ABSTRACT

Title of Dissertation	AN INFORMATION CORRECTION METHOD FOR TESTLET-BASED TEST ANALYSIS: FROM THE PERSPECTIVES OF ITEM RESPONSE THEORY AND GENERALIZABILITY THEORY
	Feifei Li, Doctor of Philosophy, 2009
Directed by	Professor Robert J. Mislevy Department of Measurement, Statistics and Evaluation

An information correction method for testlet-based tests is introduced in this dissertation. This method takes advantage of both generalizability theory (GT) and item response theory (IRT). The measurement error for the examinee proficiency parameter is often underestimated when a unidimensional conditional-independence IRT model is specified for a testlet dataset. By using a design effect ratio composed of random variances which can be easily derived from GT analysis, it becomes possible to adjust the underestimated measurement error from the unidimensional IRT models to a more appropriate level. It is demonstrated how the information correction method can be implemented in the context of a testlet design.

Through the simulation study, it is shown that the underestimated measurement errors of proficiency parameters from IRT calibration could be adjusted to the appropriate level despite the varying magnitude of local item dependence (LID), testlet length, balance of testlet length and number of the item parameters in the model. Each of the three factors (i.e., LID, testlet length and balance of testlet length) and their interactions have statistically significant effects on error adjustment. The real data example provides more details about when and how the information

correction should be used in a test analysis. Results are evaluated by comparing the measurement errors from the IRT model with those from the testlet response theory (TRT) model. Given the robustness of the variance ratio, estimation of the information correction should be adequate for practical work.

AN INFORMATION CORRECTION METHOD FOR TESTLET-BASED TEST ANALYSIS: FROM THE PERSPECTIVES OF ITEM RESPONSE THEORY AND GENERALIZABILITY THEORY

By

Feifei Li

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2009

Advisory Committee:
Professor Robert J. Mislevy, Chair
Professor Robert W. Lissitz
Professor Robert Croninger
Assistant Professor Hong Jiao
Assistant Professor Jeffrey Harring

Acknowledgments

It is difficult to put into words how grateful I am to everyone who assisted on this dissertation. I thank my advisor, Dr. Robert Mislevy, and other committee members, Dr. Robert Lissitz, Dr. Robert Croninger, Dr. Hong Jiao, and Dr. Jeffrey Harring, for your support, encouragement, patience and mentorship. Each of you has played important roles in my professional development. Your vision and wisdom have been and will continue to be the guiding light in my career life.

I am also grateful to a few people who made this dissertation possible and because of whom my graduate study has been the one that I will cherish forever: Dr. C. Mitchell Dayton, Dr. Gregory R. Hancock, Dr. George Macready, Dr. Andre Rupp, Dr. Shu Jing Yen, and Dr. Frank Rijman at ETS. Words cannot express how much your tutoring has meant. Gratitude goes to CTB McGraw-Hill Company for providing research and development grant for me to initiate this study. Gratitude also goes to my friends and administrative staff who have assisted me during my study at the University of Maryland.

To my dearest parents Manqing Cao and Yizhong Li, your encouragement and generous support have strengthened my confidence to complete this dissertation and my PhD. I attribute my efforts and persistence largely to the ethic you instilled in me.

Most importantly, thank you Hua for everything you have done to ensure my achievements. Those difficulties would not have been conquered without your continuous support.

Finally, I dedicate this thesis to my lovely son, Frank. You have been and will continue to be my motivation. I love you heart and soul.

Table of Contents

List of Tables

List of Figures

CHAPTER I: INTRODUCTION

“Testlet” indicates a set of items sharing a single common stimulus (Rosenbaum, 1988), where the performance of each item depends on both a general ability and a specific ability related to the particular content or occasion, for example, a reading passage or an information graph. Testlets help to develop a more realistic and contextualized test. They can provide insights into not only general “abilities”, but also a series of specific cognitive “information-processing” in complex tasks (Rosenbaum, 1988; Sternberg, 1977). With testlets, the time and cost for collecting additional information can be reduced. As testlets can bring beneficial consequences to educational practices (Messick, 1994), they have been seen in many large-scale tests. Examples can be found in Graduate Record Examinations (GRE), Scholastic Aptitude Test (SAT), and Iowa Test of Basic Skills (ITBS).

However, due to the particular statistical properties of testlets, issues have emerged in regard to applying unidimensional measurement models to the testlet datasets. One of the properties that has brought up many technical concerns is local item dependence (LID). Namely, the common stimulus which the set of items rely upon can introduce dependence among the responses within an individual. For example, when some students have a special interest or better prior background knowledge in a passage than others, they are likely to perform better on the items related to this passage than other items on the same difficulty level, or than other test-takers on the same general ability level.

In contrast, conditional independence (CI) or local item independence (LII) is assumed in the conventional IRT models. The CI assumption states that given the fixed ability level, an examinee’s performance on one item must not affect his or her responses to any other items in the test. Unidimensional IRT models may not be robust to the violation of the CI assumption (Hambleton & Swaminathan, 1985). In that case, analyzing testlet datasets with misspecified unidimensional IRT models could lead to undesired results. As shown in a number of previous studies, ignoring LID would result in overestimated precision of ability estimates as well as bias in item difficulty and discrimination estimates, and such biases were exacerbated when either the testlet length or the testlet effect increased (Bradlow, Wainer, & Wang, 1999; Sireci, Wainer & Thissen, 1991; Wainer, 1995; Yen, 1993).

The accuracy of ability estimates is particularly crucial under some circumstances. For example, if the test results are used in ways that have consequences for individual examinees, greater accuracy is required at the score level. When the cut scores are applied to proficiency classification, the measurement errors of proficiency estimates also need to be considered. In computer adaptive testing (CAT) for another example, overestimation of precision would present difficulties for setting stopping rules and lead to premature termination (Du, 1998; Wainer, Bradlow, & Du, 2000).

To account for LID from the response patterns of testlets, a number of non-parametric and parametric approaches have been created and employed. Generalizabilty theory (GT) has been traditionally used to model and analyze a variety of statistical dependencies on the raw score scales (Brennan, 1992; Cronbach, Linn, Brennan, & Haertel, 1997; Koretz, Stecher, Klein, & McCaffrey, 1994; Lee & Frisbie, 1999; Sireci, et al., 1991). By using GT, one does not have to demonstrate the satisfaction of strong statistical assumptions that are required by IRT. A GT approach has been regarded as a convenient method as it can easily partition the variances from different resources and provide the information about the reliabilities and errors of measurement. However, GT was originally created for continuous variables rather than for discrete item scores (Brennan, 1997). Although a hybrid approach that incorporates GT and IRT has been developed to fulfill a nonlinear transformation from the discrete raw test scores to the continuous item and person variables, this approach has currently been limited to single-facet measurement design with binary items (Briggs & Wilson, 2007).

In contrast, IRT models specify a probabilistic relationship between the item responses and the characteristics of the individual and the items. The link function makes it possible to connect the discrete responses with the continuous latent variables. Testlet models from the IRT approach generally capture the person-testlet interactions in terms of multidimensional variables modeled as random effects, for example, the Rasch testlet and random-effects facet model (Wang & Wilson, 2005a, 2005b) which are special cases of the multidimensional random coefficients multinomial logit model (MRCMLM) (Adams, Wilson, & Wang, 1997); the bi-factor model (Gibbons & Hedeker, 1992); the multilevel model (Jiao, Wang, & Katama, 2005); and testlet response theory (TRT) models (Bradlow, et al., 1999; Wainer, et al., 2000; Wainer, et al., 2007; Wang, Bradlow, & Wainer, 2002). As has been shown in a series of simulation and real data studies (Bradlow, et al., 1999; DeMars, 2006; Jiao, et al., 2005; Jiao & Wang, 2008; Wainer, et al., 2000; Wainer, et al., 2007; Wang, et al., 2002; Wang & Wilson, 2005a, 2005b), these multidimensional testlet IRT models demonstrate good model fit, small bias and satisfactory accuracy in parameter recovery on the testlet datasets, compared with their unidimensional IRT counterparts.

