulated Conditions

Manipulated Factors

The manipulated factors in the two simulation studies included sample size (500, 1000, 2000), number of testlets (3, 6), number of items per testlet (5, 10), magnitude of the testlet effects represented by standard deviations (SDs) (0, 0.5, 1, 1.5), and magnitude of the content clustering effects represented by SDs (0, 0.5, 1, 1.5). The difference between the two simulation studies was in the SD pattern. That is, for simulation study 1, equal SDs for both the testlet effects and the content clustering effects were assumed; while for simulation study 2, mixed SDs were used to generate the data. For easy of reporting, two simulation studies were conducted separately to differentiate equal SDs from mixed SDs. Table 3 details the levels for the manipulated factors.

Table 3

Simulation Design for Manipulated Factors

	Levels			
Manipulated Factors	1	2	3	4
Sample Size	500	1000	2000	
Number of Testlets	3	6		
Number of Items per Testlet	5	10		
Magnitude of Testlet Effect	0	0.5	1	1.5
Magnitude of Content Clustering Effect	0	0.5	1	1.5

Sample Size. Three levels of sample size (500, 1000, and 2000) were simulated to represent a small, medium, and large sample size. In exploring the multilevel modeling of LID caused by testlets, Jiao and her colleagues (2005, 2013) fixed the sample size at 1000; Chen (2010) used 500 and 1000 to represent smaller and larger sample sizes, and found significant impact on parameter bias. Beretvas and Walker (2012) fixed the sample size at 2000 when using their proposed two-level testlet response model to assess differential testlet functioning. Wang and Wilson (2005) used 200 and 500 to explore the Rasch testlet model. Bradlow et al. (1999) fixed the sample size at 1000 to assess their proposed 2PL-TRT model. Specifically, Jiao et al. (2009) simulated 2000 examinees to evaluate the performance of the proposed 3PL testlet model in dealing with LID caused by testlets and contents. Therefore, 500, 1000, and 2000 were selected to evaluate the parameter estimation at different levels of sample sizes.

Number of Testlets. Previous studies have either manipulated the number of testlets directly or indirectly, or fixed the number of testlets. For direct manipulation, studies treated the number of testlets as a manipulated factor. For example, Wang and Wilson (2005) used 4 and 8 testlets. For indirect manipulation, studies fixed the total number of items and manipulated the number of items per testlet, which will also

result in varied numbers of testlets. For example, Bradlow et al. (1999) fixed the total number of testlet items at 30 and chose two values, 5 and 10, as the number of items per testlet, which yielded 6 and 3 testlets, respectively; Chen (2010) set the total number of items at 50 and used two levels of testlet length, 5 and 10, and thus, the number of testlets were 10 and 5. Some other previous studies fixed the number of testlets, e.g. Jiao et al. (2005) used 6 testlets and Li et al. (2006) simulated 4 testlets. In the present study, 3 and 6 were selected to represent small and large number of testlets respectively.

Number of Items per Testlet. The number of items per testlet, or testlet length, also varied across previous simulation studies. Both Jiao (2005) and Li et al. (2006) fixed the testlet length at 5; Bradlow et al. (1999), Wang and Wilson (2005), and Chen (2005) all considered the testlet length at two levels, 5 and 10. Therefore, 5 and 10 were selected in the present study to represent smaller and larger testlet length respectively.

Since the two variables, number of testlets and number of items per testlet, have been manipulated independently, the total number of items cannot be manipulated. Fully crossing the selected two values for the number of testlets (3 and 6) and the two values for the testlet length (5 and 10) yields three different total of number of items, 15 (3×5), 30 (3×10 and 6×5), and 60 (6×10), which represents short, medium, and long test.

Magnitude of the Testlet Effects. The SD of the random testlet effects represents the magnitude of the testlet effects. This variable has always been manipulated in previous simulation studies. In Wang and Wilson's (2005) study, the

variances they used were 0.25, 0.5, 0.75, and 1. Jiao et al. (2005) specified the SD at four levels: 0, 0.5, 1 and 1.5. Jiao, Wang, and He (2013) used SDs of 0, 0.5, 0.75, and 1 to represent no, small, moderate, and large testlet effects. In Chen's (2010) study, the variances of random testlet effects were simulated as being equal across testlets for some conditions and unequal for others. For the equal variances conditions, she used variances of 0, 0.25, and 0.5 for every testlet; and for the unequal variances conditions, she set the average of the variances of the random testlet effects across testlets as 0.25 and 0.5. Chen (2010) found that the variance of random testlet effects was an influential factor in parameter estimates.

In the present study, two simulation studies were conducted with simulation study 1 generating equal SDs across testlets and simulation study 2 generating mixed SDs across testlets. Table 4 delineates the specification of SDs for the two simulation studies.

Table 4

SD of the Random Testlet Effects

	Simulation Study 1		Simulation Study 2	
	3 Testlets	6 Testlets	3 Testlets	6 Testlets
1	0-0-0	0-0-0-0-0-0	0-0-0	0-0-0-0-0-0
2	0.5-0.5-0.5	0.5-0.5-0.5-0.5-0.5-0.5	0-0.5-1	0-0-0.5-0.5-1-1
3	1-1-1	1-1-1-1-1-1	0.5-1-1.5	0.5-0.5-1-1-1.5-1.5
4	1.5-1.5-1.5	1.5-1.5-1.5-1.5-1.5-1.5	1-1.5-2	1-1-1.5-1.5-2-2

From the empirical examples used in the literature (e.g., Wainer & Wang, 2000; Wang & Wilson, 2005), the SDs of the testlet effects in real tests may range from as small as zero to as large as the SD of persons' ability. For both simulation studies, the SD of persons' ability was fixed at 1 across conditions. Therefore, in this study, the SDs, 0, 0.5, 1, and 1.5, were selected to represent small to large testlet

effects, where 0 represents no testlet effect, 0.5 represents small testlet effect, 1 and 1.5 represents large testlet effect. Even though in the empirical examples, the SD of the testlet effects may not reach 1.5, 1.5 was used in the current simulation studies for theoretical illustration.

In simulation study 1, the SDs were set equal across each testlet with values of 0, 0.5, 1, and 1.5 for the four levels respectively. When the SDs of the random testlet effects were simulated to be 0, it means there is no LID caused by testlet effects. Similar to simulation study 1, simulation study 2 also simulated four levels of SD for the random testlet effects. Even though under each level, the SDs of the testlet effects were not necessarily the same, the average SDs of the random testlet effects across testlets ranged from 0, 0.5, 1, to 1.5, which were the same values as those used in simulations study 1.

Magnitude of the Content Clustering Effects. Almost no previous simulation studies have explored LID caused by content clustering effects, except the study by Jiao et al. (2009). Jiao simulated four content areas, and used three levels of SD, 0.5, 0.75, and 1, to represent small, moderate, and large LID caused by the content clustering effect. In addition, she also constrained the content clustering effects' SD as the same across the four content areas.

In the present study, two content areas were simulated. Even though content clustering effect represents a different type of item clusters from the testlet effect, the two effects, testlet effect and content clustering effect, are essentially the same from the statistical perspective. Therefore, the same four values as the SDs of the testlet effects, 0, 0.5, 1, and 1.5, were used to represent the magnitude of the content

clustering effect. Table 5 delineates the specification of SDs for the content clustering effects for the two simulation studies. Like the testlet effects in simulation study 1, the SDs were equal across each content area with values of 0, 0.5, 1, and 1.5 for the four levels respectively; and in simulation study 2, the average SD of the random content clustering effects for the two content areas ranged from 0, 0.5, 1, to 1.5.

Table 5

SD of the Random Content Clustering Effects

	Simulation Study 1	Simulation Study 2
1	0-0	0-0
2	0.5-0.5	0.25-0.75
3	1-1	0.75-1.25
4	1.5-1.5	1.25-1.75

Fixed Factors

For each of the 384 conditions ($192 = 3 \times 2 \times 2 \times 4 \times 4$ in each simulation study) above, some common factors were set fixed across simulated study conditions. Under each condition, the true values of persons' ability parameter were randomly generated from a standard normal distribution, $N(0,1)$. As described above, the testlet length could be 5 or 10. When the testlet length was 5, the item difficulty parameters for the five items in each testlet were fixed at -2, -1, 0, 1, and 2; when the testlet length was 10, the item difficulty parameters for the ten items in each testlet were fixed at -2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2, and 2.5. These values were selected in order to generate tests that cover items with a broad range of difficulties, including easy, medium, and hard.

Data Generation

Within each condition, 50 replications were implemented. Harwell, Stone, Hsu, and Kirisci (1996) recommended a minimum of 25 replications in IRT-based research. In addition, the condition with the fewest items (15) and smallest sample size (500) was selected to justify that 50 replications were sufficient. A post hoc check of the standard errors for five item difficulty parameters under this condition indicates that the magnitude of standard errors flattened out when the number of replications was greater than 30 (see Figure 6). Therefore, 50 replications in this study were sufficient.

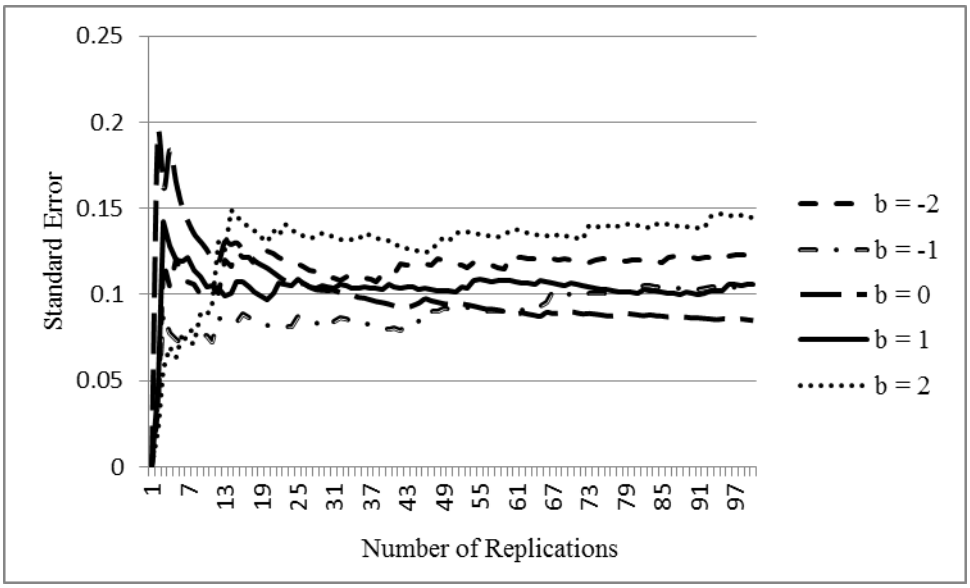


Figure 6: Standard Errors for Item Difficulty Parameters with Different Number of Replications

The free software package, R was used to generate the data for each of the 50 replications per condition (See Appendix C for the R code for data generation). As noted above, the generating model was the proposed cross-classified modeling of

DLID, which has the following probability function after combining equations from the three levels (Equation 24):

$$p_{i(t,c)j} = \frac{1}{1+\exp[-(u_{00j} + w_{0tj} + w_{0cj} - \gamma_{i00})]}.$$

The residual term, u_{00j}, which represents persons' ability, was randomly sampled from a normal distribution with a mean of zero and standard deviation of one. The other two residual terms, w_{0tj} and w_{0cj}, which represents person-specific testlet effect and person-specific content effect respectively, were both sampled from a normal distribution with a mean of zero, and the their corresponding SDs depending on the specific design condition described in the section of manipulated factors. The values of the fixed effect, γ_{i00}, which designates the item difficulty parameter, have been specified in the section of fixed factors.

Thus, the probability that person *j* responds to item *i* in testlet *t* and content area *c* correctly could be obtained by substituting the three residual terms, u_{00j}, w_{0tj}, and w_{0cj}, and the fixed term, γ_{i00}, into the equation above. This probability was compared to a random number sampled from the uniform [0, 1] distribution. A simulated item response of 1 was assigned if the random number was less than or equal to the associated probability; otherwise, 0 was assigned.

Models

For each of the 384 conditions, four models were estimated. Model 1 was the proposed cross-classified model, which was also the data generating model:

$$p_{i(t,c)j} = \frac{1}{1+\exp[-(u_{00j} + w_{0tj} + w_{0cj} - \gamma_{i00})]}. \quad (26)$$

Model 2 was a multilevel model in which items (Level-1) were nested within testlets (Level 2) which were modeled as nested within persons (Level 3). In this case, the cross-classified data structure was ignored by omitting the content clustering effect on LID. Thus, the following model was used to model the data:

$$p_{itj} = \frac{1}{1+\exp[-(u_{00j} + w_{0tj} - \gamma_{i00})]}. \quad (27)$$

Model 3 was also a multilevel model in which items (Level-1) were nested within content areas (Level 2) which were modeled as nested within persons (Level 3). The difference between Model 2 and Model 3 is that, Model 3 ignored the cross-classified structure by omitting the testlet effect on LID. The resulting model is defined as:

$$p_{icj} = \frac{1}{1+\exp[-(u_{00j} + w_{0cj} - \gamma_{i00})]}. \quad (28)$$

Finally, the data sets were analyzed using the Rasch-equivalent two-level model (Model 4), in which neither testlet effects nor content clustering effects were included in the model, which is defined as:

$$p_{ij} = \frac{1}{1+\exp[-(u_{0j} - \gamma_{i0})]}. \quad (29)$$

Model Identification

For the four estimating models described in the section above, a constant can be added to the difficulty parameters but subtracted from the ability parameters to keep the probability of a correct response the same. Therefore, the four models cannot be identified without imposition of constraints. The common approach to identifying

IRT models is to constrain the mean ability or the mean of the item difficulty parameters to be 0.

In this study, the mean of each of the random effects was constrained to be 0 for each of the four estimating models without other adjustments. This approach for scale identification is the default approach unitized in the PROC GLIMMIX macro package in SAS 9.2 software (SAS Institute Inc., 2008), which is going to be discussed in the next section.

Model Parameter Estimation

All the four models were estimated in SAS 9.2 software (SAS Institute Inc., 2008) using the PROC GLIMMIX macro package (see Appendix D for the SAS Code). The GLIMMIX procedure fits statistical models that are known as generalized linear mixed models (GLMM). Since all the four models belong to the family of GLMM, the GLIMMIX procedure is an appropriate estimation option for use in this study. Six estimation methods are available in the PROC GLIMMIX procedure, including four pseudo-likelihood methods (RSPL, MSPL, RMPL, and MMPL), maximum likelihood with Laplace approximation (LAPLACE), and maximum likelihood with adaptive quadrature (QUAD).

For the four pseudo-likelihood methods, RSPL, MSPL, RMPL, and MMPL, the first letter determines whether estimation is based on a residual likelihood ("R") or a maximum likelihood ("M"); the second letter identifies the expansion locus for the underlying approximation, either the vector of random effects solutions ("S") or the mean of the random effects ("M"); the last two letters "PL" represent pseudo-likelihood. LAPLACE estimation approximates the marginal likelihood by using

Laplace's method, in which parameter estimates are determined by minimizing twice the negative of the resulting log-likelihood approximation. If QUAD method was chosen, the GLIMMIX procedure approximates the marginal log likelihood with an adaptive Gauss-Hermite quadrature.

Chen (2010) used the residual pseudo-likelihood method (RSPL) to estimate the three multilevel models (MMMT-2f, MMMT-2r, and MMMT-3) in her study. Beretvas and Walker (2012) also used RSPL when estimating the MMMT-2r model. Jiao et al. (2013) compared three methods in estimating the MMMT-3 model, including the Markov chain Monte Carlo (MCMC) method, marginal maximum likelihood estimation (MMLE) with the expectation-maximization (EM) algorithm in ConQuest and the six-order Laplace approximation estimation in HLM6. Even though Jiao et al. (2013) found that estimation methods could have significant effects on parameter estimations, it was difficult to ascertain which estimation algorithm is preferable to use, because each method has its advantages and disadvantages, e.g., the Laplace method resulted in the best ability parameter estimation while the MCMC method produced the best item parameter estimates. Therefore, there is still no consensus about the best estimation method based on the literature. Since the purpose of this study was not to compare the estimation methods, only one method, LAPLACE, in the PROC GLIMMIX procedure was applied. LAPLACE method was chosen because the GLIMMIX Produce indicates that LAPLACE estimates typically exhibit better asymptotic behavior and less small-sample bias than pseudo-likelihood estimators. The parameters to be estimated depend on the models. For the proposed cross-classified model (Model 1), the parameters to be estimated include a difficulty

parameter (γ_{i00}) for each item, ability parameter for each person (u_{00j}), SD of testlet effects for each testlet ($\sqrt{\tau_t}$), SD of content clustering effects for each content area ($\sqrt{\tau_c}$), and SD of persons' ability ($\sqrt{\tau_\theta}$). For Model 2, the same parameters were estimated as for Model 1 except the SD of content clustering effects. In addition, for Model 3, the same parameters were estimated as for Model 1 except the SD of testlet effects. For the Rasch model (Model 4), only the item difficulty parameters (γ_{i0}), ability parameter for each person (u_{0j}), and SD of persons' ability ($\sqrt{\tau_\theta}$) need to be estimated. Table 6 details the parameters and number of parameters that were estimated under each model.

Table 6

Parameters and Number of Parameters to be Estimated

	Models			
	Model 1	Model 2	Model 3	Model 4
Parameters				
Fixed Effects				
Item difficulty	γ_{i00}	γ_{i00}	γ_{i00}	γ_{i0}
Random Effects				
Persons' ability	u_{00j}	u_{00j}	u_{00j}	u_{0j}
SD of persons' ability	$\sqrt{\tau_\theta}$	$\sqrt{\tau_\theta}$	$\sqrt{\tau_\theta}$	$\sqrt{\tau_\theta}$
SD of testlet effects	$\sqrt{\tau_t}$	$\sqrt{\tau_t}$		
SD of content clustering effects	$\sqrt{\tau_c}$		$\sqrt{\tau_c}$	
Number of Parameters				
Fixed Effects				
Item difficulty	*mk*	*mk*	*mk*	*mk*
Random Effects				
Persons' ability	*s*	*s*	*s*	*s*
SD of persons' ability	1	1	1	1
SD of testlet effects	*m*	*m*		
SD of content clustering effects	2		2	

Note: *m* represents the number of testlets; *k* represents the number of items per testlet; *s* represents sample size

In this study, the likelihood of the observed response patterns is:

$$L = P(\{R = r\} \mid \{u_{00j}, w_{0tj}, w_{0cj}, \gamma_{i00}\}) = \prod_{i=1}^{l} P_i^{r_{ij}} (1-P)_i^{(1-r_{ij})}, \tag{30}$$

where P is the probability function (Equation 26, 27, 28, 29) corresponding each estimating model. l represents the number of items in the test. R is the response pattern for person j, and r_{ij} represents person j's response to item i. In the process of LAPLACE estimation, for each model, parameter estimates are determined by minimizing twice the negative of the resulting log-likelihood approximation:

$$
\begin{aligned}
-2\ln(L) &= -2\ln\{P(\{R=r\}\,|\,\{u_{00j}, w_{0ij}, w_{0cj}, \gamma_{i00}\})\} \\
&= -2\sum_{i=1}^{l}[r_{ij}\ln P_i + (1-r_{ij})\ln(1-P_i)].
\end{aligned} \tag{31}
$$

Analyses

Parameter Estimates

The estimation results were summarized separately for the two simulation studies. For each study, parameter estimates were examined by comparing them with the true values used for data generation. The four models were compared in terms of bias, relative bias, root mean square error (RMSE), and standard error (SE) in corresponding fixed and random effect estimates. The four error indexes were selected to represent different types of errors in estimation. These indexes were examined both descriptively and using analysis of variance (ANOVA).

Bias. The error index, bias, represents the systematic error in estimation. To determine whether estimates were consistently too high or too low, the bias of the parameter estimate was calculated for each estimated parameter. It is defined as:

$$
Bias(\eta) = \frac{\sum_{r=1}^{R}\left(\eta_r - \eta\right)}{R}, \tag{32}
$$

where η represents the true value for the parameter η, η_r is the estimated value for the r^{th} replication, and R is the number of replications (in this study, R=50).

Relative Bias. The relative bias provides a measure of the magnitude of the bias. The relative bias of parameter estimates has been broadly used to estimate the difference between the average of estimated values across replications and the true

value over the true parameter value (e.g., Chen, 2010; Luo, 2007; Meyer & Beretvas, 2006; Ren, 2011). It is defined as:

$$B(\eta) = \frac{\bar{\eta} - \eta}{\eta}, \tag{33}$$

where $\bar{\eta}$ is the mean of the parameter estimation across the 50 replications. In the general statistical world, relative bias with an absolute value less than 0.05 is considered acceptable (Hoogland & Boomsma, 1998); however, this criterion may not apply to IRT models. It should be noted that, in this study, some true values were generated to be zero; under those conditions, the relative bias was not applied.

Root Mean Square Error (RMSE). RMSE represents the total error in estimation. It is used as a measure of the precision of the parameter estimates, and it is defined as:

$$RMSE(\eta) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\eta_r - \eta\right)^2}. \tag{34}$$

Since the calculation of the mean square error involves the sum of the squared bias and its variance, the RMSE captures bias and variability of estimation simultaneously (Enders, 2001).

Standard Error (SE). SE represents the random error in estimation. It provides an estimate of the standard deviation of the parameter estimates, and therefore, SE could be used to evaluate the consistency of the estimates across replications. It is defined as:

$$SE(\eta) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\eta_r - \frac{\sum_{r=1}^{R}\eta_r}{R}\right)^2} \quad . \tag{35}$$

Analysis of Variance (ANOVA). To determine the impacts of the manipulated factors, i.e. sample size, number of testlets, testlet size, magnitude of the testlet effects, and magnitude of the content clustering effects, ANOVAs were conducted with the four criteria above as dependent variables respectively and each manipulated variable and model as factors. Therefore, the ANOVAs included five between-subject factors (manipulated variables) and one within-subject factor (models). Following previous research, the 0.05 alpha-level was used to determine statistical significance (e.g., Meyers & Beretvas, 2006; Ren, 2011). In addition, for effects that resulted in statistical significance, eta-squared (η^2) effect sizes were computed as a measure of practical significance. η^2 was calculated by dividing the sum of squares for the effect by the total sum of squares. η^2 values of 0.01, 0.06, and 0.14 represent small, moderate, and large effect sizes, respectively (Cohen, 1988).

