EQUATING WITH LOCAL DEPENDENCE UNDER THE ANCHOR TEST DESIGN

by

Ting Xu

B.A., East China Normal University, 2002

M.A., The Ohio State University, 2006

Submitted to the Graduate Faculty of

School of Education in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2016

UNIVERSITY OF PITTSBURGH

SCHOOL OF EDUCATION

This dissertation was presented

by

Ting Xu

It was defended on

April 20, 2016

and approved by

Feifei Ye, Assistant Professor, Psychology in Education

Clement A. Stone, Professor, Psychology in Education

Suzanne Lane, Professor, Psychology in Education

Lan Yu, Associate Professor, Medicine and Psychiatry

Dissertation Advisor: Feifei Ye, Assistant Professor, Psychology in Education

EQUATING WITH LOCAL DEPENDENCE UNDER THE ANCHOR TEST DESIGN

Ting Xu, PhD

University of Pittsburgh, 2016

Item response theory (IRT) models are often used in test equating. The effectiveness of IRT equating depends upon how well test data meet the IRT model assumptions. When tests are composed of testlets (i.e., groups of items sharing a common stimulus), the assumption of local item independence is likely to be violated. When examinees are nested within groups (e.g., classrooms, schools, etc.), the assumption of local person independence (i.e., independence of subjects) is unlikely to hold. Multilevel models allow the flexibility of modeling item and person dependence structures simultaneously.

This research investigated the effectiveness of multilevel models as concurrent calibration models on test equating under the anchor test design with the presence of local dependence. The performance of multilevel models was compared to that of traditional IRT models and testlet response theory (TRT) model through two simulation studies. Local item dependence (LID) was considered in the first study, whereas both LID and person dependence were considered in the second study.

The first study compared the performance of four concurrent calibration approaches on equating testlet-based tests: (a) modeled LID using a three-level hierarchical generalized linear model (HGLM); (b) ignored LID and used a two-level HGLM; (c) ignored LID and used the Rasch model; and (d) used testlet scoring and applied the graded-response model (GRM). The results suggested that the two-level HGLM and the Rasch approaches were robust to the violation of the local item independence assumption, in terms of expected score recovery. In addition, the first three approaches provided better equating results than concurrent calibration using the GRM. Further research confirmed previous findings that degree of LID affected the precision of person parameter estimates.

The second study compared the performance of three models (i.e., 3PL IRT model, 3PL TRT model, and 3PL multilevel TRT model) as concurrent calibration models on equating testlet-based tests when examinees were nested within groups. The results showed that ignoring LID affected item parameter recovery. With the presence of both LID and person dependence, the 3PL multilevel TRT model provided the most accurate estimation for person parameters, especially with a high degree of person dependence.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I would like to express my deepest gratitude to my academic advisor, Dr. Feifei Ye, for her support and guidance throughout the years of my graduate studies at the University of Pittsburgh. I am indebted to her patience, kindness, and encouragement. I would like to express my appreciation to Dr. Clement A. Stone and Dr. Suzanne Lane, for their support and encouragement at various stages of my graduate study. To all my committee members, including Dr. Lan Yu, for their invaluable feedback and suggestions throughout my dissertation project. I would like to thank all staff at the Research Methodology program for creating such a positive learning environment.

Thanks to Dr. Kevin H. Kim for his kindness, support, and instruction.

I sincerely thank my managers and colleagues for providing me with the opportunity to work with this wonderful group of people, while completing this dissertation. I would like to thank my friends, classmates, relatives, and everyone else who support me on this journey.

My sincere appreciation belongs to my family, for keeping me company.

1.0 INTRODUCTION

Test equating is widely used in standardized testing in education and psychology to allow direct comparison of scores across multiple test forms. It refers to the statistical process of adjusting test scores on different forms of a test so that scores from different forms can be used interchangeably (Kolen & Brennan, 2004). Because standardized tests are administered on multiple occasions, most testing programs develop multiple test forms to ensure test security. Although different test forms are constructed to be as similar as possible in content and statistical properties, they usually differ somewhat in difficulty. Test equating is therefore intended to account for differences in difficulty among test forms (Kolen & Brennan, 2004).

Both classical test theory (CTT) procedures (e.g., linear equating and equipercentile equating) and Item Response Theory (IRT)-based procedures were developed to equate test forms. Most standardized testing programs now use IRT models to estimate an examinee's ability level. IRT-based procedures thus become a natural choice for test equating. Therefore, the present study will focus on IRT equating.

Test equating involves placing parameter estimates on a common scale. Under the anchor test design, the common items from the base and target form are used to place item parameter estimates on the same scale. Item parameters can be estimated separately for each form (i.e., linking separate calibration) or simultaneously across forms using combined data from the base and target group (i.e., concurrent calibration). When item parameters are estimated separately and the two groups differ in ability (nonequivalent groups), the item parameter estimates are not on the same scale. The reason is that when item parameters are estimated, generally the prior of the ability parameter is set to be a standard normal distribution, no matter how difficult the test items are. In this case, transformation of the scale of one form onto the scale of the other is necessary so as to establish a common scale across the forms. When the two groups of item parameters are estimated simultaneously, item parameter estimates are automatically placed on the same scale.

1.1 STATEMENT OF THE PROBLEM

When IRT models are used to equate test forms, the effectiveness of IRT equating depends upon how well test data meet the IRT model assumptions. Compared with CTT, IRT models make several strong assumptions such as a unidimensional trait, local independence of item responses, and model-data fit. The local independence assumption has two implications: local item independence and local person independence (Reckase, 2009; also see Jiao et al., 2012). The former is achieved when the probability of answering an item correctly is unaffected by the probability of answering other items correctly, conditional on the person's ability level. The latter assumes that responses to a specific item by different persons are independent to each other. If either local item dependence (LID) or local person dependence (LPD) is present, the local independence assumption is violated (Jiao et al., 2012). Therefore, it is crucial to check both facets of the local independence assumption in order to validate the use of IRT models in test equating.

Local independence assumption unlikely holds in real testing applications. LID often occurs when tests are composed of testlets (i.e., groups of items sharing a common stimulus). Potential sources of LID have been discussed in Yen (1993). LDP is likely to occur when data are collected in hierarchical settings (e.g., students nested within classrooms, classrooms nested within schools, etc.). Other potential sources of LPD have been summarized by Jiao et al. (2012). Dual local dependence may exist when testlet data are collected using a cluster sampling method (Jiao et al., 2012).

Previous studies have demonstrated that ignoring LID affects IRT model parameter estimation, test reliability, test equating, etc. (e.g., Bradlow, Wainer, & Wang, 1999; Chen & Thissen, 1997; Wainer, Bradlow, & Wang, 2007; Wang & Wilson, 2005). It has also been shown that ignoring person dependence can lead to increased classification errors (Jiao et al., 2012) and underestimation of ability variance (Jiao, Wang, & Kamata, 2005). Consequently, methods that address the issue of local dependence have been developed.

One approach to address LID is to apply a polytomous IRT model and use testlet scores as the unit of analysis (e.g., Lee, et al., 2001; Thissen, Steinberg, & Mooney, 1989; Zhang, 2007). By summing item scores within the same testlet and treating the number-correct score as a single polytomous item score, it is believed that the within-testlet item dependency can be absorbed and independence between testlet scores is achieved (Zhang, 2007). However, this approach leads to some loss of information because the response pattern within each testlet is ignored (Wainer, Bradlow, & Du, 2000; Wainer, Bradlow, & Wang, 2007).

Another approach is to fit the data with a model that explicitly models LID effects and also maintains testlet information, such as the testlet response theory (TRT) model. The TRT model handles the conditional dependence within testlets by incorporating a random testlet effect parameter into the IRT model. The family of TRT models include the Rasch TRT model (Wang & Wilson, 2005), the two-parameter (Bradlow, Wainer, & Wang, 1999) and the three-parameter TRT model (Wainer, Bradlow, & Du, 2000) for dichotomous items, the TRT model for polytomous data and mixed-format data (Wainer & Wang, 2000; Wang, Bradlow, & Wainer, 2002), and the TRT model that allows the inclusion of covariate information (Wainer, Bradlow, & Wang, 2007). These testlet models have shown to effectively account for item dependence. Similarly, models like the bifactor model (Gibbons & Hedeker, 1992), and modified testlet models (Li, Bolt, & Fu, 2006), can also be used to model item dependence structure.

Alternatively, Jiao, Wang, and Kamata (2005) proposed a hierarchical generalized linear model (HGLM) that can account for LID from the multilevel modeling perspective. According to Kamata (2001), the Rasch model can be conceptualized as a two-level HGLM in which item responses are nested within persons. Based on this, a three-level HGLM can be formulated to model the clustering of items within item clusters (testlets), where level-one is the item-level model, level-two is the item clusters model, and level-three is the person-level model (Jiao, Wang, & Kamata, 2005). This model has been proved to be mathematically equivalent to the Rasch testlet model (Jiao, Wang, & Kamata, 2005).

Similarly, the person clustering effects can also be modelled within the HGLM framework. A three-level HGLM has been formulated to model item responses nested within persons, which are further nested within groups (Kamata, 2001). Person-level and group-level covariates, as well as their interactions, can be added into the level-two and level-three models to investigate the effects of individual and group characteristics on test scores. In addition to the HGLM approach, researchers have proposed some other forms of multilevel IRT models to account for person dependence (e.g., Adams, Wilson, & Wu, 1997; Fox & Glas, 2001; Mislevy, 1987; Mislevy & Bock, 1989).

The HGLM approach has also been extended to account for dual local dependence. Jiao et al. (2012) proposed a four-level HGLM as a solution to the violation of both local item independence and local person independence assumption. To simultaneously model item and person clustering effects, a fourth level can be added to the three-level HGLM for LID to model the group effects. Jiao et al. (2013) later extended this model to a five-level HGLM to account for LID and multiple levels of person clustering effects.

When LID is present, the equating methods for the tests which are made up of independent items are no longer appropriate. Though developing testlet-based tests has become a new trend of test development (Zhang, 2010), equating testlet-based tests is a relatively new area of interest. Due to the additional random component in the TRT model, transformation of the scales across the forms becomes more complex. However, there are a few studies developing linking separate calibration methods for tests composed of testlets (Li, Bolt, & Fu, 2005; Li, 2009; Zhang, 2010). These methods are developed based on the TRT model and aim to minimize the difference in Test Characteristic Curves (TCCs) between the base and the target form. In addition, the concurrent calibration method has also been applied in equating testlet-based tests (e.g., Zhang, 2010). These studies have shown that equating methods based on the TRT model can provide more accurate equating results than those based on the traditional IRT model. However, few studies have been done to equate tests when data exhibit person dependence.

1.2 PURPOSE OF THE STUDY

The purpose of this research is to assess the performance of multilevel models as concurrent calibration models on test equating in situations where test data exhibit local dependence. To the knowledge of the author, only two relevant studies have directly applied the multilevel modelling approach on test equating (Chu & Kamata, 2000; Turhan, 2006), whereas the focus is on equating/linking tests composed of independent items. Because LID can be modeled under the multilevel framework, the current study aims to develop multilevel-based concurrent equating approach for testlet-based tests. In addition, the TRT-based equating approach cannot account for dependence between persons. The flexibility of multilevel modelling approach, however, allows equating procedures to account for item and person dependence structures simultaneously. The current study thus also aims to address such paucity in equating with dual local dependence.

In order to accomplish these goals, two simulation studies are proposed. The first simulation study considers data containing responses to testlet items collected from independent examinees. We propose to use HGLM to concurrently equate two test forms and compare this approach to concurrent calibration based on the Rasch model and the GRM. Compared to the TRT-based linking separate calibration approach, equating methods under the multilevel framework save the need of computing transformation coefficients and can be easily implemented in commercially available software such as SAS and HLM. The second simulation study examines equating procedures for testlet data collected from examinees nested within groups. We propose to use a multilevel TRT model for concurrent equating with the presence of dual local dependence, and compare this approach to concurrent calibration based on the traditional IRT model and the TRT model. Factors that are expected to affect equating results were also investigated, such as degree of local dependence, sample size, ability distribution, and number of common testlets. It was expected that the results from these studies could provide evidence as a reference for researchers interested in applying multilevel models to test equating, especially in situations where local dependence is present.

1.3 RESEARCH QUESTIONS

The research questions are as follows:

1. How well does the concurrent equating method based on the multilevel model recover model parameters when LID is present?
    a. Does the equating method based on the multilevel model perform better than the concurrent calibration method based on the dichotomous IRT model, with the presence of LID?
    b. Does the equating method based on the multilevel model perform better than the concurrent calibration method based on the polytomous IRT model, with the presence of LID?
2. How well does the concurrent equating method based on the multilevel model recover model parameters when dual local dependence is present?
    a. Does the equating method based on the multilevel model perform better than that based on the traditional IRT model when dual local dependence is present?
    b. Does the equating method based on the multilevel model perform better than that based on the TRT model when dual local dependence is present?
3. What is the impact of LID on equating results for each of the equating methods?

4. What is the impact of person dependence on equating results for each of the equating methods?
5. What are the effects of ability distribution and number of common testlets on equating results?

2.0 LITERATURE REVIEW

The purpose of this chapter is to review (1) IRT models, (2) local independence assumption of IRT models, (3) TRT models, (4) multilevel IRT models, (5) basic concepts of equating, (6) IRT equating, and (7) IRT equating with local dependence.

2.1 IRT MODELS

IRT models estimate persons' latent trait levels based on their responses to test items. The probability of a correct answer to an item is modeled mathematically as a function of item characteristics (e.g., difficulty and discrimination) and a person's ability level. There is a large family of IRT models and the primary difference among them is the number of parameters they use to describe the characteristics of test items. For example, the three-parameter logistic (3PL) model (Birnbaum, 1968) assumes that test items discriminate between high and low performers differently. In addition, it also allows the chance of pseudo-guessing to be estimated. The probability of person '*j*' getting an item '*i*' correct $p(X_{ij} = 1)$, given the *j*th examinee's ability level (θ_j) is given by:

$$p\{X_{ij} = 1 \mid \theta_j, a_i, b_i, c_i\} = c_i + (1 - c_i)\frac{\exp[a_i(\theta_j - b_i)]}{1 + \exp[a_i(\theta_j - b_i)]}, \tag{2.1}$$

where a_i, b_i, and c_i are the discrimination parameter, difficulty parameter, and guessing parameter for the i^{th} item, respectively. The item discrimination parameter (a_i) characterizes the slope of the item characteristic curve of the i^{th} item where it reaches the maximum. The item difficulty parameter (b_i) corresponds to the point at which the slope is maximized. The guessing parameter (c_i) is also called the lower asymptote or pseudo-chance level parameter. When $c = 0$, the item difficulty parameter corresponds to the trait level at which the probability of a correct answer is 0.5. When $c \neq 0$, the item difficulty value is halfway between c and 1. The ability level (θ_j) is typically assumed to follow a standard normal distribution.

2.1.1 The assumption of local independence

IRT models are widely used in many testing applications, such as test equating and scaling, item banking, test development, and adaptive testing (Kolen & Brennan, 2004). The effectiveness of IRT model applications depends on how well test data meet the model assumptions. IRT entails several assumptions and local independence is one of them. This assumption implies two facets of independence: local item independence and local person independence (Reckase, 2009; also see Jiao et al., 2012). The assumption of local independence is violated if either LID or LPD is present (Jiao et al., 2012).

2.1.1.1 Local item independence

Local item independence assumes that a person's response to an item in a test is independent to the person's response to another item in the test. It can be expressed mathematically as:

$$p(X = x \mid \theta) = \prod_{i}^{I} p(X_i = x \mid \theta), \qquad (2.2)$$

where *I* is the total number of items. This equation describes that the probability of a pattern of responses to all items in the test, conditioned on the latent trait (θ), is the product of the conditional probability of the response to each item. This equation shows that once the latent trait level has been taken into account, item responses are completely independent. It defines a strong form of local item independence.

A weak form of local item independence has been proposed by McDonald (1997). It defines that conditional on the latent trait, the pairwise covariances among test items are zero as test length approaches infinity (McDonald, 1997). When this assumption holds, the joint probability of responses to a pair of items, conditional on the latent abilities, is the product of the conditional probabilities of responses to the two items. It is mathematically expressed as

$$p\{X_i = x_i, X_{i'} = x_{i'} \mid \theta\} = p(X_i = x_i \mid \theta) p(X_{i'} = x_{i'} \mid \theta), \qquad (2.3)$$

where X_i and $X_{i'}$ are the responses for item *i* and item *i'*. This is a weaker form of local item independence because higher-order dependencies among items are allowed.

There are a variety of factors that may cause LID. These include external assistance or interference, speededness, fatigue, practice effects, item format, passage dependence, item chaining, explanation of previous answers, scoring rubrics, content areas, etc. (Yen, 1993). Among all these factors, the passage dependence is the most commonly studied factor in educational assessments. It references to the contextual effects on a group of items constructed around a common stimulus (passage).

A group of items constructed around a common stimulus (passage) are termed as an item bundle (Rosenbaum, 1988) or testlet (Wainer & Kiely, 1987). According to Wainer and Kiely, a testlet is “a group of items related to a single content area that is developed as a unit and contains a fixed number of predetermined paths that an examinee may follow” (1987, p. 190). Thus, if a particular testlet is administered to a group of examinees, the items within the testlet should be presented in the same order.

Testlets are commonly used in standardized educational tests. For example, in a reading comprehensive test, a reading passage is often used as the stem for a set of items. Similarly, a graph or a table in a science test is often used as the focus of a set of items (Wainer, Bradlow, & Wang, 2007). Because items within a testlet are connected by the common context, responses to these items tend to be conditionally dependent. As a result, the assumption of local item independence is violated.

2.1.1.2 Local person independence

Local person independence assumes that subjects are independent to each other. Mathematically, it can be expressed as

$$p(X_i = x_i \mid \theta) = \prod_{j=1}^{n} p(x_{ij} \mid \theta_j) = p(x_{i1} \mid \theta_1) p(x_{i2} \mid \theta_2) \cdots p(x_{in} \mid \theta_n), \qquad (2.4)$$

where n is the total number of persons. This equation describes that the probability of a set of responses to the i^{th} item in the test by n persons, conditional on the vector of abilities, is the product of the conditional probability of each individual person’s response to that item (Jiao et al., 2012).

Factors that may cause LPD include cluster sampling, external assistance or interference, differential opportunity to learn, and different problem-solving strategies, etc. (Jiao et al., 2012). In educational research, data are usually collected in hierarchical settings, such as students nested within classrooms, and so on. Therefore, these data would have a hierarchical structure in nature

(Jiao et al., 2010). Studies have shown that the intra-class correlation coefficient (ICC) of many achievement test data usually ranges from 0.12 to 0.49, indicating that some degree of person dependence does exist in these test data (Schochet 2005; Wang, 2006; see Jiao et al., 2010).

2.1.2 Violation of local item independence

It has been shown that ignoring LID and fitting a standard IRT model tends to result in bias in item parameter estimates, and overstatement of test information, test reliability, and precision of measures (e.g., Sireci, Thissen, & Wainer, 1991; Yen, 1993; Wainer, 1995; Wainer & Thissen, 1996; Bradlow, Wainer, & Wang, 1999; Wainer, Bradlow, & Du, 2000; Wainer & Wang, 2000). For example, Bradlow, Wainer, and Wang (1999) found that ignoring LID and fitting the traditional 2PL IRT model led to underestimation of the discrimination parameters and bias in difficulty parameter estimates. Wainer, Bradlow, and Du (2000) found that when LID existed and the traditional 3PL IRT model was fit to the data, estimation of the difficulty and person parameters were well done, whereas the discrimination and guessing parameters were better estimated using the 3PL TRT model. Bradlow, Wainer, and Wang (1999) and Wainer, Bradlow, and Du (2000) also found that ignoring LID led to overestimation of precision of both item and person parameter estimates. Subsequently, these effects may affect applications of IRT models such as test equating (Yen, 1993).

2.1.3 Violation of local person independence

LPD cannot be easily accommodated using traditional IRT models since the assumption of independence of subjects is violated (Kreft & de Leeuw, 1998). The person dependence within clusters would reduce the effective sample size, and consequently lead to biased parameter estimates and decreased measurement precision (Cochrane, 1977; Cyr & Davies, 2005; Kish, 1965; Jiao et al., 2012). It has been shown that ignoring person dependence and fitting a standard IRT model can lead to decreased accuracies of model parameter estimation, increased classification errors (Jiao et al., 2012), and underestimation of ability variance (Jiao, Wang, & Kamata, 2005).

2.2 MODELS FOR TESTLET-BASED TESTS

Different methods have been proposed to address the issue of LID. The first approach is to fit the data with a polytomous IRT model and use testlet scores. The second approach involves using a model that explicitly accounts for LID effects such as the bifactor model or the TRT model. The third approach models item-clustering effects by building multilevel models. These models will be discussed next.

2.2.1 Polytomous IRT model

One way to handle the within-testlet dependence is to treat the entire testlet as a single polytomous item instead of multiple independent dichotomous items. In this approach, number-correct scores are calculated for each testlet, and then calibrated using a polytomous item response model. The family of polytomous IRT models include the GRM (Samejima, 1969), the Partial Credit Model (Masters, 1982), the Rating Scale Model (Andrich, 1978), and the Nominal Response Model (Bock, 1972). By doing so, the assumption of local item independence is met because testlet-level scores are used for calibration and they are locally independent. This approach may be appropriate when the magnitude of LID is moderate and the proportion of independent items within a test is large (Wang & Wilson, 2005). However, because number-correct scores are used, the information in the response patterns within a testlet is missing (Wainer, Bradlow, & Wang, 2007). In situations when item parameter estimation is required, the polytomous model approach is inappropriate. In addition, this approach cannot be applied to adaptive testing (Wang & Wilson, 2005).

2.2.2 Bifactor model

The bifactor model is a constrained multidimensional model with a primary dimension and multiple secondary dimensions. Each item has a nozero factor loading on the primary dimension and one of the secondary dimensions. All these dimensions are orthogonal to each other (Gibbons & Hedeker, 1992). When applied to testlet-based tests, the testlet effects are considered as secondary traits. For a test with *K* testlets, each item response reflects one's trait level on the primary trait dimension and one of the *K* testlet dimensions. The multidimensional extension of the 3PL can be expressed as:

$$p_i(\theta) = c_i + (1 - c_i)\frac{\exp(a_i'\theta + d_i)}{1 + \exp(a_i'\theta + d_i)}\ , \qquad (2.5)$$

where $p_i(\theta)$ is the probability of answering item *i* correctly given the vector of trait θ, a_i' is a vector of discrimination parameters, d_i is the difficulty parameter, and c_i is the lower asymptote. For the bifactor model, each item has a nonzero *a*-value on the primary θ dimension and one of the secondary θ dimensions, respectively. The correlations between trait dimensions are constrained to be zero.

DeMars (2006) compared the performance of the bifactor model to the TRT model, the polytomous IRT model, and the traditional IRT model using simulated datasets. The results indicated that when the dataset was generated under the TRT model, the bifactor model provided equal or even smaller bias and RMSEs than the TRT model. This suggests that the bifactor model can be a good alternative to the TRT model.