The feasibility of the estimation of these testlet response models, to a large extent, depends on the recent increase in computational power. Marginal maximum likelihood estimation (MMLE) with the expectation-maximization (EM) algorithm has been applied in MRCMLM and the bi-factor model; penalized quasi-likelihood (PQL) estimation or Laplace approximation (Laplace) has been often used for the multilevel models; Markov Chain Monte Carlo (MCMC) is the method for estimation in TRT models. In comparison, Laplace and MCMC yielded accurate parameter recovery and appropriate precision of estimates but it took very long time to converge (Jiao & Wang, 2008, Sinharay, 2003), and thus, has been rarely implemented in operational testing. MMLE with the EM algorithm was relatively efficient and its performance in parameter estimation was adequate (Jiao & Wang, 2008; Demars, 2006). However, in these applications of MMLE, ability and testlet parameters were estimated conditional on the point estimates of the item parameters, so the uncertainty in the estimation of the item parameters has been ignored (Wainer et al., 2007).

Considering the design of testlets where items are clustered versus the design of the independent items, the downward biased estimation variance of the ability estimates as a result of misspecifying the unidimensional IRT models on the testlet data may be adjusted through the design effect. Bock, Brennan and Muraki (2002) proposed correcting the information function of multiple ratings by using a variance ratio term derived from the GT analysis. Taking account of the similarity between the testlets and multiple ratings in terms of local dependency between the responses in clusters, I extend this method to the situation of testlets to adjust the underestimated measurement error of abilities. The design effect is often used as a measure of the precision gained or lost by the use of a more complex design instead of the simple random sampling (SRS) (Cornfield, 1951). Because cluster samples usually give less precision per observation unit than an SRS, the design effect is usually larger than 1 for cluster sampling. With respect to its estimation procedure, it is relatively efficient to obtain the design effect by deriving the variance of person estimates of either design through GT, and the information of the ability estimates in independent item design through maximum likelihood estimation (MLE). Hereby, given their strengths, GT and IRT can jointly contribute to adjust the information of ability estimates in the testlet-based tests to a more appropriate level.

The purpose of this study is to propose and evaluate the information correction method that uses the design effect ratio from the GT analysis to adjust the underestimated measurement error as a result of misspecifying the unidimensional IRT models for the testlet data. To achieve this purpose, it is necessary to (1) explore a computational approach for this information correction method; (2) conduct a simulation study to evaluate the performance of the proposed information correction method under conditions with varying factors; (3) apply the information correction method to a real data case.

1.1. Significance of the Study

This research is significant in four aspects. First, testlets have been frequently used in standardized educational tests. One reason is that they can save time and cost in test development. A more important reason is that the stimuli of the testlets often address real-life problems and require the integration of the knowledge and skills which cannot be represented in simple independent multiple-choice items. For example, performance assessments that have gained increasing popularity in recent years intentionally have the examinees produce item responses that are interconnected.

Second, the conventional unidimensional IRT models do not account for LID in testlets, which leads to overestimation of the precision of measurement, substantially on some occasions. The measurement error can be critical for individuals in high-stake tests or when the cut scores are applied for the purpose of classification in proficiency. In complex computerized simulations, LID affects the IRT evidence accumulation process on the test level (Wainer, Brown, Bradlow, Wang, Skorupski, Boulet, & Mislevy, 2006). Specifically, the precision bias will jeopardize the item selection procedure. As multiple dependent observations will not have as much increase in accuracy as observations elicited by independent items, the test length that is necessary to achieve certain measurement precision is very likely to be underestimated. For example, an item stimulus in CAT may have 15 associated items. These items usually have excessive local dependence. Instead of administering all 15 items in a testlet to an examinee, an item selection algorithm is used to pick out the testlets and then the items based on the item's properties and the examinee's performance on the previous items, as well as the status of content balance. In this circumstance, it is necessary to have a relatively accurate estimate of the measurement precision or even quantify the local dependency features of testlets.

Third, although some testlet models have been proposed and demonstrated with satisfactory performance in terms of model-data fit and parameter recovery compared with their unidimensional IRT counterparts, each of them has limitations. The GT model has not been sufficiently developed to connect continuous latent values with discrete scores. Models in the IRT approach such as MRCMLM, the bi-factor model, the multilevel model and TRT are more complex and usually take long time to converge in the ways they are currently estimated.

Fourth, this study will show how GT and IRT might be combined to provide an efficient solution for the measurement problem. It is intended to provide a practical approximation to correct the measurement precision. By deriving the error variances of person estimates in a cluster design (testlets) and a SRS design (independent items) from GT analysis, the design effect of the two designs is easily available. It is hypothesized that more accurate precision estimates of the ability could be obtained by applying this design effect to adjust the measurement error from the unidimensional IRT models.

1.2. Structure of the Dissertation

The remaining of this thesis is organized as follows:

Chapter 2: Literature Review. A full explanation is provided on the concepts of CI versus LID as well as the causes of LID, followed by a synthesis on the measures and models for LID in testlets. The methods that have been used for different types of testlet models are also reviewed. Two examples of sequential use of IRT and GT are illustrated. One is generalizability in item response modeling (GIRM) and the other is the information correction in multiple ratings.

Chapter 3: Theoretical Framework. The variance terms from the generalizability analysis, the likelihood functions and the information functions from the IRT analysis for both the independent item design and the testlet design are illustrated. The computational approach of the information correction method is described. It is shown mathematically how the information correction ratio term becomes a coefficient in the information function of the conditional independent IRT models.

Chapter 4: Methodology. The simulation study is implemented to evaluate the performance of the information correction method in adjusting the measurement error of the proficiency parameters when the conditional-independent IRT models are specified to the testlet dataset. Through the real data example, the use of the information correction method is demonstrated and the results are evaluated in comparison with those from the TRT calibration.

Chapter 5: Results. Results in regard to the research questions of interest are presented and interpreted.

Chapter 6: Conclusions and Discussions. The major findings are summarized. The implications for testing practices, the limitations as well as the directions for the future research are discussed.

CHAPTER II: LITERATURE REVIEW

This study is grounded on four fields of research: local item dependency, models and measures for local item dependency, estimation methods of testlet models, and sequential uses of IRT and GT.

2.1. Local Item Dependency

2.2.1. Local Item Independence

In classical testing theory (CTT), CI implies that the errors of measurement are statistically uncorrelated among different items given an examinee's true score (Yen & Fitzpatrick, 2006). In IRT, CI assumption states that given the fixed ability level, an examinee's performance on one item must not affect his or her responses to any other items in the test. From a statistical perspective, the probability of any pattern of item scores for an examinee is the product of the probability of the scores on each test item conditional on the values along the trait scales. This conditional independence is also called local item independence (LII). Its mathematical presentation was given by Lord (1980) as

$$P(\mathbf{X} = \mathbf{x} \mid \mathbf{\Theta} = \mathbf{\theta}) = \prod_{i=1}^{n} P(X_i = x_i \mid \mathbf{\Theta} = \mathbf{\theta}), \qquad (2.1)$$

where **X** denotes a person's response pattern on a sample of test items; X_i is the score on item i; **Θ=(**Θ_1, Θ_2,..., Θ_d**)** is a vector of d-dimensional latent traits that are measured by the IRT models. McDonald (1994) has defined weak local item dependence in which conditional independence is only required for pairwise items, that is, for a pair of items,

$$P(X_1 = x_1 \; and \; X_2 = x_2 \mid \theta) = P(X_1 = x_1 \mid \theta) P(X_2 = x_2 \mid \theta). \qquad (2.2)$$

From the perspective of factor analysis, the CI is guaranteed when all factors that have systematical effect on the responses are identified and accounted for by the model, while LID occurs when these factors are not modeled and the residuals are correlated.

2.1.2. LID and Factors Causing LID

LID arises from the existence of an additional factor that consistently affects the performance of students on some items to a greater extent than on other items. LID can be positive or negative (Habing & Roussos, 2003). Positive LID between two items means that one performs better than expected (based on their overall test performance as reflected by their $\hat{\theta}$ values) on one item, he or she also performs better than expected on the other item. Negative LID means that if an examinee performs better than expected on one of the items, he or she performs worse than expected on the other item.