Model Selection

PROC GLIMMIX reports five information criteria, all of which were used to provide information about overall model fit and identify the best fitting model. The five fit indices include Akaike information criterion (AIC; Akaike, 1973), an adjusted AIC for small sample sizes (AICC; Sugiura, 1978), bayesian information criterion (BIC; Schwartz, 1978), consistent Akaike information criterion (CAIC; Bozdogan, 1987), and Hannan-Quinn information criterion (HQIC; Hannan and Quinn, 1979).

Specifically, AIC is defined as:

$$AIC_k = -2\ln L_k + 2q_k\,. \tag{36}$$

where q_k is the number of parameters and $\ln L_k$ is the log-likelihood attained by model *k*. The AICC, a version of the AIC index that is corrected for small sample sizes, is defined as:

$$AICC_k = AIC_k + \frac{2q_k(q_k+1)}{N-q_k-1}, \tag{37}$$

where the correction term strengthens the penalty for smaller sample size and approaches zero for large sample size. The BIC criterion penalizes models with additional parameters more severely than does AIC; it adjusts for the number of observations (*N*), and is defined as:

$$BIC_k = -2\ln L_k + q_k \ln(N)\,. \tag{38}$$

Similar to BIC, the CAIC criterion also tends to penalize complex model by adjusting for number of observations, and is defined as:

$$CAIC_k = -2\ln L_k + q_k \ln(N+1)\,. \tag{39}$$

Finally, the HQIC criterion is defined as:

$$HQIC_k = -2\ln L_k + 2q_k \ln(\ln(N))\,, \tag{40}$$

The $-2\ln L_k$ term appearing in each formula is an estimate of the deviance of the model fit. The coefficients for q_k in the second part of each formula show the degree to which the number of model parameters is being penalized. Taking the sample size ($N \geq 500$) in this study into consideration, the AIC and AICC are expected to have close values and have the least penalty; BIC and CAIC are close and have the largest penalty; HQIC holds the middle ground. Smaller values for these fit

indices indicate better fit. Under each condition, the proportion of replications that lead to correct model identifications is tallied.

Chapter 4: Result

In the present study, two simulation studies were conducted, with simulation study 1 generating equal SDs across testlets as well as equal SDs across content areas while simulation study 2 generating mixed SDs across testlets and mixed SDs across content areas. For both simulation studies, none of the 50 replications under each condition encountered convergence problems for each estimating model. In addition, no inadmissible estimates, such as negative variance estimates, were detected.

For both simulation studies, estimates of item difficulty, persons' ability, SD of persons' ability were obtained using four models, including the proposed cross-classified model, the multilevel model with testlet effects, the multilevel model with content clustering effects, and the Rasch model. Testlet effects' SD was obtained using the proposed model and the multilevel model with testlet effects. Content effects' SD was obtained using the proposed model and the multilevel model with content clustering effects. Bias, relative bias, RMSE, and SE for model parameter estimates were obtained based on 50 replications under each condition for both simulation studies. Analysis of variance (ANOVA) was conducted first by specifying each of the four error indexes as the dependent variable and the five manipulated simulation variables and model as six factors. Based on the results of ANOVAs, main and interaction effects that were identified to have significant impacts on bias, relative bias, RMSE and SE were identified, respectively. An alpha-level of 0.05 paired with a minimum value of 0.01 for eta-squared was used as cutoff for practical significance. Significant main effects, two-way interaction between the magnitude of the testlet effects and model, and two-way interaction between the magnitude of the

content clustering effects and model, were given further interpretations. Other identified significant two-way interactions or more complex multi-way interactions, even though reported in the significance table, were not given further interpretations by considering the research interest for this study.

For each condition, two manipulated variables, magnitude of the testlet effects (hereafter simplified as testlet effect) and magnitude of the content clustering effects (hereafter simplified as content effect), determined which model was the true model, the under-parameterized model, the over-parameterized model, and the mis-specified model among the four models (see Table 7). Mis-specified model here occurred in two situations: when the true model was the multilevel model with testlet effects, the multilevel model with content clustering effects was the mis-specified model; or when the true model was the multilevel model with content clustering effects, the multilevel model with testlet effects was the mis-specified model (see Table 7). It is expected that the rankings of the four models' performance, to a great extent, are determined by the two manipulated variables, testlet effect and content effect. Therefore, averages for the four error indexes, bias, relative bias, RMSE, and SE, were provided for each model under the aggregated sixteen conditions formed by the four levels of testlet effect and the four levels of content effect across the other three manipulated variables, sample size, number of testlets, and number of items per testlet.

This chapter is composed of four sections. Analysis of item difficulty is presented first. Then, estimates of persons' ability are compared and discussed. The recovery of three SDs, including SD of persons' ability, SD of testlet effects, and SD

of content effects, are analyzed in the third section. Finally, fit index results are presented.

Table 7

Determine the True, Over-Parameterized, Under-Parameterized, and Mis-specified Model by Testlet Effects' SD and Content Effects' SD

Testlet effects' SD	Content Effects' SD	Estimating Model			
		Cross-Classified	Multilevel with Testlet Effects	Multilevel with Content Effects	Rasch
0	0	Over	Over	Over	True
	0.5	Over	Mis	True	Under
	1	Over	Mis	True	Under
	1.5	Over	Mis	True	Under
0.5	0	Over	True	Mis	Under
	0.5	True	Under	Under	Under
	1	True	Under	Under	Under
	1.5	True	Under	Under	Under
1	0	Over	True	Mis	Under
	0.5	True	Under	Under	Under
	1	True	Under	Under	Under
	1.5	True	Under	Under	Under
1.5	0	Over	True	Mis	Under
	0.5	True	Under	Under	Under
	1	True	Under	Under	Under
	1.5	True	Under	Under	Under

Note: "True" means this model is the true model; "Over" means this model is over-parameterized; "Under" means this model is under-parameterized; and "Mis" means this model is mis-specified.

Estimation of Item Difficulty

Bias. Based on the full factorial six-way ANOVA results, for both simulation studies, none of the main effects and none of the interaction effects significantly impacted bias with a value of η^2 larger than 0.01. However, even though the impact of the calibration model on bias in item difficulty estimation was not significant, consistent patterns were observed when the magnitude of testlet effects and the magnitude of content clustering effects changed (see Table 8). Generally, the over-parameterized model and the true model had average biases that were close to zero across the sixteen aggregated conditions, while the under-parameterized model or the mis-specified model had relatively larger average biases. Since the proposed cross-classified model was the true model or an over-parameterized model across all of the sixteen aggregated conditions, it always had less average biases than the other three models. This was consistent with the expectations that the proposed model, which properly accounted for both the testlet effects and the content clustering effects, should have less systematic estimation error.

Table 8

Average Biases in Item Difficulty Estimation by Testlet Effect and Content Effect across the Other Manipulated Variables

			Estimating Model			
	Testlet Effects' SD	Content Effects' SD	Cross-Classified	Multilevel with Testlet Effects	Multilevel with Content Effects	Rasch
Simulation 1	0	0	0.0008	0.0008	0.0006	0.0005
		0.5	0.0000	-0.0035	-0.0003	-0.0039
		1	0.0007	-0.0122	0.0005	-0.0126
		1.5	-0.0022	-0.0278	-0.0023	-0.0290
	0.5	0	-0.0015	-0.0015	-0.0069	-0.0070
		0.5	-0.0008	-0.0044	-0.0060	-0.0097
		1	0.0000	-0.0134	-0.0052	-0.0177
		1.5	-0.0008	-0.0264	-0.0059	-0.0299
	1	0	0.0006	0.0005	-0.0193	-0.0194
		0.5	0.0031	-0.0001	-0.0170	-0.0191
		1	-0.0001	-0.0118	-0.0193	-0.0286
		1.5	-0.0022	-0.0287	-0.0211	-0.0407
	1.5	0	-0.0030	-0.0030	-0.0403	-0.0403
		0.5	0.0000	-0.0041	-0.0379	-0.0395
		1	0.0022	-0.0106	-0.0344	-0.0414
		1.5	0.0008	-0.0227	-0.0349	-0.0496
Simulation 2	0	0	-0.0001	-0.0002	-0.0004	-0.0005
		0.5	-0.0016	-0.0009	-0.0019	-0.0012
		1	0.0009	-0.0188	0.0006	-0.0192
		1.5	-0.0024	-0.0344	-0.0025	-0.0360
	0.5	0	0.0003	0.0002	-0.0081	-0.0083
		0.5	0.0003	-0.0031	-0.0084	-0.0112
		1	-0.0006	-0.0139	-0.0092	-0.0208
		1.5	-0.0030	-0.0266	-0.0110	-0.0330
	1	0	-0.0019	-0.0019	-0.0223	-0.0224
		0.5	-0.0008	-0.0083	-0.0208	-0.0281
		1	0.0001	-0.0129	-0.0212	-0.0297
		1.5	0.0017	-0.0237	-0.0189	-0.0388
	1.5	0	-0.0018	-0.0018	-0.0377	-0.0380
		0.5	0.0006	-0.0054	-0.0358	-0.0394
		1	-0.0016	-0.0185	-0.0362	-0.0485
		1.5	-0.0036	-0.0272	-0.0394	-0.0518

Relative Bias. For simulation study 1, the six-way ANOVA results (see Table 40 in Appendix E) indicated that three significant main effects, testlet effect (F (3, 21696) = 29124.3, $p < 0.001$), content effect (F (3, 21696) = 12110.2, $p < 0.001$), and model (F (3, 21696) = 48980.8, $p < 0.001$) were found to have a large ($\eta^2 = 0.153$), moderate ($\eta^2 = 0.064$), and large ($\eta^2 = 0.257$) impact on relative bias, respectively. The two-way interactions, including interaction between testlet effect and model (F (9, 21696) = 9458.4, $p < 0.001$) and interaction between content effect and model (F (9, 21696) = 4244.6, $p < 0.001$), were identified to have a large ($\eta^2 = 0.149$) and moderate ($\eta^2 = 0.067$) impact on relative bias, respectively.

The same main effects and two-way interactions were identified to be significant in simulation study 2, even though with different effect sizes from simulation study 1. Testlet effect (F (3, 21696) = 4113.1, $p < 0.001$), content effect (F (3, 21696) = 1929.3, $p < 0.001$), and model (F (3, 21696) = 8105.4, $p < 0.001$) each had a moderate ($\eta^2 = 0.118$), small ($\eta^2 = 0.056$), and large ($\eta^2 = 0.232$) impact on relative bias, respectively. The two-way interaction between testlet effect and model (F (9, 21696) = 1286.4, $p < 0.001$) and the two-way interaction between content and model (F (9, 21696) = 647.9, $p < 0.001$) each had a moderate ($\eta^2 = 0.110$) and a small ($\eta^2 = 0.056$) impact on relative bias.

No absolute value of the average relative bias was larger than 0.006 when the proposed cross-classified model was the estimating model across all of the sixteen aggregated conditions for both simulation studies (see Table 9). The multilevel model with testlet effects was found to have average relative biases ranging from -0.162 to -

0.075 when the data were generated with content clustering effects' SD as 1 and 1.5. The multilevel model with content clustering effects was identified to have average relative biases ranging from -0.236 to -0.118 when the data were generated with large testlet effects' SD as 1 and 1.5. The absolute values of average relative biases for the Rasch model were all larger than 0.05 except when both the testlet effects' SD and the content clustering effects' SD were 0 or 0.5. In addition, it appears that no matter whether the testlet effects' SDs and the content clustering effects' SDs were generated as being equal across testlets and across content areas (simulations study 1) or not (simulation study 2), for the multilevel model with testlet effects and the Rasch model, the larger the average of the content clustering effects' SD, the larger the magnitude of the relative bias; for the multilevel model with content clustering effects and the Rasch model, the larger the average of the testlet effects' SD, the larger the magnitude of the relative bias.

In summary, the proposed cross-classified model had relatively smaller average relative biases than the other three models. In addition, the higher the magnitude of the testlet effects and the magnitude of the content clustering effects, the larger the average relative biases for the two multilevel models and the Rasch model. This was consistent with the expectations that increasing the magnitude of the testlet effects and the content clustering effects would lead the two multilevel models and the Rasch model, which inappropriately ignored the testlet effects and/or the content clustering effects, to perform even worse by having more systematic errors.

Table 9

Average Relative Biases in Item Difficulty Estimation by Testlet Effect and Content Effect across the Other Manipulated Variables

			Estimating Model			
	Testlet Effects' SD	Content Effects' SD	Cross-Classified	Multilevel with Testlet Effects	Multilevel with Content Effects	Rasch
Simulation 1	0	0	0.0057	0.0052	0.0037	0.0031
		0.5	0.0041	-0.0186	0.0024	-0.0213
		1	0.0039	**-0.0790**	0.0024	**-0.0831**
		1.5	0.0027	**-0.1516**	0.0013	**-0.1624**
	0.5	0	0.0001	-0.0004	-0.0352	-0.0359
		0.5	-0.0017	-0.0236	-0.0357	**-0.0572**
		1	-0.0018	**-0.0851**	-0.0341	**-0.1128**
		1.5	-0.0033	**-0.1529**	-0.0330	**-0.1800**
	1	0	0.0005	0.0001	**-0.1275**	**-0.1281**
		0.5	0.0006	-0.0213	**-0.1246**	**-0.1421**
		1	-0.0006	**-0.0781**	**-0.1212**	**-0.1829**
		1.5	-0.0001	**-0.1574**	**-0.1182**	**-0.2357**
	1.5	0	0.0024	0.0021	**-0.2308**	**-0.2327**
		0.5	-0.0010	-0.0230	**-0.2355**	**-0.2450**
		1	-0.0005	**-0.0749**	**-0.2264**	**-0.2694**
		1.5	-0.0005	**-0.1472**	**-0.2215**	**-0.3085**
Simulation 2	0	0	0.0047	0.0041	0.0026	0.0019
		0.5	0.0027	-0.0246	0.0008	-0.0272
		1	0.0021	**-0.0822**	0.0006	**-0.0875**
		1.5	0.0036	**-0.1526**	0.0021	**-0.1608**
	0.5	0	0.0012	0.0007	**-0.0542**	**-0.0548**
		0.5	-0.0006	-0.0281	**-0.0562**	**-0.0793**
		1	-0.0003	**-0.0845**	**-0.0534**	**-0.1267**
		1.5	-0.0028	**-0.1569**	**-0.0530**	**-0.1964**
	1	0	0.0012	0.0008	**-0.1308**	**-0.1322**
		0.5	-0.0002	-0.0258	**-0.1306**	**-0.1524**
		1	-0.0009	**-0.0865**	**-0.1342**	**-0.1919**
		1.5	-0.0016	**-0.1525**	**-0.1262**	**-0.2449**
	1.5	0	0.0015	0.0012	**-0.2281**	**-0.2294**
		0.5	-0.0001	-0.0254	**-0.2275**	**-0.2415**
		1	-0.0010	**-0.0811**	**-0.2251**	**-0.2715**
		1.5	-0.0033	**-0.1618**	**-0.2247**	**-0.3147**

RMSE. A six-way ANOVA was also conducted by specifying the RMSE in item difficulty estimation as the dependent variable. For simulation study 1, based on the ANOVA results (see Table 41 in Appendix E), the RMSE was significantly affected by sample size (F (2, 25152) = 521.9, p < 0.001), testlet effect (F (3, 25152) = 1763.0, p < 0.001), content effect (F (3, 25152) = 644.7, p < 0.001), and model (F (3, 25152) = 1988.8, p < 0.001), each with a small (η^2 = 0.020), moderate (η^2 = 0.099), small (η^2 = 0.036), and moderate (η^2 = 0.112) effect size, respectively; the two-way interaction between testlet effect and model (F (9, 25152) = 519.9, p < 0.001) was significant with a moderate effect size (η^2 = 0.088), and the two-way interaction between content effect and model (F (9, 25152) = 197.5, p < 0.001) was significant with a small effect size (η^2 = 0.033).

For simulation study 2, the identified significant main effects and interaction effects were the same as simulation study 1, even though with different effect sizes: sample size (F (2, 25152) = 322.7, p < 0.001, η^2 = 0.015), testlet effect: (F (2, 25152) = 1039.1, p < 0.001, η^2 = 0.071), content effect: (F (2, 25152) = 440.6, p < 0.001, η^2 = 0.030), model: (F (2, 25152) = 1519.6, p < 0.001, η^2 = 0.104), interaction between testlet effect and model: (F (2, 25152) = 306.6, p < 0.001, η^2 = 0.063), and interaction between content effect and model: (F (2, 25152) = 133.4, p < 0.001, η^2 = 0.028).

For both simulation studies, as sample size increased, the average RMSEs became smaller for each model (see Table 10). This indicates that larger sample size

resulted in more accurate item parameter estimation, which is consistent with the expectations.

The proposed cross-classified model had the lowest average RMSEs when both the testlet effects' SD and the content clustering effects' SD was nonzero (see Table 11). In addition, when the testlet effects' SD and/or the content effects' SD were zero, the average RMSEs for the proposed cross-classified model were close to the true model, but smaller than the mis-specified and/or the under-parameterized model. Moreover, the differences in the average RMSEs among the four models were smaller for small magnitude of the testlet effects and the content clustering effects (SD = 0.5) than those for large magnitude of the testlet effects and the content clustering effects (SD = 1 or 1.5). This was consistent with the results from the six-way ANOVA, where the interactions between model and testlet/content effect were significant.

The lower average RMSEs in the proposed cross-classified model was consistent with the expectations that the proposed model should have less total estimation error. In addition, as the magnitude of the testlet effects and the magnitude of the content clustering effects became larger, the two multilevel models and the Rasch model should have more total estimation errors by inappropriately ignoring the testlet effects and/or content clustering effects, which was also consistent with the expectations.

Table 10

Average RMSEs in Item Difficulty Estimation by Sample Size across the Other Manipulated Variables

		Estimating Model			
	Sample Size	Cross-Classified	Multilevel with Testlet Effects	Multilevel with Content Effects	Rasch
Simulation 1	500	0.1239	0.1577	0.1863	0.2300
	1000	0.0876	0.1287	0.1626	0.2122
	2000	0.0616	0.1115	0.1476	0.2020
Simulation 2	500	0.1242	0.1629	0.1936	0.2410
	1000	0.0874	0.1342	0.1708	0.2229
	2000	0.0621	0.1173	0.1558	0.2119

Table 11

Average RMSEs in Item Difficulty Estimation by Testlet Effect and Content Effect across the Other Manipulated Variables

	Testlet Effects' SD	Content Effects' SD	Estimating Model			
			Cross-Classified	Multilevel with Testlet Effects	Multilevel with Content Effects	Rasch
Simulation 1	0	0	0.0865	0.0864	0.0860	0.0859
		0.5	0.0863	0.0884	0.0859	0.0890
		1	0.0881	0.1319	0.0878	0.1359
		1.5	0.0912	0.2086	0.0909	0.2205
	0.5	0	0.0865	0.0864	0.0953	0.0956
		0.5	0.0875	0.0916	0.0969	0.1118
		1	0.0891	0.1392	0.0973	0.1656
		1.5	0.0935	0.2129	0.1018	0.2425
	1	0	0.0892	0.0891	0.1804	0.1811
		0.5	0.0905	0.0938	0.1793	0.1975
		1	0.0919	0.1371	0.1772	0.2448
		1.5	0.0946	0.2198	0.1767	0.3072
	1.5	0	0.0929	0.0929	0.3002	0.3026
		0.5	0.0938	0.0971	0.3049	0.3159
		1	0.0963	0.1351	0.2958	0.3461
		1.5	0.0984	0.2117	0.2920	0.3937
Simulation 2	0	0	0.0851	0.0849	0.0845	0.0844
		0.5	0.0868	0.0980	0.0863	0.0984
		1	0.0887	0.1452	0.0884	0.1488
		1.5	0.0921	0.2153	0.0918	0.2238
	0.5	0	0.0875	0.0874	0.1192	0.1194
		0.5	0.0879	0.1002	0.1187	0.1384
		1	0.0896	0.1455	0.1179	0.1866
		1.5	0.0937	0.2188	0.1185	0.2616
	1	0	0.0886	0.0885	0.1907	0.1921
		0.5	0.0906	0.1002	0.1906	0.2129
		1	0.0915	0.1469	0.1926	0.2567
		1.5	0.0955	0.2166	0.1871	0.3189
	1.5	0	0.0935	0.0934	0.2987	0.3002
		0.5	0.0943	0.1011	0.2976	0.3134
		1	0.0960	0.1422	0.2962	0.3487
		1.5	0.0984	0.2261	0.2956	0.4005

SE. Item difficulty recovery was also evaluated and compared in terms of SE. For simulation study 1, three effects were found to have significant impacts on SE (see Table 42 in Appendix E). Two main effects, sample size (F (2, 25152) = 31960.5, p<0.001) and model (F (3, 25152) = 1137.6, p<0.001), were significant factors, each with a large (η^2 = 0.558) and a small (η^2 = 0.030) effect size. Two-way interaction between testlet effect and model had a small impact on SE (F (9, 25152) = 215.4, p<0.001, η^2 = 0.017).

The same patterns were detected for simulation study 2. ANOVA results (see Table 42 in Appendix E) indicated that SE was significantly affected by sample size (F (2, 25152) = 30280.5, p<0.001) and model (F (3, 25152) = 1207.2, p<0.001), each with a large (η^2 = 0.546) and a small effect size (η^2 = 0.033), respectively. The interaction between testlet effect and model (F (9, 25152) = 189.9, p<0.001) was significant with a small effect size (η^2 = 0.015).

Matching the ANOVA results which indicated that sample size had a large impact on SE, Table 12 shows that, for both simulation studies, the larger the sample size, the smaller the average SEs for each of the four models. This indicates that increasing sample size produced more stable item parameter estimates and less random estimation errors.

Generally, the proposed cross-classified model had the largest average SEs among the four models across the sixteen aggregated conditions, and the Rasch model had the smallest average SEs (Table 13). No consistent patterns were found when comparing the two multilevel models. The higher average SEs in the proposed cross-

classified model might have been due to the increased number of parameters estimated.