2.2.3 Testlet Response Theory model

The TRT model is proposed explicitly as a solution to the violation of the local item independence assumption. It is a special case of the bifactor model when the factor loadings on the specific testlet dimension are proportional to the loadings to the primary dimension (Rijmen, 2010). The TRT model handles the within-testlet dependence by incorporating a random testlet effect parameter into the IRT model. Bradlow, Wainer, and Wang (1999) first developed the two-parameter TRT model for dichotomous items. Wainer, Bradlow, and Du (2000) later extended the two-parameter TRT model into a three-parameter TRT model that allows for guessing and variation in the local dependence across testlets. The TRT model has also been extended to handle tests composed of polytomous data and mixed-format data (Wainer & Wang, 2000; Wang, Bradlow, & Wainer, 2002) and to allow the inclusion of covariate information (Wainer, Bradlow, & Wang, 2007). In this section, only the two-parameter and three-parameter TRT model will be introduced.

Let Y_{ij} be the response of examinee *j* on item *i*. The probability of a correct answer of examinee *j* on item *i* is:

$$p(Y_{ij}=1)=\frac{\exp[a_i(\theta_j-b_i-\gamma_{jd(i)})]}{1+\exp[a_j(\theta_j-b_i-\gamma_{jd(i)})]}, \qquad (2.6)$$

where θ_j, a_i, and b_i are the person trait, item discrimination, and item difficulty parameters, respectively. Compared to the standard two-parameter IRT model, this model adds a random effect parameter, $\gamma_{jd(i)}$, which models the person-specific testlet effect, or the interaction of person *j* with testlet *d*. Within a testlet, the value of $\gamma_{jd(i)}$ is constant for a person but varies across persons. It is assumed to be independent from the θ distribution and follow a normal distribution, $\gamma_{jd(i)} \sim N(0, \sigma_\gamma^2)$. By definition, the sum of $\gamma_{jd(i)}$ over examinees within the same testlet is zero (i.e., $\sum_j \gamma_{jd(i)} = 0$). If $\gamma_{jd(i)} = 0$, there is no testlet effect and the item fits the standard two-parameter model. The testlet variance, σ_γ^2, is assumed to be constant across testlets. It denotes the magnitude of local dependence. If $\sigma_\gamma^2 = 0$, there is no extra dependence among items within the same testlet and the local item independence assumption is met, while the larger the variance, the larger the amount of LID within that testlet.

Wainer, Bradlow, and Wang (2007) pointed out that the two-parameter TRT model had the following limitations (pp. 130-131). First, it does not model the chance of guessing, and therefore its application to multiple-choice items is limited. Second, the variance of the testlet effect is assumed to be constant across testlets. This is unlikely to hold for real test data, as different testlets may exhibit different amounts of LID. Third, this model considers only binary response data. Last, the model does not allow for modeling covariate effects.

Wainer, Bradlow, and Du (2000) extended this two-parameter TRT model into a three-parameter TRT model that allows for guessing and variation in the local dependence across testlets.

$$p(Y_{ij} = 1) = c_i + (1 - c_i)\frac{\exp[a_i(\theta_j - b_i - \gamma_{jd(i)})]}{1 + \exp[a_i(\theta_j - b_i - \gamma_{jd(i)})]}, \tag{2.7}$$

where as before with the two-parameter TRT model, θ_j , a_i , and b_i are the person trait, item discrimination, and item difficulty parameters, respectively, c_i is the guessing parameter, and $\gamma_{idj} \sim N\ (0, \sigma^2_{d(i)})$. The two-parameter TRT model is a special case of the three-parameter TRT model with $c_i = 0$. The testlet variance, $\sigma^2_{d(i)}$, is testlet-specific and allowed to vary across testlets.

According to Wainer and Wang (2000), the interpretation of the *a*, *b*, and *c* parameters from the three-parameter TRT models are the same as those from the standard three-parameter IRT model (p. 206). Ip (2010) pointed out that because of the way that TRT models specify the random testlet effect, the meaning of item parameters and item information functions also changes. For example, the *a*-parameter in the traditional 3PL IRT model characterizes the slope of the item response function (IRF) at the location where the slope reaches its maximum at $\theta = b$. In contrast, in the TRT model, the meaning of the *a*-parameter is conditional on the value of the person-specific testlet effect. If the testlet effect is zero, the *a*-parameter in the three-parameter TRT model characterizes the slope of the IRF at the location where $\theta = b$. However, if the testlet effect is nonzero, the slope of the IRF does not reach its maximum at the point where $\theta = b$. Accordingly, any direct comparison of the *a*-parameter in the testlet model with the *a*-parameter in the traditional IRT model is inappropriate.

2.2.4 Alternative model formulations for testlets

Li, Bolt, and Fu (2006) proposed three alternative models that account for the testlet effect. These models differ in their assumptions regarding how testlets influence item performance. The first model, denoted as the general model, follows a multidimensional IRT approach and treats the testlet effects as a secondary ability dimension. The two-parameter normal ogive (2PNO) version of this model can be written as:

$$p(y_{ij}=1)=\Phi(a_{i1}\theta_j - b_i + a_{i2}\gamma_{jd(i)}) \tag{2.11}$$

where a_{i1} is the discrimination parameter with respect to the ability dimension for item *i*, a_{i2} is the discrimination parameter with respect to the testlet dimension for item *i*, and b_i is the item difficulty parameter. The inclusion of the parameter a_{i2} allows for modeling item-specific testlet effect and thus makes no assumption that the testlet effect is constant across all items within the same testlet. This model is essentially the same as the bifactor model. The 2PNO testlet model is a special case of this general model when $a_{i2}=a_{i1}C_d$, where C_d is a constant for testlet *d*.

The second model imposes a constraint on the discrimination power for the testlet dimension such that there is an inverse relationship between the θ primary dimension and the testlet dimension regarding the item’s discrimination power. The basic idea is that if an item has a high discrimination power on the primary dimension, it tends to have a low discrimination power on the secondary dimension. The model is given by:

$$p(y_{ij}=1)=\Phi(a_{i1}\theta_i - b_i + \sqrt{MDISC^2 - a_{i1}^2}\gamma_{jd(i)}), \tag{2.12}$$

where *MDISC* is constant across all items, denoted as the multidimensional discrimination parameter. The third model assumes constant item discrimination power for the testlet dimension, which can be expressed as

$$p(y_{ij}=1)=\Phi(a_{i1}\theta_i - b_i + \gamma_{jd(i)}). \tag{2.13}$$

Li, Bolt, and Fu (2006) adopted four Bayesian model comparison criteria to compare the above three alternative models to the 2PNO testlet model using real data. Results indicated that the general model provided the best fit, followed by the 2PNO testlet model. Hence, the general model can be a very attractive candidate to model testlet-based tests.

2.3 MULTILEVEL IRT MODELS FOR LOCAL DEPENDENCE

2.3.1 Multilevel IRT models

A multilevel IRT model combines IRT and a multilevel model, taking into account both within- and between-group (such as schools) variance of the data. Accordingly, it allows estimation of latent traits at different levels (e.g., students, classrooms, schools.). Also, it offers the opportunity to model the effect of individual- and group-level covariates, as well as the cross-level interactions on the latent traits. This is very useful for school effectiveness research, which is interested in the relationship between explanatory variables and outcome measures (Fox, 2005). By simultaneously estimating all model parameters, multilevel IRT modeling can yield better estimation of the relationships between IRT latent traits and explanatory variables than the traditional two-step procedure, because the measurement errors of the latent traits are incorporated into the total variance of the model (Maier, 2001). Moreover, the flexibility of multilevel modeling approach allows one to handle latent explanatory variables, model latent individual growth, and identify clusters of respondents (Fox, 2005). Recently, the multilevel IRT approach has also been used to model LID (Jiao, Wang, & Kamata, 2005) and dual local dependence due to person and item clustering (Jiao et. al, 2012; Jiao & Zhang, 2014).

Multilevel IRT models have been formulated in a number of ways. For example, the Rasch model has been formulated as a hierarchical nonlinear model (Adams, Wilson, & Wu, 1997; Raudenbush & Sampson, 1999) or a hierarchical generalized linear model (HGLM) (Kamata, 1998, 2001). The HGLM approach has been extended to model group-level covariates (Kamata, 2001), and to handle polytomous outcomes (Beretvas & Williams, 2004) and 2PL dichotomous items (Turhan, 2006). Some other researchers have developed multilevel IRT models as a combination of IRT and a two-level hierarchical linear model (HLM) to accommodate the nested structure in the latent trait variable (e.g., Fox & Glas, 2001; Fox, 2005; Maier, 2001). In these models, the latent trait variable in the IRT model is treated as the outcome variable in the level-one model of the HLM. The current study will focus on multilevel IRT as HGLM, because it has been shown to easily accommodate dual local dependence due to person and item clustering.

2.3.2 Two-level HGLM as the Rasch model

An IRT model can be conceptualized as a multilevel model in which item responses are nested within persons. Kamata (1998, 2001) first proved the algebraic equivalence of the two-level HGLM model and the Rasch model. Under the two-level model formulation of the Rasch model, person ability parameters are treated as random and item parameters are considered as fixed (Kamata, 2001). This enables the decomposition of person parameters into a linear combination of fixed and random effects, which make them analogous to a multilevel linear model with mixed effects (Kamata, 2001).

In the two-level HGLM, the first level is the item-level model and can be specified under the GLM framework. The second level is the person-level model and is formulated under the HLM framework. At level-one, a logistic regression model is used to model the log odds of giving a correct response. Let p_{ij} be the probability that person *j* answers item *i* correctly, the item-level model is given by:

$$\begin{aligned}\log\left(\frac{p_{ij}}{1-p_{ij}}\right) &= \eta_{ij} \\ &= \beta_{0j} + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij} + \ldots.. + \beta_{(I-1)}X_{(I-1)ij}\ , \\ &= \beta_{0j} + \sum_{q=1}^{I-1}\beta_{qj}X_{qij}\end{aligned} \tag{2.14}$$

Where X_{qij} is the q^{th} dummy variable for person *i*, and X_{qij} =1 when *q*=*i* and 0 when *q*≠*i*, *I* is the number of items in the test, β_{0j} is an intercept term, and β_{qj} is the coefficient associated with X_{qij} when $q = 1,\ldots., I-1$. To achieve full rank for the design matrix, the dummy variable for the I^{th} item is dropped, thereby resulting in $I-1$ dummy variables. This dropped dummy variable is treated as the reference item, and therefore β_{0j} is interpreted as the item effect for the dropped item for person *j*, and β_{qj} is the effect of the q^{th} item compared to the reference item. The probability that person *j* answers item *i* correctly is:

$$p_{ij} = \frac{1}{1+\exp[-\eta_{ij}]}\,. \tag{2.15}$$

The level-two (person-level) model assumes that the level-one intercept β_{0j} has a random effect across persons, while the effects of level-one slopes β_{qj} are constant across persons. The level-two model for person *j* is given by,

$$\begin{cases} \beta_{oj} = \gamma_{00} + u_{0j} \\ \beta_{1j} = \gamma_{10} \\ \beta_{2j} = \gamma_{20} \\ \quad\vdots \\ \beta_{(I-1)j} = \gamma_{(I-1)0} \end{cases}, \tag{2.16}$$

where r_{00} is the effect of the reference item, and γ_{q0} is the effect of the i^{th} item (when *i*=*q*). The random component u_{0j} indicates the deviation of the ability of person *j* from the mean of u_{0j}, and $u_{0j} \sim N(0, \tau_{\beta})$.

Combining the level-one model with level-two model, the probability that person *j* gets item *i* correct is:

$$p_{ij} = \frac{1}{1 + \exp\{-[\mu_{0j} - (-\gamma_{q0} - \gamma_{00})]\}}, \tag{2.17}$$

when *i* = *q*, and $\eta_{ij} = \gamma_{00} + u_{0j} + \gamma_{q0}$. Kamata (2001) showed that this equation is algebraically equivalent to the Rasch model,

$$p_{ij} = \frac{1}{1 + \exp[-(\theta_j - \delta_i)]}, \tag{2.18}$$

where $\theta_j = u_{0j}$, and $\delta_i = -\gamma_{q0} - \gamma_{00}$. Both δ_i and $-\gamma_{q0} - \gamma_{00}$ are treated as fixed in the two formulations. In the Rasch model, the person trait parameters can be considered as either fixed or random. While in Kamata's model, u_{0j} is treated as random.

If the interest is on the effects of person characteristics on examinee responses, person-level predictors can be added to the level-two model. In situations where the item effects are considered to be fixed across persons (e.g., no differential item functioning (DIF)), person-level predictors are only included in the intercept equation. The level-two model can be written as:

$$\begin{cases} \beta_{oj} = \gamma_{00} + \gamma_{01} W_{1j} + \cdots + \gamma_{0p} W_{pj} + u_{0j} \\ \beta_{1j} = \gamma_{10} \\ \beta_{2j} = \gamma_{20} \\ \quad \vdots \\ \beta_{(I-1)j} = \gamma_{(I-1)0} \end{cases}, \tag{2.19}$$

where $W_{sj}(s=1,\cdots,p)$ are the person-level predictors and γ_{0p} are the associated coefficients. The combined model can be expressed as:

$$p_{ij} = \frac{1}{1+\exp[-\{[\gamma_{01}W_{1j}+\cdots+\gamma_{0p}W_{pj}+u_{0j}]-(-\gamma_{q0}-\gamma_{00})\}]}. \tag{2.20}$$

The overall ability for person j now becomes $\gamma_{01}W_{1j}+\cdots+\gamma_{0p}W_{pj}+u_{0j}$. The item difficulty is still denoted by $-\gamma_{q0}-\gamma_{00}$.

2.3.3 Three-level HGLM for local person dependence

The two-level HGLM can be easily extended to a three-level model that simultaneously takes into account person clustering effects. Let p_{ijg} be the probability that person *j* answers item *i* in school *g* correctly, the item-level model is given by:

$$\begin{aligned}\log\left(\frac{p_{ijg}}{1-p_{ijg}}\right) &= \eta_{ijg} \\ &= \beta_{0jg}+\beta_{1jg}X_{1ijg}+\beta_{2jg}X_{2ijg}+.....+\beta_{(I-1)jg}X_{(I-1)ijg} \\ &= \beta_{0jg}+\sum_{q=1}^{I-1}\beta_{qjg}X_{qjg}\end{aligned}. \tag{2.21}$$

The level-two model for person *j* in school *g* is given by,

$$\begin{cases} \beta_{ojg} = \gamma_{00g}+u_{0jg} \\ \beta_{1jg} = \gamma_{10g} \\ \beta_{2jg} = \gamma_{20g} \\ \vdots \\ \beta_{(I-1)jg} = \gamma_{(I-1)0g} \end{cases}, \tag{2.22}$$

where γ_{00g} is the effect of the reference item, and γ_{q0g} is the effect of the i^{th} item (when i=q). The random component u_{0jg} indicates the deviation of the ability of person j in school g from the mean of u_{0jg} within school g, and $u_{0jg} \sim N(0, \tau_{\beta})$.

The level-three (school-level) model can be added to estimate group-level abilities and the effect of group-level covariates on the latent trait. The model can be expressed as:

$$\begin{cases} \gamma_{00g} = \pi_{000} + r_{00g} \\ \gamma_{10g} = \pi_{100} \\ \gamma_{20g} = \pi_{200} \\ \quad\vdots \\ \gamma_{(I-1)0g} = \pi_{(I-1)00} \end{cases}, \tag{2.23}$$

where r_{00g} can be interpreted as the average ability of students in school g, and $r_{00g} \sim N(0, \tau_{\gamma})$. When the three-level model is combined, the probability that person j in school g answers item i correctly is given by

$$p_{ijm} = \frac{1}{1 + \exp\{-[(r_{00g} + u_{0jg}) - (-\pi_{q00} - \pi_{000})]\}}, \tag{2.24}$$

where $-\pi_{q00} - \pi_{000}$ gives the item difficulties for items $i = q$ (i=1,..., k-1) and π_{000} is the item difficulty for item I. The term $r_{00g} + u_{0jg}$ denotes the overall ability for person j in school m. Collateral variables can be added into the level-two and level-three models to explain variation in ability across persons and groups.

The three-level IRT model has been used in applied research. Kamata (2001) fitted a three-level HGLM to a sample from the Third International Mathematics and Science Studies (TIMSS) to investigate the effect of a school-level predictor (teacher’s experience) and a student-level predictor (studying science at home), as well as their interaction effect on students’ science literacy test scores. Pastor (2003) illustrated how to apply the three-level HGLM to obtain estimates for item parameters as well as person- and site-level latent trait estimates using the HLM software.

2.3.4 Three-level HGLM for local item dependence

As described earlier, a three-level HGLM can be used to model the clustering of persons within groups, with items (level-one) nested within persons (level-two), which are further nested within groups (level-three). Similarly, a three-level HGLM can be formulated to model the clustering of items within item clusters (testlets), where level-one is the item-level model, level-two is the item clusters model, and level-three is the person-level model (Jiao, Wang, & Kamata, 2005). Let p_{imj} be the probability that person *j* answers item *i* in item cluster *m* correctly, the level-one model can be expressed as:

$$\log(\frac{p_{imj}}{1-p_{imj}}) = \eta_{imj} = \beta_{0mj} + \sum_{q=1}^{I-1} \beta_{qmj} X_{qimj}, \tag{2.25}$$

where X_{qimj} is the q^{th} dummy variable for person *j*, and X_{qimj} =1 when *q*=*i* and 0 when *q*≠*i*, β_{0mj} is the intercept, and β_{qmj} is the coefficient associated with X_{qimj} where $q = 1,....,I-1$. To achieve full rank for the design matrix, one of the dummy variables in the equation is dropped, thereby resulting in $I-1$dummy variables. This dropped dummy variable is treated as the reference item, and therefore β_{0mj} is interpreted as the item effect for the dropped item in item cluster *m* for person *j*. The individual item effect β_{qmj} is the difference of the item with the q^{th} dummy variable from the effect of the reference item.

The item cluster-level (level-two) model is given by

$$\begin{cases} \beta_{omj} = \gamma_{00j} + u_{0mj} \\ \beta_{1mj} = \gamma_{10j} \\ \beta_{2mj} = \gamma_{20j} \\ \quad\vdots \\ \beta_{qmj} = \gamma_{q0j} \end{cases}, \tag{2.26}$$

where γ_{00j} is the fixed effect of the level-one intercept, γ_{q0j} is the item-specific effect for item with the q^{th} dummy variable (when i=q), and u_{0mj} is the random effect of the level-one intercept. The random effect u_{0mj} is assumed to follow a normal distribution with mean of zero and variance of σ_u^2. It can be considered as the interaction between latent trait and item cluster, which is analogous to the person-specific testlet effect in Bradlow, Wainer, and Wang's (1999) formulation of the Rasch testlet model. In both models, the variance of the interaction effect σ_u^2 indicates the magnitude of LID and is assumed to be constant across item clusters.

The person-level (level-three) model is given by

$$\begin{cases} \gamma_{00j} = \pi_{000} + r_{00j} \\ \gamma_{10m} = \pi_{100} \\ \gamma_{20m} = \pi_{200} \\ \quad\vdots \\ \gamma_{q0m} = \pi_{q00} \end{cases}, \tag{2.27}$$

where r_{00j} is the person ability and $r_{00j} \sim N(0, \sigma_r^2)$. In this model, the items are assumed having only fixed effects, while the person effect is considered to be random.

The combined three-level model for LID can be expressed as

$$p_{ijm} = \frac{1}{1 + \exp\{-[r_{00j} - (-\pi_{q00} - \pi_{000}) + u_{0mj}]\}}, \tag{2.28}$$

where r_{00j} indicates the ability for person j, and the term $-\pi_{q00} - \pi_{000}$ denotes the difficulty level of item q. In the framework of TRT, this model can be rewritten as:

$$p_{jdig} = \frac{1}{1 + \exp\{-[\theta_j - b_i + \gamma_{jd(i)})]\}}, \tag{2.29}$$

where $\theta_j = r_{00j}$, $b_i = -\pi_{q00} - \pi_{000}$, and $\gamma_{jd(i)} = u_{0mj}$. Thus, the three-level HGLM for LID is equivalent to the Rasch testlet model (Wang & Wilson, 2005).

Based on a simulation study, Jiao et al. (2005) demonstrated that when LID was present, the three-level HGLM could capture the magnitude of LID and provide more accurate estimates for the item difficulty parameters than the two-level HGLM ignoring LID. In addition, the ability variance would be underestimated if LID was ignored.

2.3.5 Four-level HLM for dual local dependence

Dual local dependence occurs when test data are composed of testlets and collected using a cluster sampling method (Jiao et al., 2010). Several recent studies have addressed such dual local dependence. For example, Jiao et al. (2012) proposed a four-level HGLM (or multilevel Rasch testlet model) that can simultaneously account for LID and local person dependence. Jiao et al. (2013) later extended this model to a five-level HGLM to account for LID and multiple levels of person clustering effects. Jiao and Zhang (2014) proposed a polytomous version of the multilevel Rasch testlet model for dual local dependence.

Jiao et al. (2012) proposed a four-level HGLM as a solution to the violation of both local item and person independence assumption. As described earlier, a three-level HGLM can be used to model the clustering of persons within groups, with items (level-one) nested within persons (level-two), which are further nested within groups (level-three). Similarly, a three-level HGLM can be formulated to model the clustering of items within item clusters (testlets), where level-one is the item-level model, level-two is the item clusters model, and level-three is the person-level model. To simultaneously model both item and person clustering, a fourth level can be added to model the group effects, resulting in a four-level HGLM.

Let p_{imjg} be the probability that person *j* in group *g* answers item *i* in item cluster *m* correctly, the level-one model can be expressed as:

$$\log(\frac{p_{imjg}}{1-p_{imjg}}) = \eta_{imjg} = \beta_{0jmg} + \sum_{q=1}^{I-1} \beta_{qmjg} X_{qimjg}, \tag{2.30}$$

where X_{qimjg} is the q^{th} dummy variable for person *j* in group *g*, and $X_{qimjg} = 1$ when *q*=*i* and 0 when *q*≠*i*, β_{0mjg} is the intercept, and β_{qmjg} is the coefficient associated with X_{qimjg} where $q = 1,....,I-1$. To achieve full rank for the design matrix, one of the dummy variables in the equation is dropped, thereby resulting in $I-1$ dummy variables. This dropped dummy variable is treated as the reference item, and therefore β_{0mjg} is interpreted as the item effect for the dropped item in item cluster *m* for person *j*. The individual item effect β_{qmjg} is the difference of the item with the q^{th} dummy variable from the effect of the reference item.

The item cluster-level (level-two) model is given by

$$\begin{cases} \beta_{omjg} = \gamma_{00jg} + u_{0mjg} \\ \beta_{1mjg} = \gamma_{10jg} \\ \beta_{2mjg} = \gamma_{20jg} \\ \quad\vdots \\ \beta_{qmjg} = \gamma_{q0jg} \end{cases}, \tag{2.31}$$

where γ_{00jg} is the fixed effect of the level-one intercept, γ_{q0jg} is the item-specific effect for item with the q^{th} dummy variable (when *i*=*q*), and u_{0mjg} is the random effect of the level-one intercept. The random effect u_{0mjg} is assumed to follow a normal distribution with mean of zero and variance of σ_u^2.