The reasons of LID are varied, as have been elaborated in some studies (Ferrara, Huynh, & Baghi, 1997; Ferrara, Huynh, & Michaels, 1999; Hoskens & De Boeck, 1997; Yen, 1993). Yen (1993) had included the following factors of LID, the external assistance of interference with some items, speededness, fatigue, practice, special item formats, variation in response format (multiple choice items vs. constructed-response items), a shared stimulus or passage, item chaining, items requiring explanation of a previous answer, cloze items (where examinees need to fill in multiple blanks in one passage), scoring rubrics or raters, unique content knowledge or abilities, and a differential opportunity to learn.

There are several categorizations depending on the causes of LID, for example, “underlying local dependence” versus “surface local dependence” (Chen & Thissen, 1997), “local dependence in the presence of multidimensionality” versus “local dependence caused by speededness" (Douglas, Kim, Habing, & Gao, 1998), and "local dependence caused by item order" versus "local dependence caused by contextual effect of a cluster of items" (Hoskens & De Boeck, 1997). In comparison, the first categorization is more exhaustive and encompassing. "Underlying local dependence" usually indicates LID that has introduced multidimensionality for the test. The typical causes are the content and the ability measured by a subset of items. "Surface local dependence", in contrast, refers to the external situations such as administration or score procedures in which examinees tend to produce similar responses to a set of items. For example, when speededness occurs, students are very likely to give the same responses to the items they have not yet reached. It is possible to avoid LID due to external causes by adjusting the administration procedures or diminishing the chances of external interferences. However, it is unrealistic to get rid of LID that is inherent in items, especially in performance assessments where LID causes such as common stimulus and item chaining are seen as necessary and desirable.

2.2. Measuring and Modeling Testlet Effects

2.2.1. Measures of LID

To account for LID from the response pattern of testlets, a number of non-parametric and parametric approaches have been created. Yen (1984, 1993) proposed Q_3 statistic as a measure of dependency between two items. It is parameterized as the correlation between the scores on two items residualized on the expected score of each student. Chen and Thissen (1997) also suggested two alternative dependence indices for item pairs, Pearson χ^2 and likelihood ratio G^2. The likelihood ratio is computed using the expected frequency and the observed frequency of all possible response patterns by a sample of examinees to a pair of test items of interest. Both indices were shown to be sensitive to the presence of local dependence, but they have been less used than Q_3 statistic. In the *item bundle model* (Rosenbaum, 1988), the conditional population covariance was used to detect LID. Ferrara, Huynh, and Baghi (1997) described a procedure to identify LID based on the raw test scores by checking the magnitude of the interitem correlations for the examinees at different intervals on the test score scale under the assumption of unidimensionality. Douglas, Kim, Habing, and Gao (1998) investigated LID using conditional covariance functions.

Among all these statistics, Yen's Q_3 is the statistic that is most frequently used for detecting and measuring the degree of local item dependence of any two test items. To compute Q_3 for any pair of binary items, each examinee's ability θ_i must be estimated from the responses to all items using the selected IRT model. Based on examinee trait parameter estimates and item parameter estimates, each examinee's expected score (i.e. the probability that the examinee will answer that item correctly) on any item is computed using the selected IRT model. The difference between an examinee's observed score and expected score, d_{ij}, is then obtained using the following equation

$$d_{ij} = y_{ij} - P(y_{ij} = 1 \mid \theta_i), \tag{2.3}$$

where y_{ij} is the observed score of examinee *i* on item *j*, and $P(y_{ij} = 1 \mid \theta_i)$ is the examinee's expected score on item *j*. Q_3 value for item *j* and *j'* is actually the correlation of d_j and $d_{j'}$ taken over all examinees. When the model is true, Fisher's transformation of Q_3 may be distributed as a normal variable with mean equal to zero and variance equal to 1/(N-3), where N indicates the number of examinees in the test.

Kingston and Dorans (1982) noted that because items scores were involved in the calibration of θ and were later used for the calculation of residuals, when there was no LID the expected value of the correlation between residuals, -1/(n-1), would tend to be slightly negative, where n is the total number of items involved in the test.

2.2.2. Polytomous IRT

Researchers once came up with a scoring approach that accounted for the overlapping item information more accurately by scoring polytomously the testlet as a single item (Thissen, Steinberg, & Mooney, 1989; Wainer, 1995). Responses to the items within the testlets were collapsed into an aggregated variable. This approach was generally accepted and had been applied to testing programs in which the testlets were involved (e.g., Sireci, et al., 1991; Wainer, 1995; Wainer & Lewis, 1990), but this approach has certain drawbacks. One drawback is that ignoring the pattern of responses would lead to the loss of information that would have been extracted from each item in the testlets (Wainer et al., 2000). Another drawback is that the parameters of the binary items which are more meaningful and interpretable than polytomous items would not remain in that case.

2.2.3. Generalizability Theory

Generalizability theory (GT) models have been traditionally used to model and analyze a variety of statistical dependencies on the raw score scales (Brennan, 1992; Cronbach, et al., 1995; Koretz, Stecher, Klein, & McCaffrey, 1994; Lee & Frisbie, 1999; Sireci et al., 1991). There is no need for demonstrating the satisfaction of strong statistical assumptions such as CI and dimensionality that are required by IRT. Normality of data is not assumed in GT either (Brennan, 2001). GT has been regarded as a convenient method as it can easily partition the variances from different resources and provide the information about the reliabilities and errors of measurement.

However, unlike the response models that connect continuous latent variables and the discrete observable variables through logit or probit links, GT was originally developed for continuous observable scores (Brennan, 1997). Some recent studies have shown with the discrete scores taken into account, GT and IRT can be combined to produce results comparable to those from conventional GT and IRT models respectively (Briggs & Wilson, 2007). However, this approach has currently been limited to single-facet measurement design with binary items and is not of immediate use for the multi-facet measurement design such as a testlet design.

The formula below represents the GT model for testlets. Assuming the testlet dataset has a univariate nested design of i x (j:d), that is, persons (i) crossed with items (j) nested in testlets (d). The mean of the item scores in a given set of items is the best linear unbiased estimator of an individual, the linear model of which can be represented as follows assuming completely random (Lee & Frisbie, 1999).

$$\begin{aligned}
X_{ij:d} = \mu & \qquad \text{(grand mean)} \\
+ \mu_i - \mu & \qquad \text{(person effect)} \\
+ \mu_d - \mu & \qquad \text{(testlet effect)} \\
+ \mu_{j:d} - \mu_d & \qquad \text{(item within testlet effect)} \\
+ \mu_{id} - \mu_i - \mu_d + \mu & \qquad \text{(person} \times \text{testlet interaction effect)} \\
+ X_{ijd} - \mu_{id} - \mu_{j:d} + \mu_d & \qquad \text{(residual effect)}
\end{aligned} \qquad (2.4)$$

In the GT model of testlets, the magnitude of the testlet effect is measured by variance of person by testlet interaction σ^2(iJ:D). A unique benefit of GT is that through “Generalizability study” and “Design study”, we would be able to obtain “generalizability coefficient” that provides information for the design of test, namely, the number of facets (e.g., items, testlets or raters) to achieve a certain level of reliability. Items instead of testlets are used as units of analysis. Additionally, an empirical study (Lee & Frisbie, 1999) reported that GT model could also lead to an accurate standard error of measurement (SEM).

2.2.4. TRT Models

To account for the dependency structure in testlets, Bradlow, Wainer and Wang (1999) proposed a modification to the traditional two-parameter logistic (2-PL) IRT model to include an additional parameter for the person-specific random effects of testlets, which is denoted as $\gamma_{id(j)}$ in the follows. Wainer, Bradlow and Du (2000) extended this modification to the 3-PL IRT model where the computation was more complex. Wang, Bradlow and Wainer (2002) further generalized this model to the situation in which the test consisted of a mixture of binary and polytomous scored items and testlets. This set of models, in correspondence to the set of binary and polytomous conventional IRT models, are termed as testlet response theory (TRT) models. TRT retains the structure of IRT, but contains an extra random testlet effect variable.