Table 12

Average SEs in Item Difficulty Estimation by Sample Size across the Other Manipulated Variables

		Estimating Model			
	Sample Size	Cross-Classified	Multilevel with Testlet Effects	Multilevel with Content Effects	Rasch
Simulation 1	500	0.1219	0.1136	0.1088	0.1020
	1000	0.0862	0.0805	0.0770	0.0721
	2000	0.0605	0.0565	0.0540	0.0506
Simulation 2	500	0.1222	0.1134	0.1083	0.1012
	1000	0.0859	0.0800	0.0761	0.0712
	2000	0.0609	0.0565	0.0540	0.0505

Table 13

Average SEs in Item Difficulty Estimation by Testlet Effect and Content Effect across the Other Manipulated Variables

	Testlet Effects' SD	Content Effects' SD	Estimating Model			
			Cross-Classified	Multilevel with Testlet Effects	Multilevel with Content Effects	Rasch
Simulation 1	0	0	0.0854	0.0853	0.0850	0.0850
		0.5	0.0851	0.0830	0.0849	0.0826
		1	0.0871	0.0798	0.0868	0.0793
		1.5	0.0898	0.0758	0.0896	0.0746
	0.5	0	0.0854	0.0853	0.0817	0.0816
		0.5	0.0864	0.0842	0.0828	0.0808
		1	0.0879	0.0803	0.0842	0.0773
		1.5	0.0919	0.0776	0.0889	0.0749
	1	0	0.0880	0.0879	0.0759	0.0758
		0.5	0.0891	0.0869	0.0771	0.0753
		1	0.0902	0.0829	0.0782	0.0726
		1.5	0.0926	0.0775	0.0808	0.0696
	1.5	0	0.0913	0.0913	0.0689	0.0687
		0.5	0.0920	0.0899	0.0686	0.0677
		1	0.0944	0.0869	0.0717	0.0674
		1.5	0.0960	0.0816	0.0736	0.0652
Simulation 2	0	0	0.0840	0.0839	0.0836	0.0835
		0.5	0.0858	0.0833	0.0855	0.0828
		1	0.0877	0.0802	0.0875	0.0795
		1.5	0.0906	0.0758	0.0904	0.0751
	0.5	0	0.0864	0.0863	0.0810	0.0809
		0.5	0.0868	0.0841	0.0812	0.0790
		1	0.0880	0.0803	0.0829	0.0761
		1.5	0.0920	0.0775	0.0868	0.0733
	1	0	0.0873	0.0873	0.0747	0.0745
		0.5	0.0889	0.0865	0.0762	0.0741
		1	0.0900	0.0820	0.0771	0.0716
		1.5	0.0933	0.0785	0.0806	0.0692
	1.5	0	0.0918	0.0918	0.0693	0.0691
		0.5	0.0925	0.0898	0.0701	0.0685
		1	0.0937	0.0859	0.0712	0.0669
		1.5	0.0956	0.0799	0.0731	0.0645

Estimation of Persons' Ability

The recovery of persons' ability was evaluated and compared in terms of bias, RMSE, and SE. Since the true values of persons' ability parameter were randomly generated from a standard normal distribution, $N(0,\ 1)$, plenty of the generated true values would be very close to 0, and then the calculated relative bias would become extremely large. Therefore, the relative bias index is not appropriate to be used in assessing persons' ability recovery.

For both simulation studies, a six-way ANOVA was conducted by specifying each of the three error indexes, bias, RMSE, and SE, as the dependent variable and the five manipulated variables and model as factors.

Bias. For both simulation studies, the six-way ANOVA results indicated that none of the main effects and none of the interactions had a significant impact on bias in the ability parameter estimation. In addition, all of the average biases in Table 14 were close to zero. A possible explanation for this result was that all of the four models were identified by constraining the mean ability to zero.

Table 14

Average Biases in Persons' Ability Estimation by Testlet Effect and Content Effect across the Other Manipulated Variables

			Estimating Model			
	Testlet effects' SD	Content Effects' SD	Cross-Classified	Multilevel with Testlet Effects	Multilevel with Content Effects	Rasch
Simulation 1	0	0	-0.0013	-0.0014	-0.0013	-0.0013
		0.5	-0.0015	-0.0015	-0.0014	-0.0014
		1	-0.0014	-0.0017	-0.0014	-0.0016
		1.5	-0.0056	-0.0023	-0.0055	-0.0020
	0.5	0	-0.0024	-0.0024	-0.0014	-0.0013
		0.5	-0.0024	-0.0025	-0.0016	-0.0014
		1	-0.0021	-0.0025	-0.0017	-0.0016
		1.5	-0.0024	-0.0027	-0.0019	-0.0019
	1	0	-0.0037	-0.0037	-0.0014	-0.0014
		0.5	-0.0041	-0.0039	-0.0019	-0.0015
		1	-0.0032	-0.0040	-0.0016	-0.0015
		1.5	-0.0031	-0.0042	-0.0020	-0.0018
	1.5	0	-0.0045	-0.0045	-0.0014	-0.0013
		0.5	-0.0040	-0.0047	-0.0010	-0.0014
		1	-0.0060	-0.0052	-0.0034	-0.0016
		1.5	-0.0027	-0.0053	-0.0008	-0.0016
Simulation 2	0	0	-0.0014	-0.0014	-0.0013	-0.0013
		0.5	-0.0034	-0.0015	-0.0033	-0.0014
		1	-0.0008	-0.0016	-0.0008	-0.0015
		1.5	-0.0053	-0.0022	-0.0054	-0.0019
	0.5	0	-0.0024	-0.0024	-0.0014	-0.0014
		0.5	-0.0036	-0.0025	-0.0022	-0.0015
		1	-0.0041	-0.0027	-0.0025	-0.0016
		1.5	-0.0029	-0.0027	-0.0025	-0.0017
	1	0	-0.0033	-0.0033	-0.0014	-0.0013
		0.5	-0.0014	-0.0034	-0.0010	-0.0014
		1	-0.0048	-0.0039	-0.0027	-0.0016
		1.5	-0.0021	-0.0039	-0.0015	-0.0017
	1.5	0	-0.0043	-0.0043	-0.0014	-0.0014
		0.5	-0.0033	-0.0044	-0.0013	-0.0014
		1	-0.0013	-0.0047	-0.0018	-0.0015
		1.5	-0.0038	-0.0049	-0.0019	-0.0016

RMSE. ANOVA results (see Table 43 in Appendix E) from both simulation studies indicated that two factors, number of testlets and content effect, significantly impacted RMSE in the ability parameter estimation; all the other effects were negligible. For simulation study 1, both number of testlets (F (1, 895,232) = 10494.3) and content effect (F (3, 895,232) = 16982.9) had small effects (η^2 = 0.010 and η^2 = 0.050, respectively) on RMSE. For simulation study 2, the impacts of number of testlets (F (1, 895,232) = 11919.1) and content effect (F (3, 895,232) = 16271.4) were also small (η^2 = 0.012 and η^2 = 0.048, respectively).

In both stimulation studies, for each of the four models, when data were generated with number of testlets as 3, the average RMSE in persons' ability parameter estimation was slightly higher than that generated with number of testlets as 6 (see Table 15). This indicates that the number of testlets influenced the accuracy of persons' ability estimation: the larger the number of testlets, the more accurate the ability parameter estimation.

Matching the ANOVA results, content effect's impact on the RMSE was reflected in Table 16. In both simulation studies, as the magnitude of the content effects increased, the average RMSEs increased for both the multilevel model with testlet effects and the Rasch model. This is due to the fact that as the magnitude of the content clustering effects becomes larger, the total estimation error for the multilevel model with testlet effects and the Rasch model is expected to become larger.

Generally, the proposed cross-classified model had the smallest average RMSEs when both the testlet effects' SD and the content clustering effects' SD were nonzero (see Table 17). This can be explained by the proper modeling of the persons'

ability, testlet effect, and content effect separately in the proposed cross-classified model. When the proposed cross-classified model was not the true model but an over-parameterized model, the average RMSEs associated with the proposed model were close to the average RMSEs associated with the true model, which were smaller than the average RMSEs associated with the other two models. This indicates that over-parameterization would not result in larger RMSE than the true model.

Table 15

Average RMSEs in Persons' Ability Estimation by Number of Testlets across the Other Manipulated Variables

		Estimating Model			
	Number of Testlets	Cross-Classified	Multilevel with Testlet Effects	Multilevel with Content Effects	Rasch
Simulation 1	3	0.6111	0.6610	0.6228	0.6814
	6	0.5377	0.5872	0.5380	0.5863
Simulation 2	3	0.6116	0.6633	0.6283	0.6880
	6	0.5336	0.5891	0.5382	0.5912

Table 16

Average RMSEs in Persons' Ability Estimation by Content Effect across the Other Manipulated Variables

		Estimating Model			
	Content Effects' SD	Cross-Classified	Multilevel with Testlet Effects	Multilevel with Content Effects	Rasch
Simulation 1	0	0.4936	0.4936	0.5015	0.5019
	0.5	0.5355	0.5433	0.5428	0.5539
	1	0.6079	0.6630	0.6134	0.6753
	1.5	0.6607	0.7964	0.6640	0.8042
Simulation 2	0	0.4907	0.4908	0.5077	0.5083
	0.5	0.5349	0.5522	0.5474	0.5698
	1	0.6045	0.6711	0.6141	0.6830
	1.5	0.6601	0.7906	0.6639	0.7973

Table 17

Average RMSEs in Persons' Ability Estimation by Testlet Effect and Content Effect across the Other Manipulated Variables

	Testlet Effects' SD	Content Effects' SD	Estimating Model			
			Cross-Classified	Multilevel with Testlet Effects	Multilevel with Content Effects	Rasch
Simulation 1	0	0	0.4163	0.4163	0.4494	0.4162
		0.5	0.4840	0.4950	0.4840	0.4951
		1	0.5791	0.6463	0.5792	0.6463
		1.5	0.6450	0.7999	0.6450	0.7984
	0.5	0	0.4527	0.4527	0.4541	0.4542
		0.5	0.5074	0.5173	0.5089	0.5201
		1	0.5904	0.6512	0.5915	0.6541
		1.5	0.6530	0.8063	0.6539	0.8003
	1	0	0.5209	0.5210	0.5305	0.5306
		0.5	0.5515	0.5582	0.5602	0.5713
		1	0.6164	0.6681	0.6230	0.6836
		1.5	0.6635	0.7947	0.6670	0.8079
	1.5	0	0.5844	0.5844	0.6052	0.6067
		0.5	0.5992	0.6025	0.6183	0.6289
		1	0.6456	0.6865	0.6601	0.7173
		1.5	0.6812	0.7848	0.6902	0.8103
Simulation 2	0	0	0.4152	0.4153	0.4152	0.4152
		0.5	0.4816	0.5076	0.4816	0.5078
		1	0.5723	0.6461	0.5723	0.6450
		1.5	0.6488	0.8119	0.6488	0.8103
	0.5	0	0.4604	0.4604	0.4729	0.4731
		0.5	0.5081	0.5288	0.5150	0.5402
		1	0.5899	0.6719	0.5947	0.6783
		1.5	0.6555	0.7904	0.6570	0.7910
	1	0	0.5141	0.5141	0.5364	0.5371
		0.5	0.5481	0.5606	0.5629	0.5819
		1	0.6146	0.6848	0.6273	0.6954
		1.5	0.6617	0.7867	0.6644	0.7941
	1.5	0	0.5732	0.5732	0.6061	0.6079
		0.5	0.6019	0.6119	0.6302	0.6495
		1	0.6413	0.6817	0.6618	0.7134
		1.5	0.6743	0.7736	0.6855	0.7939

SE. Regarding the SE of persons' ability parameters, all of the main effects except sample size were significant; the two-way interaction between content effect and model was also significant; all the other effects were negligible (see Table 44 in Appendix E). For simulation study 1, all of the identified significant effects had a moderate effect size: number of testlets (F (1, 895,232) =571286 , p<0.001, η^2 = 0.117); number of items per testlet (F (1, 895,232) = 582776, p<0.001, η^2 = 0.119); testlet effect (F (3, 895,232) =108957 , p<0.001, η^2 = 0.067); content effect (F (3, 895,232) = 115775, p<0.001, η^2 = 0.071); model (F (3, 895,232) = 182622, p<0.001, η^2 = 0.112); two-way interaction between content effect and model (F (9, 895,232) = 43113, p<0.001, η^2 = 0.079). For simulation study 2, all of the identified significant effects were moderate except that the testlet effect was small: number of testlets (F (1, 895,232) =529745 , p<0.001, η^2 = 0.128); number of items per testlet (F (1, 895,232) = 476949, p<0.001, η^2 = 0.116); testlet effect (F (3, 895,232) =67336 , p<0.001, η^2 = 0.049); content effect (F (3, 895,232) = 118717, p<0.001, η^2 = 0.087); model (F (3, 895,232) = 111089, p<0.001, η^2 = 0.087); two-way interaction between content effect and model (F (9, 895,232) = 36836, p<0.001, η^2 = 0.081).

For both simulation studies, the average SE was largest when the generated number of testlets was 3 and the testlet size was 5 (number of items per test was 15), and the average SE was smallest when the generated number of testlets was 6 and the testlet size was 10 (number of items per test was 60) (see Table 18). The other two conditions, where the number of items per test was 30 for both, had very close

average SEs. Given this pattern, it appears that, the more items in a test, the smaller the magnitude of the SE.

Generally, the proposed cross-classified model had smaller average SEs than the other three models across the aggregated sixteen conditions, no matter whether the proposed model was the true model or not (see Table 19). However, lower SE in the proposed model is inconsistent with the other findings: generally, the proposed model, which has more parameters to estimate, should have relatively larger SE than the other three simpler models. Further research should investigate whether this result generalizes to other study conditions.

Table 18

Average SEs in Persons' Ability Estimation by Number of Testlets and Testlet Length across the Other Manipulated Variables

			Estimating Model			
	Number of Testlets	Testlet Size	Cross-Classified	Multilevel with Testlet Effects	Multilevel with Content Effects	Rasch
Simulation 1	3	5	0.3137	0.3907	0.3437	0.4181
		10	0.2534	0.3168	0.2872	0.3464
	6	5	0.2676	0.3334	0.2700	0.3333
		10	0.2081	0.2594	0.2124	0.2606
Simulation 2	3	5	0.3341	0.4035	0.3517	0.4150
		10	0.2828	0.3356	0.2961	0.3445
	6	5	0.2888	0.3429	0.2797	0.3304
		10	0.2339	0.2748	0.2239	0.2587

Table 19

Average SEs in Persons' Ability Estimation by Testlet Effect and Content Effect across the other Manipulated Variables

			Estimating Model			
	Testlet Effects' SD	Content Effects' SD	Cross-Classified	Multilevel with Testlet Effects	Multilevel with Content Effects	Rasch
Simulation 1	0	0	0.3659	0.3667	0.3662	0.3670
		0.5	0.3347	0.3653	0.3349	0.3656
		1	0.2650	0.3630	0.2652	0.3626
		1.5	0.2000	0.3570	0.2001	0.3547
	0.5	0	0.3489	0.3496	0.3566	0.3575
		0.5	0.3248	0.3516	0.3309	0.3577
		1	0.2609	0.3509	0.2656	0.3536
		1.5	0.1995	0.3531	0.2028	0.3481
	1	0	0.3060	0.3065	0.3344	0.3352
		0.5	0.2855	0.3100	0.3108	0.3345
		1	0.2358	0.3201	0.2563	0.3350
		1.5	0.1838	0.3272	0.2001	0.3323
	1.5	0	0.2509	0.2511	0.2988	0.3054
		0.5	0.2386	0.2579	0.2886	0.3060
		1	0.2088	0.2797	0.2453	0.3101
		1.5	0.1624	0.2912	0.1969	0.3081
Simulation 2	0	0	0.3648	0.3657	0.3651	0.3661
		0.5	0.3550	0.3647	0.3552	0.3651
		1	0.2867	0.3617	0.2868	0.3617
		1.5	0.2028	0.3555	0.2029	0.3542
	0.5	0	0.3574	0.3581	0.3508	0.3517
		0.5	0.3523	0.3598	0.3465	0.3519
		1	0.2890	0.3588	0.2887	0.3521
		1.5	0.2153	0.3519	0.2142	0.3417
	1	0	0.3370	0.3375	0.3305	0.3318
		0.5	0.3284	0.3402	0.3179	0.3311
		1	0.2751	0.3506	0.2766	0.3340
		1.5	0.2048	0.3408	0.2021	0.3254
	1.5	0	0.2821	0.2824	0.3057	0.3075
		0.5	0.2799	0.2911	0.3003	0.3088
		1	0.2443	0.3004	0.2576	0.3080
		1.5	0.1837	0.3081	0.2047	0.3034

Estimation of Radom Effects' SD

This section presents the estimation of random effects' SD. Analysis of ability's SD is presented first, followed by an analysis of testlet effects' SD. Finally, recovery of content clustering effects' SD is presented.

Estimation of Ability's SD

The true SD of persons' ability was 1 across all conditions. The bias, RMSE, and SE in ability's SD were computed for each condition first, and then a five-way ANOVA was conducted by specifying each of the three error indexes as the dependent variable and the five manipulated variables and model as the factors. The relative bias error index was not used here, since relative bias is the same as bias when the parameter's true value is 1.

Bias. The identified effects with both statistical and practical significance are provided in Table 45 Appendix E. For both simulation studies, content effect, model, and the interaction between the two factors, were found to have large effect sizes.

Table 20 shows that the proposed cross-classified model had smaller average biases than the other three models across the sixteen aggregated conditions. This can be explained by the proper modeling of the persons' ability, testlet effects, and content effects in the proposed model. In addition, as the magnitude of the content effects became larger, the average biases for the multilevel model with testlet effects and the Rasch model became increasingly higher. This might be due to the fact that larger magnitude of the content effects led the multilevel model with testlet effects and the Rasch model to fit the data even worse by ignoring the content effects. However, the magnitude of the testlet effects was not a significant factor. It is

expected that as the magnitude of the testlet effects becomes larger, the average biases for the multilevel model with content effects and the Rasch model, both of which do not account for the testlet effects, become increasingly higher. However, such expected pattern was not observed. Further research need to be conducted to explore the reasons for this unexpected result.

Table 20

Average Biases in Ability's SD Estimation by Testlet Effect and Content Effect across the Other Manipulated Variables

	Testlet effects' SD	Content Effects' SD	Estimating Model			
			Cross-Classified	Multilevel with Testlet Effects	Multilevel with Content Effects	Rasch
Simulation 1	0	0	-0.0031	-0.0020	-0.0036	-0.0025
		0.5	0.0004	0.0374	-0.0001	0.0364
		1	-0.0092	0.1224	-0.0096	0.1201
		1.5	-0.0018	0.2456	-0.0021	0.2346
	0.5	0	-0.0032	-0.0020	-0.0108	-0.0094
		0.5	0.0044	0.0363	-0.0038	0.0268
		1	-0.0094	0.1105	-0.0163	0.0984
		1.5	0.0022	0.2518	-0.0049	0.2304
	1	0	-0.0071	-0.0061	-0.0334	-0.0320
		0.5	-0.0059	0.0306	-0.0330	-0.0017
		1	-0.0232	0.1076	-0.0489	0.0664
		1.5	-0.0363	0.2273	-0.0542	0.1801
	1.5	0	-0.0191	-0.0184	-0.0631	-0.0576
		0.5	-0.0169	0.0152	-0.0543	-0.0332
		1	0.0006	0.1340	-0.0463	0.0525
		1.5	-0.0314	0.2212	-0.0626	0.1296
Simulation 2	0	0	-0.0026	-0.0012	-0.0031	-0.0016
		0.5	-0.0092	0.0315	-0.0097	0.0309
		1	0.0004	0.1280	-0.0001	0.1229
		1.5	-0.0529	0.2151	-0.0532	0.2102
	0.5	0	-0.0085	-0.0074	-0.0208	-0.0190
		0.5	0.0011	0.0444	-0.0132	0.0291
		1	-0.0035	0.1404	-0.0124	0.1236
		1.5	-0.0034	0.2153	-0.0136	0.1914
	1	0	-0.0063	-0.0050	-0.0349	-0.0320
		0.5	-0.0067	0.0256	-0.0404	-0.0081
		1	-0.0173	0.1302	-0.0389	0.0808
		1.5	-0.0335	0.2105	-0.0622	0.1543
	1.5	0	-0.0106	-0.0098	-0.0572	-0.0538
		0.5	-0.0143	0.0303	-0.0583	-0.0216
		1	-0.0035	0.1249	-0.0504	0.0463
		1.5	-0.0175	0.1996	-0.0514	0.1144

RMSE. The effects that were identified to have both statistical and practical significant impacts on the RMSE of ability's SD are presented in Table 46 Appendix E. Same as bias, three effects, including content effect, model, and interaction between content effect and model, had large effect sizes; all the other effects were small or negligible.

The proposed cross-classified model had smaller average RMSEs than the other three models across the sixteen aggregated conditions (see Table 21). This is consistent with the ANOVA results which indicated that model was a significant factor with a large effect size. The smaller average RMSEs in the proposed cross-classified model was consistent with the expectations that a better fitting model usually has less total estimation error. In addition, Table 23 shows that as the magnitude of the content effects became larger, the average RMSEs for the multilevel model with testlet effects and the Rasch model became increasingly higher. A possible explanation is that larger magnitude of the content effects led the multilevel model with testlet effects and the Rasch model to fit the data even worse by ignoring the content effects. Similar to the results in bias, the testlet effect was not a significant factor on RMSE as the content effect. Further research need to be conducted to explain whether this result was a function of the estimation procedure.

Table 21

Average RMSEs in Ability's SD Estimation across the Other Manipulated Variables

	Testlet effects' SD	Content Effects' SD	Estimating Model			
			Cross-Classified	Multilevel with Testlet Effects	Multilevel with Content Effects	Rasch
Simulation 1	0	0	0.0211	0.0207	0.0210	0.0205
		0.5	0.0267	0.0437	0.0265	0.0428
		1	0.0337	0.1242	0.0338	0.1219
		1.5	0.0521	0.2465	0.0518	0.2355
	0.5	0	0.0231	0.0228	0.0268	0.0262
		0.5	0.0277	0.0433	0.0295	0.0355
		1	0.0417	0.1136	0.0432	0.1015
		1.5	0.0712	0.2527	0.0710	0.2312
	1	0	0.0296	0.0294	0.0538	0.0533
		0.5	0.0341	0.0395	0.0570	0.0448
		1	0.0506	0.1105	0.0679	0.0756
		1.5	0.0702	0.2286	0.0782	0.1812
	1.5	0	0.0455	0.0452	0.0852	0.0803
		0.5	0.0408	0.0381	0.0758	0.0663
		1	0.0560	0.1366	0.0907	0.0788
		1.5	0.0829	0.2231	0.0863	0.1326
Simulation 2	0	0	0.0208	0.0204	0.0206	0.0202
		0.5	0.0284	0.0439	0.0282	0.0430
		1	0.0481	0.1300	0.0479	0.1249
		1.5	0.0897	0.2161	0.0897	0.2112
	0.5	0	0.0255	0.0250	0.0314	0.0300
		0.5	0.0279	0.0532	0.0300	0.0427
		1	0.0341	0.1421	0.0406	0.1253
		1.5	0.0539	0.2164	0.0571	0.1925
	1	0	0.0271	0.0268	0.0552	0.0535
		0.5	0.0366	0.0437	0.0618	0.0518
		1	0.0454	0.1325	0.0655	0.0861
		1.5	0.0719	0.2118	0.0898	0.1555
	1.5	0	0.0327	0.0324	0.0786	0.0768
		0.5	0.0511	0.0516	0.0879	0.0818
		1	0.0523	0.1278	0.0823	0.0789
		1.5	0.0746	0.2013	0.0856	0.1161

SE. The effects that were identified to have both statistical and practical significant impacts on the SE of ability's SD are presented in Table 47 Appendix E. For both simulation studies, four main effects, including sample size, number of testlets, number of items, and model, had a moderate or large impact on the SE.