The person-level (level-three) model is given by

$$\begin{cases} \gamma_{00jg} = \pi_{000g} + r_{00jg} \\ \gamma_{10jg} = \pi_{100g} \\ \gamma_{20jg} = \pi_{200g} \\ \quad\vdots \\ \gamma_{q0jg} = \pi_{q00g} \end{cases}, \tag{2.32}$$

where r_{00jg} is the person ability and $r_{00jg} \sim N(0, \sigma^2_{rg})$. In this model, the items are assumed having only fixed effects, while the person effect is considered to be random.

If person clustering exists, the fourth level can be added to model the group effects. It is given by

$$\begin{cases} \pi_{000g} = \varphi_{0000} + r_{000g} \\ \pi_{100g} = \varphi_{1000} \\ \pi_{200g} = \varphi_{2000} \\ \quad\vdots \\ \pi_{q00g} = \varphi_{q000} \end{cases}, \tag{2.33}$$

where φ_{0000} is the effect of the reference item, φ_{q000} is the item-specific effect relative to the reference item. The group-specific ability r_{000g} follows a normal distribution with mean of zero and variance of σ^2_g. Again, the item effect is considered to be fixed across groups, while the average ability varies across groups.

The combined four-level model can be expressed as

$$p_{ijm} = \frac{1}{1 + \exp\{-[(w_{00jg} + r_{000g}) - (-\varphi_{0000} - \varphi_{q000}) + u_{0djg}]\}}, \tag{2.34}$$

where w_{00jg} is the person-specific ability for person *j* in group *g* and r_{000g} is the group-specific ability for group *g*. The item difficulty parameter (b_i) is decomposed into the reference item effect (φ_{0000}) and the item-specific effect (φ_{q000}). In the IRT framework, it can be written as

$$p_{jdig} = \frac{1}{1+\exp\{-[(\theta_j + \theta_g) - b_i + \gamma_{jd(i)})]\}}, \tag{2.35}$$

where $\theta_j + \theta_g = w_{00jg} + r_{000g}$, $b_i = -\varphi_{0000} - \varphi_{q000}$, and $\gamma_{jd(i)} = u_{0djg}$.

Using simulated data sets, Jiao et al. (2012) showed that when both item and person clustering were present, ignoring one or both clustering effects could reduce the estimation accuracy of the item difficulty and person ability parameters, as well as classification accuracy. They suggested applying this model to analyze test data from testlet-based assessment collected using a cluster sampling method.

2.4 EQUATING

2.4.1 Basic concepts of equating

For many standardized educational tests, it is a common practice to administer a test on multiple occasions so that examinees can have the flexibility in choosing a test date. By administering a test on different dates, researchers can also explore changes in student achievement over time (Kolen & Brennan, 2004). The use of the same test form on different occasions could cause test security problems, because examinees who are administered the test later may have the advantage of knowing some test items prior to testing from their previous examinations or from earlier examinees.

To ensure test security, most testing programs develop multiple forms of the same test so that some examinees will not be advantaged on the test by their prior knowledge. Although these test forms are constructed to be as similar as possible in content and statistical properties, they usually differ slightly in difficulty. As a result, scores from different test forms cannot be compared directly because the difference in test scores reflects not only the difference in examinees' trait levels but also the difficulty levels of the test form (Kolen & Brennan, 2004).

Test equating is a technique that takes into account differences in difficulty among different test forms. It refers to the statistical process of adjusting test scores on alternate forms of a test so that scores from the forms can be used interchangeably (Kolen & Brennan, 2004). Thus, in standardized testing programs when different forms of a test are used, equating is important and necessary.

2.4.2 IRT equating

Both CTT procedures (e.g., mean, linear, and equipercentile equating) and IRT procedures were developed for test equating. One major advantage of IRT procedures over CTT procedures is that the former treats the test item as the unit of analysis, whereas the latter usually focus on the entire test. IRT procedures are therefore more flexible and can be used in situations where CTT procedures typically are not used, such as equating to an item pool (Kolen & Brennan, 2004). Compared to CTT procedures, IRT procedures may also provide better equating results at the upper ends of score scales, and offer greater flexibility in choosing a plan for equating (Cook & Eignor, 1991). In addition, most standardized testing programs now use IRT models to calibrate an examinee's ability level. IRT procedures thus become a natural choice for test equating. The focus of this paper is on IRT equating. A recent review of IRT equating can be found in Kolen and Brennan (2004).

IRT equating procedures adjust item parameter estimates from different forms of a test so as to be on the same scale. It involves three steps: selecting an equating design, placing parameter estimates on a common scale, and equating test scores (Cook & Eignor, 1991). There are three commonly used data collection designs for equating: (a) single group design; (b) random groups design; and (c) common-items nonequivalent groups design (Kolen & Brennan, 2004) or Non-equivalent groups with Anchor Test (NEAT) design (von Davier, Holland, & Thayer, 2004). In the single group design, each examinee is administered two test forms. In the random groups design, examinees are randomly assigned one of the test forms. In the third design, different groups of examinees are administered different test forms, which share a set of common test items. Because the set of common items is also termed an "Anchor Test", this design is also called NEAT design. Since the first two designs use samples from a common population, the task in equating is simply to adjust differences in difficulty between the two test forms. In the third design, however, the two groups of examinees are from different populations and usually not considered to be equivalent. Hence, it is necessary to separate group differences in ability from form-to-form differences in difficulty (Kolen & Brennan, 2004). The anchor test items are used to control for differences in ability between two examinee groups.

There are two major types of procedures to put the estimates of item parameters from different test forms onto a common scale: the linking separate calibration and the concurrent calibration.

2.4.2.1 Scale linking methods

When two test forms are administered to two different ability groups, the parameter estimates obtained in separate calibration runs are not on the same scale. Because the two scales (*I* and *J*) are assumed to have a linear relationship ($\theta_J = A\theta_I + B$), a simple linear transformation can be used to place the two sets of parameter estimates on the same scale. The transformation parameters (*A* and *B*) can be calculated using the common items. For the 3PL model, the relations of item discrimination, difficulty, and guessing (*a*-, *b*-, and *c*-) parameters of common item *i* between the two scales can be expressed as:

$$a_{Ji} = \frac{a_{Ij}}{A}, \tag{2.36}$$

$$b_{Ji} = Ab_{Ii} + B, \tag{2.37}$$

$$C_{Ji} = C_{Ii}. \tag{2.38}$$

In real testing situations, there are multiple common items and not all of them satisfy the above linear relationship. Hence, there is a need to estimate the values of *A* and *B*. A variety of methods have been proposed to calculate the transformation parameters, including the mean/mean method (Loyd & Hoover, 1980), mean/sigma method (Marco, 1977), and the Test Characteristic Curves (TCC) methods. The mean/sigma method uses the first two moments of the difficulty parameter estimates for the common items to calculate the transformation parameters. The mean/mean method is based on the means of the discrimination and difficulty parameter estimates for the common items. A major disadvantage of the mean/sigma and mean/mean methods is that they do not consider all of the item parameters simultaneously (Kolen & Brennan, 2004). To compensate for this problem, the TCC methods have been developed to search the transformation coefficients that minimize the difference in TCCs between the base and the target test. The Haebara (1980) method and the Stocking-Lord (1983) method are two commonly used TCC methods. The Haebara method minimizes the squared differences between the sets of item characteristic curve for the common items between the base and target test. Hence, it is also called the item characteristic curve method. For the 3PL IRT model, the Haebara function can be expressed as follows,

$$H(\theta_j) = \sum_{i:V} [p_{ij}(\theta_{Jj}; \hat{a}_{Ji}, \hat{b}_{Ji}, \hat{c}_{Ji}) - p^*_{ij}(\theta_{Jj}; \frac{\hat{a}_{Ii}}{A}, A\hat{b}_{Ii} + B, \hat{c}_{Ii})]^2 , \tag{2.39}$$

where *i:V* represents the set of common items, $(\hat{a}_{Ji}, \hat{b}_{Ji}, \hat{c}_{Ji})$ and $(\hat{a}_{Ii}, \hat{b}_{Ii}, \hat{c}_{Ii})$ are the two sets of item parameter estimates for the base and target test from separate calibration runs, *J* and *I* indicate the scale for the base and target groups, respectively. And $p_{ij}(\theta_{Jj}; \hat{a}_{Ji}, \hat{b}_{Ji}, \hat{c}_{Ji})$ is the probability of answering item *i* correctly. The criterion function is defined by either integrating over θ or summing up *F* (as mentioned below) over all examinees. The transformation coefficients *A* and *B* can be obtained through minimizing the criterion function.

The Stocking-Lord method minimizes the squared differences between TCCs for the common items. The loss function to be minimized is given by

$$F = \frac{1}{N} \sum_{n=1}^{N} [\tau(\theta_j) - \tau*(\theta_j)]^2 , \tag{2.40}$$

where *n*=1, 2, …, *N*, and *N* is the number of arbitrary points over the latent trait scale; τ and τ^* are the estimated true scores on the common items for the base form and rescaled true scores on the common items for the target form, respectively.

Researchers have compared the performance of different linking separate calibration methods. It has been shown that the TCC methods provide more accurate estimates of transformation parameters compared to the mean/sigma and mean/mean methods (Baker & Al-

Karni, 1991; Hanson & Beguin, 2002; Way & Tang, 1991). It has also been shown that the Haebara method and the Stocking-Lord method produce similar equating results for dichotomous IRT model (Way & Tang, 1991) and polytomous IRT model (Li & Yin, 2008).

2.4.2.2 Concurrent calibration

As an alternative to scale linking methods, concurrent calibration simultaneously estimates all item parameters in both test forms in a single calibration run. Parameter estimation is based on the combined data of the base and target groups. Items not taken by any particular group are treated as not reached or missing (Lord, 1980). For the single group and random groups design, parameter estimates are automatically placed on the same scale, even if they are obtained from separate calibrations rather than concurrent calibration. For the NEAT design, if item parameters are not estimated using the concurrent calibration method, a linking procedure is necessary to put all parameter estimates on the same scale. Software packages such as MULTILOG (Thissen, 1991) or BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 2003) can be used to perform concurrent calibration.

2.4.2.3 Comparison between linking separate and concurrent calibration

Researchers have compared the equating performance between linking separate calibration and concurrent calibration (Béguin & Hanson, 2001; Béguin, Hanson, & Glas, 2000; Hanson & Béguin, 2002; Kim & Cohen, 1998; and Kim & Cohen, 2002). Kim and Cohen (2002) showed that concurrent calibration yielded smaller equating errors than linking separate calibration, while the difference was very small. Hanson and Béguin (2002) showed that when groups were equivalent, concurrent calibration tended to provide more accurate equating results than linking separate calibration. Béguin and Hanson (2001) and Béguin, Hanson, and Glas (2000) showed that when test data did not meet the unidimensionality assumption, linking separate calibration provided more accurate equating results compared to concurrent calibration in the non-equivalent group conditions. Kerkee et al. (2003) compared concurrent calibration and the Stocking-Lord method for non-equivalent groups in vertical scaling using a real data set with a complex form of violations to the IRT model assumptions. Their results showed that the Stocking-Lord method in general resulted in better fits for items compared to concurrent calibration. Based on previous studies, Kolen and Brennan (2004) conclude that concurrent calibration tends to be more accurate than linking separate calibration when the IRT assumptions are met, whereas it is also less robust to assumption violations of the IRT models.

2.4.3 Factors that affect equating results

Factors that are relevant to the item and person characteristics may affect the accuracy of test equating. These include number of common items, test length, item type, item calibration models, equating procedures, and distributions of item difficulty and ability parameters, etc. (Zhang, 2010; Skaggs & Lissitz, 1986). Angoff (1968) developed guidelines for constructing the set of common items in the NEAT equating design. He recommended at least 20% of the items in a test or 20 items to form the set of common items, whichever is larger. Zhu (1998) suggested that the number of common items should exceed 20 to 25 percent of the number of total items on either of the test forms. Some other researchers stated that 5 to 15 common items were enough to produce acceptable equating results (Smith & Kramer, 1992; Wright & Master, 1982; Wright & Stone, 1979). Hills, Subhiyah, and Hirsch (1988) found that for the three-parameter model, 10 common items randomly selected from a set of 30 common items were enough to produce acceptable equating results.

As group difference in ability increases, the equating accuracy might decrease. There are at least two possible explanations for why this may happen (Powers, 2011). First, the equating relationship may vary for subgroups of examinees. Second, as the group differences increase, the statistical assumptions underlying the equating methods are more likely to be violated.

2.5 IRT EQUATING WITH LOCAL ITEM DEPENDENCE

When LID is present, the use of a standard IRT equating procedure is inappropriate and may lead to biased equating results. It has been shown that when LID exists, estimation of linking coefficients can be biased if a standard IRT model is used for linking calibrations (Li, Bolt, & Fu, 2005). Lee et al. (2001) also found that for equating testlet-based tests, using testlet scores and fitting a polytomous IRT model provided better equating results than ignoring LID and using the standard 3PL IRT model.

Due to the within-testlet dependency, linking testlet-based tests is less straightforward than linking tests which are made up of independent items. Although testlet-based tests have been widely used, research on equating procedures for testlet-based tests is limited. To the knowledge of the author, only three studies have addressed the issue directly for the NEAT design (Li, Bolt, & Fu, 2005; Li, 2009; Zhang, 2010). In this section, equating procedures for testlet-based tests proposed in these studies are reviewed. Research design, results and limitations are discussed.

2.5.1 Test Characteristic Curve equating methods for the TRT model

2.5.1.1 Stocking-Lord characteristic curve method

Li, Bolt, and Fu (2005) applied the Stocking-Lord method to the 2PNO TRT model. An essential step in the Stocking-Lord method is to compute the true scores. Due to the additional random effect parameter $\gamma_{jd(i)}$ in the TRT model, computing the true scores is more complicated than for traditional IRT models. Li, Bolt, and Fu adopted the alternative model formulation proposed by Glas, Wainer, and Bradlow (2000), in which the term $a_i(\theta_j - b_i - \gamma_{jd(i)})$ in the TRT model is reparameterized as $a_i(\xi_{jd} - \gamma_{jd(i)})$, and $\xi_{jd} = \theta_j - b$. Hence, $\xi_{jd} \mid \theta$ is distributed as $N(\theta, \sigma^2_{\gamma d})$. The probability of person *j* answering item *i* within testlet *d* for the 2PNO TRT model can be written as

$$p(y_{jdi} = 1 \mid \xi_{jd}) = \Phi[a_i(\xi_{jd} - b_i)]. \qquad (2.41)$$

Then, the probability of answering item *i* within testlet *d* correctly conditional on θ can be written as

$$p(y_{di} = 1 \mid \theta; \sigma_{\xi_d}) = \int p(y_{di} = 1 \mid \xi_d) h(\xi_d \mid \theta; \sigma_{\xi_d}) d\xi_d, \qquad (2.42)$$

where *h* is the distribution of ξ_{jd} given θ.The item parameters and σ_{ξ_d} are assumed known. The integral can be approximated using the Gaussian quadrature method

$$\int p(y_{di} = 1 \mid \xi_d) h(\xi_d \mid \theta; \sigma_{\xi_d}) d\xi_d = \sum_{p=1}^{P} p(X_p) W_p, \qquad (2.43)$$

where X_p and W_p represent the p^{th} quadrature point and its associated weight. The true score for the whole test is the sum of the testlet true scores and can be calculated as follows:

$$\tau(\theta) = \sum_{d=1}^{D}\sum_{i=1}^{I}\int p(y_{di} = 1 \,|\, \xi_d) h(\xi_d \,|\, \theta; \sigma_{\xi_d}) d\xi_d = \sum_{d=1}^{D}\sum_{i=1}^{I}\sum_{p=1}^{P} p(X_p) W_p \,. \tag{2.44}$$

where D is the total number of testlets in a test, and I is the number of items within each testlet.

Then, the transformation coefficients can be obtained by solving the Stocking-Lord loss function. Because the function is a nonlinear function of transformation coefficients, it may have multiple saddle points. A quasi-Newton method such as the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method (Dennis & Schnabel, 1996) can be used to search the estimates of the transformation coefficients.

2.5.1.2 Haebara method

Researchers also develop methods based on the Haebara approach to compute the transformation coefficients for testlet-based tests (Li, 2009; Zhang, 2010). Zhang (2010) applied the Haebara approach to the two-parameter (2PL) logistic TRT model. The quadratic loss function can be expressed as follows,

$$F = \sum_{J}\sum_{D}\sum_{i(d)} [p_{jdi}(\xi_{jJ}) - p^*_{jdi}(\xi_{jJ})] \tag{2.45}$$

where

$$p_{jdi}(\xi_{jJ}) = \frac{\exp[1.7 a_{iJ} (\xi_{jJ} - b_{iJ})]}{1 + \exp[1.7 a_{iJ} (\xi_{jJ} - b_{iJ})]} \,, \tag{2.46}$$

and

$$p^*_{jdi}(\xi_{jJ}) = \frac{\exp\{1.7 \frac{a_{iI}}{A} [\xi_{jJ} - A(b_{iI} - \mu_{\gamma d}) - B]\}}{1 + \exp\{1.7 \frac{a_{iI}}{A} [\xi_{jJ} - A(b_{iI} - \mu_{\gamma d}) - B]\}} \,. \tag{2.47}$$

The estimates of the transformation coefficients can be obtained by solving the above nonlinear minimization problem. Specifically, the BFGS method can be employed to find the minimum value of the function.

Li (2009) extended the Haebara method to the 3PL TRT model. The Haebara function for the 3PL TRT model can be expressed as follows,

$$H(\theta_j) = \sum_{i:V}[p_{ij}(\theta_{Jj};\hat{a}_{Ji},\hat{b}_{Ji},\hat{c}_{Ji},\gamma_{Jd(i)}) - p_{ij}^{*}(\theta_{Jj};\frac{\hat{a}_{Ii}}{A},A\hat{b}_{Ii}+B,\hat{c}_{Ii},A\gamma_{Id(i)})]^2 , \qquad (2.48)$$

where

$$p_{ij}(\theta_{Jj};\hat{a}_{Ji},\hat{b}_{Ji},\hat{c}_{Ji},\gamma_{Jd(i)}) = \int\{\hat{c}_{Ji} + (1-\hat{c}_{Ji})\frac{\exp[\hat{a}_{Ji}(\theta_{Jj}-\hat{b}_{Ji}-\gamma_{Jjd(i)})]}{1+\exp[\hat{a}_{Ji}(\theta_{Jj}-\hat{b}_{Ji}-\gamma_{Jjd(i)})}\}\psi(\gamma_{Jjd(i)})d\gamma_{Jjd(i)} , \qquad (2.49)$$

where $\psi(\gamma_{Jjd})$ is the estimated distribution of γ_{Jjd}. The distribution of the testlet effect γ_{Jjd} can be approximated with a discrete distribution on a finite number of equally spaced quadratic points so that

$$p_{ij}(\theta_{Jj};\hat{a}_{Ji},\hat{b}_{Ji},\hat{c}_{Ji},\gamma_{Jd(i)}) \cong \sum_{k}((\hat{c}_{Ji} + (1-\hat{c}_{Ji})\frac{\exp[\hat{a}_{Ji}(\theta_{Jj}-\hat{b}_{Ji}-p_{k(\gamma_{Jd})})]}{1+\exp[\hat{a}_{Ji}(\theta_{Jj}-\hat{b}_{Ji}-p_{k(\gamma_{Jd})})})W_{k(\gamma_{Jd})}) , \qquad (2.50)$$

where $p_{k(\gamma_{Jd})}$ and $W_{k(\gamma_{Jd})}$ are the k^{th} quadrature point and associated weight, respectively. Hence, the true score for each item given θ can be computed. The loss function is obtained by integrating over θ

$$F = \sum_{J} H(\theta_j) . \qquad (2.51)$$

The Newton Raphson method can be used to find the values for *A* and *B* that minimize the above function.

2.5.1.3 The testlet characteristic curve method

Zhang (2010) developed a new TCC-based approach of finding the transformation coefficients for the testlet-based test. Different from the Stocking-Lord approach and the Haebara approach, this approach was used to search the transformation coefficients that minimize the squared differences between the sets of testlet characteristic curves for the same testlet between the base and the target tests. The quadratic loss function can be expressed as follows,

$$F = \sum_{J}\sum_{D}\left(\sum_{i\in d} p_{jdi}(\xi_{jJ}) - \sum_{i\in d} p^{*}_{jdi}(\xi_{jJ})\right)^{2}, \qquad (2.52)$$

where

$$p_{jdi}(\xi_{jJ}) = \frac{\exp[1.7a_{iJ}(\xi_{jJ} - b_{iJ})]}{1+\exp[1.7a_{iJ}(\xi_{jJ} - b_{iJ})]}, \qquad (2.53)$$

and

$$p^{*}_{jdi}(\xi_{jJ}) = \frac{\exp\{1.7\frac{a_{iI}}{A}[\xi_{jJ} - A(b_{iI} - \mu_{\gamma d}) - B]\}}{1+\exp\{1.7\frac{a_{iI}}{A}[\xi_{jJ} - A(b_{iI} - \mu_{\gamma d}) - B]\}}. \qquad (2.54)$$

The BFGS method can be used to find the minimum value of the above loss function.

2.5.2 Concurrent calibration under the TRT model

Concurrent calibration estimates item and person ability parameter simultaneously using the combined data from the base and target group. For the TRT model, Bayesian estimation can be employed to put item parameters on the same scale through a single Markov chain Monte Carlo (MCMC) run. The concurrent calibration method has been applied to equate testlet-based tests for dichotomous items and mixed-format items (Zhang, 2010).

2.5.3 Comparisons of different methods for equating testlet-based tests

Researchers have conducted studies to compare different equating procedures for equating testlet-based tests. Using real data, Lee et al. (2001) compared two polytomous IRT models (the nominal response model and the GRM) to the dichotomous 3PL IRT model in the context of equating testlet-based tests. Their results showed that with the presence of LID, equating based on the two polytomous IRT models produced results that were more consistent with the results of traditional equipercentile method.

Li, Bolt, and Fu (2005) evaluated the performance of the Stocking-Lord method for the 2PNO testlet model using simulated data. The equating results were compared to those obtained when using 2PNO IRT model. The study manipulated two factors: (a) number of common testlets (2 and 4); and (b) testlet variance (0, 0.5, and 1). The WinBUGS program (Spiegelhalter, Thomas, Best, & Lunn, 2003) was employed to implement the MCMC method for estimation of the 2PNO testlet model. The computer program BILOG (Mislevy & Bock, 1983) was used to estimate the 2PNO IRT model. The evaluation criterion was absolute differences between the estimated transformation parameters and the true transformation parameters. The results showed that when LID was present, the TRT-based Stocking-Lord method provided more accurate estimates of the transformation coefficients than did the IRT-based method. The results also indicated that for the 30-item tests, two common five-item testlets (10 common items) were enough to obtain accurate linking coefficients.