Assume a testlet dataset in which each of the I (i=1,…I) examinees takes a linear test of J (j=1,…, J) binary-scored items. The 2-PL IRT model (Lord, 1952) is presented as

$$P_{ij}(1) = P(y_{ij} = 1 \mid \theta_i, \alpha_j, \beta_j) = \text{logit}^{-1}(t_{ij}), \tag{2.5}$$

where t_{ij} is a linear score predictor

$$t_{ij} = \alpha_j(\theta_i - \beta_j) \tag{2.6}$$

α_j is the item discrimination on item j;

β_j is the item difficulty of item j;

θ_i is the proficiency of person i;

y_{ij} is the binary score of examinee i on item j, which is distributed as a Bernoulli:

$$y_{ij} = \begin{cases} 1 & \text{if } t_{ij} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{2.7}$$

Given J items nested in D (d=1,…,D) independent testlets, the linear score predictor t_{ij} is extended from its standard form to its modified form by including a random interaction term $\gamma_{id(j)}$ that parameterizes the random effect for person i on testlet d(j). It is independent of the ability (θ_i) and item parameters (β_j):

$$t_{ij} = \alpha_j(\theta_i - \beta_j - \gamma_{id(j)}) \tag{2.8}$$

d(j) denotes the testlet in which item j is nested. This parameterization allows for greater dependence of items within testlets compared to item dependency across testlets.

TRT models can be embedded within hierarchical Bayesian frameworks (Bradlow et al., 1999). The inferences on the unknown parameters can then be obtained by Bayesian estimation. The prior distributions for the parameters are specified by convention as:

$$\begin{aligned} &\theta_i \sim N(0,\sigma_\theta^2) \\ &\alpha_j \sim \mathrm{logN}(\mu_\alpha,\sigma_\alpha^2) \\ &\beta_j \sim N(\mu_\beta,\sigma_\beta^2) \\ &\gamma_{id(j)} \sim N(0,\sigma_{\gamma d(j)}^2) \end{aligned} . \tag{2.9}$$

The variance of the testlet effect $\sigma_{\gamma d(j)}{}^2$ indicates the strength of the testlet effect for testlet d(j). To estimate the strength of the testlet effects, a hyperprior for $\sigma_{\gamma d(j)}{}^2$ is specified as an inverse gamma distribution with the shape parameter g and the scale parameter ξ.

$$\sigma_\gamma^2 \sim Gamma^{-1}(g,\xi) \tag{2.10}$$

In the case of the 3-PL item response model

$$p_{ij}(1) = P(y_{ij} = 1 \mid \theta_i, \alpha_j, \beta_j, \omega_j) = \omega_j + (1-\omega_j)\mathrm{logit}^{-1}(t_{ij}) \tag{2.11}$$

ω_j is the guessing parameter for item j. 3-PL TRT model is achieved by retaining the 3-PL IRT structure yet modifying logit's linear predictor (Equation 2.6) to include a random effect as we did with the 2-PL TRT model (Equation 2.8). Specifically, the

3-PL TRT model is presented as

$$P(y_{ij}=1)=\omega_j+(1-\omega_j)\frac{\exp(\alpha_j(\theta_i-\beta_j-\gamma_{id(j)}))}{1+\exp(\alpha_j(\theta_i-\beta_j-\gamma_{id(j)}))} \qquad (2.12)$$

Its prior distributions except that of guessing parameter are the same as in the 2-PL TRT model, and the guessing parameter is distributed as

$$\omega_j \sim N(\mu_\omega,\sigma_\omega^2). \qquad (2.13)$$

Then TRT was further extended to the testlets consisting of polytomous items (Wang et al., 2002). A general polytomous IRT model (Samejima, 1969) is given by

$$p_{ij}(r)=P(y_{ij}=r\,|\,\theta_i,\alpha_j,\beta_j,\omega_j,g_r)=\Phi(g_r-t_{ij})-\Phi(g_{r-1}-t_{ij}), \qquad (2.14)$$

where r is the observed score number; g_r is the latent cutoff for the polytoumous item such that observed score y_{ij}=r if the latent score s_{ij} satisfies g_{r-1}<s_{ij}<g_r; Φ denotes the normal cumulative density function. Like the 2-PL and 3-PL IRT models, the TRT model of polytomous items is formulated by extending the linear score predictor t_{ij} from its standard form (Equation 2.6) to its testlet random effect form (Equation 2.8).The prior distributions of parameters are identical to those in the 3-PL TRT setting.

TRT models can explicitly detect, model and assess the magnitude of item dependency within each testlet through a parametric approach. Items rather than testlets are the units of analysis in TRT, and the operational item-level parameters such as loadings and locations are retained. Such a modification facilitates an easy transformation from the conventional IRT to TRT models mathematically and conceptually (Bradlow et al., 1999; Wainer et al., 2000; Wang et al., 2002). Importantly, the variance of testlet effect $\sigma_{\gamma d(j)}^2$ quantifies the magnitude of testlet effects in circumstances of test construction and ad hoc parameter estimation.

The simulation study on the 2-PL TRT model showed that the 2-PL TRT model provided more accurate parameter estimates from testlet data than the 2-PL IRT model and equally accurate estimates from independent item responses as the 2-PL IRT model. The advantages of TRT were even salient in the situation with substantial LID (Bradlow et al. 1999). Simulations on 3-PL and polytomous TRT models also yielded consistent results (Wainer et al., 2000; Wang et al., 2002). Analyses of operational test datasets (e.g., SAT, GRE, TSE and North Carolina Test of Computer Skills) using TRT further confirmed that this testlet modeling approach was required, trustable, adequate and had great potential in application. It has been demonstrated that the estimates of magnitudes of testlet effects provided by TRT models can be critical in evaluating the testing programs (Bradlow et al., 1999; Wainer, 1995; Wainer et al., 2000; Wang et al., 2002).

TRT has been applied to test equating, scaling and linking (Lee, Kolen, Frisbie, & Ankenmann, 2001; Li, Bolt, & Fu 2005). For instance, Li, et al. (2005) applied TRT models to link the calibrations between two tests with common testlets in a non-equivalent group design. Their results suggested that the scale transformation coefficients were accurately recovered and were superior to those obtained by using the IRT model in linking calibrations when LID was present. In addition, TRT model was regarded as having potential to be used in other situations with LID such as multiple ratings (Wang et al., 2002).

2.2.5. Hierarchical Models

Jiao, Wang and Katama (2005) modeled LID using a three-level hierarchical generalized linear model (HGLM). HGLM combined item response model and multilevel modeling. Thus, like TRT, HGLM would allow the outcome variables to be discrete response variables. The three-level HGLM for testlets is formulated as follows.

Level-1 models the item effect

$$\log(\frac{p_{idj}}{1-p_{idj}}) = \beta_{0di} + \sum_{q}^{k-1} \beta_{qdi} X_{qidj} \qquad \mathrm{X}_{\mathrm{qidj}} = \begin{cases} 0 & \text{when q} \neq j \\ 1 & \text{when q} = j \end{cases}, \tag{2.15}$$

where

p_{idj} is the probability that person i answers item j in testlet d correctly;

X_{qidj} is the qth dummy variable for person i;

β_{0di} is an intercept term;

β_{qdi} is a coefficient associated with X_{qidj}. It corresponds to the individual item effect β_j in TRT.

Level-2 models the testlet-level effect. It is

$$\beta_{0di} = \gamma_{00i} + \upsilon_{0di}, \text{ and } \beta_{qdi} = \gamma_{q0i}, \quad \upsilon_{0di} \sim N(0, \sigma_{\upsilon}^2) \;, \tag{2.16}$$

where

γ_{00i} is the fixed effect of the level-1 intercept;

υ_{0di} is a random effect of the level-l intercept, which can be conceptualized as an interaction between person and testlet; it is analogous to $\gamma_{id(j)}$ in TRT;

γ_{q0i} is the individual item effect for item with the qth dummy variable;

σ_{υ}^2 provides the magnitude of LID, which is analogous to $\sigma_{\gamma d(j)}^2$ in TRT.

Level-3 models the person-level effects by further decomposing coefficients at level-2,

$$\begin{aligned} \gamma_{00i} &= \pi_{000} + w_{00i} \qquad w_{00i} \sim N(0, \sigma_w^2) \\ \gamma_{q0i} &= \pi_{q00} \end{aligned}, \tag{2.17}$$

where q=0,…,k-1;

W_{00i} is the person effect, which is equivalent to the person's ability θ_i in TRT, except that W_{00i} is a random variable and the other is a fixed variable;

σ_w^2 is the variance of the ability distribution.