To better understand the identified significant effects, average SEs under each level of sample size for the four models (Table 22), average SEs under each level of testlet number and testlet size for the four models (Table 23), and average SEs under each level of content effect and testlet effect for the four models (Table 24), are calculated and provided.

For each model, as the sample size increased, a smaller magnitude of SE was found (Table 22). As expected, larger sample size would increase the stability in ability's SD estimation.

For each model, the average SE was largest when the generated number of testlets was 3 and the testlet size was 5 (number of items per test was 15), and the average SE was smallest when the generated number of testlets was 6 and the testlet size was 10 (number of items per test was 60) (see Table 23). The other two conditions, where the number of items per test was 30 for both, had close average SEs. Given this pattern, it appears that, the more items in a test, the smaller the magnitude of the SE in ability's SD estimate.

Among the sixteen aggregated conditions, the proposed cross-classified model had the largest average SEs, and the Rasch model had the smallest average SEs (see Table 24). No consistent patterns were found when comparing between the multilevel model with testlet effects and the multilevel model with content effects. Again, the

higher average SEs in the proposed model might have been due to the increased number of parameters estimated.

Table 22

Average SEs in Ability's SD Estimation by Sample Size across the Other Manipulated Variables

		Estimating Model			
	Sample Size	Cross-Classified	Multilevel with Testlet Effects	Multilevel with Content Effects	Rasch
Simulation 1	500	0.0403	0.0299	0.0336	0.0253
	1000	0.0270	0.0206	0.0228	0.0175
	2000	0.0193	0.0147	0.0162	0.0125
Simulation 2	500	0.0402	0.0298	0.0326	0.0246
	1000	0.0271	0.0205	0.0224	0.0171
	2000	0.0195	0.0150	0.0161	0.0125

Table 23

Average SEs in Ability's SD Estimation by Number of Testlets and Testlet Size across the Other Manipulated Variables

			Estimating Model			
	Number of Testlets	Testlet Size	Cross-Classified	Multilevel with Testlet Effects	Multilevel with Content Effects	Rasch
Simulation 1	3	5	0.0427	0.0321	0.0360	0.0272
		10	0.0276	0.0209	0.0225	0.0175
	6	5	0.0265	0.0200	0.0225	0.0171
		10	0.0187	0.0140	0.0158	0.0120
Simulation 2	3	5	0.0421	0.0320	0.0345	0.0266
		10	0.0280	0.0211	0.0224	0.0173
	6	5	0.0267	0.0200	0.0226	0.0169
		10	0.0189	0.0140	0.0153	0.0116

Table 24

Average SEs in Ability's SD Estimation by Testlet Effect and Content Effect across the Other Manipulated Variables

	Testlet effects' SD	Content Effects' SD	Estimating Model			
			Cross-Classified	Multilevel with Testlet Effects	Multilevel with Content Effects	Rasch
Simulation 1	0	0	0.0203	0.0200	0.0201	0.0199
		0.5	0.0224	0.0194	0.0223	0.0194
		1	0.0269	0.0192	0.0268	0.0191
		1.5	0.0344	0.0195	0.0343	0.0189
	0.5	0	0.0207	0.0206	0.0195	0.0194
		0.5	0.0235	0.0213	0.0223	0.0203
		1	0.0277	0.0200	0.0262	0.0191
		1.5	0.0343	0.0191	0.0328	0.0180
	1	0	0.0224	0.0224	0.0180	0.0180
		0.5	0.0253	0.0219	0.0209	0.0176
		1	0.0291	0.0205	0.0237	0.0170
		1.5	0.0413	0.0218	0.0334	0.0187
	1.5	0	0.0285	0.0283	0.0185	0.0181
		0.5	0.0294	0.0264	0.0187	0.0172
		1	0.0332	0.0243	0.0219	0.0170
		1.5	0.0425	0.0232	0.0278	0.0172
Simulation 2	0	0	0.0195	0.0192	0.0193	0.0190
		0.5	0.0221	0.0200	0.0219	0.0198
		1	0.0285	0.0209	0.0285	0.0201
		1.5	0.0360	0.0188	0.0358	0.0186
	0.5	0	0.0220	0.0218	0.0198	0.0191
		0.5	0.0238	0.0211	0.0214	0.0189
		1	0.0283	0.0200	0.0254	0.0183
		1.5	0.0354	0.0204	0.0321	0.0186
	1	0	0.0238	0.0235	0.0183	0.0176
		0.5	0.0251	0.0224	0.0194	0.0174
		1	0.0312	0.0212	0.0244	0.0176
		1.5	0.0378	0.0205	0.0295	0.0170
	1.5	0	0.0251	0.0249	0.0163	0.0159
		0.5	0.0290	0.0255	0.0189	0.0167
		1	0.0352	0.0254	0.0225	0.0177
		1.5	0.0403	0.0228	0.0257	0.0169

Estimation of Testlet Effects' SD

Since testlet effects were modeled under the proposed cross-classified model and the multilevel model with testlet effects, the discussion of testlet effects' SD would refer only to the two models. A five-way ANOVA was conducted on each of the four error indexes, bias, relative bias, RMSE, and SE by including the five manipulated variables and model as the factors.

Bias. For both simulation studies, only testlet effect had a large impact on the bias in testlet effects' SD estimation, all the other effects were small or negligible (see Table 48 in Appendix E).

Table 25 provides the average biases in the testlet effects' SD estimation for the proposed cross-classified model and the multilevel model with testlet effects under the sixteen aggregated conditions formed by the four levels of testlet effect and the four levels of content effect. Under the conditions with no testlet effect, both models had positive average biases; while, under the conditions with non-zero testlet effects, both models had negative average biases. Table 25 shows that, the cross-classified model had relatively smaller average biases than the multilevel model with testlet effects, which can be explained by the proper modeling of the content effects in the proposed cross-classified model.

Table 25

Average Biases in Testlet Effect's SD Estimation by Testlet Effect and Content Effect across the Other Manipulated Variables

Testlet effects' SD	Content Effects' SD	Simulation 1		Simulation 2	
		Cross-Classified	Multilevel with Testlet Effects	Cross-Classified	Multilevel with Testlet Effects
0	0	0.0591	0.0597	0.0624	0.0634
	0.5	0.0509	0.0701	0.0537	0.0683
	1	0.0428	0.0839	0.0436	0.0971
	1.5	0.0373	0.1466	0.0411	0.0917
0.5	0	-0.0428	-0.0428	-0.0052	-0.0054
	0.5	-0.0549	-0.0565	-0.0092	-0.0193
	1	-0.0689	-0.1094	-0.0325	-0.0680
	1.5	-0.0907	-0.1457	-0.0419	-0.0864
1	0	-0.0347	-0.0351	-0.0428	-0.0431
	0.5	-0.0417	-0.0556	-0.0505	-0.0692
	1	-0.0549	-0.1129	-0.0581	-0.1532
	1.5	-0.0621	-0.2134	-0.0820	-0.2212
1.5	0	-0.0454	-0.0457	-0.0544	-0.0548
	0.5	-0.0488	-0.0792	-0.0680	-0.0972
	1	-0.0616	-0.1513	-0.0724	-0.1766
	1.5	-0.0723	-0.2637	-0.0802	-0.2997

RMSE. The effects that were identified to have both statistical and practical significant impacts on the RMSE of testlet effects' SD are presented in Table 49 Appendix E. For both simulation studies, the number of items per testlet, content effect, and model each had a moderate impact on RMSE. The impact of the interaction between content effect and model was moderate in simulation study 1 and small in simulation study 2.

For both simulation studies, when the number of items per testlet was generated to be five, the average RMSE was larger as compared with that with ten items per testlet for both the proposed cross-classified model and the multilevel model with testlet effects (see Table 26). A possible explanation is that longer testlet provides more information in testlet effects' SD estimation, which then reduces the total estimation error.

Table 27 shows that the proposed cross-classified model had relatively smaller average RMSEs than the multilevel model with testlet effects across simulation conditions. This is consistent with the expectations that the proposed model, which appropriately accounted for the content effects, is expected to have less total estimation error than the multilevel model with testlet effects, which inappropriately ignored the content effects.

In addition, Table 27 demonstrates that as the magnitude of the content effects became larger, the average RMSEs for the multilevel model with testlet effects became larger. Again, this might due to the increasingly worse fit of the multilevel model with testlet effects by ignoring the content effects.

Table 26

Average RMSEs in Testlet Effects' SD Estimation by Testlet Size across the Other Manipulated Variables

Testlet Size	Simulation 1		Simulation 2	
	Cross-Classified	Multilevel with Testlet Effects	Cross-Classified	Multilevel with Testlet Effects
5	0.1279	0.1894	0.1288	0.1861
10	0.0701	0.1191	0.0757	0.1242

Table 27

Average RMSEs in Testlet Effects' SD Estimation by Testlet Effect and Content Effect across the Other Manipulated Variables

Testlet effects' SD	Content Effects' SD	Simulation 1		Simulation 2	
		Cross-Classified	Multilevel with Testlet Effects	Cross-Classified	Multilevel with Testlet Effects
0	0	0.1116	0.1123	0.1144	0.1153
	0.5	0.1018	0.1161	0.1051	0.1137
	1	0.0922	0.1140	0.0934	0.1193
	1.5	0.0857	0.1553	0.0902	0.1021
0.5	0	0.0990	0.0989	0.1015	0.1015
	0.5	0.1097	0.1169	0.0978	0.1148
	1	0.1228	0.1879	0.1061	0.1829
	1.5	0.1450	0.3057	0.1145	0.2302
1	0	0.0750	0.0751	0.0861	0.0860
	0.5	0.0796	0.0883	0.0941	0.1115
	1	0.0906	0.1639	0.1022	0.1810
	1.5	0.0971	0.2633	0.1222	0.2941
1.5	0	0.0842	0.0843	0.0925	0.0926
	0.5	0.0898	0.1068	0.0993	0.1212
	1	0.0969	0.1750	0.1047	0.1985
	1.5	0.1036	0.3038	0.1117	0.3179

Relative Bias. The same significant main effects were identified for both simulation studies, even though with different magnitudes of effect sizes (see Table 50 in Appendix E). For simulations study 1, the magnitude for the four main effects, testlet size, testlet effect, content effect, and model, were small, moderate, moderate, and small, respectively; however, for simulation study 2, the magnitude were moderate, small, moderate, and small, respectively.

Table 28

Average Relative Biases in Testlet Effects' SD Estimation by Testlet Size across the Other Manipulated Variables

Testlet Size	Simulation 1		Simulation 2	
	Cross-Classified	Multilevel with Testlet Effects	Cross-Classified	Multilevel with Testlet Effects
5	-0.1110	-0.1542	-0.1088	-0.1912
10	-0.0324	-0.0935	-0.0331	-0.0970

For both simulation studies, when the number of items per testlet was generated to be five, the average relative bias was larger when compared to that with ten items per testlet for both the proposed cross-classified model and the multilevel model with testlet effects (Table 28). This might be explained by the larger information provided by longer testlet.

Table 29 shows that the proposed cross-classified model had relatively smaller average relative biases than the multilevel model with testlet effects across simulation conditions. This is consistent with the expectations that the proposed model has less systematic estimation error than the multilevel model with testlet effects, which ignored the content effects. In addition, Table 29 demonstrates that as the magnitude of the content effects became larger, the average relative biases for the multilevel

model with testlet effects became larger. This might due to the increasingly worse fit of the multilevel model with testlet effects by ignoring the content effects.

Table 29

Average Relative Biases in Testlet Effects' SD Estimation by Testlet Effect and Content Effect across the Other Manipulated Variables

Testlet effects' SD	Content Effects' SD	Simulation 1		Simulation 2	
		Cross-Classified	Multilevel with Testlet Effects	Cross-Classified	Multilevel with Testlet Effects
0	0	NA	NA	NA	NA
	0.5	NA	NA	NA	NA
	1	NA	NA	NA	NA
	1.5	NA	NA	NA	NA
0.5	0	-0.0856	-0.0857	-0.0708	-0.0713
	0.5	-0.1099	-0.1131	-0.0708	-0.1092
	1	-0.1379	-0.2189	-0.1167	-0.2698
	1.5	-0.1815	-0.2914	-0.1400	-0.3181
1	0	-0.0347	-0.0351	-0.0528	-0.0530
	0.5	-0.0417	-0.0556	-0.0646	-0.0850
	1	-0.0549	-0.1129	-0.0754	-0.1847
	1.5	-0.0621	-0.2134	-0.1109	-0.2771
1.5	0	-0.0303	-0.0305	-0.0353	-0.0356
	0.5	-0.0325	-0.0528	-0.0460	-0.0645
	1	-0.0411	-0.1009	-0.0502	-0.1208
	1.5	-0.0482	-0.1758	-0.0562	-0.2041

SE. Table 51 in Appendix E provides the identified effects that were both statistical and practical significant. For both simulation studies, sample size and testlet size were found to have large impacts on the SE of testlet effects' SD. Testlet effect had a moderate effect size in simulation study 1, but a small effect size in simulation study 2.

For both the proposed cross-classified model and the multilevel model with testlet effects, as sample size increased, the magnitude of SE decreased (see Table 30). In addition, as the number of items per testlet increased, the magnitude of SE also decreased (see Table 31). Generally, the proposed cross-classified model had relatively larger average SEs than the multilevel model with testlet effects (see Table 30, 31, 32), which can be explained by the increased number of parameter estimated for the proposed model.

Table 30

Average SEs in Testlet Effects' SD Estimation by Sample Size across the Other Manipulated Variables

Sample Size	Simulation 1		Simulation 2	
	Cross-Classified	Multilevel with Testlet Effects	Cross-Classified	Multilevel with Testlet Effects
500	0.1049	0.0959	0.1054	0.0919
1000	0.0767	0.0681	0.0767	0.0642
2000	0.0526	0.0457	0.0539	0.0456

Table 31

Average SEs in Testlet Effects' SD Estimation by Testlet Size across the Other Manipulated Variables

Testlet Size	Simulation 1		Simulation 2	
	Cross-Classified	Multilevel with Testlet Effects	Cross-Classified	Multilevel with Testlet Effects
5	0.0967	0.0867	0.0949	0.0810
10	0.0594	0.0530	0.0624	0.0535

Table 32

Average SEs in Testlet Effects' SD Estimation by Testlet Effect and Content Effect across the Other Manipulated Variables

Testlet effects' SD	Content Effects' SD	Simulation 1		Simulation 2	
		Cross-Classified	Multilevel with Testlet Effects	Cross-Classified	Multilevel with Testlet Effects
0	0	0.0943	0.0947	0.0954	0.0958
	0.5	0.0877	0.0889	0.0898	0.0854
	1	0.0812	0.0608	0.0821	0.0484
	1.5	0.0765	0.0308	0.0797	0.0265
0.5	0	0.0873	0.0872	0.0843	0.0843
	0.5	0.0929	0.0900	0.0811	0.0801
	1	0.0991	0.0908	0.0832	0.0676
	1.5	0.1097	0.0814	0.0853	0.0548
1	0	0.0619	0.0619	0.0709	0.0707
	0.5	0.0626	0.0611	0.0740	0.0728
	1	0.0672	0.0620	0.0793	0.0752
	1.5	0.0692	0.0639	0.0856	0.0616
1.5	0	0.0624	0.0624	0.0659	0.0659
	0.5	0.0647	0.0633	0.0654	0.0631
	1	0.0657	0.0612	0.0676	0.0631
	1.5	0.0664	0.0578	0.0688	0.0604

Estimation of Content Effects' SD

The analysis of the content effects' SD recovery is discussed in this section. Models mentioned in this section refer to the proposed cross-classified model and the multilevel model with content effects, which are the only two models that have random content effects. Again, five-way ANOVA was conducted first, followed by some descriptive statistics.

Bias. Table 52 in Appendix E contains the effects that were identified to have both statistical and practical significant impacts on bias of content effects' SD recovery. For both simulation studies, content effect was significant with a large effect size; testlet effect, model, and interactions among testlet effect, content effect, and model were significant with moderate effect sizes.

The proposed cross-classified model in general had relatively smaller average biases compared with the multilevel model with content effects (see Table 33). This is consistent with the expectations that the appropriate modeling of the testlet effects in the proposed model is expected to produce smaller systematic estimation error than the multilevel model with content effects, which inappropriately ignored the testlet effects.

In addition, as the magnitude of the testlet effects became larger, the multilevel model with content effects had increasingly higher average biases in content effects' SD estimation (see Table 33). This might be due to the increasingly worse fit of the multilevel model with content effects when fitting data with increasingly larger testlet effects.

Table 33

Average Biases in Content Effects' SD Estimation by Testlet Effect and Content Effect across the Other Manipulated Variables

Testlet effects' SD	Content Effects' SD	Simulation 1		Simulation 2	
		Cross-Classified	Multilevel with Content Effects	Cross-Classified	Multilevel with Content Effects
0	0	0.0374	0.0386	0.0402	0.0415
	0.5	-0.0142	-0.0142	-0.0316	-0.0317
	1	-0.0090	-0.0102	-0.0227	-0.0228
	1.5	-0.0160	-0.0177	-0.0145	-0.0165
0.5	0	0.0353	0.0422	0.0342	0.0402
	0.5	-0.0348	-0.0446	-0.0497	-0.0705
	1	-0.0252	-0.0509	-0.0313	-0.0826
	1.5	-0.0321	-0.0685	-0.0375	-0.1005
1	0	0.0254	0.0351	0.0275	0.0721
	0.5	-0.0489	-0.0562	-0.0538	-0.0671
	1	-0.0440	-0.1351	-0.0321	-0.1681
	1.5	-0.0335	-0.1881	-0.0302	-0.1818
1.5	0	0.0176	0.1009	0.0184	0.0704
	0.5	-0.0739	-0.1855	-0.0633	-0.1489
	1	-0.0464	-0.2260	-0.0604	-0.2618
	1.5	-0.0430	-0.3293	-0.0377	-0.3379

Relative Bias. Table 53 in Appendix E provides the effects that were identified to have both statistical and practical significant impacts on relative bias of content effects' SD recovery. For simulation study 1, testlet effect was found to have a large impact on the relative bias, model had a moderate impact, and the interaction between testlet and model was also moderate; all the other effects were small or negligible. However, no moderate or large effect was identified in simulation study 2.

Similar to the results in bias, the proposed cross-classified model in general had relatively smaller average relative biases compared with the multilevel model with content effects (see Table 34). As the magnitude of the testlet effects became larger, the multilevel model with content effects had increasingly higher average relative biases in content effects' SD estimation (see Table 34). This can be explained by the appropriate modeling and inappropriate modeling of the testlet effects in the proposed model and the multilevel model with content effects, respectively.

Table 34

Average Relative Biases in Content Effects' SD Estimation by Testlet Effect and Content Effect across the Other Manipulated Variables

Testlet effects' SD	SD of Content Effects	Simulation 1		Simulation 2	
		Cross-Classified	Multilevel with Content Effects	Cross-Classified	Multilevel with Content Effects
0	0	NA	NA	NA	NA
	0.5	-0.0284	-0.0284	**-0.0938**	**-0.0936**
	1	-0.0090	-0.0102	-0.0271	-0.0269
	1.5	-0.0107	-0.0118	-0.0080	-0.0094
0.5	0	NA	NA	NA	NA
	0.5	**-0.0697**	**-0.0891**	**-0.1628**	**-0.1995**
	1	-0.0252	**-0.0509**	-0.0354	**-0.0893**
	1.5	-0.0214	-0.0457	-0.0252	**-0.0666**
1	0	NA	NA	NA	NA
	0.5	**-0.0978**	**-0.1123**	**-0.1452**	**-0.1112**
	1	-0.0440	**-0.1351**	-0.0398	**-0.1805**
	1.5	-0.0224	**-0.1254**	-0.0188	**-0.1193**
1.5	0	NA	NA	NA	NA
	0.5	**-0.1478**	**-0.3711**	**-0.1615**	**-0.2506**
	1	-0.0464	**-0.2260**	**-0.0682**	**-0.2688**
	1.5	-0.0287	**-0.2196**	-0.0244	**-0.2254**

RMSE. Effects that were both statistically and practically significant on RMSE are presented in Table 54 Appendix E. Testlet effect had a large and a moderate impact in simulation study 1 and simulation study 2, respectively. Model and the interaction between model and testlet effect were both of moderate effects.

Generally, the multilevel model with content effects had relatively larger average RMSEs compared with the proposed cross-classified model across the sixteen aggregated conditions (see Table 35). In addition, as the magnitude of the testlet effects became larger, the average RMSEs associated with the multilevel model with content effects became larger. This is consistent with the expectations that the proposed model is expected to have less total estimation error than the multilevel model with content effects. It is also expected that the multilevel model with content effects becomes increasingly worse fit as the magnitude of testlet effects becomes larger, which results in larger RMSEs.