Li (2009) compared the performance of the Haebara method for the 3PL TRT model with the scaling linking methods for the 3PL IRT model and the GRM. The simulation study had three conditions of testlet variances (0, 1, and 2), representing no, moderate, and strong testlet effect, respectively. The three models were estimated using different computer programs. The

WinBUGS program was employed to implement the MCMC method for estimation of the 3PL testlet model. The computer program BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996) was used to estimate the 3PL IRT model. The GRM was estimated in the computer program PARSCALE (Muraki & Bock, 1996). The evaluation criteria included: (a) Mean Squared Error (MSE) for the transformation coefficients; (b) Root Mean Squared Deviation (RMSD) and Mean Absolute Difference (MAD) for item and ability parameters; and (c) Test Information Functions (TIFs) inflation ratios for the GRM-estimated TIFs vs. the 3PL IRT model-estimated TIFs; and for the testlet model-estimated TIFs vs. the 3PL IRT model-estimated TIFs. The results indicated that when LID was present, the TRT-based equating method in general provided more accurate estimation of linking coefficients and item parameters than the IRT-based equating method, especially when the degree of LID was high. The TRT-based, IRT-based, and GRM equating procedures yielded comparable person parameter estimates, with the former two performed slightly better than the latter in some conditions. The equating method based on the TRT model also yielded better reliability statistics than that based on the 3PL IRT model, as indicated by the TIF inflation ratios. The GRM and TRT model produced similar reliability statistics, whereas the latter generated more accurate person parameter estimates.

Zhang (2010) compared the performance of separate and concurrent estimation under the 2PNO testlet response model using simulated data. Three linking separate calibration methods were considered: (i) Stocking-Lord; (ii) Haebara; and (iii) TCC method. The conditions of the study were: (a) number of common testlets (2 and 4); (b) testlet variance (0.25, 0.50, 0.75, and 1.00); and (c) calibration models (traditional IRT model vs. TRT model). The WinBUGS program was employed to implement the MCMC method for estimation of the 2PNO testlet model. The computer program BILOG-MG was used to estimate the 2PNO IRT model. The evaluation criteria were: (a) RMSD and mean Euclidean distance for item discrimination and difficulty parameters, and (b) mean differences for the means of testlet effects. The results indicated that the Haebara approach and the TCC approach under the 2PNO testlet model performed similar or even better than did the Stocking-Lord method and the concurrent calibration method. The concurrent calibration method performed as well as the TCC methods, especially when the number of common testlet is not small (n=4). The results also suggested that ignoring LID could substantially increase equating errors. The equating errors increased as the testlet variance increased. In general, equating errors decreased as the number of common testlets increased. However, the difference was small, especially for the large sample size conditions (n=1,000).

Using real data, He et al. (2012) compared the equating performance of the TRT and bifactor model under the NEAT design for testlet-based tests. Chen (2014) explored the equating performance of these two models under the random groups design using both real and simulated data. The results from both studies showed that these two models generated similar equating results, and they both performed better than the dichotomous IRT model when LID was present. Moreover, the results from Chen’s (2014) study showed that the GRM produced more stable equating results than the TRT and bifactor models, even with a large degree of LID.

Cao, Lu, and Tao (2014) conducted a simulation study to compare the performance of three IRT models on number-correct equating under the random groups design using linking separate calibration: the 2PL IRT model, the GRM, and the 2PL TRT model. The results show that with the presence of LID, the results from the 2PL IRT model and the 2PL TRT model more closely agreed with those from the baseline equipercentile method, compared to those from the

GRM. In addition, the 2PL IRT model was quite robust to the violation of the local item independence assumption.

In sum, the results from the above mentioned studies in general suggest that with the presence of LID, the TRT model and the bifactor model tend to generate better equating results than traditional dichotomous IRT models. These are some inconsistent findings as well, which may be caused by different evaluation criteria used in these studies. For example, the accuracy of number-correct score estimates was examined in Cao, Lu, and Tao’s (2014) study, while in Lee et al. (2001) and He et al. (2012) the traditional (e.g., linear and equipercentile) equating results were used as baselines. In addition, the results from above mentioned studies are inconsistent with regard to the relative performance of the TRT model and polytomous IRT model. Therefore, it is the interest of the current study to also compare the equating performance between the TRT model and polytomous IRT model.

2.6 IRT EQUATING WITH LOCAL PERSON DEPENDENCE

It has been shown that with the presence of LID, equating methods based on the testlet model can provide more accurate linking coefficients and model parameter estimation than those based on standard IRT models (e.g., Li, Bolt, & Fu, 2005; Li, 2009; Zhang, 2010). However, equating test forms when data exhibit person dependence remains an area in which there has been little research. A related study conducted by Wang et al., (2010) investigates the impact of ignoring person dependence on accuracy of vertical scaling. The results in general suggest that ignoring person dependence would lead to reduced effective sample size of IRT models and increased vertical scaling errors. The current study aims to address such paucity in equating with dual local dependence.

3.0 METHODOLOGY

The essential goal of this study was to explore the effectiveness of multilevel models on test equating under the anchor test design in situations where test data exhibit local dependence. The equating results from using multilevel models were to be compared to those from using traditional IRT model and TRT model. In addition, factors that are expected to affect equating results were also investigated, such as degree of local dependence, sample size, ability distribution, and number of common testlets.

In order to accomplish these goals, two simulation studies were conducted. Two types of test data were considered: data containing responses to testlet items collected from independent examinees and those collected from examinees nested within groups. The use of two types of multilevel models on test equating was demonstrated in these studies. It was expected that the results from these studies could provide evidence as a reference for researchers interested in applying multilevel models to test equating, especially in situations where item and person dependence are present.

This chapter is organized in two sections. The first section reports the methodology for simulation study 1, which evaluated the effectiveness of HGLM as a concurrent calibration model on equating testlet-based tests. The second section presents the methodology for simulation study 2, which investigated the performance of a 3PL multilevel testlet model as a concurrent calibration model on test equating in situations where dual local dependence was present in test data. Each section describes the design and factors of the simulation study, data generation procedures, equating methods, and evaluation criteria to evaluate and compare the equating results.

3.1 SIMULATION STUDY 1

3.1.1 Research questions

The purpose of the first simulation study was to explore the effectiveness of a three-level HGLM as a concurrent calibration model on equating testlet-based tests under the anchor-test design. The equating accuracies under the three-level HGLM were also compared with those under a two-level HGLM which ignored the testlet effects. In addition, the HGLM approach was also compared to concurrent calibration using MULTILOG based on the Rasch model or GRM. The study attempted to answer the following research questions:

1. How well does the proposed HGLM concurrent calibration method recover model parameters with the presence of LID?
    1.1. How well does the two-level HGLM recover model parameters, compared to the Rasch concurrent calibration with the presence of LID?
    1.2. Does the proposed three-level HGLM, which accounts for the testlet effects, provide more accurate equating results than the two-level HGLM with the presence of LID?

1.3. For concurrent calibration, does polytomous scoring using the GRM provide more accurate equating results than Rasch concurrent calibration with the presence of LID?

2. For each of the four investigated equating methods, what is the impact of degree of LID on equating results?

3.1.2 HGLM as a concurrent calibration model

The Rasch model has been widely used in educational measurement. For example, many state assessments programs are using the Rasch model to measure students' performance. As described above, the Rasch testlet model can be reformulated as a three-level HGLM. The effects of test items, item clusters, and persons are modeled in the level-one, level-two, and level-three models, respectively. Because HGLM allows for missing data, it can be directly applied in test equating as a concurrent calibration model (Chu & Kamata, 2000). When the anchor-test design is used, the item parameters across test forms are estimated on the same scale through the anchor test items (Chu & Kamata, 2000).

The three-level HGLM model for testlet effects described earlier assumes that examinees are from one underlying latent distribution. Let p_{imj} be the probability that person *j* answers item *i* in item cluster *m* correctly, the level-one model can be expressed as:

$$\log(\frac{p_{imj}}{1-p_{imj}}) = \eta_{imj} = \beta_{0mj} + \sum_{q=1}^{I-1} \beta_{qmj} X_{qimj}, \tag{3.1}$$

where X_{qimj} is the q^{th} dummy variable for person *j*, and $X_{qimj} = 1$ when *q*=*i* and 0 when *q*≠*i*, β_{0mj} is the intercept, and β_{qmj} is the coefficient associated with X_{qimj} where $q = 1,....,I-1$. To achieve full rank for the design matrix, one of the dummy variables in the equation is dropped, thereby resulting in $I-1$ dummy variables. This dropped dummy variable is treated as the reference item, and therefore β_{0mj} is interpreted as the item effect for the dropped item in item cluster *m* for person *j*. The individual item effect β_{qmj} is the difference of the item with the q^{th} dummy variable from the effect of the reference item.

The item cluster-level (level-two) model is given by

$$\begin{cases} \beta_{omj} = \gamma_{00j} + u_{0mj} \\ \beta_{1mj} = \gamma_{10j} \\ \beta_{2mj} = \gamma_{20j} \\ \qquad \vdots \\ \beta_{qmj} = \gamma_{q0j} \end{cases}, \tag{3.2}$$

where γ_{00j} is the fixed effect of the level-one intercept, γ_{q0j} is the item-specific effect for item with the q^{th} dummy variable (when $i=q$), and u_{0mj} is the random effect of the level-one intercept. The random effect u_{0mj} is assumed to follow a normal distribution with mean of zero and variance of σ_u^2. It can be considered as the interaction between latent trait and item cluster, which is analogous to the person-specific testlet effect in Bradlow, Wainer, and Wang's (1999) formulation of the one-parameter testlet model. In both models, the variance of the interaction effect σ_u^2 indicates the magnitude of LID and is assumed to be constant across item clusters.

The person-level (level-three) model is given by

$$\begin{cases} \gamma_{00j} = \pi_{000} + r_{00j} \\ \gamma_{10m} = \pi_{100} \\ \gamma_{20m} = \pi_{200} \\ \qquad \vdots \\ \gamma_{q0m} = \pi_{q00} \end{cases}, \tag{3.3}$$

where r_{00j} is the person ability and $r_{00j} \sim N(0, \sigma_r^2)$. In this model, the items are assumed having only fixed effects, while the person effect is considered to be random.

The combined three-level model for LID can be expressed as

$$p_{ijm} = \frac{1}{1 + \exp\{-[r_{00j} - (-\pi_{q00} - \pi_{000}) + u_{0mj}]\}}, \tag{3.4}$$

where r_{00j} indicates the ability for person *j*, and the term $-\pi_{q00} - \pi_{000}$ denotes the difficulty level of item *q*. In the framework of TRT, this model can be rewritten as:

$$p_{jdig} = \frac{1}{1 + \exp\{-[\theta_j - b_i + \gamma_{jd(i)})]\}}, \tag{3.5}$$

where $\theta_j = r_{00j}$, $b_i = -\pi_{q00} - \pi_{000}$, and $\gamma_{jd(i)} = u_{0mj}$. Thus, the three-level HGLM for LID is equivalent to the Rasch testlet model (Wang & Wilson, 2005).

For equivalent groups equating, it can be directly applied for item and ability calibration. However, for non-equivalent groups equating, the base and target groups have different distribution characteristics (e.g., means and standard deviations). To differentiate the base and target groups, a group indicator can be added into the person-level model. Hence, it can model groups with different means. Equation (3.3) becomes:

$$\begin{cases} \gamma_{00j} = \pi_{000} + \pi_{010}(Group)_j + w_{00j} \\ \gamma_{10j} = \pi_{100} \\ \gamma_{20j} = \pi_{200} \\ \quad\vdots \\ \gamma_{q0j} = \pi_{q00} \end{cases}, \tag{3.6}$$

where group is the person-level variable, indicating whether the examinee is from the base group or the target group. The value of the variable *group* equals one if the examinee is from the base group, and zero otherwise. The term π_{010} denotes the difference in ability between the base and target groups. The combined model for equating testlet-based tests can be expressed as

$$p_{ijm} = \frac{1}{1+\exp[-\{[w_{00j}+\pi_{010}(Group)_j]-(-\pi_{q00}-\pi_{000})+u_{0mj}\}]}, \tag{3.7}$$

where the term $(-\pi_{q00}-\pi_{000})$ denotes the difficulty level for item *i* for *i*=*q*(*i*=1, …, *k*-1), and π_{000} is the item difficulty for item *k*. The overall ability for person *j* in this formulation is $w_{00j}+\pi_{010}(Group)_j$. For examinees in the base group, the ability can be expressed as $\theta_j = w_{00j}$. For examinees in the target group, it can be calculated as $\theta_j = w_{00j}+\pi_{010}$. It should be noted that this adjusted model still assumes equal variance of the two latent distributions. Therefore, this model is appropriate for equating situations where the distributions of the base and target groups are similar in variance.

The performance of the three-level HGLM on equating testlet-based test was also to be compared to that of a two-level HGLM ignoring the item clustering effect. As described above, the Rasch model can be reformulated as a two-level HGLM. Let p_{ij} be the probability that person *j* answers item *i* correctly, the item-level model is given by:

$$\begin{aligned}\log\left(\frac{p_{ij}}{1-p_{ij}}\right) &= \eta_{ij}\\ &= \beta_{0j}+\beta_{1j}X_{1ij}+\beta_{2j}X_{2ij}+\dots..+\beta_{(I-1)}X_{(I-1)ij},\\ &= \beta_{0j}+\sum_{q=1}^{I-1}\beta_{qj}X_{qij}\end{aligned} \tag{3.8}$$

where X_{qij} is the q^{th} dummy variable for person *i*, and $X_{qij}=1$ when *q*=*i* and 0 when *q*≠*i*, *I* is the number of items in the test, β_{0j} is an intercept term, and β_{qj} is the coefficient associated with X_{qij} when $q=1,\dots.,I-1$. To achieve full rank for the design matrix, the dummy variable for the

I^{th} item is dropped, thereby resulting in $I-1$ dummy variables. This dropped dummy variable is treated as the reference item, and therefore β_{0j} is interpreted as the item effect for the dropped item for person *j*, and β_{qj} is the effect of the q^{th} item compared to the reference item.

The person-level model can be written as:

$$\begin{cases} \beta_{0j} = \gamma_{00} + \gamma_{01}(Group)_j + \mu_{0j} \\ \beta_{1j} = \gamma_{10} \\ \beta_{2j} = \gamma_{20} \\ \quad \vdots \\ \beta_{qj} = \gamma_{q0} \end{cases} . \tag{3.9}$$

The combined model can be expressed as

$$p_{ij} = \frac{1}{1+\exp[-\{u_{0j} + [r_{01}(Group)_j] - (-\gamma_{q0} - \gamma_{00})\}]}, \tag{3.10}$$

where the term $(-\gamma_{q0} - \gamma_{00})$ denotes the difficulty level for item *i* for *i*=*q*(*i*=1, …, *I*-1), and γ_{00} is the item difficulty for item *i*. The overall ability for person *j* in this formulation is $\gamma_{01}(Group)_j + u_{0j}$.

3.1.3 Design and data generation

This simulation study employed the anchor test equating design and concurrent calibration method. The equating results based on the four investigated models – the three-level HGLM, the two-level HGLM, the Rasch model, and the GRM model – were compared. Two estimation procedures were used to simultaneously put the item parameters for the target form onto the base form scale: Laplace approximation estimation in SAS for the HGLM approach and MMLE in Multilog for Rasch and GRM concurrent calibration.

An anchor test data collection design was used for data generation. It was assumed that two forms of a 30-item multiple-choice test were given to two groups of examinees. There were 6 testlets in each test form and each testlet contained 5 items. The test length is fixed at 30 because it is the common quantity among the reviewed literature on equating testlet-based test. A testlet size of 5 is commonly used in actual reading comprehension tests (Li, Bolt, & Fu, 2005).

The two-parameter TRT model (Wainer, Bradlow, & Du, 2000) was used for data generation. Let Y_{ij} be the response of examinee *j* on item *i*. The probability of a correct answer of examinee *j* on item *i* is:

$$p(Y_{ij}=1)=\frac{\exp[a_i(\theta_j-b_i+\gamma_{jd(i)})]}{1+\exp[a_j(\theta_j-b_i+\gamma_{jd(i)})]}, \tag{3.11}$$

Where θ_j, a_i, and b_i are the person trait, item discrimination, and item difficulty parameters, respectively. Compared to the standard two-parameter IRT model, this model adds a random effect parameter, $\gamma_{jd(i)}$, which models the person-specific testlet effect, or the interaction of person *j* with testlet *d*. Within a testlet, the value of $\gamma_{jd(i)}$ is constant for a person but varies across persons. It is assumed to be independent from the θ distribution and follow a normal distribution, $\gamma_{jd(i)} \sim N(0,\sigma_\gamma^2)$. By definition, the sum of $\gamma_{jd(i)}$ over examinees within the same testlet is zero (i.e.,$\sum_j \gamma_{jd(i)} = 0$). If $\gamma_{jd(i)} = 0$, there is no testlet effect and the item fits the standard two-parameter model. The testlet variance, σ_γ^2, is assumed to be constant across testlets. It denotes the magnitude of LID: the larger the variance, the larger the amount of item dependence within that testlet. If $\sigma_\gamma^2 = 0$, there is no extra dependence among items within the same testlet and the local item independence assumption is met.

The true values of the item difficulty parameter were randomly generated from a standard normal distribution. The true values of the item discrimination parameter were generated from a uniform [0.8, 1.2] distribution to reflect the discrimination values of test items used for Rasch calibration in real testing situations. For each examinee, a primary trait and six testlet-specific traits were generated. The testlet-specific traits were randomly drawn from a multivariate normal distribution with means of 0 and variances as specified below. For examinees in the base group, the primary trait was randomly generated from a standard normal distribution. For examinees in the target group, the primary trait was randomly generated from a normal distribution with means and variances as specified below. Response data were generated for base and target groups separately based on the simulated item and person parameter values. To simulate a large-scale testing situation, 1,000 examinees' responses were simulated for both base and target groups. For the GRM concurrent calibration, the item scores within each testlet were summed to form a single testlet score. Response data were generated using the statistical software SAS 9.4.

3.1.4 Manipulated factors

This study examined whether variations in ability distribution for the target group, number of common testlets, and degree of LID affect equating results.

The first factor examined was the ability distribution for the target group. Two conditions were considered: *N* (0, 1), or *N* (1, 1). The *N* (0, 1) target group condition simulated an equivalent group equating situation, and the *N* (1, 1) target group condition reflected a nonequivalent group equating situation.

The second factor examined was the number of common testlets. Two different lengths of common testlets were used: 2 and 3. In conditions with 2 common testlet, there were 10 unique testlets across the two test forms. Whereas in conditions with 3 common testlets, there were 9 unique testlets across the two test forms. Larger number of common items generally produces better equating results.

The third factor examined was the degree of LID. The degree of LID was manipulated by varying the level of testlet variance. Four levels of LID were simulated: $\sigma_{\gamma}^{2} = 0$, 0.5, 1, and 1.5, reflecting zero, low, moderate, and high degree of testlet effect. These values (i.e., 0, 0.5, 1 and 1.5) were similar to those specified in Li, Bolt, and Fu (2006) ($\sigma_{\gamma}^{2} = 0.2$, 0.5, 1, and 1.5) and Bradlow, Wainer, and Wang (1999) ($\sigma_{\gamma}^{2} = 0$, 0.5, 1, and 2). Many previous studies used the values of testlet variance within the range between 0 and 1 (e.g., Jiao, Wang, & He, 2013; Wang & Wilson, 2005; Wang et al., 2002), or 0 and 1.5 (e.g., Chen, 2014; Li, Bolt, & Fu, 2006). Therefore, the current study chose the values within the range between 0 and 1.5 to specify the degree of LID.

With the combination of ability distribution for the target group, number of common testlets, and levels of item dependence, 16 simulation conditions resulted. Within each simulated condition, 200 replications were run, resulting in a total of 3,200 datasets.

With the combination of ability distribution for the target group, number of common testlets, and levels of item dependence, 16 simulation conditions resulted. Within each simulated condition, 200 replications were run, resulting in a total of 3,200 datasets.

3.1.5 Model calibration and estimation methods

For the concurrent calibration, the combined data sets from the base and target groups for each simulation condition were used. After the datasets were simulated, the three-level HGLM for equating testlet-based tests was fitted to estimate model parameters. For comparison purpose, the same combined data set was also fitted to the two-level HGLM, which is equivalent to a standard Rasch model. Thus, a single combined data set was analyzed twice under the HGLM framework. Altogether, 2400 concurrent calibration runs were performed under the HGLM framework. The SAS 9.4 PROC GLIMMIX procedure was used for model calibration.

The most commonly used method for estimating the parameter of HGLMs is the penalized quasi-likelihood (PQL; Breslow & Clayton, 1993) procedure. This method is less computationally intensive and can provide reasonable estimation in many cases. However, it tends to yield considerable downward bias for the variances of the random components in HGLM with binary outcomes. On the contrary, the Laplace approximation estimation option is computationally intensive but can reduce the estimation bias in the variance of random components. In *Laplace 6* approximation, parameters were estimated by a sixth order approximation to the likelihood for the model based on a Laplace transformation (Raudenbush, Yang, & Yosef, 2000). The recent versions of SAS, including SAS 9.3 and SAS 9.4, allow for Laplace approximation. Since the recovery of the variance of random components is to the interest of this study, *Laplace 6* approximation was chosen for the parameter estimation.

For comparison purposes, concurrent calibration based on the Rasch and a polytomous IRT model was considered in this study. The calibration model considered was the GRM. The computer program MULTILOG was used to estimate item and person parameters based on the Rasch model and GRM. In total, this study conducted equating four times for each simulated dataset, two using HGLM, and two using MULTILOG. As a result, 6,400 HGLM concurrent calibration runs and 6,400 MULTILOG concurrent calibration runs were performed.

3.1.6 Evaluation criteria

The effectiveness of the concurrent calibration method under HGLM can be evaluated by examining how well the model parameters are recovered. Since item parameters of dichotomous IRT and TRT models cannot be directly compared to those of polytomous IRT models, the focus of the comparison in this study is on the person parameter. Because the estimated theta obtained from different software packages may not be on the same scale, expected scores were computed and compared across the four equating procedures. The expected score (*E(X)*) or true score (τ) is obtained as:

$$E(X) = \tau = \sum_{i=1}^{I} E(U_i)\,, \tag{3.12}$$

where U_i is the response of the ith item, and i=1,…I. For dichotomous item,

$$E(U_i) = 1 \times P(u=1) + 0 \times P(u=0) = P(u=1)\,. \tag{3.13}$$

Therefore,

$$E(X) = \sum_{i=1}^{I} E(U_i) = \sum_{i=1}^{I} P_i(u=1)\,. \tag{3.14}$$

Similarly, for five-category polytomous items, the expected score is calculated as:

$$E(X) = \sum_{i=1}^{I} E(U_i) = \sum_{i=1}^{I} \begin{matrix}[0 \times P_i(u=0) + 1 \times P_i(u=1) + 2 \times P_i(u=2) + 3 \times P_i(u=3) \\ + 4 \times P_i(u=4) + 5 \times P_i(u=5)].\end{matrix} \tag{3.15}$$

RMSDs were calculated for the expected true scores to assess the accuracies of equating. For each of the simulation conditions, the RMSD was calculated for the expected scores in a replication, and then averaged across replications.