The HGLM of testlets captures the testlet effect through the variance of the random testlet effect variable, σ_{υ}^{2}. Like the item response models, the person ability and item location can be parameterized and aligned on the same scale. Another advantage of HGLM is that the covariates related to the characteristics of the items or person abilities can be easily incorporated into the linear predictor, including the higher level variables.

2.2.6. Factor Analysis

The bi-factor model (Gibbons & Hedeker, 1992) and the correlated measurement error analysis (Reddy, 1992) modeled LID from the perspective of factor analysis. Reddy examined the effects of ignoring the correlated measurement error in general, including repeated measures and testlets. The results of the simulation studies on the effects of misspecifications for the bi-factor model were aligned with those of TRT. For example, the model fit indices showed significant misfit of the misspecified models in contrast to correctly specified models; ignoring the correlated error would lead to biases in parameter estimates (Reddy, 1992).

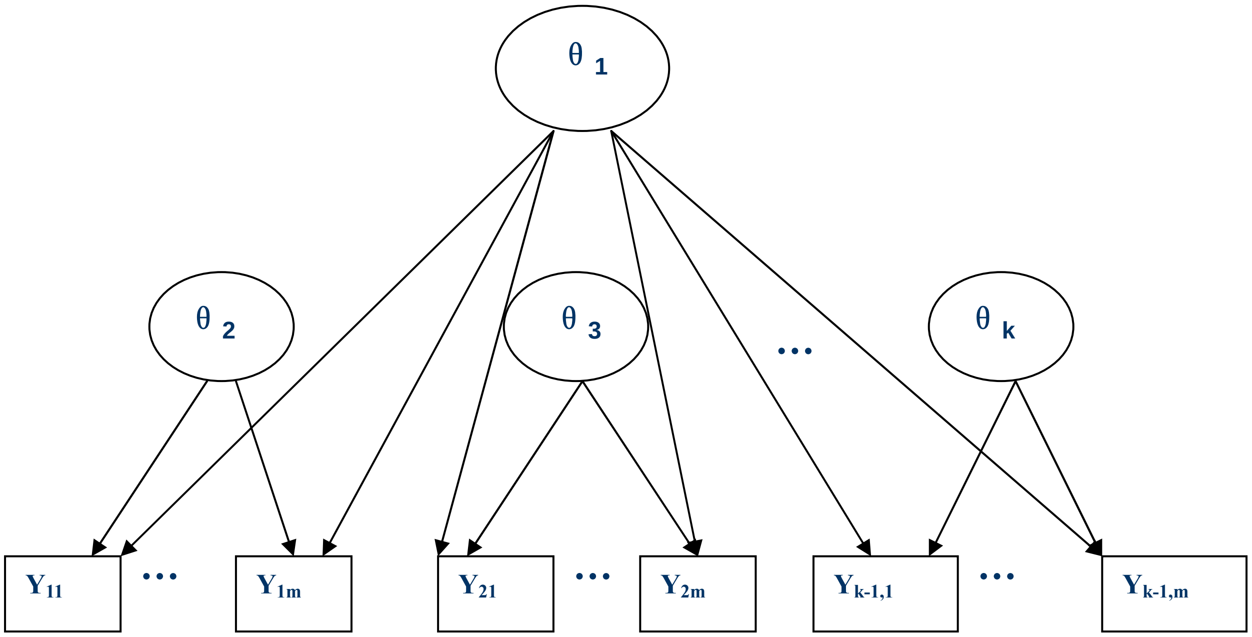
In contrast, the bi-factor model (Gibbons & Hedeker, 1992) represented a structural framework that is similar to TRT models. Given an s-factor matrix consisting of one primary factor and s-1 group factors, the bi-factor model constrains each item j to have a nonzero loading on the primary dimension and a secondary loading (α_{jk}, k=2,…, s) on not more than one of the s-1 group factors. Shown below is the bi-factor pattern matrix for four items

$$\boldsymbol{\alpha} = \begin{bmatrix} \alpha_{11} & \alpha_{12} & 0 \\ \alpha_{21} & \alpha_{22} & 0 \\ \alpha_{31} & 0 & \alpha_{33} \\ \alpha_{41} & 0 & \alpha_{43} \end{bmatrix} \tag{2.18}$$

In a context of testlets, the primary dimension can be conceptualized as the primary ability (θ_1), and the additional factors ($\theta_2, \theta_3, ...\theta_k$) as the content area knowledge associated with each testlet, as illustrated in Figure 2.1. In this context, items are conditionally independent between testlets, but are conditionally dependent within testlets, which are exactly the same assumptions in the TRT framework. The bi-factor model permits analysis of models with large numbers of group factors (e.g. testlets). Besides, it provides more parsimonious solution than the unrestricted factor analysis (Gibbons & Hedeker, 1992). The major difference between the bi-factor model and TRT is that the bi-factor model allows separate discrimination parameters for the primary and testlet (secondary) dimensions, and these discrimination parameters can be independent. Thus, the 2-PL TRT model can be regarded as a constrained version of the bi-factor model (DeMars, 2006).

The bi-factor model is often used when it is necessary to apply different discriminations on the testlet dimensions than the primary dimension. Otherwise, TRT is more parsimonious. The bi-factor would better be used for testlets with large conditional dependencies, while for tests with independent items, fitting bi-factor would introduce more error (DeMars, 2006). In addition, when the testlet effect is small, the bi-factor model is very likely to be unidentifiable (Li & Rijmen, 2009).Thus, it is recommended to detect the presence of LID before using the bi-factor model.

Figure 2.1 Graphical illustration of bi-factor model with k-1 testlets



2.3. Estimation of Testlet Models

Table 2.1: Estimation Methods for Testlet Models

Estimation method	Testlet models	Parameter recovery	Estimation time
Marginal maximum likelihood (MML) with expectation –maximization (EM) algorithm	Bi-factor model (Gibbons & Hedeker, 1992) Radom-effects testlet model (Wang & Wilson, 2005a) Rasch testlet model (Wang & Wilson, 2005b)	Accurate	Efficient
Penalized quasi-likelihood (PQL)	Hierarchical generalized linear models (Jiao, et al., 2005)	Less accurate	Less efficient
Laplace approximation		More accurate	
Bayesian inference with Markov Chain Monte Carlo (MCMC) methods	Testlet response theory models (Bradlow, et al., 1999; Wainer, et al., 2000; Wang, et al., 2002; Du, 1998)	More accurate	Less efficient

Comparing the four estimation methods that are often used for the testlet models, estimates resulted form PQL are least accurate. Laplace and MCMC yield accurate parameter recovery and appropriate precision of estimates but it takes very long time to converge, and thus, has been rarely implemented in operational testing. MML estimation with EM algorithm is relatively efficient and its performance in parameter estimation is sufficient. However, for the dataset with very small testlet effects, the convergence could be very slow (Li & Rijmen, 2009).

The MML estimation procedure with EM algorithm is implemented in two stages. In the first stage, the likelihood function is integrated over the ability and testlet parameters. Since the integration cannot be done in a closed mathematical form, it is usually performed either through a numerical simulation of "plausible values" for the ability and testlet variables (Mislevy, Beaton, Kaplan, & Sheehan, 1992) or through a summation function over a grid of quadrature points. After the ability and testlet variables were integrated out, maximizing the likelihood function for the item parameters has to be conducted by using numerical optimization techniques such as the Newton-Raphson method, because the first-order likelihood equations cannot be solved directly. In the second stage, the ability and testlet parameters are estimated conditional on the assumed known point estimates of items parameters obtained from the first stage. Again the numerical method is employed to maximize the likelihood function. The MML procedure has to be used with caution in some operational test settings, because by inserting the point estimates of item parameters, we have ignored the uncertainty with which they are obtained, especially when the sample size of examinees is not large enough to neglect the uncertainty reduction elicited by the plug-in method (Wainer et al., 2007).

The hierarchical Bayesian computational method is often used for estimation of

TRT models. Through the Markov Chain Monte Carlo (MCMC) procedure, samples are drawn from the posterior distributions of model parameters along a Markov chain. After sufficient iterations, the sample distribution converges to the posterior distribution of interest.