Table 35

Average RMSES in Content Effects' SD Estimation by Testlet Effect and Content Effect across the Other Manipulated Variables

Testlet effects' SD	Content Effects' SD	Simulation 1		Simulation 2	
		Cross-Classified	Multilevel with Content Effects	Cross-Classified	Multilevel with Content Effects
0	0	0.0780	0.0793	0.0837	0.0849
	0.5	0.0741	0.0737	0.1000	0.0997
	1	0.0535	0.0534	0.0625	0.0615
	1.5	0.0610	0.0611	0.0716	0.0714
0.5	0	0.0744	0.0816	0.0757	0.0758
	0.5	0.0806	0.0825	0.1047	0.1177
	1	0.0582	0.0702	0.0674	0.0990
	1.5	0.0684	0.0885	0.0697	0.1125
1	0	0.0640	0.0669	0.0663	0.0934
	0.5	0.1014	0.1183	0.1113	0.1369
	1	0.0804	0.1483	0.0713	0.1777
	1.5	0.0707	0.1947	0.0716	0.1884
1.5	0	0.0521	0.1237	0.0532	0.0833
	0.5	0.1254	0.2056	0.1267	0.2200
	1	0.0882	0.2328	0.0969	0.2711
	1.5	0.0786	0.3328	0.0753	0.3409

SE. For both simulation studies, sample size, number of testlets, number of items per testlet, and content effect, were all found to have a moderate or a large impact on the SE in content effects' SD estimation (see Table 55 in Appendix E). Model did not have a significant impact on SE in simulation study 1, but a small impact in simulation study 2.

For both the cross-classified model and the multilevel model with content effects, as sample size increased, the magnitude of SE became smaller (Table 36). This indicates that increasing the sample size might improve the stability in content effects' SD estimation.

For both models, the average SE was largest when the generated number of testlets was 3 and the testlet size was 5 (number of items per test was 15), and the average SE was smallest when the generated number of testlets was 6 and the testlet size was 10 (number of items per test was 60) (see Table 37). The other two conditions, where the number of items per test was 30 for both, had very close average SEs. Given this pattern, it appears that, the more items in a test, the smaller the magnitude of the SE in content effects' SD estimation. A possible explanation is that longer test improves the stability in content effects' SD estimation.

For both simulation studies, the proposed cross-classified model had relatively larger average SEs compared with the multilevel model with content effects across simulation conditions (see Table 38). This might have been due to the increased number of parameters estimated for the proposed model.

Table 36

Average SEs in Content Effects' SD Estimation by Sample Size across the Other Manipulated Variables

Sample Size	Simulation 1		Simulation 2	
	Cross-Classified	Multilevel with Content Effects	Cross-Classified	Multilevel with Content Effects
500	0.0790	0.0723	0.0855	0.0738
1000	0.0569	0.0525	0.0602	0.0532
2000	0.0390	0.0377	0.0428	0.0375

Table 37

Average SEs in Content Effects' SD Estimation by Number of Testlets and Testlet Size across the Other Manipulated Variables

Number of Testlets	Testlet Size	Simulation 1		Simulation 2	
		Cross-Classified	Multilevel with Content Effects	Cross-Classified	Multilevel with Content Effects
3	5	0.0874	0.0810	0.0916	0.0795
	10	0.0548	0.0485	0.0611	0.0496
6	5	0.0548	0.0535	0.0584	0.0555
	10	0.0362	0.0335	0.0402	0.0347

Table 38

Average SEs in Content Effects' SD Estimation by Testlet Effect and Content Effect across the Other Manipulated Variables

Testlet effects' SD	Content Effects' SD	Simulation 1		Simulation 2	
		Cross-Classified	Multilevel with Content Effects	Cross-Classified	Multilevel with Content Effects
0	0	0.0682	0.0691	0.0731	0.0737
	0.5	0.0653	0.0650	0.0777	0.0775
	1	0.0447	0.0446	0.0520	0.0511
	1.5	0.0467	0.0466	0.0480	0.0478
0.5	0	0.0650	0.0692	0.0670	0.0623
	0.5	0.0670	0.0645	0.0848	0.0789
	1	0.0453	0.0435	0.0515	0.0500
	1.5	0.0462	0.0449	0.0469	0.0444
1	0	0.0584	0.0537	0.0600	0.0514
	0.5	0.0761	0.0639	0.0829	0.0610
	1	0.0520	0.0449	0.0551	0.0517
	1.5	0.0511	0.0447	0.0480	0.0404
1.5	0	0.0488	0.0551	0.0495	0.0361
	0.5	0.0867	0.0691	0.0957	0.0703
	1	0.0606	0.0486	0.0626	0.0416
	1.5	0.0507	0.0388	0.0506	0.0391

Model Fit Indices

Five indices, AIC, AICC, BIC, CAIC, and HQIC, produced by SAS PROC GLIMMIX were used to assess model fit. Appendix F shows the percentages of replications in which the correct model was identified by using each of the five indices for each condition.

Generally, the five indices performed equally well in correctly identifying the proposed cross-classified model as the best fitting model when both the magnitude of the testlet effects and the magnitude of the content clustering effects were large (SD = 1 or 1.5) (see Appendix F). A possible explanation is that, the proposed cross-classified model performed much better than the three under-parameterized models when the magnitude of the two random effects were large, which makes the five indices easy to identify the proposed model as the best fitting model.

The five fit indices, especially BIC and CAIC, did not perform well in correctly identifying the proposed cross-classified model as the best fitting model when equal testlet effects' SDs and/or equal content effects' SDs were generated with small magnitudes (SD = 0.5) (see Appendix F). Table 39 contains the conditions that have small percentages of replications in which the proposed cross-classified model was correctly identified as the best fitting model using each of the five indices. Under those conditions, the percentages of replications were small for both BIC and CAIC, while AIC, AICC, and HQIC performed better than BIC and CAIC by having relatively larger percentages. An identified common characteristic of those conditions with small percentages was that either the magnitude of the testlet effects was small (SD = 0.5) or the magnitude of the content clustering effects was small (SD = 0.5). A

possible explanation is that, when the magnitude of the testlet effects and/or the content clustering effects was small, the proposed model did not perform significantly better than the under-parameterized models, which made the fit indices hard to identify the proposed model as the best fitting model.

However, if the testlet effects' SDs and the content effects' SDs were generated to be unequal across testlets and content areas, even though the average SDs were small (average SD = 0.5), the five fit indices still performed well in identifying the proposed model as the best fitting model under most conditions (see Table 39). A possible explanation is that, when the testlet effects' SDs and the content effects' SDs were generated to be unequal, even though the average SD was small, one or more testlet effects' SD and content clustering effects' SD was large, which made the proposed model fit the generated data much better than the three under-parameterized models.

Table 39

Conditions with Low Percentages of replications in which the proposed cross-classified model was correctly identified as the best fitting model using each of the five indices

Condition		Simulation Study 1					Simulation Study 2				
		AIC%	AICC%	BIC%	CAIC%	HQIC%	AIC%	AICC%	BIC%	CAIC%	HQIC%
6	1-1-1-2-2	72	72	2	0	34	100	100	86	72	100
7	1-1-1-2-3	64	64	10	2	28	100	100	84	80	92
8	1-1-1-2-4	38	38	8	4	24	100	100	96	90	100
10	1-1-1-3-2	78	78	26	22	60	92	92	66	54	84
14	1-1-1-4-2	74	74	28	26	48	96	96	52	38	76
24	1-1-2-2-4	98	98	84	80	98	100	100	100	100	100
38	1-2-1-2-2	98	98	22	8	66	100	100	100	100	100
39	1-2-1-2-3	94	94	8	0	62	100	100	100	100	100
40	1-2-1-2-4	80	80	0	0	32	100	100	100	100	100
70	2-1-1-2-2	96	96	26	20	74	100	100	94	86	98
71	2-1-1-2-3	92	92	20	12	64	100	100	100	100	100
72	2-1-1-2-4	80	80	12	6	46	100	100	100	98	100
74	2-1-1-3-2	98	98	70	56	92	100	100	92	88	98
78	2-1-1-4-2	96	96	42	32	78	100	100	88	78	98
102	2-2-1-2-2	100	100	78	50	100	100	100	100	100	100
103	2-2-1-2-3	98	98	46	32	88	100	100	100	100	100
104	2-2-1-2-4	100	100	6	4	68	100	100	100	100	100
134	3-1-1-2-2	100	100	82	66	100	100	100	100	100	100
135	3-1-1-2-3	98	98	48	32	92	100	100	100	100	100
136	3-1-1-2-4	96	96	18	14	76	100	100	100	100	100
142	3-1-1-4-2	96	96	58	48	90	96	96	68	60	86
167	3-2-1-2-3	100	100	94	86	100	100	100	100	100	100
168	3-2-1-2-4	100	100	64	50	96	100	100	100	100	100

Note: Condition a-b-c-d-e, where a = sample size, b = number of testlets, c = number of items per testlet, d = magnitude of the testlet effect, e = magnitude of the content effects. Refer Table 3 to get the corresponding levels for each manipulated variable.

Chapter 5: Summary and Discussions

This chapter includes three sections. The first section summarizes the results for both simulation studies. The second section presents the contributions. The third section discusses limitations of this study and recommendations for future research.

Summary of Results

The present study proposed a cross-classified model to account for local item dependence (LID) that is caused by two factors simultaneously, which is named as dual local item dependence (DLID) in this study. It demonstrated that the proposed cross-classified model accounting for DLID is algebraically equivalent with a constrained version of the testlet model accounting for two types of LID (Jiao et al., 2009).

Two simulation studies were designed and conducted with the primary purpose of evaluating the performance of the proposed cross-classified model. Data sets with DLID were simulated with both testlet effects and content clustering effects. The second purpose of this study was to investigate the potential factors affecting the need to use the more complex cross-classified model over the simplified multilevel modeling of LID by ignoring cross-classification structure.

For both simulation studies, five factors were manipulated, including sample size (500, 1000, and 2000), number of testlets (3 and 6), number of items per testlet (5 and 10), magnitude of testlet effects represented by standard deviation (SD) (0, 0.5, 1, and 1.5), and magnitude of content clustering effects represented by SD (0, 0.5, 1, and

1.5). The difference between the two simulation studies was that, simulation study 1 constrained the testlet effects' SDs as well as the content clustering effects' SDs as the same across the testlets and content areas, respectively; simulation study 2 released this constraint by having mixed testlet effects' SDs and mixed content clustering effects' SDs.

Bias, relative bias, root mean square error (RMSE), and standard error (SE) for parameter estimates were investigated by conducting analysis of variance (ANOVA) and providing descriptive statistics. The 0.05 alpha-level was used to determine statistical significance first, and then a minimum cutoff for practical significance of $\eta^2 = 0.01$ was used, which resulted in the detection of effect with at least small practical significance.

Estimation of Item Difficulty

Sample size significantly impacted the RMSE and the SE in item difficulty estimation. As sample size increased, both RMSE and SE became smaller. This indicates that, larger samples yielded more accurate parameter estimations by having less total estimation errors and more stable estimations across replications.

Both of the two manipulated variables, number of testlets and number of items per testlet, had no significant impact on any of the four error indexes, bias, relative bias, RMSE, and SE, in item difficulty estimation.

For both simulation studies, magnitude of the testlet effects, magnitude of the content clustering effects, and model, were all found to have significant impacts on relative bias, RMSE, and SE. In addition, the interaction between model and testlet

effect and the interaction between model and content effect also had significant impacts on relative bias, RMSE, and SE.

Generally, the proposed cross-classified model had smaller bias, smaller relative bias, smaller RMSE, and larger SE than the other three estimating models across simulated conditions. This result is consistent with what was found by Jiao et al. (2012). Lower bias, relative bias, and RMSE were found in the proposed model, which appropriately accounted for both the testlet effects and the content clustering effects. However, the proposed cross-classified model had slightly higher SE due to the increased number of parameters estimated.

Estimation of Persons' Ability

Persons' ability was evaluated and compared in terms of three error indexes, bias, RMSE, and SE. Relative bias was not appropriately used in analyzing ability, because when ability was randomly generated from $N(0, 1)$, plenty of values would be very close to 0, which made the relative bias became extremely large.

Sample size had no significant impact on any of the three error indexes. This is consistent with the expectations that the precision of person parameters should not be affected by the number of persons, which was also found by previous research (e.g. Kamata, 2001).

Number of testlets had a significant impact on RMSE: the larger the number of testlets, the smaller the RMSE in persons' ability estimation. Both number of testlets and number of items per testlet significantly impacted the SE in the ability estimation: the more items in a test, the smaller the SE in persons' ability estimation.

The proposed cross-classified model had relatively smaller RMSE and SE than the other three models. The lower RMSE in the proposed model can be explained by the proper modeling of persons' ability, testlet effects, and content clustering effects separately. However, lower SE in the proposed model is inconsistent with the expectations: the proposed model, which has more parameters to estimate, is expected to have relatively larger SE than the other three simpler models. Further research should investigate whether this result generalizes across other conditions.

Estimation of Ability's SD

Sample size had a large impact on the SE in ability's SD estimation. Larger sample size resulted in smaller SE in ability's SD estimation. Both factors, the number of testlets and the number of items per testlet, had large impacts on the SE of ability's SD: the more items in a test, the smaller the SE in ability's SD recovery.

The model factor significantly impacted bias, RMSE, and SE. Generally, the proposed cross-classified model had smaller bias and RMSE in ability's SD estimation than the other three models. Therefore, the proposed model was more effective in recovering ability's SD. However, the proposed model had larger SE than the other three estimating models. Again, this can be explained by the increased difficulty in separating the ability, testlet effects, and content clustering effects when estimating the proposed model.

As the magnitude of the content clustering effects became larger, the bias and RMSE for the multilevel model with testlet effects and the Rasch model became increasingly higher. A possible explanation is that larger magnitude of the content

clustering effects led the multilevel model with testlet effects and the Rasch model to perform even worse by inappropriately ignoring the content clustering effects. However, the magnitude of the testlet effects was not a significant factor. It is expected that as the magnitude of the testlet effects becomes larger, the bias and RMSE for the multilevel model with content effects and the Rasch model, both of which do not account for the testlet effects, become increasingly higher. Further research should explain whether this result was a function of the estimation procedure.

Estimation of Testlet Effects' SD and Content Effects' SD

The recovery of testlet effects' SD was evaluated and compared between the proposed cross-classified model and the multilevel model with testlet effects. Similarly, the recovery of content effects' SD was evaluated and compared between the proposed cross-classified model and the multilevel model with content effects.

Sample size was found to have a large impact on both the SE of testlet effects' SD and the SE of content effects' SD. As sample size increased, the SEs of the two random effects' SDs became smaller. Testlet length had significant impacts on all of the four error indexes. As testlet length increased, the magnitude of each of the four error indexes became smaller. For the multilevel model with testlet effects, the magnitude of the content clustering effects had a large impact on the recovery of testlet effects' SD: as the magnitude of the content clustering effects increased, the bias, RMSE and relative bias became larger for the multilevel model with testlet effects. Similarly, for the multilevel model with content effects, the magnitude of the testlet effects had a large impact on the recovery of content effects' SD; as the

magnitude of the testlet effects increased, the bias, RMSE and relative bias became larger for the multilevel model with content effects.

Generally, the proposed cross-classified model had smaller bias, relative bias, and RMSE than the two multilevel models when recovering the two random effects' SD. However, the proposed model had larger SE than the two multilevel models. This can be explained by the increased difficulty in separating the general ability, testlet effect, and content effect when estimating the proposed model.

Model Fit Indices

In this study, five indices, AIC, AICC, BIC, CAIC, and HQIC, produced by SAS PROC GLIMMIX were used to assess model fit. Percentages of replications in which the correct model was identified by using each of the five indices for each condition were tallied.

Generally, the five indices performed equally well in identifying the proposed cross-classified model as the best fitting model when the magnitude of the testlet effects and the content clustering effects was large (SD = 1 or 1.5). The five fit indices, especially BIC and CAIC, did not perform well in identifying the proposed model as the best fitting model when equal testlet effects' SDs and equal content effects' SDs were generated with small magnitude (SD = 0.5). However, if the testlet effects' SDs and the content effects' SDs were generated to be unequal across testlets and content areas, even though the average SDs were small (average SD = 0.5), the five fit indices still performed well.

Since BIC and CAIC penalize for additional model parameters more severely than the other three, it is expected that both BIC and CAIC performed not well in

identifying the proposed model as the best fitting model when the magnitude of the testlet effect and/or the content effect was small. Therefore, generally, the performance of the five information criteria was consistent with the expectations.

Conclusion

In summary, when the data sets were generated with large magnitude of testlet effects and content clustering effects, the proposed cross-classified model yielded more accurate parameter estimation, including item difficulty, persons' ability, and random effects' SD, with smaller bias, relative bias, and RMSE than the two multilevel models and the Rasch model. When the data sets were generated with small magnitude of testlet effects and/or small magnitude of content clustering effects, the proposed cross-classified model still produced smaller bias, relative bias, and RMSE than the other three under-parameterized models, even though the differences were not substantial. When the data sets were generated with no testlet effect and/or no content clustering effect, even though the cross-classified model was an over-parameterized model, the bias, relative bias, and RMSE were about the same for the cross-classified model and the true model. The lower bias, relative bias, and RMSE in the cross-classified model was consistent with the expectations that the proposed model should have smaller systematic and total estimation errors in parameter estimation by appropriately accounting for both the testlet effects and the content clustering effects. However, it should be noted that even though, in this study, the proposed cross-classified model worked well even when it was an over-parameterized model, practitioners should ensure the sample size is large enough when putting this model into real application.

Generally, the cross-classified model had slightly higher SE in parameter estimation, including item difficulty estimation and random effects' SD estimation, compared with the other three models. A possible explanation is that the proposed cross-classified model has more parameters to estimate. One exception is persons' ability recovery, for which, the cross-classified model had slightly smaller SE than the other three models. Future research should investigate the reason for this result.

Among the five manipulated factors, sample size was a significant factor in item difficulty estimation and random effects' SD estimation. Generally, in this study, increasing sample size resulted in more accurate item difficulty estimates, and more stable item difficulty estimates and random effects' SD estimates. Larger number of testlets gave more accurate estimation for persons' ability. In addition, testlet length played a role in testlet effects' SD recovery; it appears that, increasing the testlet length would reduce the error in testlet effects' SD estimation. Moreover, test length impacted persons' ability estimation and random effects' SD estimation; longer tests produced more stable estimates for persons' ability and random effects' SD. The magnitude of the two variables, testlet effect and content effect, determined the necessity of using the more complex cross-classified model over the simplified multilevel models and the Rasch model. In summary, the larger the magnitude of the testlet effects and the content clustering effects, the better the proposed cross-classified model than the other three models in parameter estimation.

Contributions

Local item independence is one of the important assumptions underlying the IRT models. Violation of this assumption might be caused by various factors, like the testlet effect and the content effect described in this study. Ignoring the violation of this assumption might have negative impacts, e.g. inaccurate estimation of both item and person parameters, over-estimation of test reliability, and equating errors. Previous studies have mainly focused on investigating one source of local item dependence (LID). However, in some cases, such as scenario-based science assessments, LID might be caused by two possible sources simultaneously, one is testlet effect and the other is content clustering effect. Such kind of LID that is caused by two factors simultaneously is named as dual local item dependence (DLID).

Researchers have used multilevel parameterization of IRT models to incorporate the clustering of items, such as testlets (Jiao et al., 2005). However, such multilevel models with testlet effects fail to model item response data structures that have DLID. The primary contribution of this study is that a cross-classified model is proposed to deal with the issue of DLID by accounting for two types of LID simultaneously.

When the item response data structures were generated to have two sources of LID, the simulation studies demonstrated that the proposed cross-classified model produced more accurate estimation of both item and person parameters than the multilevel models and the Rasch model. In other words, the proposed cross-classified model is more appropriate to be used when the true nature of the data structure has DLID. Given that many assessments have the issue of DLID, like the scenario-based

science assessment, this cross-classified modeling approach will improve the accuracy of both item and person analysis.

Limitations and Directions for Future Research

This study proposed a new model to account for local item dependence caused by two factors simultaneously from a cross-classified multilevel modeling framework. It was a preliminary investigation designed to evaluate the performance of the proposed cross-classified model and explore the potential factors affecting use of the more complex cross-classified model over other simpler multilevel models. Given that it is a preliminary study, this study has several limitations.

In terms of the model assumptions, one of the primary limitations is that both the testlet effects and the content clustering effects were assumed to be random in this study. Future research might consider relaxing this restriction. Instead of modeling both testlet effect and content effect as person-specific, some of the testlet effects and/or content effects could be modeled as fixed.

In addition, the proposed cross-classified model was extended from the Rasch model, which is the simplest model in the IRT field. Theoretically, this proposed model can be extended to two parameter and three parameter IRT models, as well as polytomous IRT models; however, studies should be conducted to evaluate the performance of those more complicated models.

The third limitation of the proposed cross-classified model is that it assumes independence among persons' ability, person-specific testlet effect, and person-specific content effect. Future research need to be conducted to explore the impact of the violation of this assumption.

With regards to the process of data generation, one of the primary limitations of the two simulations studies is that each simulated data set was generated with two

content areas. However, in real world testing scenarios, more than two content areas could be covered, such as the science assessment. Future research should be conducted to manipulate the number of content areas instead of fixing it as two.

In addition, in this study, when data sets were generated, each of the two content areas was assumed to be assessed by about 50% of the items. However, in real world tests, one content area could be assessed by more items than others, which forms a complex cross-classified matrix between the testlet factor and the content factor. Future research should consider the complexity imbedded in the cross-classifications of the two factors.

Moreover, equal numbers of items (either 5 or 10) per testlet were generated. Future research should consider investigating a more realistic design with different numbers of items for different testlets. Furthermore, in this study, all items were generated to be part of a testlet, which might be also unrealistics in real world testing scenarios. Therefore, tests that contain both single items and testlets should be investigated in the future.

As to the model estimation approach used in this study, the Laplace method in the PROC GLIMMIX procedure was used in model estimation. However, it was found that estimation methods could have significant impacts on parameter estimations when estimating models like the multilevel model with testlet effects (Jiao et al., 2013). Therefore, future research should consider other estimation procedures and investigate whether consistent results could be produced by using a different estimation method.

Some limitations also existed in the result. First, when analyzing the recovery of random effects' SD, since many high-way interaction effects were identified to have both statistical and practical significance, to simplify the complexity of interpreting the ANOVA results, only effects associated with a moderate or a large effect size were given further interpretations. However, in this way, some important findings might be missed. Even though a factor might be found to have a small effect size, some interesting patterns could be observed when examining the averages across the levels of this factor.

Generally, the results showed that the proposed cross-classified model had slightly larger SEs compared with the other three models when estimating item difficulty and three random effects' SDs. A possible explanation is that the cross-classified model has more parameters to estimate. However, it was surprising to see that the cross-classified model yielded smaller SEs than did the other three models when estimating the persons' ability. It is unclear why this happened, and future research should explore whether the results generalize to additional design conditions.