RMSD is defined as the squared root of the average squared differences between estimated and true parameter values. The RMSD of the expected scores is obtained as

$$RMSD_{\hat{\theta}} = \sqrt{\frac{1}{n}\sum_{j=1}^{J}(\hat{X}_j - X_j)^2} \, , \tag{3.16}$$

where X_j is the true value of the expected score, $\hat{X}_j$ is the estimated value of the expected score, and *n* is the number of persons.

3.2 SIMULATION STUDY 2

3.2.1 Research questions

The purpose of this simulation study was to explore the effectiveness of concurrent calibration using a multilevel 3PL testlet model on equating testlet-based tests in situations where respondents were nested in groups under the NEAT design. A 3PL multilevel testlet model was used to model item and person dependence structures simultaneously. The equating accuracies under the 3PL multilevel testlet model were also compared with those under the 3PL testlet model which ignored the testlet effects, and those under the traditional 3PL IRT model which ignored both testlet effects and person clustering effects. The study attempted to answer the following research questions:

1. How well does the 3PL multilevel TRT model recover the item and person parameters under the NEAT design, compared to the 3PL TRT model and the traditional 3PL IRT model, with the presence of LID and person dependence?
2. For each of the three investigated models, what is the impact of degree of LID on equating results?

3. For each of the three investigated models, what is the impact of degree of person dependence on equating results?

3.2.2 A 3PL multilevel testlet model for item and person dependence

The 3PL testlet model (Wainer, Bradlow, & Du, 2000) can be modified to include the dependence structure of persons. The probability of person ‘*j*’ in school ‘*g*’ getting an item ‘*i*’ correct $P(Y_{jgi} = 1)$ is given by a logistic function:

$$p(Y_{jgi} = 1 \mid \theta_{j(g)}, \theta_g, a_i, b_i, c_i, \gamma_{jm(i)}) = c_i + (1 - c_i)\frac{\exp[a_i(\theta_{j(g)} + \theta_g - b_i + \gamma_{jm(i)})]}{1 + \exp[a_i(\theta_{j(g)} + \theta_g - b_i + \gamma_{jm(i)})]}, \quad (3.17)$$

where θ_g is the average ability of persons in school ‘*g*’ and $\theta_g \sim N(\mu_1, \sigma_1^2)$, $\theta_{j(g)}$ is the difference between the latent ability of person ‘*j*’ in school ‘*g*’ and the average ability of the school and $\theta_{j(g)} \sim N(\mu_2, \sigma_2^2)$. The term $(\theta_g + \theta_{j(g)})$ is thus the latent ability of person ‘*j*’ in school ‘*g*’, and follows a $N(\mu_\theta, \sigma_\theta^2)$ distribution, where $\mu_\theta = \mu_1 + \mu_2$ and $\sigma_\theta^2 = \sigma_1^2 + \sigma_2^2$.

The degree of person dependence is measured by ICC, which is the proportion of total variance explained by the group variance, $\sigma_1^2 / (\sigma_1^2 + \sigma_2^2)$. As with the 3PL testlet model, a_i, b_i, c_i, and $\gamma_{jm(i)}$ are the item discrimination, difficulty, guessing, and testlet parameters, respectively. The random effect parameter, $\gamma_{jm(i)}$, models the person-specific testlet effect, or the interaction of person *j* with testlet *m(i)* (i.e., the testlet *m* that contains item *i*), and $\gamma_{jm(i)} \sim N(0, \sigma_{m(i)}^2)$. The testlet variance, $\sigma_{m(i)}^2$, denotes the magnitude of LID.

This model is also a natural extension to the four-level IRT model proposed by Jiao et al. (2012). Compared to the four-level IRT model, this model allows the chance of pseudo-guessing and variation in item discrimination to be estimated.

3.2.3 Design and data generation

In this simulation study, the NEAT design was used for data generation. It was assumed that two test forms (the base form and the target form) with common items were administered to two groups of examinees, the base group and the target group. For each group, there were 20 students in each of 50 classrooms, resulting in a sample size of 1,000. Both test forms consisted of 30 items, with 6 testlets in each test form and each testlet contained 5 items.

The 3PL multilevel testlet model was used to generate response data. The item parameter values were simulated according to Li's (2009) specifications. The item discrimination values were randomly chosen from a lognormal distribution *LN*(-0.3, 0.35^2) ; the item difficulty values were obtained from the standard normal distribution *N*(0, 1); and the guessing parameters were simulated from a normal distribution *N*(0.2, 0.05^2). The testlet parameters were generated from a normal distribution with a constant variance, $\sigma^2_{m(i)} \sim N(0, \sigma^2_{\gamma})$. The true values of the person ability parameter for the base group were randomly generated from a standard normal distribution.

3.2.4 Manipulated factors

The manipulated factors included the following: (a) ability distribution for the target group; (b) degree of item dependence; and (c) degree of person dependence.

The first factor examined was the ability distribution for the target group. Two different ability distributions were generated for the target group: N (.25, 1.1^2) and N (.5, 1.2^2) (see Kang & Petersen, 2009). The different means and standard deviations between the base and target group ability distributions reflect the nonequivalent nature of the two groups. According to Kang and Petersen (2009), the N (.25, 1.1^2) target group condition simulated an equating condition where the base and target groups differ somewhat in ability. The N (.5, 1.2^2) target group condition reflected an equating situation where the two groups differ significantly in ability.

The second factor examined was the degree of LID, which was manipulated by varying the level of testlet variance. Two levels of item dependence were simulated: σ_{γ}^{2} = 0.5 and 1, reflecting moderate and high degree of testlet effect. The third factor examined was the degree of person dependence, which was manipulated by varying the level of ICC. Two levels of ICC were used: ¼ and ½, representing moderate and high level of person dependence.

With the combination of the ability distribution of the target group, number of common testlets, levels of item dependence, and levels of person dependence, 8 simulation conditions resulted. Within each simulation condition, 10 replications were run, resulting in a total of 80 datasets. The simulation was programmed in SAS 9.4.

3.2.5 Equating and calibration method

Concurrent calibration method could put the item and person parameters for the base and target groups onto the same scale by simultaneously calibrating a combined data set. For purposes of consistency, the same computer package and the same estimation procedure were used to run all concurrent calibrations. The MCMC method in WinBUGS 1.4 (Spiegelhalter, Thomas, Best, & Lunn, 2004) was used to conduct the concurrent calibration under the 3PL multilevel testlet model, the 3PL testlet model, and the 3PL IRT model. Thus, a single combined data set for the base and target groups was analyzed three times using the MCMC method in WinBUGS 1.4, once for each of the three calibration method conditions. Altogether, 480 WinBUGS concurrent calibration runs were performed. The MCMC method implemented in WinBUGS 1.4 was used because none of the current available IRT or multilevel software can analyze the 3PL multilevel testlet model.

In the Bayesian estimation framework, a first step is to specify a prior distribution for each unknown parameter. In this study, the priors for the *a*, *b*, and *c* parameters were set by a log normal (0, 0.25) distribution, a normal (0, 4) distribution, and a beta (5, 17) distribution, respectively, in all three calibration models. These are the priors commonly adopted for the 3PL IRT model estimation. Since the 3PL testlet model and the 3PL multilevel testlet model are the extensions of the 3PL IRT model, the same set of the priors for the item parameters were also adopted for the two more complex models. Following Bradlow, Wainer, and, Wang (1999), the prior for the testlet variance σ_{γ}^{2} was set to be an inverse-gamma (0.5, 1) distribution. The prior for the ability distribution of the base group was set to be the normal (0, 1) distribution, while the prior for the ability distribution of the target group was freely estimated. Therefore, the final estimates of item and person parameters were all expressed on the metric of the base group.

The Metropolis sampling method was used to simulate the posterior distributions. Two chains with indispersed initial values were used. The convergence of the MCMC samples was assessed by examining the Gelman-Rubin convergence statistic and history plots. For each of the calibration models, 5,000 iterations were run after the burn-in iteration phase, resulting in a total of 10,000 MCMC samples. The number of burn-in iterations differed by calibration model. For the 3PL multilevel TRT model, a total of 20,000 iterations for each chain were run and the first 15,000 iterations were discarded.

3.2.6 Evaluation criteria

The effectiveness of the concurrent calibration method under each calibration model can be evaluated by examining how well the model parameters are recovered. Correlation and RMSD were used to assess the accuracies of item and person parameter recovery. The two indices were calculated for each simulation condition and each calibration model to evaluate the accuracy of equating.

4.0 RESULTS

This chapter presents the results of the two simulation studies and is divided into three sections. The first section presents the results from simulation study 1, which compares the HGLM approach and MULTILOG in equating testlet-based tests. The RMSDs of estimated expected scores versus true expected scores were compared across the four investigated methods (i.e., two-level HGLM, three-level HGLM, Rasch and GRM concurrent calibration using MULTILOG). The second section presents the results from simulation study 2, which compares the equating results obtained from each of the IRT models (i.e., 3PL IRT model, 3PL TRT model, and 3PL multilevel TRT model). The estimated item and person parameter values were compared against true values under the three IRT models. The third section provides a summary of the results from the two simulation studies.

4.1 RESULTS FROM SIMULATION STUDY 1

Simulation study 1 aimed to explore the performance of the HGLM approach in equating testlet-based tests. This section presents equating results for testlet-based tests across four equating methods. The design factors included target group ability distribution (*N* (0, 1), *N* (1, 1)), degree of LID ($\sigma_{\gamma}^{2} =$0, 0.5, 1, 1.5), and number of common testlets (2, 3), resulting in a total of 16 simulation conditions. To investigate the performance of the various investigated models on concurrent equating of testlet-based tests, average RMSDs of estimated expected score were calculated across the 200 replications within each of the 16 simulated conditions and for each of the four investigated methods. These average RMSDs are presented in Tables 1 through 4.

4.1.1 Analysis of variance

Tables 1 and 2 present the average RMSDs of the base group for the two-common-testlet and three-common-testlet equating design, respectively. Tables 3 and 4 present the average RMSDs of the target group for the two-common-testlet and three-common-testlet equating design, respectively. Lower RMSDs indicate better estimation performance of the equating method.

Table 1. RMSDs of expected scores over 200 replications for the two-common-testlet equating design

(Base group)

Target Group Ability	Testlet variance	2-level HGLM	3-level HGLM	Multilog_Rasch	Multilog_GRM
N (0, 1)	0	2.141	2.141	2.141	2.171
	0.5	2.096	2.096	2.096	2.169
	1	2.049	2.046	2.049	2.202
	1.5	2.025	2.020	2.025	2.271
N (1, 1)	0	2.152	2.152	2.152	2.180
	0.5	2.099	2.098	2.099	2.170
	1	2.055	2.052	2.055	2.210
	1.5	2.025	2.019	2.025	2.276

Table 2. RMSDs of expected scores over 200 replications for the three-common-testlet equating design

(Base group)

Target Group Ability	Testlet Variance	2-level HGLM	3-level HGLM	Multilog_Rasch	Multilog_GRM
N (0, 1)	0	2.145	2.145	2.145	2.175
	0.5	2.096	2.096	2.096	2.164
	1	2.052	2.049	2.052	2.202
	1.5	2.017	2.010	2.016	2.269
N (1, 1)	0	2.141	2.141	2.141	2.168
	0.5	2.093	2.092	2.092	2.163
	1	2.050	2.047	2.049	2.202
	1.5	2.019	2.013	2.019	2.275

Table 3. RMSDs of expected scores over 200 replications for the two-common-testlet equating design

(Target group)

Target Group Ability	Testlet Variance	2-level HGLM	3-level HGLM	Multilog_Rasch	Multilog_GRM
N (0, 1)	0	2.136	2.136	2.136	2.167
	0.5	2.093	2.092	2.092	2.164
	1	2.055	2.052	2.055	2.205
	1.5	2.019	2.012	2.019	2.272
N (1, 1)	0	2.011	2.011	2.011	2.044
	0.5	1.976	1.975	1.976	2.051
	1	1.955	1.952	1.954	2.117
	1.5	1.937	1.931	1.937	2.205

Table 4. RMSDs of expected scores over 200 replications for the three-common-testlet equating design (Target group)

Target Group Ability	Testlet Variance	2-level HGLM	3-level HGLM	Multilog_Rasch	Multilog_GRM
N (0, 1)	0	2.146	2.146	2.146	2.175
	0.5	2.097	2.096	2.097	2.167
	1	2.050	2.047	2.050	2.202
	1.5	2.022	2.015	2.021	2.275
N (1, 1)	0	2.013	2.013	2.013	2.045
	0.5	1.980	1.979	1.979	2.054
	1	1.956	1.953	1.955	2.113
	1.5	1.929	1.923	1.929	2.190

A mixed ANOVA was performed on average RMSDs as a function of three between-subject factors and one within-subject factor for the base and target group, separately. As the research questions involve the impact of equating method, target group ability distribution, number of common testlets, and testlet variance, the main effects of these factors as well as their 2-way, 3-way, and 4-way interactions were examined. Table 5 shows the between-subject and within-subject factors along with the levels of each factor.

Table 5. Mixed ANOVA factors and levels

Simulated Factors		Levels
Between-Subject Factors	Target group ability distribution	*N*(0, 0), *N*(0, 1)
	Number of common testlets	2, 3
	Testlet variance	0, 0.5, 1, and 1.5
Within-Subject Factor	Equating method	2-level HGLM, 3-level HGLM, Multilog_Rasch, Multilog_GRM

Table 6 presents the partial eta squared (η_p^2) for the ANOVA performed for the base and target group, respectively. As can be seen from Table 6, none of the three-way and four-way interaction terms are practically significant. For the base group, testlet variance and equating method, as well as their interaction term, had statistically and practically significant impact on the average RMSDs. The effect of testlet variance was moderate ($\eta_p^2 = .186$, p <.001), while the effects of equating method and the interaction term were large ($\eta_p^2 = .954$, p <.001 and $\eta_p^2 = .904$, p <.001, respectively). For the target group, in addition to the above factors, target group ability distribution and the interaction between target group ability distribution and testlet variance, also had statistically and practically significant impact on the average RMSDs. Equating method, target group ability distribution, and the interaction between equating method and testlet variance had a large effect ($\eta_p^2 = .955$, p <.001; $\eta_p^2 = .433$, p <.001; and $\eta_p^2 = .907$, p <.001, respectively), while testlet variance and the interaction between testlet variance and target group ability distribution had a small effect ($\eta_p^2 = .079$, p <.001; and $\eta_p^2 = .020$, p <.001). For both base and target group, the number of common testlets was found to have no significant impact on the average RMSDs. Because the factors that significantly impacted the average RMSDs were slightly different for the base and target group, the discussion will be based on the base and target group, separately.

Table 6. η_p^2 for mixed ANOVAs

Factor		Base group	Target group
Main Effects	Target group ability distribution (A)	-	*.433*
	Number of common testlets (N)	.001	-
	Testlet variance (V)	***.186***	***.079***
	Equating method (M)	***.954***	***.955***
Two-way interactions	A*N	-	-
	A*V	-	***.020***
	N*V	-	-
	A*M	-	.013
	N*M	-	.001
	V*M	***.904***	***.907***
Three-way interactions			
	A*N*V	-	-
	A*N*M	-	-
	A*V*M	.002	.003
	N*V*M	.004	-
Four-way interactions	A*N*V*M	-	.001

Note: Partial Eta-Square (η_p^2) is reported in the table.

-: indicates that the $\eta_p^2 < .001$

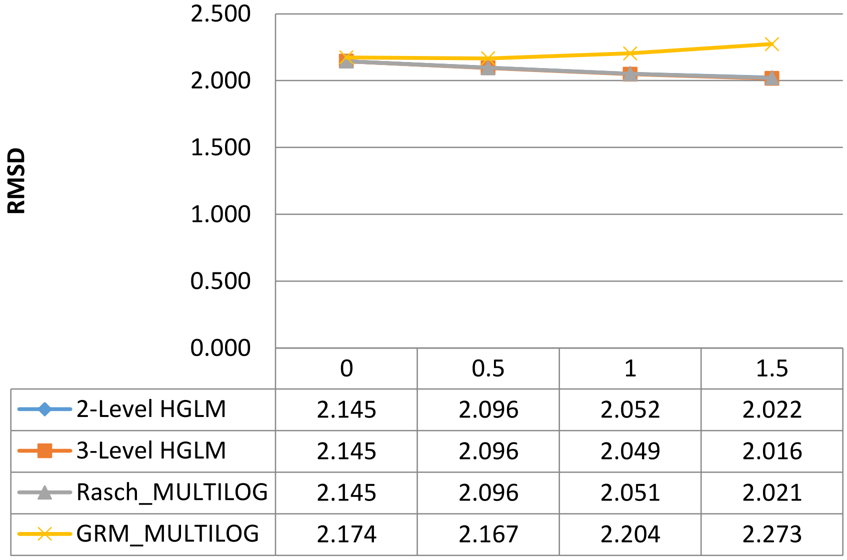
4.1.2 RMSDs for the base group

Figure 1 shows the two-way interaction between testlet variance and equating method for the base group. As can be seen from Figure 1, the patterns of differences on the average RMSDs across the four equating methods were different across the LID conditions. When there was no testlet effect ($\sigma_\gamma^2 = 0$), the two-level HGLM, three-level HGLM, and the Rasch calibration using MULTILOG yielded equivalent results. This finding suggests that the two HGLMs performed as well as the Rasch calibration using MULTILOG when tests were composed of independent items. In the low LID (i.e., $\sigma_{\gamma}^{2} = 0.5$) test conditions, the average RMSDs obtained using the three-level HGLM were almost identical to those obtained using the two-level HGLM and the Rasch calibration using MULTILOG. This finding indicates that when LID was present, the Rasch model tended to perform as well as the Rasch testlet model when the testlet variance was less or equal to 0.5.

As the degree of LID increased from low to high (i.e., $\sigma_{\gamma}^{2} = 0.5$ through 1.5), the differences in RMSDs between the three-level HGLM and the other methods gradually increased, with the three-level HGLM consistently yielding smaller RMSDs. However, the differences in RMSDs among the two-level, three-level HGLM, and Rasch calibration using MULTILOG were very small (less than 3% of the RMSDs). For example, as shown in Table 1, when $\sigma_{\gamma}^{2} = 1.5$ and the target group ability followed a standard normal distribution, the RMSDs for the two-level HGLM, Rasch calibration using MULTILOG, and the three-level HGLM were 2.025, 2.025, and 2.020, respectively. Overall, the above results indicate that when the Rasch model was applied for concurrent calibration, the HGLM approach and the MULTILOG estimation performed similarly, regardless of LID levels. Moreover, the Rasch model was quite robust to the violation of the local item independence assumption.

Across all LID conditions, the GRM calibration using MULTILOG consistently displayed the largest RMSDs among the four investigated methods. When LID was not present (i.e., $\sigma_{\gamma}^{2} = 0$), the average RMSDs produced by the GRM calibration using MULTILOG were slightly larger than those produced by the other three methods. The differences increased gradually as the degree of LID increased from low to high (i.e., $\sigma_{\gamma}^{2} = 0.5$ through 1.5), regardless of the target group ability distribution. The highest average RMSDs were found in the test with large testlet effects (i.e., $\sigma_{\gamma}^{2} = 1.5$) and when the GRM was used for calibration. These findings indicate that the equating errors were larger when the GRM versus Rasch model was used for calibration, with respect to ability recovery.

On comparing the average RMSDs across LID levels, the values of average RMSD slightly decreased for the HGLM approach and Rasch calibration using MULTILOG as σ_{γ}^{2} increased from 0 to 1.5. However, a different trend was observed for the other method. For the GRM calibration using MULTILOG, the average RMSDs dropped a little as σ_{γ}^{2} increased from 0 to 0.5 and then increased gradually as the degree of LID increased from low to high (i.e., $\sigma_{\gamma}^{2} = 0.5$ through 1.5).

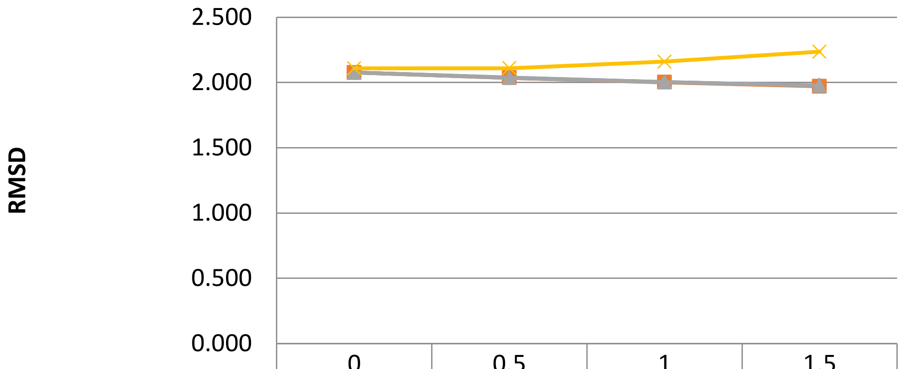


	0	0.5	1	1.5
2-Level HGLM	2.145	2.096	2.052	2.022
3-Level HGLM	2.145	2.096	2.049	2.016
Rasch_MULTILOG	2.145	2.096	2.051	2.021
GRM_MULTILOG	2.174	2.167	2.204	2.273

Figure 1. Two-way interaction between testlet variance and equating method (base group)

4.1.3 RMSDs for the target group

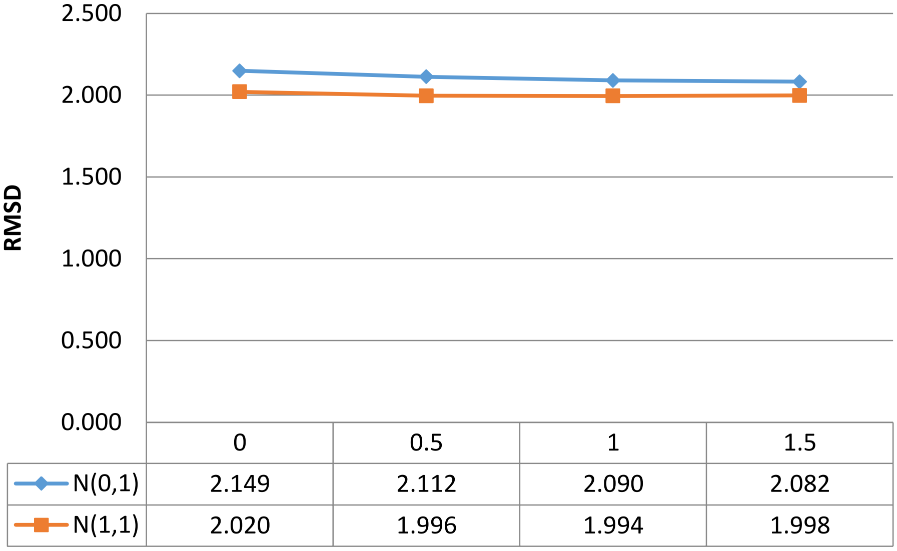
Figure 2 shows the two-way interaction between testlet variance and equating method for the target group. The patterns were similar to those shown in Figure 1: (1) across all LID conditions, the GRM calibration using MULTILOG consistently displayed the largest RMSDs among the four investigated methods, and the differences increased as the LID levels increased from 0 to 1.5; (2) for tests with zero or low degree of LID (i.e., $\sigma_{\gamma}^{2}=0$ or 0.5), the two-level HGLM, three-level HGLM, and the Rasch concurrent calibration using MULTILOG yielded almost equivalent results; (3) for tests with medium or high degree of LID (i.e., $\sigma_{\gamma}^{2}=1$ or 1.5), the three-level HGLM produced slightly smaller RMSDs than the two-level HGLM and Rasch concurrent calibration, but the differences were trivial; (4) the values of average RMSD decreased as σ_{γ}^{2} increased from 0 to 1.5 for the HGLM approaches and Rasch calibration using MULTILOG.