The Bayesian MCMC method is appropriate when a number of unknown parameters need to be calibrated simultaneously, for example, in IRT models where structural as well as incidental parameters have to be estimated (Zellner, 1971). This is particularly true when the prior information about the parameters is available, since the incorporation of such information will certainly increase the meaningfulness and the "accuracy" of the estimates (Gifford & Swaminathan, 1990). Glas, Wainer, and Bradlow (2000) compared marginal maximum likelihood estimation (MMLE) and expected a posteriori (EAP) estimation with MCMC and pointed out that MMLE and MCMC estimates for the testlet model were highly correlated but MCMC provided more accurate interval and point estimate results. However, the prior distributions will have influence on the posterior estimates. In that case, the prior distribution should be carefully selected before the estimation is implemented.

Swaminathan and Gifford compared the estimates from the Bayesian procedure and the joint maximum likelihood (JML) procedure by conducting a set of simulation studies in the context of 1-PL (Swaminathan & Gifford, 1982), 2-PL (Swaminathan & Gifford, 1985), and 3-PL (Swaminathan & Gifford, 1986) IRT models. Results had consistently shown that despite the "shrunken" estimators, the Bayesian procedure produced better estimates than the JML procedure under the criterion of the mean squared differences between estimates and true values.

By using MCMC methods, the inference can be easily drawn based on the posterior samples that have been obtained, while for maximum likelihood (ML) methods inference may not always be available because on some occasions the estimated variable does not converge. Unlike ML in which the standard error of estimate (SEE) has to rely on asymptotic theory, MCMC methods allow finite-sample inference. In addition, by virtue of the specialized Bayesian software such as WinBUGS (Spiegelhalter, Thomas, & Best, 2003), the full conditional distributions can be constructed for any user specified models in an automatic way (Wainer, et al., 2007). The computation capacity related to MCMC has been enhanced to a large extent with the increase in computing power and the creation of methods such as the Metropolis-Hastings algorithm (Hastings, 1970), data augmentation applied to MCMC methods (Tanner & Wong, 1987), Griddy-Gibbs sampling methods (Ritter & Tanner, 1992), adaptive rejection sampling (Gilks, 1992), and slice sampling (Damien, Wakefield, & Walker, 1999).

Simulation studies showed that with well-identified parameters, different specifications of prior distributions had relatively minor effects on the Bayesian estimates as long as the prior distributions were not too extreme (Gifford & Swaminathan, 1990). If the parameters of the prior distribution are specified at extreme values, the numerical procedure often results in non-convergence. In addition, vague or diffuse priors seem to improve the quality of estimation, namely, they provide less biased estimates and higher correlations between true values and estimates. In that case, priors that are not too tight are preferable on these occasions (Gifford & Swaminathan, 1990; Swaminathan & Gifford, 1982, 1985, 1986). For example, in the context of the Rasch model, the priors of the variances of the person ability and item difficulty are often specified as the inverse chi-square distributions. As the degrees of freedom increase, the prior distributions become more concentrated, reflecting increasingly stronger beliefs about the values of variances, and the accuracy of estimation steadily increases. To prevent extreme bias, relatively small values of degrees of freedom between 5 and 15 are specified so that the prior distributions produce similar results (Swaminathan & Gifford, 1982).

The Bayesian procedure is reasonably robust to changes in the specification of the prior distribution for different sample sizes and test lengths. If the sample size is small, or available data provide only indirect information about the parameters of interest, the prior distribution becomes more important, but when the sample size is large enough, reasonable choices of prior distributions will have modest effects on posterior inferences. For example, in the case of the Rasch model, when the test length reaches 50 items and the sample size attains 500, the accuracy of estimates remains unchanged despite the increase in the degrees of freedom (Swaminathan & Gifford, 1982).

Prior distributions in the hierarchical Bayesian framework (Lindley & Smith, 1972) are specified based on the assumptions for a normal population. The values of the guessing parameters usually range between 0.1 and 0.3, item discriminations range between 1/3 and 3, and person abilities range between -3 and 3 (Mislevy, 1986). It is also assumed that on the first level the person ability and item difficulty parameters are independently, identically and normally distributed and this assumption of normality has little effect on the outcomes (Swaminathan & Gifford, 1982). The chi-square distribution or the normal approximation to the chi-square distribution is chosen to indicate prior belief on item discrimination and item difficulty. The prior distribution for guessing parameters may be taken as the beta distribution (Swaminathan & Gifford, 1985). On the second level, the hyper-priors are specified assuming that the mean is uniform and the variance follows an inverse chi-square distribution (Novick & Jackson, 1974; Swaminathan & Gifford, 1985).

The information on these parameters is exchangeable (Swaminathan & Gifford, 1986), that is, when very little knowledge is available about each parameter, identical prior distributions can be used for each parameter. However, different priors for each item can be specified to improve the "meaningfulness" of the estimates. These assumptions can be incorporated directly into the specifications of the priors.

2.4. Sequential Use of IRT and GT

It has been noticed that both the GT model and IRT models can be conceptualized as instances of multilevel models (Goldstein, 1995; Patz, Junker, Johnson, & Mariano, 2002; Verhelst & Verstralen, 2001). GT represents a linear model of main effects and interaction effects from the object of measurement and measurement facets. In the case of IRT models, the first level describes how the item effects and the person abilities shape the log-odds of a correct response. The second level defines how the abilities vary over the population of examinees (Raudenbush & Bryk, 2002, pp.365-371). Given the conceptual similarity between GT and IRT as multilevel random effects models, the sampling model of GT can be incorporated into the scaling model of IRT.

One example of this sequential use of IRT and GT can be seen in Briggs and Wilson's generalizability in item response modeling (GIRM) (2007), where GT and IRT models are combined by making distributional assumptions about the relevant measurement facets, so that variance components and generalizability coefficient estimates that are central to GT can be derived within an IRT framework. Another prominent example of this approach is to correct the item response information function and the standard errors for conditional dependence of multiple ratings by using a ratio of random variances derived from the GT analysis (Bock, et al., 2002)

2.4.1. Generalizability in Item Response Modeling

By specifying a random effect measurement model and using the MCMC estimation method, it becomes possible to estimate GT variance components simultaneously within the traditional IRT models. The paper of Briggs and Wilson (2007) demonstrated how GT and IRT could be linked together in the context of a single-facet measurement design with binary items. GIRM is essentially a GT analysis on a matrix of expected rather than observed item responses. The underlying model of the GIRM comes from IRT, but the results are used as the basis for a GT analysis.

The main effects and interaction effects from item and person as well as their variances are estimated by making distributional assumptions on these variables. Both simulated and empirical data analyses provided evidence that GIRM would lead to the same estimates of variance components and generalizability coefficients as would be reached through GT. Results from the simulation studies also showed that GIRM seemed to be robust to misspecification of distributional assumptions and the parametric form of the item response function.

Advantages of GIRM include (1) variance components for error and facet interaction effects can be estimated separately; (2) because the variances are estimated as a function of expected rather than observed responses, all measurement designs can be treated as if they were complete and balanced, and thus GIRM can provide answers to these designs where the use of GT may not be applicable; (3) most importantly, results from both a GT and IRT analysis are available within a single modeling framework. However, it is unknown whether the GIRM approach can be extended to the context of more complex measurement design as many of these designs may not be so easily expressed in the GT notation. In addition, the statistical assumptions on GIRM may not be met or need to be further justified. For example, the random sampling assumption in GT is not compatible with IRT, since the items included in the instrument do not represent a random sample from the universe of possible items. It is also necessary to further explore whether the GIRM approach is sensitive to violations of the standard IRT assumptions of unidimensionality and local independence.

2.4.2. Information Correction Method

In performance assessments, the open-ended responses are often read and scored by multiple raters so as to gain as much information as possible. However, as the multiple ratings on the same examinee's response are conditionally dependent given the ability level, they provide overlapping information about ability and attenuate random errors if the scores are treated conditionally independent, which presents a similar situation with the testlets. IRT does not provide a straightforward approach to the estimation of ability under conditions with LID as conditional independence is assumed in IRT. Therefore, Bock, Brennan and Muraki (2002) introduced using the variance ratio from generalizability analysis to correct the overestimated information function in IRT analysis for multiple ratings that were conditionally dependent. In their article, the procedure and the didactic examples were presented in a balanced design with items nested in raters, administered to all examinees. In correspondence with the situation in IRT analysis, the same assumption was taken with generalizability analysis, that is, the main effects of item difficulty and rater severity were fixed while the main effect of person ability was random.