Another unexpected result occurred in the analysis of ability's SD estimation. The magnitude of the content clustering effects was found to have large impacts on both the bias and RMSE in ability's SD estimation. However, the magnitude of the testlet effects was not a significant factor. Further research need to be conducted to explain whether this result was a function of the estimation procedure.

Finally, one major limitation of this study relates to the model fit indices. The five fit indices performed well in identifying the proposed model as the best fitting model when the magnitude of both the testlet effects and the content clustering effects

was large. However, the five fit indices, especially BIC and CAIC, performed not well when equal testlet effects' SDs and/or equal content effects' SDs were generated with small magnitude. Future research should explore other fit indices that have the potential to identify the proposed model as the best fitting model under such conditions that have small magnitude of random effects.

Appendix A: Example of Science Assessment

Use the information below to answer questions 17 through 20.

Zebra mussels arrived in Lake St. Clair, near Detroit, by accident. Mussels are in the same family as oysters, and they form hard, protective outer shells. Scientists believe zebra mussels were transported by large ships from Europe and spread rapidly throughout the Great Lakes. They consume large quantities of tiny parts and animals and have a high reproductive rate.

17. Within the Great Lakes ecosystem, scientists refer to the zebra mussel as a

A) parasite.
B) producer.
C) non-native species.
D) single-celled organism.

This is a question assessing related to **population dynamics**.

18. Zebra mussels reproduce and spread quickly, reducing food resources and crowding native species. This result in

A) an increase in native producer populations
B) a decrease in native consumer populations
C) higher reproductive rates for native species
D) mutually beneficial relationships with native species

This is a question assessing related to **population dynamics**.

19. The zebra mussel would best be classified as a(n)

A) mammal.
B) producer.
C) amphibian.
D) invertebrate.

This is a question assessing related to **classify organisms**.

Appendix B: Example of Reading Assessment in TOEFL

Read the following passage. Then answer the questions and check your answers.

Most people can remember a phone number for up to thirty seconds. When this short amount of time **elapses**, however, the numbers are erased from the memory. How did the information get there in the first place? Information that makes its way to the short term memory (STM) does so via the sensory storage area. The brain has a filter which only allows stimuli that is of immediate interest to pass on to the STM, also known as the working memory……

Reading Comprehension questions:

1. According to the passage, how do memories get transferred to the STM?

A) They revert from the long term memory.
B) They are filtered from the sensory storage area.
C) They get chunked when they enter the brain.
D) They enter via the nervous system.

This is a **factual** question.

2. The word **elapses** in paragraph 1 is closest in meaning to:

A) passes
B) adds up
C) appears
D) continues

This is a **vocabulary** question.

3. All of the following are mentioned as places in which memories are stored EXCEPT the:

A) STM
B) long term memory
C) sensory storage area
D) maintenance area

This is a **negative factual** question.

4. How do theorists believe a person can remember more information in a short time?

A) By organizing it
B) By repeating it
C) By giving it a name
D) By drawing it

This is a **factual** question.

5. The author believes that rote rotation is:

A) the best way to remember something
B) more efficient than chunking
C) ineffective in the long run
D) an unnecessary interruption

This is a **factual** question.

6. The word **it** in the last paragraph refers to:

A) encoding
B) STM
C) semantics
D) information

This is a **reference** question.

Appendix C: R Code for Data Generation

```
# This code is generating cross-classified data with mixed variance patterns
# The number of content areas is fixed at 2
C=2

N=c(500,1000,2000)       # number of examinees: n
T=c(3,6)                 # number of testlets: t
I=c(5,10)                # number of items per testlet: i
b1=c(-2,-1,0,1,2)        # the item difficulty value when the number of items per testlet is 5
b2=seq(-2,2.5,length=10) # the item difficulty value when the number of items per testlet is
10

SD_T3=matrix(c(
0,0,0,
0,0.5,1,
0.5,1,1.5,
1,1.5,2
),4,3,byrow=T)     # pattern of variance for testlets: sd_t3,the number of testlets is 3

SD_T6=matrix(c(
0,0,0,0,0,0,
0,0,0.5,0.5,1,1,
0.5,0.5,1,1,1.5,1.5,
1,1,1.5,1.5,2,2
),4,6,byrow=T)     # pattern of variance for testlets: sd_t6,the number of testlets is 6

SD_C=matrix(c(
0,0,
0.25,0.75,
0.75,1.25,
1.25,1.75
),4,2,byrow=T)     # pattern of variance for contents: sd_c

for (n in 3:3){
for (t in 1:2){
for (i in 1:2){
for (sd_t in 1:4){
for (sd_c in 1:4){

# create folders for data generation and change working directory
dir_name=paste("K:\\Study\\Dissertation\\Simulation Study 2\\",n,"-",t,"-",i,"-",sd_t,"-
",sd_c,sep="")
dir.create(dir_name)
setwd(dir_name)

# create item difficulty correspond with the number of items per testlet
```

```
if (i==1){
b=rep(b1,T[t])
write.table(b,"b.dat",quote = FALSE, sep = "",row.names = FALSE,col.names = FALSE)
}
else{
b=rep(b2,T[t])
write.table(b,"b.dat",quote = FALSE, sep = "",row.names = FALSE,col.names = FALSE)
}

# The total number of items
TI=T[t]*I[i]

# Assign testlet number for each item
testlet=matrix(NA,TI,1)
for (ti in 1:TI){
        testlet[ti]=ceiling(ti/I[i])
}
write.table(testlet,"testlet.txt",quote = FALSE, sep = "",row.names = FALSE,col.names =
FALSE)

# Assign content number for each item
content=sample(c(1,2),TI,replace=TRUE,prob=c(0.5,0.5))
write.table(content,"content.txt",quote = FALSE, sep = "",row.names = FALSE,col.names =
FALSE)

item=cbind(b,testlet,content)

# generate person ability
theta <- rnorm(N[n],0,1)
theta=(theta-mean(theta))/sd(theta)
write.table(theta,"theta.dat",quote = FALSE, sep = "",row.names = FALSE,col.names =
FALSE)

# generate person specific testlet effect
if (t==1){
th1=matrix(NA,N[n],T[t])
for (d in 1:T[t]){
th1[,d]=rnorm(N[n],0,SD_T3[sd_t,d])
if (SD_T3[sd_t,d]!=0){th1[,d]=(th1[,d]-mean(th1[,d]))/sd(th1[,d])*SD_T3[sd_t,d]}
}
write.table(th1,"th1.dat",quote = FALSE, sep = " ",row.names = FALSE,col.names =
FALSE)}
else{
th1=matrix(NA,N[n],T[t])
for (d in 1:T[t]){
th1[,d]=rnorm(N[n],0,SD_T6[sd_t,d])
if (SD_T6[sd_t,d]!=0){th1[,d]=(th1[,d]-mean(th1[,d]))/sd(th1[,d])*SD_T6[sd_t,d]}
}
write.table(th1,"th1.dat",quote = FALSE, sep = " ",row.names = FALSE,col.names =
FALSE)
}
```

```
# generate person specific content effect
th2=matrix(NA,N[n],C)
for (d in 1:C){
th2[,d]=rnorm(N[n],0,SD_C[sd_c,d])
if (SD_C[sd_c,d]!=0){th2[,d]=(th2[,d]-mean(th2[,d]))/sd(th2[,d])*SD_C[sd_c,d]}
}
write.table(th2,"th2.dat",quote = FALSE, sep = " ",row.names = FALSE,col.names =
FALSE)

# generate item response
replication<-100
totmatrix<-N[n]*TI
          for (r in 1:replication){
                    res <- matrix(rep(NA, totmatrix), N[n], TI)
         for(j in 1:N[n]){
                              for (ti in 1:TI){
                                        prob<-1/(1 + exp(-
(theta[j]+th1[j,item[ti,2]]+th2[j,item[ti,3]]- b[ti])))
                                        rini<-runif(1)
                              if(rini>prob){res[j,ti]<-0}
                                        if(rini<prob){res[j,ti]<-1}
                                        }
                                        }
                    filename <- paste("1p",r,".txt",sep="")
                      write.table(res,filename,sep=" ",row.names=F,col.names=F,na=" ",quote=F)
                      }

# create folders for result in each condition
result_name=paste("K:\\Study\\Dissertation\\Simulation Study 2\\",n,"-",t,"-",i,"-",sd_t,"-
",sd_c,"\\result",sep="")
dir.create(result_name)
}
}
}
}
}
```

Appendix D: SAS Code for Parameter Estimation

```
OPTIONS nonumber nodate nocenter pagesize=MAX linesize=120
formdlim='-';
TITLE;
%GLOBAL filesave;
%let n=60; /*number of items*/

%macro condi (ss,nt,ni,tsd,csd);

%do ss=1 %to 3;
%do nt=2 %to 2; *this is fixed;
%do ni=1 %to 2;
%do tsd=1 %to 4;
%do csd=1 %to 4;

%LET filesave=K:\Study\Dissertation\Simulation Study 1\&ss-&nt-&ni-
&tsd-&csd;
LIBNAME result "&filesave\result";

%macro analysis();
%do r=1 %to 50; * iterations 250;
data resp;
infile "&filesave\1p&r..txt";
input item1-item&n;
person=_N_;
run;

DATA vresp; SET resp;
ARRAY aitem(&n) item1-item&n;
DO i=1 TO &n;
item=i; response=aitem(i); OUTPUT;
END;
RUN;


DATA vresp; SET vresp;
ARRAY dummy (&n) i1-i&n;
DO d=1 TO &n;
IF item=d THEN dummy(d)=1; ELSE dummy(d)=0;
END;
DROP i d item1-item&n i1-i&n;
RUN;

data testlet;
infile "&filesave\testlet.txt";
input testl;
id=_N_;
run;

data content;
infile "&filesave\content.txt";
```

```
input con;
id=_N_;
run;


proc sql;
create table vresp2 as
select person, response, item, testl
from vresp , testlet
where vresp.item=testlet.id;
quit;


data vresp2;
set vresp2;
if testl=1 then testl1=1;
else testl1=0;
if testl=2 then testl2=1;
else testl2=0;
if testl=3 then testl3=1;
else testl3=0;
if testl=4 then testl4=1;
else testl4=0;
if testl=5 then testl5=1;
else testl5=0;
if testl=6 then testl6=1;
else testl6=0;
run;

proc sql;
create table vresp3 as
select person, response, item, testl,testl1, testl2, testl3,testl4,
testl5, testl6,con
from vresp2 , content
where vresp2.item=content.id;
quit;

data vresp3;
set vresp3;
if con=1 then con1=1;
else con1=0;
if con=2 then con2=1;
else con2=0;
run;

/* testlet and content */
ods listing close;
ods output FitStatistics =result.tc_fit_&r CovParms=result.tc_cov_&r
ParameterEstimates = result.tc_fixed_&r SolutionR =
result.tc_rand_&r;
proc glimmix data= vresp3 method=laplace ic=pq noclprint noitprint ;
class item person ;
model response  = item
/cl dist = binary link=logit covb noint solution ddfm=bw;
random intercept testl1 testl2 testl3 testl4 testl5 testl6 con1
con2/ subject=person type=simple solution;
run;
```

```
ods listing;

/* testlet only */
ods listing close;
ods output FitStatistics =result.t_fit_&r CovParms=result.t_cov_&r
ParameterEstimates = result.t_fixed_&r SolutionR = result.t_rand_&r;
proc glimmix data= vresp3 method=laplace ic=pq noclprint noitprint ;
class item person ;
model response  = item
/cl dist = binary link=logit covb noint solution ddfm=bw;
random intercept testl1 testl2 testl3 testl4 testl5 testl6 /
subject=person type=simple solution;
run;
ods listing;

/* content only */
ods listing close;
ods output FitStatistics =result.c_fit_&r CovParms=result.c_cov_&r
ParameterEstimates = result.c_fixed_&r SolutionR = result.c_rand_&r;
proc glimmix data= vresp3 method=laplace ic=pq noclprint noitprint ;
class item person ;
model response  = item
/cl dist = binary link=logit covb noint solution ddfm=bw;
random intercept con1 con2/ subject=person type=simple solution;
run;
ods listing;

/* without testlet */
ods listing close;
ods output FitStatistics =result.fit_&r CovParms=result.cov_&r
ParameterEstimates = result.fixed_&r SolutionR = result.rand_&r;
proc glimmix data= vresp3 method=laplace ic=pq noclprint noitprint ;
class item person ;
model response  = item
/cl dist = binary link=logit covb noint solution ddfm=bw;
random intercept / subject=person type=simple solution;
run;
ods listing;


%end;
%mend analysis;

%analysis;

%end;
%end;
%end;
%end;
%end;

%mend;
%condi;
```

Appendix E: Identified Significant Effects on Error Indexes

Table 40

Identified Significant Impacts on Relative Bias in Item Difficulty Estimation

	Significant Effects	df	SS	MS	F	*p*	Eta-squared
Simulation 1	vt	3	28.5309	9.5103	29124.30	<.0001	0.1529
	vc	3	11.8635	3.9545	12110.20	<.0001	0.0636
	model	3	47.9828	15.9943	48980.80	<.0001	0.2572
	vt*model	9	27.7970	3.0886	9458.40	<.0001	0.1490
	vc*model	9	12.4743	1.3860	4244.59	<.0001	0.0669
	Error	21696	7.0846	0.0003			
	Corrected Total	22463	186.5476				
Simulation 2	vt	3	26.8697	8.9566	4113.13	<.0001	0.1176
	vc	3	12.6033	4.2011	1929.28	<.0001	0.0552
	model	3	52.9495	17.6498	8105.35	<.0001	0.2318
	vt*model	9	25.2115	2.8013	1286.43	<.0001	0.1104
	vc*model	9	12.6974	1.4108	647.89	<.0001	0.0556
	Error	21696	47.2442	0.0022			
	Corrected Total	22463	228.4166				

Note: nt = number of testlets; vt = testlet effect; vc = content effect; ni = number of items per testlet; ss = sample size.

Table 41

Identified Significant Impacts on RMSE in Item Difficulty Estimation

	Significant Effects	df	SS	MS	F	*p*	Eta-squared
Simulation 1	ss	2	6.8826	3.4413	521.86	<.0001	0.0195
	vt	3	34.8768	11.6256	1763.00	<.0001	0.0991
	vc	3	12.7540	4.2513	644.71	<.0001	0.0362
	model	3	39.3442	13.1147	1988.82	<.0001	0.1118
	vt*model	9	30.8522	3.4280	519.85	<.0001	0.0876
	vc*model	9	11.7211	1.3023	197.50	<.0001	0.0333
	Error	25152	165.8580	0.0066			
	Corrected Total	25919	352.0645				
Simulation 2	ss	2	6.5486	3.2743	322.70	<.0001	0.0148
	vt	3	31.6313	10.5438	1039.14	<.0001	0.0714
	vc	3	13.4113	4.4704	440.58	<.0001	0.0303
	model	3	46.2574	15.4191	1519.63	<.0001	0.1044
	vt*model	9	28.0032	3.1115	306.65	<.0001	0.0632
	vc*model	9	12.1782	1.3531	133.36	<.0001	0.0275
	Error	25152	255.2083	0.0101			
	Corrected Total	25919	442.9457				

Note: nt = number of testlets; vt = testlet effect; vc = content effect; ni = number of items per testlet; ss = sample size.

Table 42

Identified Significant Impacts on SE in Item Difficulty Estimation

	Significant Effects	df	SS	MS	F	*p*	Eta-squared
Simulation 1	ss	2	10.9833	5.4917	31906.50	<.0001	0.5578
	model	3	0.5874	0.1958	1137.64	<.0001	0.0298
	vt*model	9	0.3337	0.0371	215.44	<.0001	0.0169
	Error	25152	4.3291	0.0002			
	Corrected Total	25919	19.6905				
Simulation 2	ss	2	10.7750	5.3875	30280.50	<.0001	0.5463
	model	3	0.6444	0.2148	1207.19	<.0001	0.0327
	vt*model	9	0.3041	0.0338	189.89	<.0001	0.0154
	Error	25152	4.4751	0.0002			
	Corrected Total	25919	19.7241				

Note: nt = number of testlets; vt = testlet effect; vc = content effect; ni = number of items per testlet; ss = sample size.

Table 43

Identified Significant Impacts on RMSE in Persons' Ability Estimation

	Significant Effects	df	SS	MS	F	*p*	Eta-squared
Simulation 1	nt	1	1104.0631	1104.0631	10494.30	<.0001	0.0103
	vc	3	5360.1385	1786.7128	16982.90	<.0001	0.0501
	Error	895232	94184.0618	0.1052			
	Corrected Total	895999	107022.5268				
Simulation 2	nt	1	1203.5005	1203.5005	11919.10	<.0001	0.0117
	vc	3	4928.9147	1642.9716	16271.40	<.0001	0.0478
	Error	895232	90394.0921	0.1010			
	Corrected Total	895999	103030.4946				

Note: nt = number of testlets; vt = testlet effect; vc = content effect; ni = number of items per testlet; ss = sample size.

Table 44

Identified Significant Impacts on SE in Persons' Ability Estimation

	Significant Effects	df	SS	MS	F	*p*	Eta-squared
Simulation 1	nt	1	717.7431	717.7431	571286.00	<.0001	0.1167
	ni	1	732.1786	732.1786	582776.00	<.0001	0.119
	vt	3	410.6700	136.8900	108957.00	<.0001	0.0668
	vc	3	436.3651	145.4550	115775.00	<.0001	0.0709
	model	3	688.3194	229.4398	182622.00	<.0001	0.1119
	vc*model	9	487.4920	54.1658	43113.10	<.0001	0.0792
	Error	895232	1124.7378	0.0013			
	Corrected Total	895999	6152.3219				
Simulation 2	nt	1	719.4614	719.4614	529745.00	<.0001	0.1287
	ni	1	647.7576	647.7576	476949.00	<.0001	0.1159
	vt	3	274.3532	91.4511	67336.10	<.0001	0.0491
	vc	3	483.6975	161.2325	118717.00	<.0001	0.0865
	model	3	452.6183	150.8728	111089.00	<.0001	0.081
	vc*model	9	450.2598	50.0289	36836.60	<.0001	0.0805
	Error	895232	1215.8405	0.0014			
	Corrected Total	895999	5590.3531				

Note: nt = number of testlets; vt = testlet effect; vc = content effect; ni = number of items per testlet; ss = sample size.

Table 45

Identified Significant Impacts on Bias in Ability's SD Estimation

	Significant Effects	df	SS	MS	F	*p*	Eta-squared
Simulation 1	nt	1	0.1514	0.1514	628.49	<.0001	0.0221
	vt	3	0.1949	0.0650	269.62	<.0001	0.0284
	vc	3	1.4197	0.4732	1964.26	<.0001	0.2072
	model	3	1.9827	0.6609	2743.12	<.0001	0.2894
	nt*vt	3	0.1289	0.0430	178.32	<.0001	0.0188
	nt*model	3	0.1643	0.0548	227.37	<.0001	0.0240
	vt*model	9	0.0805	0.0089	37.13	<.0001	0.0118
	vc*model	9	1.6432	0.1826	757.83	<.0001	0.2399
	ss*vt*vc	18	0.0837	0.0046	19.30	<.0001	0.0122
	nt*vt*model	9	0.1422	0.0158	65.58	<.0001	0.0208
	ss*ni*vt*vc	18	0.0830	0.0046	19.15	<.0001	0.0121
	Error	54	0.0130	0.0002			
	Corrected Total	767	6.8509				
Simulation 2	nt	1	0.2089	0.2089	467.28	<.0001	0.0325
	vt	3	0.1188	0.0396	88.56	<.0001	0.0185
	vc	3	1.0617	0.3539	791.70	<.0001	0.1652
	model	3	1.9948	0.6649	1487.50	<.0001	0.3104
	nt*vt	3	0.1594	0.0531	118.87	<.0001	0.0248
	nt*model	3	0.1735	0.0578	129.37	<.0001	0.0270
	vt*model	9	0.0906	0.0101	22.51	<.0001	0.0141
	vc*model	9	1.4331	0.1592	356.20	<.0001	0.2230
	ss*vt*vc	18	0.0864	0.0048	10.74	<.0001	0.0134
	nt*vt*model	9	0.1105	0.0123	27.46	<.0001	0.0172
	ss*nt*vt*vc	18	0.0883	0.0049	10.98	<.0001	0.0137
	ss*ni*vt*vc	18	0.1116	0.0062	13.87	<.0001	0.0174
	ss*nt*ni*vt*vc	18	0.1455	0.0081	18.08	<.0001	0.0226
	Error	54	0.0241	0.0004			
	Corrected Total	767	6.4270				

Note: nt = number of testlets; vt = testlet effect; vc = content effect; ni = number of items per testlet; ss = sample size.

Table 46

Identified Significant Impacts on RMSE in Ability's SD Estimation

	Significant Effects	df	SS	MS	F	*p*	Eta-squared
Simulation 1	vc	3	1.3787	0.4596	1101.80	<.0001	0.3747
	model	3	0.5322	0.1774	425.32	<.0001	0.1446
	ss*model	6	0.0403	0.0067	16.09	<.0001	0.0109
	nt*model	3	0.0547	0.0182	43.71	<.0001	0.0149
	vt*vc	9	0.0683	0.0076	18.21	<.0001	0.0186
	vt*model	9	0.0745	0.0083	19.85	<.0001	0.0203
	vc*model	9	0.7237	0.0804	192.78	<.0001	0.1967
	ss*vc*model	18	0.0449	0.0025	5.98	<.0001	0.0122
	nt*vt*model	9	0.0676	0.0075	18.01	<.0001	0.0184
	nt*vc*model	9	0.0471	0.0052	12.54	<.0001	0.0128
	vt*vc*model	27	0.0812	0.0030	7.21	<.0001	0.0221
	nt*vt*vc*model	27	0.0530	0.0020	4.70	<.0001	0.0144
	Error	54	0.0225	0.0004			
	Corrected Total	767	3.6797				
Simulation 2	vc	3	1.1071	0.3690	357.23	<.0001	0.3543
	model	3	0.4510	0.1503	145.51	<.0001	0.1443
	nt*model	3	0.0717	0.0239	23.14	<.0001	0.0229
	vt*vc	9	0.0774	0.0086	8.33	<.0001	0.0248
	vt*model	9	0.0595	0.0066	6.40	<.0001	0.0190
	vc*model	9	0.4786	0.0532	51.48	<.0001	0.1532
	nt*vt*model	9	0.0577	0.0064	6.20	<.0001	0.0185
	nt*vc*model	9	0.0503	0.0056	5.41	<.0001	0.0161
	vt*vc*model	27	0.0657	0.0024	2.36	0.0037	0.0210
	ss*nt*vt*vc	18	0.0351	0.0020	1.89	0.0373	0.0112
	Error	54	0.0558	0.0010			
	Corrected Total	767	3.1244				

Note: nt = number of testlets; vt = testlet effect; vc = content effect; ni = number of items per testlet; ss = sample size.