	0	0.5	1	1.5
◆ 2-Level HGLM	2.077	2.037	2.004	1.977
■ 3-Level HGLM	2.077	2.036	2.001	1.970
▲ Rasch_MULTILOG	2.077	2.036	2.004	1.977
× GRM_MULTILOG	2.108	2.109	2.159	2.236

Figure 2. Two-way interaction between testlet variance and equating method (target group)

Figure 3 shows the two-way interaction between testlet variance and target group ability distribution for the target group. As can be seen in Figure 3, the *N* (0, 1) target group ability distribution condition generated smaller average RMSDs than the *N* (1, 1) condition across all LID conditions. In addition, the difference in RMSDs slightly decreased as degree of LID increased from zero to high (i.e., $\sigma_{\gamma}^{2} = 0$ to 1.5). These findings in general indicate that the investigated equating methods tended to produce relatively better equating results for the target group if the base and target group differed somewhat in average ability, compared to the conditions where there was no group difference.



	0	0.5	1	1.5
N(0,1)	2.149	2.112	2.090	2.082
N(1,1)	2.020	1.996	1.994	1.998

Figure 3. Two-way interaction between testlet variance and target group ability distribution (target group)

As can be seen from Tables 1-4, similar values of average RMSD were observed in test conditions with two common testlets and those with three common testets for both base and target groups. This finding indicates that increasing the number of common items from 10 to 15 did not lead to more accurate estimation for ability parameters.

On comparing the average RMSDs across LID levels, different patterns were observed for the GRM approach under different ability distribution conditions. In condition with *N* (0, 1) target group ability distribution, similar patterns were found as those observed in the base group. The average RMSDs dropped a little as the testlet variance increased from 0 to 0.5 and then increased gradually as the degree of LID increased from low to high (i.e., $\sigma_{\gamma}^{2} = 0.5$ through 1.5). However, in conditions with *N* (1, 1) ability distribution, the average RMSDs gradually increased as the degree of LID increased from zero to high (i.e., $\sigma_{\gamma}^{2} = 0$ through 1.5).

4.2 RESULTS FROM SIMULATION STUDY 2

Simulation study 2 aimed to evaluate the performance of three IRT models (i.e., 3PL IRT model, 3PL TRT model, and 3PL multilevel TRT model) in equating testlet-based tests when examinees are nested within groups. This section compares the equating results from each of the three IRT models. The design factors included target group ability distribution (*N* (0.25, 1.1^2), *N* (0.5, 1.2^2)), degree of LID ($\sigma_{\gamma}^{2} = 0.5$, 1), and degree of LPD (ICC=0.25, 0.5), resulting in a total of 8 simulation conditions. To investigate the performance of the various IRT models on concurrent equating of testlet-based tests with person dependence, average correlations between estimated and true item and person parameter values and RMSDs were calculated across the 10 replications within each of the 8 simulation conditions and for each of the three investigated models. The average correlations and RMSD are presented in Tables 7 and 8 respectively for the item parameters and in Tables 9 and 10 for the person parameters.

4.2.1 Correlations of item parameters

The performance of each investigated model on test equating with the presence of item and person dependence can be evaluated by examining how well it recovers the true item parameter values. The comparison was done on the discrimination parameter estimates, difficulty parameter estimates, and the guessing parameter estimates. Specifically, correlations between estimated and true item parameter values and RMSDs of the item parameter estimates were calculated and averaged across the 10 replications within each of the 8 simulation conditions and for each of the three investigated models.

Table 7 presents the average correlations between the estimated and true item parameter values across 10 replications. Higher correlations indicate better estimation performance of the investigated model. In general, the estimates of item discrimination parameters from the 3PL multilevel TRT model and the 3PL TRT model were more highly correlated with the true discrimination parameters than the 3PL IRT model. The only exception was that in the low ICC and LID condition (i.e., ICC=0.25 and $\sigma_{\gamma}^{2} = 0.5$), the 3PL multilevel TRT model and the 3PL IRT model produced similar average correlations. This is consistent with the findings from Li (2009), which showed that the 3PL TRT model yielded discrimination parameter estimates that were more closely correlated to the true discrimination parameters if LID was ignored. As σ_{γ}^{2} increased from 0.5 to 1, the discrepancies between the 3PL IRT model the other two models also increased. This indicates that the degree of LID impacted the estimation of item discrimination parameters. Across all simulation conditions, the average correlations between the true and estimated item discrimination parameters from the 3PL multilevel TRT model and the 3PL TRT model were very similar. This indicates that ignoring person dependence did not impact the rank-order of item discrimination parameters. There was no noticeable difference on average correlations between the *N* (0.25, 1.1^2) and *N* (0.5, 1.2^2) ability condition.

Table 7. Average correlations between the estimated and true item parameter values across 10 replications

				Multilevel	Testlet	3PL
	Target Group Ability	ICC	σ_γ^2			
a	*N*(.25, 1.1^2)	.25	.50	.849	.867	.847
			1	.862	.870	.818
		.50	.50	.867	.868	.854
			1	.855	.862	.803
	N(.50, 1.2^2)	.25	.50	.866	.880	.861
			1	.847	.851	.806
		.50	.50	.881	.871	.852
			1	.860	.862	.826
b	*n*(.25, 1.1^2)	.25	.50	.970	.970	.969
			1	.968	.969	.961
		.50	.50	.969	.969	.967
			1	.968	.970	.967
	n(.50, 1.2^2)	.25	.50	.968	.967	.966
			1	.966	.965	.960
		.50	.50	.967	.965	.963
			1	.963	.965	.962
c	*n*(.25, 1.1^2)	.25	.50	.338	.332	.285
			1	.340	.337	.268
		.50	.50	.377	.360	.359
			1	.354	.371	.355
	n(.50, 1.2^2)	.25	.50	.332	.301	.263
			1	.289	.274	.286
		.50	.50	.320	.303	.311
			1	.254	.247	.227

The estimates of the difficulty parameters from the three models were all highly correlated with the true difficulty parameters ($r > .96$). In addition, these correlation values were very similar across simulation conditions and estimation models. These results indicate that the estimation of the item difficulty parameters was unaffected by the degree of LID, person dependence, as well as the target group ability distribution.

None of the investigated models was good at recovering the rank-order of the guessing parameters. The average correlations from the three models were low across all simulation conditions, ranging from 0.227 to 0.377. This is not surprising, given the small range of the guessing parameter values. The 3PL multilevel TRT model tended to produce guessing parameter estimates that were better correlated with the true parameter values than those of the 3PL IRT model. The N (0.25, 1.1^2) ability condition tended to result in slightly higher average correlations than the N (0.5, 1.2^2) ability condition. There was no consistent pattern of difference on average correlations across ICC and LID levels for all three models.

4.2.2 RMSDs of item parameters

Table 8 presents the average RMSDs of item parameters across 10 replications. For item discrimination parameters, the 3PL TRT model consistently produced the largest RMSDs among the three investigated models, while the 3PL multilevel TRT models tended to yield the smallest RMSDs among the three models. This finding indicates that the 3PL multilevel TRT model performed better than the 3PL TRT model and 3PL IRT model in estimating the discrimination parameters. Compared to the 3PL IRT model, the 3PL TRT model preserved better rank-order of the item discrimination parameters as shown in Table 7, while it also tended to result in more estimation errors.

Table 8. Average RMSDs of the item parameters across 10 replications

				Multilevel	Testlet	3PL
	Target Group Ability	ICC	σ_γ^2			
a	$N(.25, 1.1^2)$	.25	.50	.191	.225	.188
			1	.179	.209	.186
		.50	.50	.165	.214	.179
			1	.176	.209	.195
	$N(.50, 1.2^2)$	.25	.50	.170	.229	.190
			1	.186	.250	.197
		.50	.50	.165	.245	.196
			1	.185	.246	.191
b	$N(.25, 1.1^2)$	.25	.50	.275	.280	.295
			1	.313	.315	.361
		.50	.50	.285	.298	.312
			1	.291	.290	.311
	$N(.50, 1.2^2)$	.25	.50	.314	.333	.346
			1	.300	.309	.343
		.50	.50	.296	.320	.333
			1	.298	.307	.334
c	$N(.25, 1.1^2)$	.25	.50	.058	.056	.058
			1	.055	.054	.058
		.50	.50	.057	.056	.057
			1	.057	.054	.057
	$N(.50, 1.2^2)$	.25	.50	.057	.057	.058
			1	.059	.057	.058
		.50	.50	.058	.056	.056
			1	.059	.058	.059

The average RMSDs for the item difficulty parameters from the 3PL IRT model were slightly larger than those from the other two models. The average RMSDs from the 3PL multilevel TRT model were very similar to, if not smaller than, the 3PL TRT model. The *N* (0.25, 1.1^2) ability condition tended to result in slightly higher average RMSDs than the *N* (0.5, 1.2^2) ability condition for all three models. This finding indicates that smaller ability differences between the base and target group leads to better estimation of the item difficulty parameters. There were no consistent patterns of difference on average RMSDs across ICC and LID levels for all three models.

For guessing parameters, similar RMSDs were observed across all simulation conditions and estimation models, ranging from 0.054 to 0.059. This finding indicates that the estimation errors of the guessing parameters was unaffected by the degree of LID, person dependence, as well as the target group ability distribution.

4.2.3 Correlations of person parameters

The recovery of the ability parameters was evaluated in terms of correlations between the estimated and true ability parameter values and RMSDs. Table 9 presents the average correlations between the estimated and true ability parameters across 10 replications for the base and target groups. Figures 4-7 present the average correlations in graphical formats. Figures 4 and 5 show the average correlations for the base group under the N (0.25, 1.1^2) and the N (0.5, 1.2^2) ability condition, respectively. Figures 6 and 7 demonstrate the average correlations for the target group under each of the two ability conditions, respectively. From these figures we can see that the patterns of difference on average correlations across the ICC and LID levels were the same for the base and target group at each of the two ability conditions.

The average correlations for the ability parameters were generally high, ranging from .779 to .900. For both base and target group, the estimates of ability parameters from the 3PL multilevel TRT model were more closely correlated to the true ability parameters than the other two models, ranging from .807 to .900. The average correlations between the estimated and true ability parameters for the 3PL TRT model and 3PL IRT model were very similar, ranging from .779 to .875.

Table 9. Average correlations between the estimated and true person parameter values across 10 replications

				Multilevel	Testlet	3PL
	Target Group Ability	ICC	σ_{γ}^{2}			
Base group	N(.25, 1.1^2)	.25	.50	.840	.825	.824
			1	.812	.792	.789
		.50	.50	.843	.805	.803
			1	.838	.786	.783
	N(.50, 1.2^2)	.25	.50	.832	.817	.815
			1	.807	.788	.784
		.50	.50	.847	.809	.808
			1	.835	.788	.779
Target group	N(.25, 1.1^2)	.25	.50	.860	.851	.850
			1	.839	.827	.826
		.50	.50	.886	.855	.855
			1	.867	.828	.826
	N(.50, 1.2^2)	.25	.50	.876	.868	.868
			1	.853	.843	.841
		.50	.50	.900	.875	.875
			1	.884	.857	.852

In conditions with low degree of person dependence (i.e., ICC=0.25), the differences on average correlations among the three models were relatively small. For example, in condition with low item and person dependence (i.e., ICC=0.25, $\sigma_{\gamma}^{2} = 0.5$), and large ability group difference (i.e., N (0.5, 1.2^2) ability condition), the average correlations of the target group from the 3PL multilevel TRT model, 3PL TRT model, and the 3PL IRT model were .860, .851, and .850, respectively. As the degree of person dependence increased from moderate to high (i.e., ICC=0.25 to 0.5), the average correlations decreased for the 3PL TRT model and the 3PL IRT model. As a result, the differences on average correlations between the 3PL multilevel TRT model and the other two models also increased. These results indicate that ignoring person dependence led to ability estimates that were less closely correlated to the true ability parameters. In addition, the degree of person dependence impacted the rank-order of persons if person dependence was not taken into account.

As σ_{γ}^{2} increased from 0.5 to 1, the average correlations decreased for all three models. This finding indicates that the degree of LID impacted the estimation of ability parameters for all three models, regardless of whether or not the model takes into account LID. The average correlations from the 3PL TRT model and 3PL IRT model were very similar across LID levels, indicating that the 3PL IRT model was quite robust to the violation of local item independence assumption.

On comparing the average correlations between the base and target group, the three models tended to generate slightly higher correlations for the target group across all simulation conditions. For example, in condition with low item and person dependence (i.e., ICC=0.25, $\sigma_{\gamma}^{2} = 0.5$) and small ability group difference (i.e., the *N* (0.25, 1.1^2) ability condition), the average correlations ranged from .824 to .840 for the base group, and from .850 to .860 for the target group.

On comparing the average correlations between the two ability conditions, there was not much difference for the base group. However, for the target group, the *N* (0.5, 1.2^2) ability condition tended to result in higher average correlations than the *N* (0.25, 1.1^2) ability condition. This finding indicates that as the base and target group differed more on ability distribution, the investigated models tended to perform better in terms of ability parameter recovery for the target group.

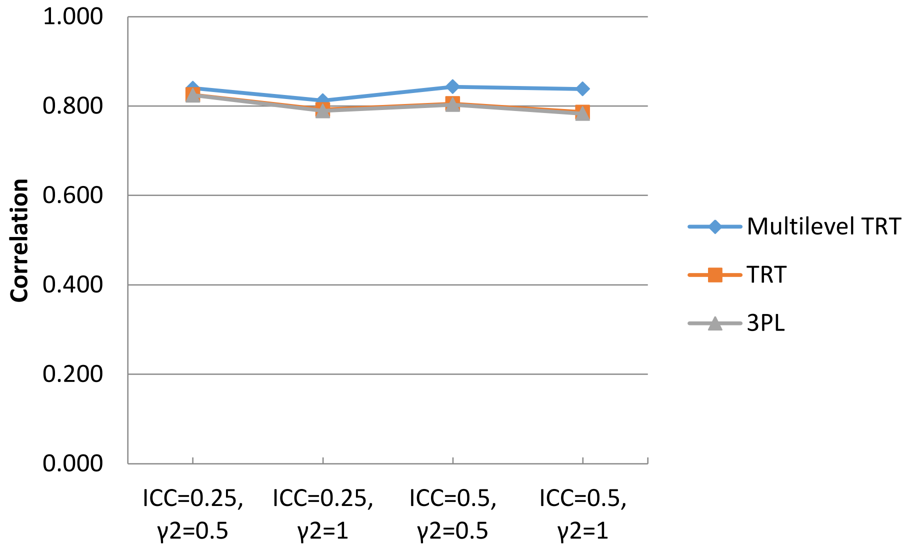


Figure 4. Average correlations between the estimated and true person parameters across 10 replications (Base group, *N* (0.25, 1.1^2) condition)

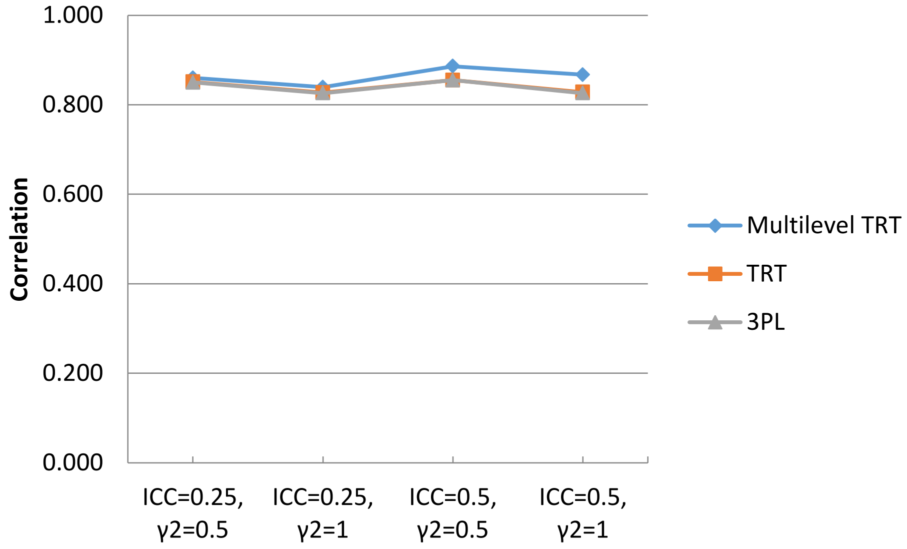


Figure 5. Average correlations between the estimated and true person parameters across 10 replications (Target group, *N* (0.25, 1.1^2) condition)

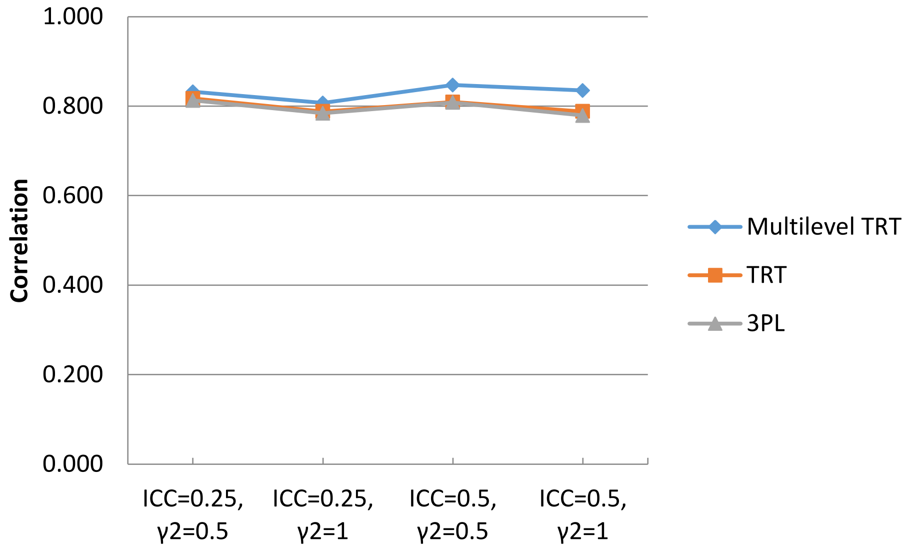


Figure 6. Average correlations between the estimated and true person parameters across 10 replications (Base group, *N* (0.5, 1.2^2) condition)

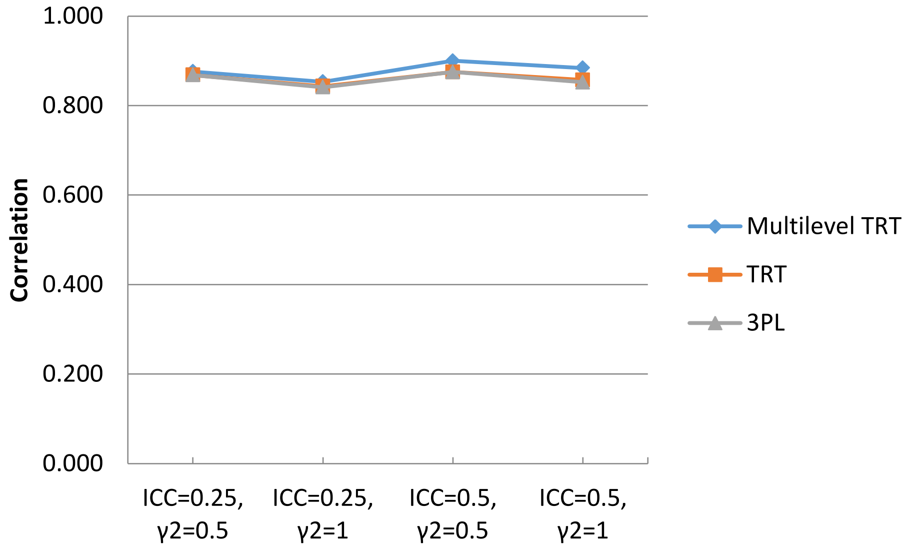


Figure 7. Average correlations between the estimated and true person parameters across 10 replications (Target group, *N* (0.5, 1.2^2) condition)

4.2.4 RMSDs of person parameters

Table 10 presents the average RMSDs of the ability parameters across 10 replications for the base and target group. Figures 8-11 present the average RMSDs in graphical format for the base and target group in each of the two ability conditions, respectively. From these figures we can see that the patterns of difference on average RMSDs across the ICC and LID levels were the same for the base and target group at each of the two ability conditions.

Table 10. Average RMSDs of the person parameters across 10 replications

				Multilevel	Testlet	3PL
	Target Group Ability	ICC	σ_{γ}^{2}			
Base Group	$N(.25, 1.1^2)$	.25	.50	.550	.575	.576
			1	.594	.623	.628
		.50	.50	.521	.573	.576
			1	.550	.620	.626
	$N(.50, 1.2^2)$	.25	.50	.559	.584	.586
			1	.585	.610	.618
		.50	.50	.513	.567	.570
			1	.554	.617	.631
Target Group	$N(.25, 1.1^2)$	.25	.50	.568	.587	.586
			1	.599	.621	.618
		.50	.50	.542	.605	.602
			1	.570	.644	.641
	$N(.50, 1.2^2)$	.25	.50	.585	.613	.610
			1	.623	.656	.655
		.50	.50	.556	.634	.630
			1	.578	.658	.666

For both base and target groups, the average RMSDs from the 3PL multilevel TRT model were smaller than those from the other two models, ranging from .513 to .623. The average RMSDs from the 3PL TRT model and 3PL IRT model were very similar, ranging from .567 to

.656. This finding indicates that with the presence of item and person dependence, the 3PL multilevel TRT model resulted in more accurate estimation of the ability parameters than the other two models.