The measurement error variances are estimated using GT in the design with the usual assumption of conditional independent ratings and in the design of nested multiple ratings respectively. Multiplying the conventional IRT error variance by the ratio of error variances in nested design to those in independent design will correct the underestimated IRT random errors. This method provides reliability estimates in terms of per-item generalizability and per-rater consistency in addition to the variance components and the traditional generalizability coefficients. These reliability estimates make it possible to show how trade-off between the number of items and the number of raters affects the reliability of test scores.

Bock, Brennan and Muraki (2002) contributed to the existing LID literature by proposing the information correction method. However, it is yet known whether this method is a better way to cope with LID compared with other models such as bi-factor analysis (Gibbons, 2001; Gibbons & Hedeker, 1992) or TRT analysis (Bradlow, et al., 1999; Wainer, et al., 2000; Wainer, et al., 2007; Wang, et al., 2002). Therefore, in the current study, it will be of interest to explore through simulations whether the information correction method will result in standard errors that are comparable to those of the Bayes estimates of examinee scores on the general factor under TRT models.

CHAPTER III: THEORETICAL FRAMEWORK

Since the purpose of the present study is to correct the measurement error of proficiency estimates from the testlet-based tests, we start with a simple situation where the main effects of items and testlets are assumed to be fixed. In that case, among all the partitioned variance terms above (Equation 2.4), the person effect, the person by testlet interaction, and the person by item within-testlet confounded with residuals are random effects to be involved in the generalizability analysis of our study. It happens that in the linear predictors of either IRT or TRT models, they are also regarded as facets with random effects, while the item difficulties are fixed effects.

In the GT framework, total variances could be expressed as $\sigma^2(X)=\sigma^2(\tau^2)+\sigma^2(\delta^2)$. $\sigma^2(\tau)$ represents the variance of the universe score which is based only on the variance of the person variable, $\sigma^2(i)$, while the relative error variance is composed of variance of the person by testlet interaction $\sigma^2(iD)$ and variance of persons by items-within-testlet interaction $\sigma^2(iJ{:}D)$, which can be expressed as $\sigma^2(\delta^2)= \sigma^2(iD)+ \sigma^2(iJ{:}D)$.

The proficiency of examinee i can be estimated by the mean scores on j items nested within d testlets.

$$X_{ij:d} = \tau_i + \varepsilon_{id} + \xi_{ij:d}\,, \tag{3.1}$$

where j=1,…,n; and d=1,…m.

τ_i denotes the grand mean;

ε_{id} denotes the person by testlet interaction effect;

$\xi_{ij:d}$ denotes the person by item within testlet interaction effect.

σ^2(i), σ^2(iD), and σ^2(iJ:D) are variance terms on τ_i, ε_{id}, and $\xi_{ij:d}$.

$n = \sum_{d=1}^{m} k_d$, where k_d indicates the number of items in each testlet. k_d=k=n/m when the test has a balanced design with equal number of items in each testlet. Then, to estimate the mean score of each examinee as the best estimator of each person's ability,

$$\begin{aligned} X_{i..} &= \frac{1}{n}\sum_{d}^{m}\sum_{j:d}^{k_d} X_{ij:d} \\ &= \tau_i + \frac{1}{m}\sum_{d}^{m}\varepsilon_{id} + \frac{1}{n}\sum_{d}^{m}\sum_{j:d}^{k_d}\xi_{ij:d} \end{aligned} \tag{3.2}$$

The variance of the mean score estimate is

$$\sigma^2(\mathrm{X}_{i..}) = \sigma^2(\mathrm{i}) + \frac{1}{m}\sigma^2(\mathrm{iD}) + \frac{1}{\mathrm{n}}\sigma^2(\mathrm{iJ{:}D})\,. \tag{3.3}$$

And the random error variance component is

$$\frac{1}{m}\sigma^2(\mathrm{iD}) + \frac{1}{\mathrm{n}}\sigma^2(\mathrm{iJ{:}D})\,. \tag{3.4}$$

Suppose the testlet facet is ignored as if all items are independent. The variance other than the true variance is the error variance which, in this case, is composed of the variance of the interaction between persons and items.

$$\begin{aligned} X_{i.} &= \frac{1}{n}\sum_{j}^{n} X_{ij} \\ &= \tau_i + \frac{1}{n}\sum_{d}^{m}\sum_{j:d}^{k_d}(\varepsilon_{id} + \xi_{ij:d}) \end{aligned} \tag{3.5}$$

The variance of the mean score estimate should be

$$\begin{aligned} \sigma^2(\mathrm{X}_{i.}) &= \sigma^2(\mathrm{i}) + \frac{1}{n}\sigma^2(\mathrm{iJ}) \\ &= \sigma^2(\mathrm{i}) + \frac{1}{n}(\sigma^2(\mathrm{iD}) + \sigma^2(\mathrm{iJ{:}D})) \end{aligned}\,. \tag{3.6}$$

The random error variance is

$$\frac{1}{n}(\sigma^2(\text{iD})+\sigma^2(\text{iJ:D}))\,. \tag{3.7}$$

The random error variance of testlet data is $\sigma^2(iD)/m+\sigma^2(iJ:D)/n$ (Equation 3.4), while the random error variance of the conventional model assuming CI is $(\sigma^2(iD)+\sigma^2(iJ:D))/n$ (Equation 3.7).

Therefore, $\dfrac{\sigma^2(iD)/m+\sigma^2(iJ:D)/n}{(\sigma^2(iD)+\sigma^2(iJ:D))/n}$ is the correction for the underestimated error variances as a result of model misspecification on the testlet data in the situation described above. We would suggest the ratio of the random error variances under testlet design and independent item design as a practical approximation term to correct the standard error variance when the testlet data are treated as independent responses and fit with conventional IRT models.

In the context of IRT, the information function is the reciprocal of the standard error variance. Thus, the ratio to correct the testlet-specific information is

$$t_d=\frac{\sigma^2(iD)+\sigma^2(iJ:D)}{k_d\sigma^2(iD)+\sigma^2(iJ:D)} \tag{3.8}$$

k_d is the number of items in each testlet.

In a conventional IRT model that can be generalized to the categorical responses, the likelihood function for the (N x n) vector u of the responses of N examinees on n items is

$$\begin{aligned}L(\mathbf{u}\,|\,\boldsymbol{\theta})&\equiv L(\mathbf{u}_1,\mathbf{u}_2,...,\mathbf{u}_N\,|\,\theta_1,\theta_2,...,\theta_N)=\prod_{i=1}^{N}L(\mathbf{u}_i\,|\,\theta_i)\\&=\prod_{i=1}^{N}\prod_{j=1}^{n}P(u_{ij}\,|\,\theta)=\prod_{i=1}^{N}\prod_{j=1}^{n}\prod_{h}^{q_j}[P_{hij}(\theta)]^{x_{hij}}\end{aligned}\quad, \tag{3.9}$$

where i is the index of examinee, (i=1,…,N);

j is the index of items;

$\boldsymbol{\theta}$ denotes the ability of examinees;

u is the response vector consisting of x_{hij};

x_{hij} is the indicator variable taking on the value 1 if the response of examinee i to item j is assigned to category h, and the value 0 otherwise, $(h=1,\ldots, q_j)$;

P_{hij} is the categorical response function for examinee i and item j.