Table 47

Identified Significant Impacts on SE in Ability's SD Estimation

	Significant Effects	df	SS	MS	F	*p*	Eta-squared
Simulation 1	ss	2	0.0360	0.0180	10168.80	<.0001	0.3100
	nt	1	0.0191	0.0191	10780.40	<.0001	0.1643
	ni	1	0.0170	0.0170	9639.24	<.0001	0.1469
	vc	3	0.0046	0.0015	874.76	<.0001	0.0400
	model	3	0.0111	0.0037	2095.57	<.0001	0.0958
	ss*nt	2	0.0024	0.0012	665.57	<.0001	0.0203
	ss*ni	2	0.0023	0.0012	655.95	<.0001	0.0200
	ss*model	6	0.0012	0.0002	114.53	<.0001	0.0105
	nt*ni	1	0.0017	0.0017	968.17	<.0001	0.0148
	vt*model	9	0.0029	0.0003	183.37	<.0001	0.0252
	vc*model	9	0.0067	0.0007	419.52	<.0001	0.0575
	Error	54	0.0001	0.0000			
	Corrected Total	767	0.1160				
Simulation 2	ss	2	0.0336	0.0168	4734.38	<.0001	0.3037
	nt	1	0.0183	0.0183	5149.55	<.0001	0.1651
	ni	1	0.0160	0.0160	4504.82	<.0001	0.1445
	vc	3	0.0046	0.0015	431.70	<.0001	0.0415
	model	3	0.0118	0.0039	1106.39	<.0001	0.1064
	ss*nt	2	0.0027	0.0013	376.99	<.0001	0.0242
	ss*ni	2	0.0018	0.0009	255.20	<.0001	0.0164
	ss*model	6	0.0013	0.0002	60.62	<.0001	0.0117
	nt*ni	1	0.0012	0.0012	337.16	<.0001	0.0108
	vt*model	9	0.0026	0.0003	81.00	<.0001	0.0234
	vc*model	9	0.0062	0.0007	193.12	<.0001	0.0557
	Error	54	0.0002	0.0000			
	Corrected Total	767	0.1106				

Note: nt = number of testlets; vt = testlet effect; vc = content effect; ni = number of items per testlet; ss = sample size.

Table 48

Identified Significant Impacts on Bias in Testlet Effects' SD Estimation

	Significant Effects	df	SS	MS	F	*p*	Eta-squared
Simulation 1	ni	1	0.6347	0.6347	66.67	<.0001	0.0215
	vt	3	6.5922	2.1974	230.83	<.0001	0.2238
	vc	3	1.1007	0.3669	38.54	<.0001	0.0374
	model	1	0.3243	0.3243	34.06	<.0001	0.0110
	ni*vt	3	0.3953	0.1318	13.84	<.0001	0.0134
	vt*vc	9	0.8038	0.0893	9.38	<.0001	0.0273
	vt*model	3	0.7234	0.2411	25.33	<.0001	0.0246
	vc*model	3	0.3127	0.1042	10.95	<.0001	0.0106
	vt*vc*model	9	0.7016	0.0780	8.19	<.0001	0.0238
	Error	1362	12.9658	0.0095			
	Corrected Total	1727	29.4617				
Simulation 2	ni	1	0.7125	0.7125	67.58	<.0001	0.0225
	vt	3	7.1336	2.3779	225.54	<.0001	0.2256
	vc	3	1.4278	0.4759	45.14	<.0001	0.0452
	model	1	0.4950	0.4950	46.95	<.0001	0.0157
	ni*vt	3	0.4653	0.1551	14.71	<.0001	0.0147
	vt*vc	9	0.5911	0.0657	6.23	<.0001	0.0187
	vt*model	3	0.7243	0.2414	22.90	<.0001	0.0229
	vc*model	3	0.4970	0.1657	15.71	<.0001	0.0157
	vt*vc*model	9	0.5605	0.0623	5.91	<.0001	0.0177
	Error	1362	14.3594	0.0105			
	Corrected Total	1727	31.6222				

Note: nt = number of testlets; vt = testlet effect; vc = content effect; ni = number of items per testlet; ss = sample size.

Table 49

Identified Significant Impacts on RMSE in Testlet Effects' SD Estimation

	Significant Effects	df	SS	MS	F	*p*	Eta-squared
Simulation 1	ss	2	0.4636	0.2318	51.10	<.0001	0.0269
	ni	1	1.6104	1.6104	355.04	<.0001	0.0936
	vt	3	0.4374	0.1458	32.14	<.0001	0.0254
	vc	3	2.0073	0.6691	147.51	<.0001	0.1166
	model	1	1.1474	1.1474	252.95	<.0001	0.0667
	ni*vt	3	0.3411	0.1137	25.06	<.0001	0.0198
	vt*vc	9	0.6036	0.0671	14.79	<.0001	0.0351
	vc*model	3	1.3411	0.4470	98.56	<.0001	0.0779
	ss*ni*vt*vc	18	0.2508	0.0139	3.07	<.0001	0.0146
	Error	1362	6.1779	0.0045			
	Corrected Total	1727	17.2103				
Simulation 2	ss	2	0.5752	0.2876	49.74	<.0001	0.0323
	ni	1	1.3851	1.3851	239.53	<.0001	0.0778
	vt	3	0.3651	0.1217	21.04	<.0001	0.0205
	vc	3	1.2251	0.4084	70.62	<.0001	0.0688
	model	1	1.0672	1.0672	184.55	<.0001	0.0599
	ni*vt	3	0.1908	0.0636	11.00	<.0001	0.0107
	vt*vc	9	0.8682	0.0965	16.68	<.0001	0.0487
	vt*model	3	0.2480	0.0827	14.30	<.0001	0.0139
	vc*model	3	0.8650	0.2883	49.86	<.0001	0.0486
	vt*vc*model	9	0.3870	0.0430	7.44	<.0001	0.0217
	Error	1362	7.8756	0.0058			
	Corrected Total	1727	17.8095				

Note: nt = number of testlets; vt = testlet effect; vc = content effect; ni = number of items per testlet; ss = sample size.

Table 50

Identified Significant Impacts on Relative Bias in Testlet Effects' SD Estimation

	Significant Effects	df	SS	MS	F	*p*	Eta-squared
Simulation 1	ni	1	1.5648	1.5648	65.85	<.0001	0.0412
	vt	2	2.2853	1.1426	48.08	<.0001	0.0602
	vc	3	2.3563	0.7854	33.05	<.0001	0.0620
	model	1	0.8076	0.8076	33.99	<.0001	0.0213
	ni*vt	2	0.4788	0.2394	10.07	<.0001	0.0126
	vc*model	3	0.7904	0.2635	11.09	<.0001	0.0208
	Error	1020	24.2390	0.0238			
	Corrected Total	1295	37.9836				
Simulation 2	ni	1	2.1009	2.1009	110.15	<.0001	0.0722
	vt	2	1.0014	0.5007	26.25	<.0001	0.0344
	vc	3	2.9270	0.9757	51.15	<.0001	0.1006
	model	1	1.4572	1.4572	76.40	<.0001	0.0501
	ni*vc	3	0.3125	0.1042	5.46	0.0010	0.0107
	vc*model	3	1.1023	0.3674	19.26	<.0001	0.0379
	Error	876	16.7087	0.0191			
	Corrected Total	1151	29.1084				

Note: nt = number of testlets; vt = testlet effect; vc = content effect; ni = number of items per testlet; ss = sample size.

Table 51

Identified Significant Impacts on SE in Testlet Effects' SD Estimation

	Significant Effects	df	SS	MS	F	*p*	Eta-squared
Simulation 1	ss	2	0.6897	0.3449	1302.57	<.0001	0.2757
	ni	1	0.4880	0.4880	1843.22	<.0001	0.1950
	vt	3	0.2441	0.0814	307.37	<.0001	0.0976
	model	1	0.0273	0.0273	102.98	<.0001	0.0109
	ss*ni	2	0.0475	0.0238	89.72	<.0001	0.0190
	nt*vt	3	0.0278	0.0093	35.00	<.0001	0.0111
	ni*vt	3	0.1100	0.0367	138.43	<.0001	0.0439
	vt*vc	9	0.0974	0.0108	40.86	<.0001	0.0389
	vc*model	3	0.0320	0.0107	40.28	<.0001	0.0128
	Error	1362	0.3606	0.0003			
	Corrected Total	1727	2.5022				
Simulation 2	ss	2	0.6576	0.3288	569.44	<.0001	0.2688
	nt	1	0.0246	0.0246	42.55	<.0001	0.0100
	ni	1	0.3520	0.3520	609.63	<.0001	0.1439
	vt	3	0.0369	0.0123	21.32	<.0001	0.0151
	vc	3	0.0480	0.0160	27.69	<.0001	0.0196
	model	1	0.0498	0.0498	86.23	<.0001	0.0203
	ss*ni	2	0.0311	0.0155	26.89	<.0001	0.0127
	ni*vt	3	0.0264	0.0088	15.22	<.0001	0.0108
	vt*vc	9	0.0851	0.0095	16.38	<.0001	0.0348
	vc*model	3	0.0543	0.0181	31.34	<.0001	0.0222
	Error	1362	0.7865	0.0006			
	Corrected Total	1727	2.4470				

Note: nt = number of testlets; vt = testlet effect; vc = content effect; ni = number of items per testlet; ss = sample size.

Table 52

Identified Significant Impacts on Bias in Content Effects' SD Estimation

	Significant Effects	df	SS	MS	F	*p*	Eta-squared
Simulation 1	vt	3	1.0414	0.3471	114.74	<.0001	0.1229
	vc	3	1.9875	0.6625	218.98	<.0001	0.2346
	model	1	0.4849	0.4849	160.29	<.0001	0.0572
	ni*vc	3	0.1257	0.0419	13.85	<.0001	0.0148
	vt*vc	9	0.5864	0.0652	21.54	<.0001	0.0692
	vt*model	3	0.4378	0.1459	48.24	<.0001	0.0517
	vc*model	3	0.5493	0.1831	60.52	<.0001	0.0648
	ni*vt*vc	9	0.0891	0.0099	3.27	0.0007	0.0105
	vt*vc*model	9	0.5448	0.0605	20.01	<.0001	0.0643
	ss*ni*vt*vc	18	0.1110	0.0062	2.04	0.0075	0.0131
	ss*vt*vc*model	18	0.0925	0.0051	1.70	0.0369	0.0109
	ss*nt*ni*vt*vc	18	0.0959	0.0053	1.76	0.0279	0.0113
	Error	402	1.2162	0.0030			
	Corrected Total	767	8.4729				
Simulation 2	ni	1	0.1676	0.1676	32.75	<.0001	0.0158
	vt	3	0.9174	0.3058	59.75	<.0001	0.0864
	vc	3	2.3227	0.7742	151.27	<.0001	0.2187
	model	1	0.6368	0.6368	124.42	<.0001	0.0599
	ni*vc	3	0.1907	0.0636	12.42	<.0001	0.0180
	vt*vc	9	0.4904	0.0545	10.65	<.0001	0.0462
	vt*model	3	0.4694	0.1565	30.57	<.0001	0.0442
	vc*model	3	0.6931	0.2310	45.14	<.0001	0.0653
	ni*vt*vc	9	0.1074	0.0119	2.33	0.0144	0.0101
	vt*vc*model	9	0.4973	0.0553	10.80	<.0001	0.0468
	ss*nt*ni*vt	6	0.1236	0.0206	4.02	0.0006	0.0116
	ss*nt*vt*vc	18	0.1533	0.0085	1.66	0.043	0.0144
	Error	402	2.0575	0.0051			
	Corrected Total	767	10.6227				

Note: nt = number of testlets; vt = testlet effect; vc = content effect; ni = number of items per testlet; ss = sample size.

Table 53

Identified Significant Impacts on Relative Bias in Content Effects' SD Estimation

	Significant Effects	df	SS	MS	F	*p*	Eta-squared
Simulation 1	ni	1	0.2225	0.2225	25.54	<.0001	0.0227
	vt	3	1.9710	0.6570	75.40	<.0001	0.2009
	vc	2	0.3727	0.1863	21.38	<.0001	0.0380
	model	1	0.7641	0.7641	87.69	<.0001	0.0779
	vt*vc	6	0.2607	0.0435	4.99	<.0001	0.0266
	vt*model	3	0.8391	0.2797	32.10	<.0001	0.0855
	ss*vt*vc*model	12	0.2103	0.0175	2.01	0.0231	0.0214
	Error	300	2.6141	0.0087			
	Corrected Total	575	9.8103				
Simulation 2	ni	1	0.3878	0.3878	11.54	0.0008	0.0165
	ss*vc	4	0.5155	0.1289	3.84	0.0047	0.0219
	ni*vt	3	0.3819	0.1273	3.79	0.0108	0.0162
	vt*model	3	0.5139	0.1713	5.10	0.0019	0.0218
	ss*nt*vc	4	0.6212	0.1553	4.62	0.0012	0.0264
	ss*ni*vc	4	0.4741	0.1185	3.53	0.0078	0.0201
	ni*vt*vc	6	0.6774	0.1129	3.36	0.0032	0.0288
	ss*nt*vt*vc	12	0.8049	0.0671	2.00	0.0243	0.0342
	Error	300	10.0810	0.0336			
	Corrected Total	575	23.5522				

Note: nt = number of testlets; vt = testlet effect; vc = content effect; ni = number of items per testlet; ss = sample size.

Table 54

Identified Significant Impacts on RMSE in Content Effects' SD Estimation

	Significant Effects	df	SS	MS	F	*p*	Eta-squared
Simulation 1	ss	2	0.1360	0.0680	33.32	<.0001	0.0253
	nt	1	0.0922	0.0922	45.19	<.0001	0.0171
	ni	1	0.3041	0.3041	148.97	<.0001	0.0565
	vt	3	0.9109	0.3036	148.76	<.0001	0.1692
	vc	3	0.1818	0.0606	29.69	<.0001	0.0338
	model	1	0.4855	0.4855	237.88	<.0001	0.0902
	vt*vc	9	0.3097	0.0344	16.86	<.0001	0.0575
	vt*model	3	0.5638	0.1879	92.07	<.0001	0.1047
	vc*model	3	0.1916	0.0639	31.29	<.0001	0.0356
	ss*vt*vc	18	0.1239	0.0069	3.37	<.0001	0.0230
	vt*vc*model	9	0.1748	0.0194	9.52	<.0001	0.0325
	ss*nt*vt*vc	18	0.0825	0.0046	2.25	0.0026	0.0153
	ss*ni*vt*vc	18	0.0896	0.0050	2.44	0.0009	0.0166
	ss*vt*vc*model	18	0.0820	0.0046	2.23	0.0028	0.0152
	ni*vt*vc*model	9	0.0573	0.0064	3.12	0.0012	0.0106
	ss*nt*ni*vt*vc	18	0.0805	0.0045	2.19	0.0034	0.0150
	Error	402	0.8205	0.0020			
	Corrected Total	767	5.3851				
Simulation 2	ss	2	0.2305	0.1153	29.80	<.0001	0.0333
	nt	1	0.2139	0.2139	55.30	<.0001	0.0309
	ni	1	0.4029	0.4029	104.16	<.0001	0.0581
	vt	3	0.7075	0.2358	60.98	<.0001	0.1021
	vc	3	0.3113	0.1038	26.83	<.0001	0.0449
	model	1	0.6434	0.6434	166.35	<.0001	0.0928
	ni*vt	3	0.0709	0.0236	6.11	0.0004	0.0102
	vt*vc	9	0.3907	0.0434	11.22	<.0001	0.0564
	vt*model	3	0.5594	0.1865	48.21	<.0001	0.0807
	vc*model	3	0.2511	0.0837	21.64	<.0001	0.0362
	vt*vc*model	9	0.2241	0.0249	6.44	<.0001	0.0323
	ss*nt*ni*vt*vc	18	0.1750	0.0097	2.51	0.0006	0.0252
	Error	402	1.5548	0.0039			
	Corrected Total	767	6.9307				

Note: nt = number of testlets; vt = testlet effect; vc = content effect; ni = number of items per testlet; ss = sample size.

Table 55

Identified Significant Impacts on SE in Content Effects' SD Estimation

	Significant Effects	df	SS	MS	F	*p*	Eta-squared
Simulation 1	ss	2	0.1792	0.0896	416.72	<.0001	0.2321
	nt	1	0.1054	0.1054	489.98	<.0001	0.1364
	ni	1	0.1289	0.1289	599.32	<.0001	0.1669
	vc	3	0.0714	0.0238	110.63	<.0001	0.0924
	ss*ni	2	0.0185	0.0093	43.03	<.0001	0.0240
	nt*ni	1	0.0084	0.0084	39.19	<.0001	0.0109
	nt*vc	3	0.0163	0.0054	25.23	<.0001	0.0211
	ni*vc	3	0.0091	0.0030	14.03	<.0001	0.0117
	vt*vc	9	0.0179	0.0020	9.25	<.0001	0.0232
	ss*vt*vc	18	0.0132	0.0007	3.40	<.0001	0.0171
	ss*nt*vt*vc	18	0.0100	0.0006	2.58	0.0004	0.0129
	ss*ni*vt*vc	18	0.0099	0.0006	2.57	0.0005	0.0129
	ss*nt*ni*vt*vc	18	0.0086	0.0005	2.23	0.0028	0.0112
	Error	402	0.0864	0.0002			
	Corrected Total	767	0.7723				
Simulation 2	ss	2	0.2009	0.1004	126.53	<.0001	0.1803
	nt	1	0.1033	0.1033	130.19	<.0001	0.0927
	ni	1	0.1188	0.1188	149.63	<.0001	0.1066
	vc	3	0.1175	0.0392	49.34	<.0001	0.1054
	model	1	0.0123	0.0123	15.47	<.0001	0.0110
	ss*nt	2	0.0132	0.0066	8.31	0.0003	0.0118
	ss*ni	2	0.0220	0.0110	13.86	<.0001	0.0198
	nt*vc	3	0.0127	0.0042	5.34	0.0013	0.0114
	vt*vc	9	0.0222	0.0025	3.11	0.0012	0.0200
	Error	402	0.3191	0.0008			
	Corrected Total	767	1.1142				

Note: nt = number of testlets; vt = testlet effect; vc = content effect; ni = number of items per testlet; ss = sample size.

Appendix F: The Percentage of Replications of Identifying Correct Models

		Simulation Study 1					Simulation Study 2				
Condition		AIC%	AICC%	BIC%	CAIC%	HQIC%	AIC%	AICC%	BIC%	CAIC%	HQIC%
1	1-1-1-1-1	96	96	100	100	100	86	86	100	100	96
2	1-1-1-1-2	**94**	**94**	**66**	**56**	**88**	**88**	**88**	**74**	**70**	**86**
3	1-1-1-1-3	96	96	100	100	100	92	92	98	98	96
4	1-1-1-1-4	98	98	100	100	100	98	98	100	100	100
5	1-1-1-2-1	**82**	**82**	**20**	**14**	**56**	**80**	**80**	**90**	**90**	**90**
6	1-1-1-2-2	**72**	**72**	**2**	**0**	**34**	**100**	**100**	**86**	**72**	**100**
7	1-1-1-2-3	**64**	**64**	**10**	**2**	**28**	**100**	**100**	**84**	**80**	**92**
8	1-1-1-2-4	**38**	**38**	**8**	**4**	**24**	**100**	**100**	**96**	**90**	**100**
9	1-1-1-3-1	98	98	100	100	98	94	94	98	98	96
10	1-1-1-3-2	**78**	**78**	**26**	**22**	**60**	**92**	**92**	**66**	**54**	**84**
11	1-1-1-3-3	100	100	100	100	100	100	100	100	100	100
12	1-1-1-3-4	100	100	100	100	100	100	100	100	100	100
13	1-1-1-4-1	98	98	100	100	98	94	94	100	100	98
14	1-1-1-4-2	**74**	**74**	**28**	**26**	**48**	**96**	**96**	**52**	**38**	**76**
15	1-1-1-4-3	100	100	100	100	100	100	100	100	100	100
16	1-1-1-4-4	100	100	100	100	100	100	100	100	100	100
17	1-1-2-1-1	92	92	100	100	98	86	86	100	100	98
18	1-1-2-1-2	88	88	100	100	98	92	92	98	98	98
19	1-1-2-1-3	94	94	100	100	98	92	92	100	100	98
20	1-1-2-1-4	92	92	100	100	98	96	96	100	100	100
21	1-1-2-2-1	96	96	100	98	100	80	80	86	86	84
22	1-1-2-2-2	100	100	98	96	100	100	100	100	100	100
23	1-1-2-2-3	100	100	98	98	100	100	100	100	100	100
24	1-1-2-2-4	**98**	**98**	**84**	**80**	**98**	100	100	100	100	100
25	1-1-2-3-1	96	96	100	100	96	96	96	100	100	100
26	1-1-2-3-2	100	100	100	100	100	100	100	100	100	100
27	1-1-2-3-3	100	100	100	100	100	100	100	100	100	100
28	1-1-2-3-4	100	100	100	100	100	100	100	100	100	100
29	1-1-2-4-1	98	98	100	100	98	98	98	100	100	98
30	1-1-2-4-2	100	100	100	100	100	100	100	100	100	100
31	1-1-2-4-3	100	100	100	100	100	100	100	100	100	100
32	1-1-2-4-4	100	100	100	100	100	100	100	100	100	100
33	1-2-1-1-1	92	92	100	100	96	86	86	100	100	100
34	1-2-1-1-2	90	90	96	96	94	94	94	100	100	98
35	1-2-1-1-3	88	88	100	100	98	96	96	100	100	100