In conditions with low degree of person dependence (i.e., ICC=0.25), the differences on average RMSDs among the three models were relatively small. As the degree of person dependence increased from moderate to high (i.e., ICC=0.25 to 0.5), the average RMSDs also increased for the 3PL TRT model and the 3PL IRT model. As a result, the differences on average RMSDs between the 3PL multilevel TRT model and the other two models also increased. For example, in condition with low degree of item and person dependence (i.e., ICC=0.25, $\sigma_{\gamma}^{2} = 0.5$) and large ability group difference (i.e., *N* (0.5, 1.2^2) ability condition), the average RMSDs for the target group from the 3PL multilevel TRT model, 3PL TRT model, and the 3PL IRT model were .568, .587, and .586, respectively. As ICC increased from 0.25 to 0.5, the average RMSDs for the target group from these three models were 0.542, 0.605, and 0.602, respectively. These results indicated that ignoring person dependence led to higher estimation errors for the ability parameters. In addition, the degree of person dependence impacted the person parameter estimation if the person clustering effects were not taken into account.

As the testlet variance increased from 0.5 to 1, the average RMSDs increased for all three models. This finding indicates that the degree of LID impacted ability estimation for all three models, regardless of whether or not LID was taken into consideration. The average RMSDs from the 3PL TRT model and 3PL IRT model were similar across LID levels, indicating that the 3PL IRT model was quite robust to the violation of local item independence assumption, in terms of ability parameter recovery.

On comparing the average RMSDs between the base and target group, the three models tended to generate slightly higher average RMSDs for the target group across all simulation conditions. For example, in condition with low item and person dependence (i.e., ICC=0.25, $\sigma_{\gamma}^{2}=0.5$) and small ability group difference (i.e., the *N* (0.25, 1.1^2) ability condition), the average RMSDs ranged from .550 to .576 for the base group, and from .568 to .587 for the target group.

Similar to the patterns observed in Table 9, there was not much difference on average RMSDs across the two ability conditions for the base group. However, for the target group, the *N* (0.5, 1.2^2) ability condition tended to result in higher average RMSDs than the *N* (0.25, 1.1^2) ability condition. This finding indicates that as the group difference on ability distribution increased, the three models yielded more estimation errors, especially for the target group.

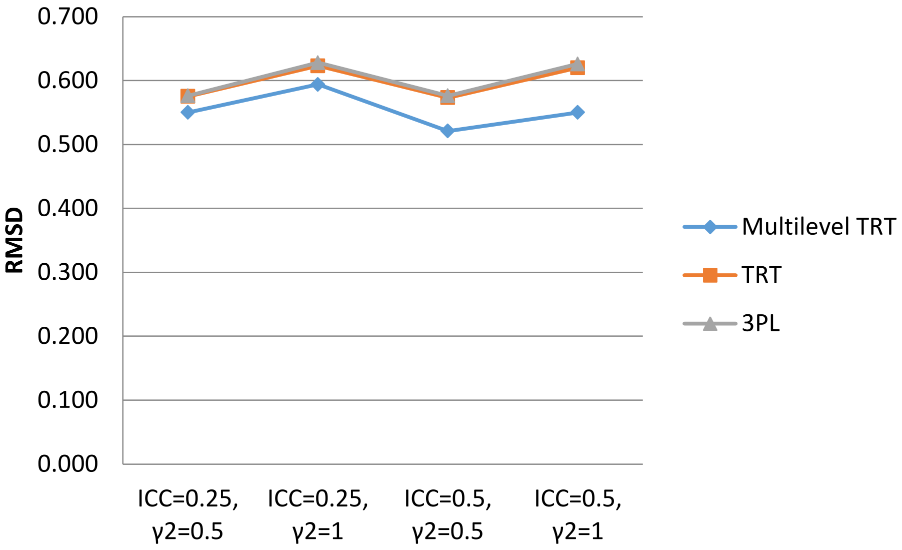


Figure 8. Average RMSDs of the person parameters across 10 replications (Base group, *N* (0.25, 1.1^2) condition)

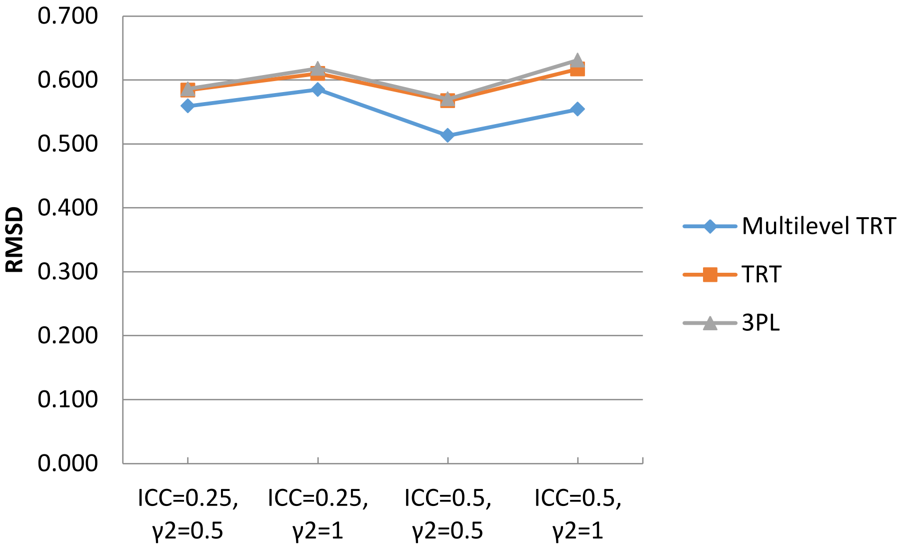


Figure 9. Average RMSDs of the person parameters across 10 replications (Target group, *N* (0.25, 1.1^2) condition)

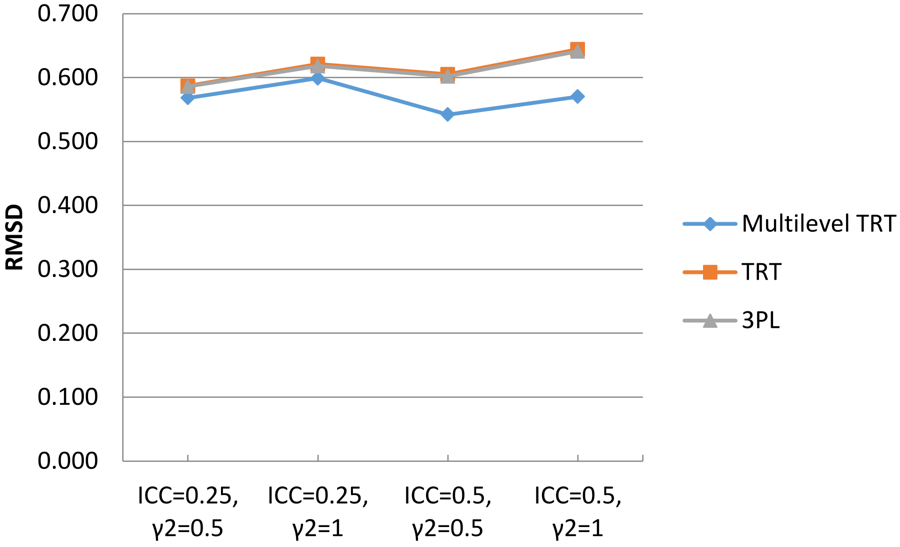


Figure 10. Average RMSDs of the person parameters across 10 replications (Base group, *N* (0.5, 1.2^2) condition)

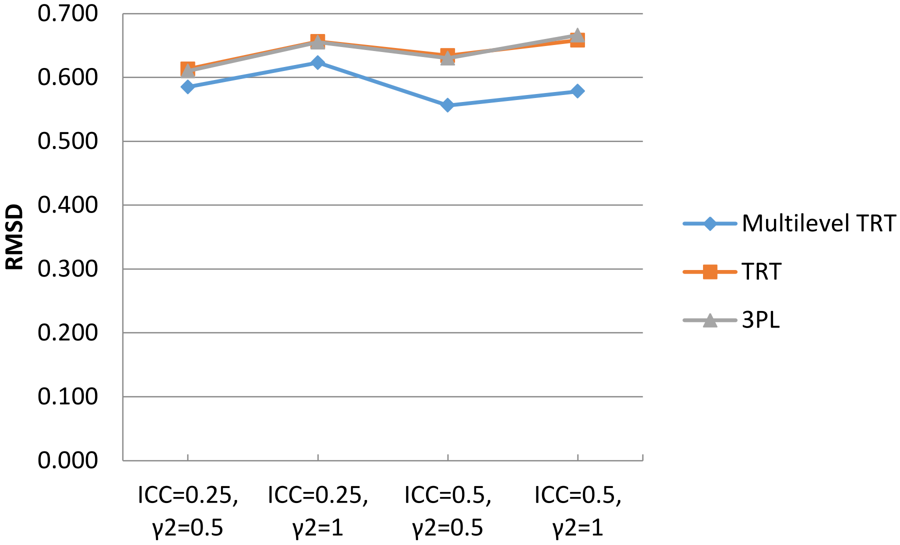


Figure 11. Average RMSDs of the person parameters across 10 replications (Target group, *N* (0.5, 1.2^2) condition)

5.0 DISCUSSION

The current work evaluates the performance of multilevel models on test equating under the anchor test design with the presence of varying degree of item and person dependence. The equating performance of the multilevel model was also compared to the dichotomous IRT model, polytomous IRT model, and TRT model through two simulation studies. The first section of this chapter revisits the research questions and summarizes the major findings of this work. The second section discusses some of the issues in the findings of this work. The last section addresses the limitation of the current work and provides future research directions.

5.1 SUMMARY OF MAJOR FINDINGS

5.1.1 Simulation Study 1

The first simulation study considers testing situations where responses to testlet items were collected from independent examinees. A three-level HGLM was proposed as a concurrent equating model to accommodate potential LID existing in testlet response data. HGLM was found to be useful in modeling hierarchical data structure due to item dependency. Simulation study 1 extended previous research to the use of test equating for testlet-based tests with the presence of varying degree of LID. The equating accuracy of the three-level HGLM was also compared with a two-level HGLM that ignores LID, the Rasch concurrent calibration using MULTILOG, and the GRM concurrent calibration using MULTILOG. Three factors are examined in this simulation study: (a) target group ability distribution (*N* (0, 1) and *N* (1, 1)); (b) degree of LID ($\sigma_{\gamma}^{2} =$0, 0.5, 1, and 1.5); and (3) number of common testlets (2 and 3), resulting in a total of 16 simulation conditions. Concurrent equating was conducted for each of these 16 simulation conditions when each of the four investigated methods was applied. The average RMSDs of expected scores across replications were computed. A summary of the findings are presented below.

5.1.1.1 Research Question 1

How well does the proposed HGLM concurrent equating method recover model parameters with the presence of LID?

In the first research question, the performance of the proposed HGLM concurrent equating method is compared with the concurrent calibration using MULTILOG, in terms of person parameter recovery. Two HGLM concurrent equating models are examined: (1) the three-level HGLM that models LID, and (2) the two-level HGLM that ignores LID. Two MULTILOG concurrent calibration models are included: (1) the Rasch model; and (2) the GRM. Conclusions on the comparison of these four equating methods are discussed in the following three sections.

Research Question 1.1: *How well does the proposed two-level HGLM concurrent equating method recover model parameters, compared to the Rasch concurrent calibration with the presence of LID?*

Because the two-level HGLM is based on the Rasch model, this research question focuses on the comparison between the use of multilevel modeling approach and the traditional IRT modeling on test equating. The results from Study 1 show that these two approaches provide equivalent equating results across all simulation conditions. This indicates that the two-level HGLM concurrent equating method is an effective alternative to the Rasch concurrent calibration using MULTILOG.

Research Question 1.2: *Does the proposed three-level HGLM concurrent equating method, which accounts for the testlet effects, provide more accurate results than the two-level HGLM concurrent equating method with the presence of LID?*

This research question focuses on the comparison between the two HGLMs as concurrent equating models. The results from the first simulation study suggest that the two HGLMs perform very similarly across all simulation conditions. When the degree of LID is zero or low (i.e., $\sigma_{\gamma}^{2} = 0$ or 0.5), the two HGLMs provide equivalent equating results. When the testlet variance is as high as 1 or above, the three-level HGLM tends to produce slightly smaller RMSDs than does the two-level HGLM. However, the difference is negligible (only in the third decimal place). Therefore, compared with the two-level HGLM, the three-level HGLM that accommodates item dependency within a testlet does not seem to produce better equating results with the presence of LID.

Research Question 1.3: *For concurrent calibration using MULTILOG, does the use of polytomous scoring based on the GRM provide more accurate equating results than Rasch concurrent calibration with the presence of LID?*

This research question focuses on the comparison between the uses of dichotomous versus polytomous scoring with the presence of LID. Although suggested in some previous studies (e.g., Chen, 2014; Lee et al, 2001), the superiority of polytomous scoring over dichotomous scoring with the presence of LID is not observed in the current study for expected score recovery. Instead, the dichotomous scoring approach is found to provide more accurate equating results. Across all LID conditions, the GRM concurrent calibration using MULTILOG consistently yield larger average RMSDs of expected scores than the Rasch concurrent calibration using MULTILOG. When tests are composed of independent items (i.e., $\sigma_{\gamma}^{2} = 0$), the difference on average RMSDs between the two approaches is found to be small. As the testlet variance increases, the average RMSDs for the GRM concurrent calibration is found to be increasingly higher than the corresponding values for the Rasch concurrent calibration.

In summary, the HGLM concurrent equating approach performs as well as, if not better than, the Rasch concurrent calibration using MULTILOG. This suggests that the HGLM is an effective way of conducting concurrent calibration. The following are the answers to research question 1: (1) The two-level HGLM concurrent equating method and the Rasch concurrent calibration using MULTILOG perform equally well at different LID conditions; (2) the two-level HGLM and the three-level HGLM perform similarly at different LID conditions; (3) the GRM concurrent calibration using MULTILOG provides the less accurate estimates of expected scores than the Rasch concurrent calibration using MULTILOG across different LID conditions.

5.1.1.2 Research Question 2

For each of the four investigated equating methods, what is the impact of degree of LID on equating results?

The second research question investigates the robustness of the four equating methods to the violation of local item independence assumption. The results from simulation study 1 show that the average RMSDs from the two HGLM methods and the Rasch concurrent calibration using MULTILOG slightly decrease as the degree of LID increases, whereas the differences are negligible across LID levels. When LID is not present or low (i.e., $\sigma_{\gamma}^{2} = 0$ or 0.5), the average RMSDs from these three methods are found to be almost identical. With the presence of medium or high degree of LID (i.e., $\sigma_{\gamma}^{2} = 1$ or 1.5), the average RMSDs from the three-level HGLM concurrent equating are found to be slightly smaller than the other two methods. However, the differences on the average RMSDs are trivial. These findings imply that the three equating methods based on dichotomous IRT models all perform well with the presence of LID, in terms of expected score recovery.

The results from the GRM concurrent calibration using MULTILOG are mixed. When the degree of LID is low (i.e., $\sigma_{\gamma}^{2} \leq$ 0.5), concurrent calibration based on the GRM is robust to the violation of local item independence assumption. This finding is illustrated by that in most test conditions, the average RMSDs decrease as the testlet variance increases from 0 to 0.5. However, when the degree of LID is moderate or high (i.e., $\sigma_{\gamma}^{2} \geq 1$), the average RMSDs tend to increase as the degree of LID increases.

Thus, the answers to the second research question are: (1) the two HGLM methods and the Rasch concurrent calibration using MULTILOG are robust to the violation of the local independence assumption, in terms of recovery of the expected scores; (2) for the GRM concurrent calibration using MULTILOG, the increase in testlet variance tends to reduce or have little impact on equating errors when the degree of LID is low (i.e., $\sigma_{\gamma}^{2} \leq 0.5$), but tends to increase equating errors when the LID level is moderate or high (i.e., $\sigma_{\gamma}^{2} \geq 1$).

5.1.2 Simulation Study 2

The second simulation study considers testing situations where responses to testlet items were collected from examinees nested within groups. A 3PL multilevel TRT model was proposed to simultaneously account for item and person dependence. The equating accuracy of the 3PL multilevel TRT model was also compared with the 3PL TRT model that ignores person dependence, and the traditional 3PL IRT model that ignores both item and person dependence. Three factors were examined in this simulation study: (a) target group ability distribution (*N* (0.25, 1.1^2) and *N* (0.5, 1.2^2)); (b) degree of LID ($\sigma_{\gamma}^{2} = 0.5$ and 1); (3) degree of person dependence (ICC=0.25 and 0.5), resulting in a total of 8 simulation conditions. Concurrent calibration using Bayesian estimation was conducted for each of these 8 simulation conditions when each of the three investigated models was applied. The average correlations and RMSDs of the item and person parameters across replications were computed within each of these 8 simulation conditions and for each of the three investigated models. A summary of the findings are presented below.

5.1.2.1 Research Question 1

How well does the 3PL multilevel TRT model recover the item and person parameters under the NEAT design, compared to the 3PL TRT model and the traditional 3PL IRT model, with the presence of LID and person dependence?

To answer the first research question, the average correlations and RMSDs of the item and person parameters from the 3PL multilevel TRT model are compared with those from the 3PL TRT model and the 3PL IRT model. Discussion on the comparison of the performance of these three models is based on the items and person parameters, respectively.

Item Parameters

The results from the second simulation study suggest that the 3PL multilevel TRT model provides more accurate estimation for the discrimination parameters, compared to the other two models. This finding is illustrated by the following: (1) the estimates of the discrimination parameters from the 3PL multilevel TRT models are more closely correlated to the true discrimination parameters than the 3PL IRT model; (2) the average RMSDs from the 3PL multilevel TRT model are smaller than those from the other two models; (3) even though the average correlations from the 3PL TRT model and the 3PL multilevel TRT model are similar, the former generates the largest RMSDs among the three models across all simulation conditions.

As to the item difficulty parameters, both the 3PL multilevel TRT model and the 3PL TRT model provide more accurate estimation than the 3PL IRT model. Although the average correlations from the three models are similar, the average RMSDs from the 3PL multilevel TRT model and the 3PL TRT model are slightly smaller than those from the 3PL IRT model.

The three models provide similar performance in terms of guessing parameter recovery. This finding is illustrated by the following: (1) none of the models provides guessing parameter estimates that are close to the true guessing parameters, which is likely due to the small range of the guessing parameter values; (2) in each of the 8 simulation conditions, the three models generate similar average RMSDs.

Person Parameters

Among the three models, the 3PL multilevel TRT model provides the most accurate estimation for the person parameters. This finding is illustrated by the following: (1) the average correlations between the estimated and true person parameters from the 3PL multilevel TRT model are higher than those from the other two models; (2) the average RMSDs of the person parameters from the 3PL multilevel TRT model are smaller than those of the other two models. The 3PL TRT model and 3PL IRT model perform similarly in terms of average correlations and RMSDs.

In summary, the results from the second simulation study suggest that with the presence of both item and person dependence: (1) the 3PL multilevel TRT model provides more accurate estimation for the item discrimination parameter and person parameter than the other two models; (2) the 3PL IRT model provides less accurate estimation for the item difficulty parameter than the other two models; (3) the three models perform similarly in terms of guessing parameter recovery.

5.1.2.2 Research Question 2

For each of the three investigated models, what is the impact of degree of LID on equating results?

The second research question investigates the robustness of the three investigated models to the violation of local item independence assumption. The results from the second simulation demonstrate that for all three models, the estimation of person parameters are affected by the degree of LID. As σ_{γ}^{2} increases from 0.5 to 1, the average correlations decrease and the average RMSDs increase for each of the three models. Moreover, the degree of LID also impacts the accuracy of discrimination parameter estimates for the 3PL IRT model, with a higher degree of LID (i.e., $\sigma_{\gamma}^{2} = 1$) associated with lower average correlations.

In summary, as the testlet variance increases from 0.5 to 1, the accuracy of person parameter estimates decreases for all three models. In addition, increasing the degree of LID decreases the correlations between the true and estimated discrimination parameters for the 3PL model.

5.1.2.3 Research Question 3

For each of the three investigated models, what is the impact of degree of person dependence on equating results?

The third research question investigates the robustness of the three investigated models to the violation of local person independence assumption. The results from simulation study 2 are mixed. On the one hand, as the ICC increases from 0.25 to 0.5, the discrepancies on average correlations and RMSDs between the 3PL multilevel TRT model and the other two models increase, suggesting that the two models that do not account for person dependence (i.e., the 3PL

TRT model and the 3PL IRT model) are not robust to the violation of local person independence assumption. On the other hand, slightly higher average correlations and lower average RMSDs are observed in conditions with high person dependence (i.e., ICC=0.5) for the 3PL multilevel TRT model. And no consistent patterns are observed for the other two models. The estimation of item parameters, however, is not affected by the level of person dependence.

5.2 DISCUSSION OF FINDINGS

5.2.1 Impact of LID on equating accuracy

The results from the second simulation study demonstrate that the difficulty and discrimination parameters were better estimated if LID was taken into consideration. Compared to the 3PL IRT model, the 3PL multilevel TRT model and the 3PL TRT model yield smaller RMSDs of the difficulty parameters and higher average correlations between the estimated and true discrimination parameters. Moreover, a higher degree of LID leads to lower average correlations between the estimated and true item discrimination parameters for the 3PL IRT model.

In regards to person parameter recovery, the results from the two simulation studies are consistent. The findings from the first simulation study show that ignoring LID does not impact the estimation of expected scores for the dichotomous models. The equating method that accounts for LID (i.e., the three-level HGLM concurrent equating) provides very similar average RMSDs of expected scores, compared to the methods that ignore LID (i.e., the two-level HGLM concurrent equating and the Rasch concurrent calibration). Consistent with Li (2009), the findings from the second simulation study also suggest that person parameters could be well estimated if LID is ignored. The 3PL IRT model and the 3PL TRT model perform similarly at different LID conditions, in terms of average correlations and RMSDs of person parameter estimates.

The two simulation studies provide different findings regarding the impact of LID levels on equating results. In the first simulation study, degree of LID has little impact on RMSDs of expected scores for the HGLMs and Rasch concurrent calibration using MULTILOG. In contrast, a higher degree of LID leads to slightly lower average correlations and higher RMSDs of the person parameters for all three models in the second simulation study.

This inconsistency could be due to the different evaluation criteria used in the two simulation studies: equating accuracy is evaluated based on the RMSDs of expected scores in the first simulation study, whereas correlations and RMSDs of person parameters are examined in the second simulation study. As explained in Chapter 3, expected scores were computed and compared across the four equating procedures in the first simulation study because the person parameter estimates obtained from two different software packages (i.e., SAS and MULTILOG) may not be on the same scale. In the second simulation study, all the three models were calibrated in WinBUGS, and therefore the person parameter estimates were compared directly between models. As presented in Chapter 3, the calculation of expected scores is based on both item and person parameter estimates. Therefore, the equating process in simulation study 1 actually involves two stages: (1) concurrent calibration of the item and person parameters; and (2) calculation of the expected scores based on the item and person parameter estimates obtained from the first stage. For the Rasch model, the estimation of expected scores depends on the accuracy of $(\theta - b)$. Therefore, what really matters is not the accuracy of individual difficulty or person parameter estimates, but the magnitude and direction of the bias in the difficulty and person parameter estimates. Future studies may compare these equating methods on recovering the true difficulty parameters and person parameters separately by referring to the true values used for data generation.