36	1-2-1-1-4	94	94	100	100	98	80	80	98	98	96
37	1-2-1-2-1	**90**	**90**	**22**	**10**	**76**	76	76	80	80	76
38	1-2-1-2-2	**98**	**98**	**22**	**8**	**66**	100	100	100	100	100
39	1-2-1-2-3	**94**	**94**	**8**	**0**	**62**	100	100	100	100	100
40	1-2-1-2-4	**80**	**80**	**0**	**0**	**32**	100	100	100	100	100
41	1-2-1-3-1	96	96	98	100	96	92	92	96	96	94
42	1-2-1-3-2	100	100	100	100	100	100	100	100	100	100
43	1-2-1-3-3	100	100	100	100	100	100	100	100	100	100
44	1-2-1-3-4	100	100	100	100	100	100	100	100	100	100
45	1-2-1-4-1	100	100	100	100	100	98	98	100	100	100
46	1-2-1-4-2	100	100	96	90	98	100	100	100	100	100
47	1-2-1-4-3	100	100	100	100	100	100	100	100	100	100
48	1-2-1-4-4	100	100	100	100	100	100	100	100	100	100
49	1-2-2-1-1	92	92	100	100	98	86	86	98	98	96
50	1-2-2-1-2	96	96	100	100	100	88	88	100	100	100
51	1-2-2-1-3	88	88	100	100	100	94	94	100	100	100
52	1-2-2-1-4	96	96	100	100	98	98	98	100	100	100
53	1-2-2-2-1	98	98	100	100	98	84	84	86	86	84
54	1-2-2-2-2	100	100	100	100	100	100	100	100	100	100
55	1-2-2-2-3	100	100	100	100	100	100	100	100	100	100
56	1-2-2-2-4	100	100	100	100	100	100	100	100	100	100
57	1-2-2-3-1	94	94	100	100	98	98	98	100	100	100
58	1-2-2-3-2	100	100	100	100	100	100	100	100	100	100
59	1-2-2-3-3	100	100	100	100	100	100	100	100	100	100
60	1-2-2-3-4	100	100	100	100	100	100	100	100	100	100
61	1-2-2-4-1	96	96	100	100	98	98	98	100	100	100
62	1-2-2-4-2	100	100	100	100	100	100	100	100	100	100
63	1-2-2-4-3	100	100	100	100	100	100	100	100	100	100
64	1-2-2-4-4	100	100	100	100	100	100	100	100	100	100
65	2-1-1-1-1	94	94	100	100	100	88	88	100	100	100
66	2-1-1-1-2	96	96	90	86	100	86	86	88	86	90
67	2-1-1-1-3	96	96	100	100	100	96	96	100	100	100
68	2-1-1-1-4	98	98	100	100	100	100	100	100	100	100
69	2-1-1-2-1	**100**	**100**	**58**	**42**	**92**	**78**	**78**	**78**	**78**	**78**
70	2-1-1-2-2	**96**	**96**	**26**	**20**	**74**	100	100	94	86	98
71	2-1-1-2-3	**92**	**92**	**20**	**12**	**64**	100	100	100	100	100
72	2-1-1-2-4	**80**	**80**	**12**	**6**	**46**	100	100	100	98	100
73	2-1-1-3-1	98	98	100	100	100	98	98	100	100	100
74	2-1-1-3-2	98	98	70	56	92	100	100	92	88	98
75	2-1-1-3-3	100	100	100	100	100	100	100	100	100	100
76	2-1-1-3-4	100	100	100	100	100	100	100	100	100	100
77	2-1-1-4-1	100	100	100	100	100	98	98	100	100	98
78	2-1-1-4-2	**96**	**96**	**42**	**32**	**78**	**100**	**100**	**88**	**78**	**98**
79	2-1-1-4-3	100	100	100	100	100	100	100	100	100	100
80	2-1-1-4-4	100	100	100	100	100	100	100	100	100	100

81	2-1-2-1-1	90	90	98	100	96	82	82	100	100	98
82	2-1-2-1-2	96	96	100	100	100	100	100	100	100	100
83	2-1-2-1-3	96	96	100	100	100	96	96	100	100	96
84	2-1-2-1-4	98	98	100	100	100	98	98	100	100	98
85	2-1-2-2-1	98	98	98	100	98	80	80	84	84	84
86	2-1-2-2-2	100	100	100	100	100	100	100	100	100	100
87	2-1-2-2-3	100	100	100	100	100	100	100	100	100	100
88	2-1-2-2-4	100	100	100	100	100	100	100	100	100	100
89	2-1-2-3-1	100	100	100	100	100	100	100	100	100	100
90	2-1-2-3-2	100	100	100	100	100	100	100	100	100	100
91	2-1-2-3-3	100	100	100	100	100	100	100	100	100	100
92	2-1-2-3-4	100	100	100	100	100	100	100	100	100	100
93	2-1-2-4-1	100	100	100	100	100	96	96	100	100	100
94	2-1-2-4-2	100	100	100	100	100	100	100	100	100	100
95	2-1-2-4-3	100	100	100	100	100	100	100	100	100	100
96	2-1-2-4-4	100	100	100	100	100	100	100	100	100	100
97	2-2-1-1-1	90	90	100	100	96	88	88	98	98	94
98	2-2-1-1-2	96	96	100	100	98	92	92	96	96	94
99	2-2-1-1-3	98	98	100	100	100	96	96	100	100	98
100	2-2-1-1-4	94	94	100	100	100	96	96	100	100	100
101	2-2-1-2-1	**90**	**90**	**78**	**74**	**92**	72	72	74	74	74
102	2-2-1-2-2	**100**	**100**	**78**	**50**	**100**	100	100	100	100	100
103	2-2-1-2-3	**98**	**98**	**46**	**32**	**88**	100	100	100	100	100
104	2-2-1-2-4	**100**	**100**	**6**	**4**	**68**	100	100	100	100	100
105	2-2-1-3-1	100	100	100	100	100	100	100	100	100	100
106	2-2-1-3-2	100	100	100	100	100	100	100	100	100	100
107	2-2-1-3-3	100	100	100	100	100	100	100	100	100	100
108	2-2-1-3-4	100	100	100	100	100	100	100	100	100	100
109	2-2-1-4-1	100	100	100	100	100	100	100	100	100	100
110	2-2-1-4-2	100	100	100	100	100	100	100	100	100	100
111	2-2-1-4-3	100	100	100	100	100	100	100	100	100	100
112	2-2-1-4-4	100	100	100	100	100	100	100	100	100	100
113	2-2-2-1-1	86	86	100	100	98	86	86	100	100	100
114	2-2-2-1-2	92	92	100	100	100	96	96	100	100	100
115	2-2-2-1-3	98	98	100	100	100	92	92	100	100	98
116	2-2-2-1-4	94	94	98	100	98	94	94	100	100	98
117	2-2-2-2-1	90	90	100	100	98	90	90	90	90	90
118	2-2-2-2-2	100	100	100	100	100	100	100	100	100	100
119	2-2-2-2-3	100	100	100	100	100	100	100	100	100	100
120	2-2-2-2-4	100	100	100	100	100	100	100	100	100	100
121	2-2-2-3-1	96	96	100	100	98	98	98	100	100	98
122	2-2-2-3-2	100	100	100	100	100	100	100	100	100	100
123	2-2-2-3-3	100	100	100	100	100	100	100	100	100	100
124	2-2-2-3-4	100	100	100	100	100	100	100	100	100	100
125	2-2-2-4-1	100	100	100	100	100	100	100	100	100	100

126	2-2-2-4-2	100	100	100	100	100	100	100	100	100	100
127	2-2-2-4-3	100	100	100	100	100	100	100	100	100	100
128	2-2-2-4-4	100	100	100	100	100	100	100	100	100	100
129	3-1-1-1-1	86	86	100	100	96	98	98	100	100	98
130	3-1-1-1-2	98	98	100	100	100	**78**	**78**	**82**	**82**	**82**
131	3-1-1-1-3	100	100	100	100	100	100	100	100	100	100
132	3-1-1-1-4	100	100	100	100	100	100	100	100	100	100
133	3-1-1-2-1	100	100	94	86	100	94	94	94	94	94
134	3-1-1-2-2	**100**	**100**	**82**	**66**	100	100	100	100	100	100
135	3-1-1-2-3	**98**	**98**	**48**	**32**	**92**	100	100	100	100	100
136	3-1-1-2-4	**96**	**96**	**18**	**14**	**76**	100	100	100	100	100
137	3-1-1-3-1	100	100	100	100	100	100	100	100	100	100
138	3-1-1-3-2	100	100	90	84	100	100	100	100	100	100
139	3-1-1-3-3	100	100	100	100	100	100	100	100	100	100
140	3-1-1-3-4	100	100	100	100	100	100	100	100	100	100
141	3-1-1-4-1	100	100	100	100	100	100	100	100	100	100
142	3-1-1-4-2	**96**	**96**	**58**	**48**	**90**	**96**	**96**	**68**	**60**	**86**
143	3-1-1-4-3	100	100	100	100	100	100	100	100	100	100
144	3-1-1-4-4	100	100	100	100	100	100	100	100	100	100
145	3-1-2-1-1	92	92	100	100	100	96	96	100	100	98
146	3-1-2-1-2	96	96	100	100	100	86	86	90	90	90
147	3-1-2-1-3	98	98	100	100	100	100	100	100	100	100
148	3-1-2-1-4	100	100	100	100	100	98	98	100	100	100
149	3-1-2-2-1	98	98	100	100	100	86	86	86	86	86
150	3-1-2-2-2	100	100	100	100	100	100	100	100	100	100
151	3-1-2-2-3	100	100	100	100	100	100	100	100	100	100
152	3-1-2-2-4	100	100	100	100	100	100	100	100	100	100
153	3-1-2-3-1	100	100	100	100	100	98	98	100	100	100
154	3-1-2-3-2	100	100	100	100	100	100	100	100	100	100
155	3-1-2-3-3	100	100	100	100	100	100	100	100	100	100
156	3-1-2-3-4	100	100	100	100	100	100	100	100	100	100
157	3-1-2-4-1	100	100	100	100	100	98	98	100	100	100
158	3-1-2-4-2	100	100	100	100	100	100	100	100	100	100
159	3-1-2-4-3	100	100	100	100	100	100	100	100	100	100
160	3-1-2-4-4	100	100	100	100	100	100	100	100	100	100
161	3-2-1-1-1	86	86	96	96	94	82	82	98	98	96
162	3-2-1-1-2	100	100	100	100	100	96	96	98	98	98
163	3-2-1-1-3	98	98	100	100	100	100	100	100	100	100
164	3-2-1-1-4	98	98	100	100	100	100	100	100	100	100
165	3-2-1-2-1	100	100	100	100	100	76	76	76	76	76
166	3-2-1-2-2	100	100	100	100	100	100	100	100	100	100
167	3-2-1-2-3	**100**	**100**	**94**	**86**	**100**	100	100	100	100	100
168	3-2-1-2-4	**100**	**100**	**64**	**50**	**96**	100	100	100	100	100
169	3-2-1-3-1	100	100	100	100	100	100	100	100	100	100
170	3-2-1-3-2	100	100	100	100	100	100	100	100	100	100

171	3-2-1-3-3	100	100	100	100	100	100	100	100	100	100
172	3-2-1-3-4	100	100	100	100	100	100	100	100	100	100
173	3-2-1-4-1	96	96	100	100	98	100	100	100	100	100
174	3-2-1-4-2	100	100	100	100	100	100	100	100	100	100
175	3-2-1-4-3	100	100	100	100	100	100	100	100	100	100
176	3-2-1-4-4	100	100	100	100	100	100	100	100	100	100
177	3-2-2-1-1	82	82	98	98	94	90	90	100	100	100
178	3-2-2-1-2	90	90	98	98	96	98	98	100	100	100
179	3-2-2-1-3	98	98	100	100	100	92	92	100	100	100
180	3-2-2-1-4	96	96	100	100	100	98	98	100	100	100
181	3-2-2-2-1	98	98	100	100	98	96	96	98	98	98
182	3-2-2-2-2	100	100	100	100	100	100	100	100	100	100
183	3-2-2-2-3	100	100	100	100	100	100	100	100	100	100
184	3-2-2-2-4	100	100	100	100	100	100	100	100	100	100
185	3-2-2-3-1	100	100	100	100	100	100	100	100	100	100
186	3-2-2-3-2	100	100	100	100	100	100	100	100	100	100
187	3-2-2-3-3	100	100	100	100	100	100	100	100	100	100
188	3-2-2-3-4	100	100	100	100	100	100	100	100	100	100
189	3-2-2-4-1	100	100	100	100	100	100	100	100	100	100
190	3-2-2-4-2	100	100	100	100	100	100	100	100	100	100
191	3-2-2-4-3	100	100	100	100	100	100	100	100	100	100
192	3-2-2-4-4	100	100	100	100	100	100	100	100	100	100

Note: Condition a-b-c-d-e, where a = sample size, b = number of testlets, c = number of items per testlet, d = magnitude of the testlet effect, e = magnitude of the content effects. Refer Table 3 to get the corresponding levels for each manipulated variable.

References

Ackerman, T. (1987). *The robustness of LOGIST and BILOG IRT estimation programs to violations of local independence.* ACT Research Report Series, 87-14, ACT, Iowa City.

Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics, 22*, 47-76.

Ainsworth, J.W. (2002). Why does it take a village? The mediation of neighborhood effects on educational achievement. *Social Forces, 81*, 117–152.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Proceedings, 2nd International Symposium on Information Theory* (pp. 267-281). Budapest: Akademiai Kiado.

Beretvas, S. N. (2010). Cross-classified and multiple membership models. In J. Hox & J. K. Roberts (Eds.), *The handbook of advanced multilevel analysis* (pp. 313-334). New York: NY: Routledge.

Beretvas, S. N., Cawthon, S. W., Lockhart, L. L., & Kaye, A. D. (2012). Assessing impact, DIF, and DFF in accommodated item scores: a comparison of multilevel measurement model parameterizations. *Educational and Psychological Measurement, 72*(5), 754-773.

Beretvas, S. N., & Kamata, A. (2005). The multilevel measurement model: introduction to the special issue. *Journal of Applied Measurement, 6*(3), 247-254.

Beretvas, S. N., & Walker, C. M. (2012). Distinguishing differential testlet functioning from differential bundle functioning using the multilevel measurement model. *Educational and Psychological Measurement, 72*(2), 200-223.

Beretvas, S. N., & Williams, N. J. (2004). The use of hierarchical generalized linear model for item dimensionality assessment. *Journal of Educational Measurement, 31*, 379-395.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord, & M.R. Novick, *Statistical Theories of Mental Test Scores* (chapter 17-29). Reading, MA: Addison-Wesley.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52*, 345-370.

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

Browne, W.J., Goldstein, H. & Rasbash, J. (2001). Multiple membership multiple classification (MMMC) models, *Statistical Modeling, 1*, 103–124.

Chen, T. A. (2010). *Random or fixed testlet effects: a comparison of two multilevel testlet models.* Ph.D. dissertation, University of Texas at Austin.

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

Cheong, Y. F. (2001). Detecting ethnic differences in externalizing problem behavior items via multilevel and multidimensional Rasch models. Paper presented at

the annual meeting of the American Educational Research Association, Seattle, WA.

Chu, K.-l., & Kamata, A. (2005). Test equating in the presence of DIF items. *Journal of Applied Measurement. Special Issue: The Multilevel Measurement Model, 6*(3), 342.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

De Boeck (2008). Random item IRT models. *Psychometrika, 73*(4), 533-559.

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

Enders, C. K. (2001). The impact of nonnormality on full information maximum likelihood estimation for structural equation models with missing data. *Psychological Methods, 6*(4), 352-370.

Ferrara, S., Huynh, H., & Baghi, H. (1997). Contextual characteristics of locally dependent open-ended item clusters in a large-scale performance assessment. *Applied Measurement in Education, 10*(2), 123-144.

Ferrara, S., Huynh, H., & Michaels, H. (1999). Contextual explanations of local dependence in item clusters in a large scale hands-on science performance assessment. *Journal of Educational Measurement, 36*(2), 119-140.

Fielding, A., & Goldstein, H. (2006). Cross-classified and multiple membership structures in multilevel models: An introduction and review. Research Report No. 791 for DfES, London.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*, 359–374.

Fox, J. P. (2003). Stochastic em for estimating the parameters of a multilevel irt model. *British Journal of Mathematical & Statistical Psychology, 56*(1), 65.

Fox, J. P., & Glas, C. A. W. (1998). *A multi-level IRT model with measure-ment error in the predictor variables*. (Research Report 98-16). Department of Educational Measurement and Data Analysis, University of Twente, the Netherlands.

Fox, J. P., & Glas, C. A. W. (2001). Bayesian estimation of a multilevel irt model using gibbs sampling. *Psychometrika, 66*(2), 271.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative general-ised least squares. *Biometrika, 73*, 43-56.

Goldstein, H. (2003). *Multilevel Statistical Models, 3rd Edition.* London, Arnold.

Grady, M., & Beretvas, S. N. (2010). Incorporating student mobility in achievement growth modeling: A cross-classified multiple membership growth curve model. *Multivariate Behavioral Research, 45*, 393-419.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 147-200). New York: American Council on Education ; Macmillan Publishing Company.

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

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

Hannan, E. J., and Quinn, B. G. (1979). The determination of the order of an autoregression, *Journal of the Royal Statistical Society, B, 41*, 190-195.

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

Hively, W., Patterson, H.L., & Page, S.H. (1968). A "universe-defined" system of arithmetic achievement tests. *Journal of Educational Measurement*, *5*, 275–290.

Hoogland, J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: an overview and a meta-analysis. *Sociological Methods and Research*, 26(3), 329-367.

Hoskens, M., & De Boeck, P. (1997). A parametric model for local dependence among test items. *Psychological Methods, 2*, 261-277.

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

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

Jiao, H., Wang S., & Kamata, A. (2005). Modeling local item dependence with the hierarchical generalized linear model. *Journal of Applied Measurement, 6(3),* 311-321.

Jiao, H., Wang, S., Wan, L., & Lu, R. (2009). *Investigation of local item dependence in scenario-based science assessments*. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.

Kamata, A. (1998). *Some generalizations of the Rasch Model: An application of the hierarchical generalized linear model*. Doctoral dissertation, Michigan State University, East Lansing.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement, 38*, 79-93.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer.

Lee, V. E. (2000). Using hierarchical linear modeling to study social contexts: The case of school effects. *Educational Psychologist, 35*(2), 125-141.

Lee, Y. W. (2004). Examining passage-related local item dependence (LID) and measurement construct using Q3 statistics in an EFL reading comprehension test, *Language Testing* (Vol. 21, pp. 74-100). Princeton, NJ: Educational Testing Service.

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

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental teat scores.* Reading, MA: Addison-Wesley.

Lu, R. (2010). *Impacts of local item dependence of testlet items with the multistage tests for pass-fail decisions*. Doctoral dissertation. University of Maryland, College Park, MD.

Luo, W. (2007). *The impact of misspecifying cross-classified random effects models in cross- sectional and longitudinal multilevel data: A monte carlo study*. (Doctoral dissertation, Texas A & M University, 2007). ProQuest.

Luppescu, S. (2002). DIF detection in HLM. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Ma, X., & Wilkins, J. L. M. (2002). The development of science achievement in middle and high school—Individual differences and school effects. *Evaluation Review, 26,* 395–417.

McCoach, D. B., O'Connell, A. A., Reis, S. M., & Levitt, H. A. (2006). Growing readers: A hierarchical linear model of children's reading growth during the first 2 years of school. *Journal of Educational Psychology, 98*(1), 14-28.

Meulders, M., & Xie, Y. (2004). Person-by-item predictors. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 213-240). New York, NY: Springer.

Meyers, J. L. (2004). *The impact of the inappropriate modeling of cross-classified data structures.* Ph.D. dissertation, University of Texas at Austin. (UMI No. 3145342).

Meyers, J. L. & Beretvas, S.N. (2006). The impact of inappropriate modeling of cross-classified data structures. *Multivariate Behavioral Research, 41*(4), 473-497.

Noble, J. P., & Schnelker, D. (2007). *Using hierarchical modeling to examine course work and ACT score relationships across high schools* (ACT Research Report No. 2007-2). Iowa City, IA: ACT, Inc.

Popham, W.J. (1978). *Criterion-referenced measurement*. Englewood Cliffs: Prentice-Hall.

Rasbash, J. & Goldstein, H. (1994). Efficient analysis of mixed hierarchical and cross classified random structures using a multilevel model. *Journal of Educational and Behavioral Statistics, 19*, 337-350.

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

Raudenbush, S. W. (1993). A crossed random effects model for unbalanced data with applications in cross-sectional and longitudinal research. *Journal of Educational Statistics, 18*, 321-349.

Raudenbush, S., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Reckase, M. (1997). A linear logistic multidimensional model for dichotomous item response data. In W. J. Linden, & R. K. Hambleton, *Handbook of Modern Item Response Theory* (pp. 271-286). New York: Springer-Verlag.

Reckase, M. (2009). *Multidimensional Item Response Theory*. New York: Springer.

Ren, W. (2011). *Impact of design features for cross-classified logistic models when the cross-classification structure is ignored*. (Doctoral dissertation, The Ohio State University). ProQuest

SAS Institute Inc. (2008). SAS (Version 9.2) [Computer software]. Cary, NC: Author.

Schwartz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461-464.

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

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Boca Raton: Chapman & Hall/CRC.

Snijders, T.A.B., & Bosker, R.J. (1999). *Multilevel Analysis: An introduction to basic and advanced multilevel modeling*. London: Sage Publications.

Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics, Theory and Methods, A7*, 13-26.

Thissen, D., Bender, R., Chen, W., Hayashi, K., & Wiesen, C. A. (1992). *Item response theory and local dependence: A preliminary report* (Research Memorandum 92-2). Chapel Hill: L. L. Thurstone Laboratory, University of North Carolina at Chapel Hill.

Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: A use of multiple-Categorical Response models. *Journal of Educational Measurement, 26*, 247-260.

Tuerlinckx, F., & De Boeck, P. (2001). The effects of ignoring item interactions on the estimated discrimination parameters in item response theory. *Psychological Methods, 6,* 181-195.

Van den Noortgate, W., & De Boeck, P. (2005). Assessing and explaining differential item functioning using logistic mixed models. *Journal of Educational and Behavioral Statistics, 30*, 443-464.

Van den Noortgate, W., De Boeck, P., & Meulders, M. (2003). Cross-classification multilevel logistic models in psychometrics. *Journal of Educational and Behavioral Statistics, 28*, 369–386.

Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 Law School Admissions Test as an example. *Applied Measurement in Education, 8*(2), 157-187.

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

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

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice , 15*(1), 22-29.

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

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

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8,* 125-145.

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