This inconsistency might also be caused by the model difference: the Rasch model was used in the first simulation study whereas the 3PL model was used in the second simulation study. It has been shown in previous studies (e.g., Bradlow, Wainer, & Wang, 1999; Wainer, Bradlow, & Du, 2000) and the second simulation of the current work that the degree of LID impacts the discrimination parameter estimation. HGLM models only difficulty parameters, and thus may be less sensitive to the ignoring of modelling testlet effects.

5.2.2 The performance of GRM

It is often considered that fitting the testlet-based tests with a polytomous IRT model bypasses the assumption of local item independence (Lee et al., 2001), and therefore may lead to better measurement outcomes. However, by using the testlet score as the unit of analysis, item-level information is lost (Zenisky, Hambleton, & Sireci, 2002; Sireci, Thissen, & Wainer, 1991). Moreover, when the degree of LID is high, the testlet scoring approach may not be appropriate (Wainer, 1995). The results from the first simulation study of this work suggest that the GRM concurrent calibration does not perform as well as the other three methods based on the dichotomous IRT models, especially with high degree of LID. This coincides with what was found in Cao, Lu, and Tao (2014). Using the random groups design, Cao, Lu, and, Tao (2014) conducted number-correct score equating by linking separate calibration to compare the performance of three measurement models: the 2PL IRT model, the 2PL TRT model, and the GRM. The results show that the 2PL IRT model and the 2PL TRT model performed similarly across various LID conditions, in terms of raw-to-raw score conversions, suggesting that 2PL IRT model is quite robust to the violation of local item independence. In addition, the 2PL IRT model and the 2PL TRT model produce equating results that are more consistent with those of the equipercentile method than does the GRM.

On the contrary, some other studies (e.g., Lee's et al., 2001) demonstrated that fitting the testlet-based tests with a polytomous IRT model in general outperformed the dichotomous IRT model in IRT true and observed score equating. This inconsistency might be due to different testlet lengths, degree of LID, as well as evaluation criteria used in the two studies. In Lee's et al. (2001) study, the polytomous equating approach and the dichotomous equating approach were compared in terms of their agreement to the results from three baseline equating methods (i.e., mean, linear, and equipercentile methods) on equated score distribution moments. In the current work, the GRM concurrent calibration is compared to the other three methods, in terms of the recovery of expected scores. This inconsistency might also be caused by different models used in these two studies: the Rasch model was used in the first simulation study of the current work whereas the 3PL IRT model was used in Lee et al.'s (2001) study.

5.2.3 Precision of measures

The effects of LID levels on equating results were further investigated by examining the accuracy of proficiency estimates. The average standard errors of person parameter estimates from the first simulation study were compared across the four equating procedures. Consistent with previous studies (e.g., Bradlow, Wainer, & Wang, 1999; Wainer, Bradlow, & Du, 2000), the results from this follow-up study also demonstrate that treating testlet items as independent items overestimates the precision of ability estimates. Moreover, the magnitude of the overstatement increases as the degree of LID increases. Tables 11 and 12 present the average standard errors of person parameter estimates from one replication within each of the 16 simulation conditions for the base and target groups, respectively. As illustrated in these two tables, when there is no testlet effect, each of the four equating approaches provide reasonable and comparable estimates of precision of person parameter estimates. As the degree of LID increases, the average standard errors for the two approaches that accounts for LID (i.e., three-level HGLM and GRM concurrent calibration) also increase monotonically, whereas they remain relatively stable for the two approaches that ignore LID (i.e., two-level HGLM and Rasch concurrent calibration using MULTILOG).

Table 11. Average standard errors of the person parameter estimates from one replication (Base group)

Target Group Ability	# of Common Testlet	Testlet Variance	2-Level HLM	3-Level HLM	Multilog_ Rasch	Multilog _GRM
N (0, 1)	2	0	0.391	0.391	0.410	0.411
		0.5	0.392	0.455	0.401	0.459
		1	0.387	0.500	0.416	0.506
		1.5	0.373	0.535	0.415	0.577
	3	0	0.401	0.401	0.389	0.390
		0.5	0.390	0.459	0.403	0.457
		1	0.379	0.485	0.427	0.529
		1.5	0.380	0.528	0.432	0.556
N (1, 1)	2	0	0.391	0.391	0.400	0.402
		0.5	0.396	0.449	0.402	0.450
		1	0.376	0.483	0.433	0.533
		1.5	0.382	0.551	0.415	0.550
	3	0	0.390	0.390	0.404	0.393
		0.5	0.394	0.458	0.406	0.458
		1	0.381	0.498	0.407	0.518
		1.5	0.382	0.551	0.412	0.550

5.2.4 The performance of multilevel models

This research demonstrates the performance of multilevel models on test equating under the anchor test design when local dependence is present. As mentioned earlier, data in educational research frequently have a hierarchical structure, while traditional IRT models do not take the clustering effects into consideration. The multilevel modeling approach takes into account both item and person dependence, and therefore can appropriately analyze such data without violating the local independence assumption.

Table 12. Average standard errors of the person parameter estimates from one replication (Target group)

Target Group Ability	# of Common Testlet	Testlet Variance	2-Level HLM	3-Level HLM	Multilog_ Rasch	Multilog_ GRM
N (0, 1)	2	0	0.390	0.390	0.409	0.391
		0.5	0.395	0.458	0.404	0.458
		1	0.381	0.495	0.409	0.518
		1.5	0.377	0.538	0.420	0.540
	3	0	0.397	0.397	0.385	0.375
		0.5	0.391	0.460	0.404	0.471
		1	0.377	0.484	0.426	0.527
		1.5	0.377	0.526	0.429	0.571
N (1, 1)	2	0	0.419	0.419	0.429	0.410
		0.5	0.414	0.465	0.421	0.458
		1	0.387	0.491	0.447	0.553
		1.5	0.406	0.566	0.441	0.576
	3	0	0.400	0.400	0.414	0.410
		0.5	0.427	0.486	0.440	0.492
		1	0.401	0.513	0.429	0.517
		1.5	0.395	0.560	0.426	0.559

The results from the two simulation studies show that the multilevel modelling approach is an effective and flexible tool for test equating, allowing both item and person dependence structure to be accommodated. Specifically, the findings from the first simulation study demonstrate that the HGLM approach performs as well as the IRT concurrent equating approach. Because the HGLM approach can be conducted in statistical software packages such as SAS and HLM, it is an alternative method of Rasch equating for researchers and practitioners who are not familiar with IRT. The findings from the second simulation study suggest that multilevel models can be a better choice than the traditional IRT model or TRT model in test equating situations where a relatively high level of person dependence (i.e., ICC=0.5) is present.

5.3 LIMITATION AND FUTURE RESEARCH DIRECTIONS

This research used two simulation studies to address the proposed research questions. Hence, the results may not be generalizable to other situations not considered in the current study as the findings are limited to the specific conditions in these simulation studies. For example, in both simulation studies, testlet length is a fixed factor and not manipulated. Previous research (e.g., Chen, 2014) suggests that the effect of LID on equating results varies by the number of items within a testlet. Future studies may include testlet length as a design factor.

In the second simulation study, only two levels of LID (i.e., $\sigma_{\gamma}^{2} = 0.5$ and 1) and two levels of person dependence (i.e., ICC=0.25 and 0.5) were simulated. It would be very interesting and meaningful in future studies to include more LID and ICC conditions, especially those with zero and low levels of LID and person dependence.

In both simulation studies, testlet variance was generated to be constant across testlets. However, for real test data, different testlets may exhibit different amounts of within-testlet dependence. Future simulation studies may consider varying the degree of LID across testlets to reflect the dependence structure of testlet items in real testing situations.

Due to the limitations on computer capacity and time-consuming analyses using the Bayesian estimation procedure, only ten replications were run for the second simulation study. Previous studies (e.g., Jiao et al., 2012; Jiao, Wang, & He, 2013; Jiao & Zhang, 2014), however, suggest more replications might be desirable for estimating multilevel IRT models. In Jiao, Wang, & He (2013), twenty-five replications were found to be adequate to estimate the Rasch testlet model. The twenty-five replications were also implemented in Jiao et al. (2012) to estimate one-parameter TRT models that accounts for person clustering effects, and in Jiao and Zhang (2014) to estimate polytomous multilevel testlet models that accounts for person clustering effects. For future simulation studies, it is desired to run more replications to more accurately evaluate the equating performance of the multilevel TRT models. Additional analyses suggest that ten replications may be adequate for the current study. Tables 13 and 14 present the variances of posterior estimates for the item and person parameters, respectively, across ten replications from one simulation condition as an example. It can be seen from the tables that the variance of posterior estimates stays quite stable across the ten replications.

In the current work, the concurrent calibration method was used for placing the item and person parameters on the same scale for all investigated equating models. In practice, linking separate calibration has been widely used. The linking separate calibration procedures have been extended to the TRT models (e.g., Li, Bolt, & Fu, 2005; Li, 2009; Zhang, 2010), but not to any multilevel models yet. Using the concurrent calibration method might simplify the equating process by avoiding the need of scale transformation. However, it has been shown to be less robust to the violation of IRT model assumptions. Therefore, the results from the current work may not be generalizable to situations using linking separate calibration. Future research may compare the effectiveness of concurrent calibration methods investigated in the current work to linking separate calibration when different IRT models are applied.

Table 13. The standard errors of posterior item parameter estimates from one simulation condition (ICC=0.5, $\sigma_{\gamma}^{2} = 1$, N (0.25, 1.1^2) target group ability distribution)

		Model		
	Replication	Multilevel	Testlet	3PL
a	1	.235	.264	.194
	2	.218	.235	.177
	3	.219	.228	.178
	4	.216	.231	.181
	5	.223	.234	.179
	6	.203	.226	.178
	7	.233	.240	.190
	8	.220	.242	.194
	9	.217	.228	.184
	10	.225	.251	.198
b	1	.360	.316	.303
	2	.343	.305	.300
	3	.367	.337	.338
	4	.374	.321	.320
	5	.383	.358	.356
	6	.360	.317	.312
	7	.341	.316	.312
	8	.349	.300	.294
	9	.344	.312	.309
	10	.377	.329	.319
c	1	.069	.070	.073
	2	.068	.069	.071
	3	.069	.070	.073
	4	.070	.069	.073
	5	.071	.071	.075
	6	.070	.070	.073
	7	.068	.070	.074
	8	.069	.069	.072
	9	.068	.069	.072
	10	.071	.071	.073

Table 14. The standard errors of posterior person parameter estimates from one simulation condition

(ICC=0.5, $\sigma_{\gamma}^{2} = 1$, *N* (0.25, 1.1^2) target group ability distribution)

		Model		
	Replication	Multilevel	Testlet	3PL
Base Group	1	.543	.581	.510
	2	.552	.593	.524
	3	.559	.610	.543
	4	.577	.602	.529
	5	.577	.628	.552
	6	.574	.616	.549
	7	.526	.591	.519
	8	.535	.589	.526
	9	.570	.604	.535
	10	.576	.613	.555
Target Group	1	.578	.563	.492
	2	.590	.576	.494
	3	.581	.586	.511
	4	.583	.582	.492
	5	.613	.617	.515
	6	.596	.581	.493
	7	.595	.589	.501
	8	.607	.564	.478
	9	.576	.574	.492
	10	.587	.573	.493

APPENDIX

WINBUGS CODE USED FOR CONCURRENT CALIBRATION BASED ON A 3PL MULTILEVEL TRT MODEL

```
Model
{
# Specify the 3PL multilevel TRT model
 for (m in 1:50) {
         r[m] ~ dnorm(0, tau1.btw);
    for (i in 1:20) {
        u[m,i] ~ dnorm(0, tau1.with);
                     }
    }

for (m in 51:100) {
     r[m] ~ dnorm(mu.btw, tau2.btw);
       for (i in 1:20) {
          u[m,i] ~ dnorm(mu.with, tau2.with);
        }
    }

for (m in 1:100) {
     for (i in 1:20) {
        for (j in 1:50){
           resp[20*(m-1)+i,j] ~ dbern(prob[m, i, j])
           logit(prob.star[m, i, j]) <- alpha[j] * (u[m, i] + r[m] -delta[j] +gamma[m, i, d[j]])
           prob[m, i, j] <- eta[j] + (1 - eta[j])*prob.star[m, i, j]
    }

     theta[m,i] <- r[m]+u[m,i]

    for (k in 1:10){
      gamma[m, i, k] ~ dnorm(0.0, pr.gamma)
    }
    gamma[m, i, 11] <- 0.0
   }
 }

# Specify priors

   for (j in 1:50){
     alpha[j] ~ dlnorm(0, 4)
     delta[j] ~ dnorm(0, .25)
     eta[j] ~ dbeta(5, 17)
   }


pr.gamma ~ dgamma(0.5, 1)
sigsq.gamma <- 1.0/pr.gamma

tau2.btw ~ dgamma(0.5, 1)
sigsq.tau2.btw <- 1.0/tau2.btw
tau2.with ~ dgamma(0.5, 1)
sigsq.tau2.with <- 1.0/tau2.with

mu.btw ~ dnorm(0, 1)
mu.with ~ dnorm(0,1)

rho ~ dunif(0,1)
tau1.btw<-1/rho
tau1.with<-1/(1-rho)
}
```

BIBLIOGRAPHY

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*(1), 47-76.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561-573.

Angoff, W. H. (1968). How we calibrate College Board scores. *College Board Review, 68*, 11-14.

Baker, F. B. (1992). Equating tests under the graded response model. *Applied Psychological Measurement*, *16*(1), 87.

Baker, F. B., & Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. *Journal of Educational Measurement, 28*(2), 147-162.

Béguin, A. & Hanson, B. (2001). *Effect of noncompensatory multidimensionality on separate and concurrent estimation in IRT observed score equating*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Seattle, WA.

Béguin, A., Hanson, B., & Glas, C. (2000). *Effect of multidimensionality on separate and concurrent estimation in IRT equating*. Paper presented at the Annual Meeting of the National Council of Measurement in Education, New Orleans, LA.

Beretvas, S. N., & Williams, N. J. (2004). The use of hierarchical generalized linear model for item dimensionality assessment. *Journal of Educational Measurement*, *41*, 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 (Eds.), *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison-Wesley.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29-51.

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

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88, 9-25.

Cao, Y., Lu, R., & Tao, W. (2014). *Effect of item response theory (IRT) model selection on testlet-based test equating* (ETS Research Report No. RR-14-19). Princeton, NJ: Educational Testing Service. Doi: 10.1002/ets2.12017.

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.

Chu, W-L, & Kamata, A. (2000). *Nonequivalent group equating via 1-P HGLM.* Paper presented at the Annual Meeting of American Educational Research Association,New Orleans, LA.

Cochrane, W. (1977). Sampling techniques. New York, NY: Wiley.

Cook, L., & Eignor, D. (1991). An NCME instructional module on IRT equating methods. *Educational Measurement: Issues and Practice, 10*, 37-45.

Cyr. A., & Davies, A. (2005, November). *Item response theory and latent variable modeling for surveys with complex sampling design: The case of the National Longitudinal Survey of Children and Youth in Canada.* Paper presented at the conference of the Federal Committee on Statistical Methodology, Office of Management and Budget, Arlington, VA.

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of test equating*. New York: Springer Verlag.

DeMars, C. E. (2006). Application of the bi-factor multidimensional item response theory model to thestlet-based tests. *Journal of Educational Measurement*, *43*(2), 145-168.

Dennis, J. E., & Schnabel, R. B. (1996). *Numerical methods for unconstrained optimization and nonlinear equations*. Prentice Hall, Englewood Cliffs, NJ.

Fox, J.-P., & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibb sampling. *Psychometrika, 66*, 269-286.

Fox, J.-P. (2005). Multilevel IRT using dichotomous and polytomous response data. *British Journal of Mathematical and Statistical Psychology, 58*, 145-172.

Glas, C. A. W., Wainer, H., & Bradlow, E. T. (2000). MML and EAP estimation in testlet-based adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 271-287). Boston: Kluwer-Nijhoff.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information bi-factor analysis. *Psychometrika, 57*, 423-436.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research, 22*, 144-149.

Hanson, B. & Beguin, A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement, 26*(1), 3.

He, W., Li, F., Wolfe, E. W., & Mao, X. (2012). *Model selection for equating testlet-based tests in the NEAT design: an empirical study*. Paper presented at the 2012 Annual NCME Conference.

Hills, J. R., Subhiyah, R. G., & Hirsch T. M. (1988). Equating minimum-competency test: Comparison of methods. *Journal of Educational Measurement, 25,* 221-231.

Jiao, H., Kamata, A., Wang, S., & Jin, Y. (2010). *Simultaneous modeling of item and person dependence using multilevel Rasch measurement model*. Paper presented at the Annual Meeting of the American Educational Research Association (AERA), Denver, CO.

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

Jiao, H**.,** Kamata, A., Van Wie, A., & Luo, Y. (2013). *A multilevel testlet model for multiple hierarchical levels of person clustering effects*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

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., & He, W. (2013). Estimation methods for one-parameter testlet models. *Journal of Educational Measurement, 50*(2), 186-203.

Jiao, H. & Zhang, Y. (2014). Polytomous multilevel testlet models for testlet-based assessments with complex sampling designs. *British Journal of Mathematical and Statistical Psychology, 68(1),* 65-83.

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

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement, 38*(1), 79-93.

Kang, T., & Petersen, N. S. (2009). Linking item parameters to a base scale. *ACT Research Report Series,* 2009-2.

Kerkee, T., Lewis, D. M., Hoskens, M., Yao, L., & Haug, C. (2003). *Separate versus concurrent calibration methods in vertical scaling.* Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL).

Kim, S. & Cohen, A. (1998). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement, 22*(2), 131-143.

Kim, S. & Cohen, A. (2002). A comparison of linking and concurrent calibration under the graded response model. *Applied Psychological Measurement, 26*(1), 25-41.

Kish, L. (1965). *Survey sampling*. New York, NY: Wiley.

Kolen, M. & Brennan, R. L. (2004). *Test equating, Scaling, and linking: Methods and practices (Second ed.)*. New York: Springer.

Kreft, I., & de Leeuw, J. (1998). *Introduction to multilevel modeling*. London: Sage.

Lee, G., Kolen, M. J., Frisbie, D. A., & Ankenmann, R. D. (2001). Comparison of dichotomous and polytomous item response models in equating scores from tests composed of testlets. *Applied Psychological Measurement, 25,* 357-372.

Li., D. (2009). *Developing a common scale for testlet model parameter estimates under the common-item nonequivalent groups design*. Doctoral dissertation at the University of Maryland.

Li, Y., Bolt, D. M., & Fu, J. (2005). A test characteristic curve linking method for the testlet model. *Applied Psychological Measurement*, *29*, 340-356.

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: Erlbaum.

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17*(3), 179-193.

Maier, K. S. (2001). A Rasch hierarchical measurement model. *Journal of Educational and Behavioral Statistics, 26*, 307-330.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement, 14(2),* 139-160.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

McDonald, R. P. (1997). Normal-ogive multidimensional model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 257-269). New York: Springer.

Mislevy, R. J. (1987). Exploiting auxiliary information about examinees in the estimation of item parameters. *Applied Psychological Measurement, 11*(1), 81-91.

Mislevy, R. J., & Bock, R. D. (1989). A hierarchical item-response model for educational testing. In R. D. Bock (Ed.), *Multilevel analysis of educational data* (pp. 57-74). San Diego, CA: Academic Press.

Mislevy, R. J., & Bock, R. D. (1983). BILOG: Item and test scoring with binary logistic models [Computer program]. Mooresville IN: Scientific Software.

Muraki, E., & Bock, R. D. (1996). *PARSCALE: IRT analysis and scoring of rating scale data (Version 3.0)*. Chicago: Scientific Software International.

Pastor, D. A. (2003). The use of multilevel item response theory modeling in applied research: An illustration. *Applied Measurement in Education, 16*(3), 223-243.

Powers, S. (2011). *Impacts of group differences on equating accuracy and the adequacy of equating assumptions.* Papers presented at the annual meeting of the National Council of Measurement in Education, New Orleans, LA.

Raudenbush, S. W., & Sampson, R. J. (1999). Ecometrics: Towards a science of assessing ecological settings, with application to the systematic social observation of neighborhoods. *Sociological Methodology, 29*, 1-41.

Raudenbush, S. W., Yang, M., & Yosef, M. (2000). Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. *Journal of Computational and Graphical Statistics, 9*, 141-157.

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

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

Rosenbaum, P. R. (1988). A note on item bundles. *Psychometrika, 53*, 349-360.

Samejima, F. (1969). Estimating of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, No. 17.

Schochet, P. (2005). *Statistical power for random assignment evaluations of education programs*. Mathematic Policy Research, Inc. Princeton, NJ.

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

Skaggs, G., & Lissitz, R. W. (1986). An exploration of the robustness of four test equating models. *Applied Psychological Measurement*, *10*, 303-317.

Smith, R. M., & Kramer, G. A. (1992). A comparison of two methods of test equating in the Rasch model. *Educational and Psychological Measurement*, *52*, 835-846.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in Item Response Theory. *Applied Psychological Measurement*, *7*(2), 201.

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.

Thissen, D. (1991). *Multilog user's guide: Multiple, categorical item analysis and test scoring using item response theory*. Computer program.

Turhan, A. (2006). *Multilevel 2PL item response model vertical equating with the presence of differential item functioning*. Unpublished doctoral dissertation, Florida State University, Tallahassee.

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,* 157-186.

Wainer, H., Bradlow, E. T., & Du, Z. (2000). Testlet response theory: An analogy for the 3-PL useful in testlet-based adaptive testing. In W. J. van der Linder, C. A. W. Glas(Eds.) *Computerized adaptive testing, theory and practice* (pp. 246-270). Boston, MA: Kluwer-Nijhoff.

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

Wainer, H., & Kiely, G. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24,* 185-202.

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, S., Jiao, H., Jin, Y., Thun, Y. M. (2010). *Investigating Effect of Ignoring Hierarchical Data Structures on Accuracy of Vertical Scaling Using Mixed-Effects Rasch Model*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Denver, CO.

Wang, S. (2006). *Brief study of impact of equating sample size on measurement error for catalog products. Research report.* Harcourt Assessment Inc.

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

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

Way, W. D., & Tang, K. L. (1991). *A comparison of four logistic model equating methods.* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: MESA Press.

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

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

Zhang, J. (2007). *Dichotomous or polytomous model? Equating of testlet-based tests in light of conditional item pair correlations*. (Doctoral dissertation). University of Iowa, IA. Available from ProQuest Dissertations and Theses database. (UMI No. 3290677)

Zhang, Z. (2010). *Comparison of different equating methods and an application to link testlet-based tests.* Doctoral dissertation at the Chinese University of Hong Kong.

Zhu, W. (1998). Test equating: What, why, how? *Research Quarterly for Exercise and Sport, 69*, 11-13.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). *BILOG-MG 3* [computer program]. Chicago, IL: Scientific Software.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items* [computer program] Chicago, IL: Scientific Software International, Inc.