The information function is given by the following expression

$$I(\theta)=\sum_{j}^{n}\sum_{h}^{q_j}\left\{\frac{x_{hij}}{[P_{hij}(\theta)]^2}\left[\frac{\partial P_{hij}(\theta)}{\partial\theta}\right]^2-\frac{x_{hij}}{P_{hij}(\theta)}\frac{\partial^2 P_{hij}(\theta)}{\partial\theta^2}\right\}. \tag{3.10}$$

For a testlet dataset where item j is nested in testlet d, in order to estimate θ, the likelihood function adjusted by the correction term t_d is

$$L_i^*=\prod_{d}^{m}(\prod_{j:d}^{k_d}\prod_{h}^{q_j}[P_{hj:d}(\theta)]^{x_{hij:d}})^{t_d}. \tag{3.11}$$

The quantity to be maximized under the restriction $\sum_h^{q_j}P_{hij:d}(\theta)=1$ is

$$\log L_i^*(\theta)-\lambda[\sum_{h}^{q_j}P_{hj:d}(\theta)-1]. \tag{3.12}$$

The first order condition is

$$\sum_{d}^{m}t_d\sum_{j:d}^{k_d}\sum_{h}^{q_j}\frac{x_{hij:d}}{P_{hj:d}(\theta)}\cdot\frac{\partial P_{hj:d}(\theta)}{\partial\theta}-\lambda\sum_{h}^{q_j}\frac{\partial P_{hj:d}(\theta)}{\partial\theta}=0, \tag{3.13}$$

from which we can derive

$$\lambda=\frac{\sum_{d}^{m}t_d\sum_{j:d}^{k_d}\sum_{h}^{q_j}\frac{x_{hij:d}}{P_{hj:d}(\theta)}\cdot\frac{\partial P_{hj:d}(\theta)}{\partial\theta}}{\sum_{h}^{q_j}\frac{\partial P_{hj:d}(\theta)}{\partial\theta}}. \tag{3.14}$$

The second derivative of the log-likelihood function is:

$$\frac{\partial^2\ln L}{\partial\theta^2}=\sum_{d}^{m}t_d\sum_{j:d}^{k_d}\sum_{h}^{q_j}\left\{-\frac{x_{hij:d}}{[P_{hj:d}(\theta)]^2}\left[\frac{\partial P_{hj:d}(\theta)}{\partial\theta}\right]^2+\frac{x_{hij:d}}{P_{hj:d}(\theta)}\frac{\partial^2 P_{hj:d}(\theta)}{\partial\theta^2}\right\}. \tag{3.15}$$

Therefore, the information matrix is

$$I(\theta)=\sum_{d}^{m} t_d \sum_{j:d}^{k_d} \sum_{h}^{q_j} \left\{ \frac{x_{hij:d}}{[P_{hj:d}(\theta)]^2} \left[\frac{\partial P_{hj:d}(\theta)}{\partial \theta} \right]^2 - \frac{x_{hij:d}}{P_{hj:d}(\theta)} \frac{\partial^2 P_{hj:d}(\theta)}{\partial \theta^2} \right\}. \qquad (3.16)$$

Comparing Equation (3.16) with (3.10), we can conclude that t_d can be used as a coefficient for correction.

For dichotomous items, q_j=2, $P_{2j:d}(\theta)$=1-$P_{1j:d}(\theta)$, and $x_{2ij:d}$= 1-$x_{1ij:d}$, the information function is

$$I(\theta)=\sum_{d}^{m} t_d \sum_{j:d}^{k_d} [P_{j:d}(\theta)]^{x_{ij:d}} [1-P_{j:d}(\theta)]^{1-x_{ij:d}} \quad , \qquad (3.17)$$

where $P_{j:d}$ is the response function for the correct response.

The following steps are implemented to obtain the correction from the generalizability analysis (shown in Table 3.1 through Table 3.3).

Table 3.1 Analogous T Terms for Unbalanced i*(j:d)

Source of variation	T
Examinees	TE=$n\sum_i \bar{X}_i^2$
testlets	TT=$N\sum_d (n_{j:d}\bar{X}_d^2)$
Items:Testlets	TIT=$N\sum_d \sum_{j:d} \bar{X}_{j:d}^2$
Testlets x Examinees	TTE=$\sum_d (n_{j:d} \sum_i \bar{X}_{id}^2)$
(Items: Testlets) x Examinees	TITE=$\sum_i \sum_j X_{ij}^2$
Mean(μ)	Tmean=$n\bar{X}^2$

Note: $\bar{X}_i$ is the mean across all items for each person;

$\bar{X}_d$ is the mean across a cluster of items and all persons for each testlet;

$\bar{X}_{j:d}$ is the mean across all persons for each item;

$\bar{X}_{id}$ is the mean across a cluster of items for each testlet and each person;

$\bar{X}$ is the grand mean across all items and all persons.

Table 3.2: Sum of Squares

Source of variation	Sum of Squares
Examinees	SSE=TE-Tmean
Testlets x Examinees	SSTE=TTE-TT-TE-Tmean
(Items: Testlets) x Examinees	SSITE=TITE-TIT-TTE+TT

Table 3.3: Expected Random–effect Variances

Source of variation	Examinees	Testlets x Examinees	(Items: Testlets) x Examinees
Mean Squares	MSE=SSE/(N-1)	MSTE=SSTE/(N-1)(m-1)	MSITE=SSI:TE/(N-1)(n-m)
Expected Mean Squares	$n\sigma^2(i)+r_d\sigma^2(iD)+\sigma^2(iJ{:}D)$	$t_d\sigma^2(iD)+ \sigma^2(iJ{:}D)$	$\sigma^2(iJ{:}D)$
Estimated Variance components	$\sigma^2(i)=(MSE-r_dMSTE/t_d+$ $(r_d-t_d)MSITE/t_d)/n$	$\sigma^2(iD)=(MSIE-MSITE)/t_d$	$\sigma^2(iJ{:}D)=MSITE$

N is the total number of examinees, n is the total number of items, and m is the number of testlets. k is the number of items in each testlet. For balanced design, r=t=k; for unbalanced design, $r=\sum_d \frac{k_d^2}{n}$ and $t=\frac{n-r}{m-1}$ (Brennan, 2001). The correction term for item-specific information is obtained through Equation 3.8.

In GT models, the mean of the set of item scores given to an examinee is the linear unbiased estimator of the ability of that individual, and the error variance is composed of random error variances of the facets and their interactions. TRT models produce relatively more accurate estimates from data with certain magnitude of LID. Thus, we may conjecture that the variance of the ability parameter from the IRT model corresponds to the random error variance of estimates from an independent item design in GT, while the variance of the primary ability parameter from the TRT model corresponds to the random error variance of estimates from a testlet design in GT. Therefore, we can use the ratio of random error variances of the ability parameter in an independent design and a testlet design by generalizability analysis to adjust the estimated measurement error to a more appropriate level. This is the reasoning about the relationship between these error variances. However, it is necessary to further obtain the quantitative evidence in regard to the performance of the information correction method through manipulating factors in simulation studies.

CHAPTER IV: METHODOLOGY

The section consists of two parts: a simulation study and a real data analysis. The simulation study is intended to examine the performance of the information correction method in adjusting measurement error in 1-PL, 2-PL, and 3-PL IRT models by manipulating three factors, namely, the magnitude of LID (indicated by the variance of testlet effect variable), the length of testlet and the balance of testlet length. In the real data analysis, the information correction method is applied and evaluated by using a real dataset from a testlet-based achievement test.

4.1. Simulation Study

4.1.1. Design

The purpose of this simulation study is to evaluate the performance of the information correction method in adjusting the measurement error of proficiency parameters by specifying 1-PL, 2-PL and 3-PL IRT models to testlet datasets. The results are compared with the expected error variances from TRT models with the same number of parameters. The research questions of interest in this study are

1. What factors have significant effects on the performance of the information correction method, which is presented in the difference between SEEs of proficiency from IRT models adjusted by the information correction ratio and SEEs from TRT models? In this study, this criterion variable is named the standard error increase discrepancy (SEID) and formulated as

$$\frac{SEE_{TRT} - SEE_{IRT}}{SEE_{IRT}} - \frac{SEE_{IRT-t_d} - SEE_{IRT}}{SEE_{IRT}} = \frac{SEE_{TRT} - SEE_{IRT-t_d}}{SEE_{IRT}} \tag{4.1}$$

2. How do the distributional characteristics of the proficiency estimates change across the simulation conditions?

It has been presented in the previous studies that LID and testlet length are factors that would affect the parameter recovery from the testlet datasets as a result of model misspecification. Accuracy in proficiency and item parameter estimates deteriorates when either the testlet length or the magnitude of LID increases (Bradlow et al., 1999; Sireci et al., 1991; Wainer, 1995; Yen, 1993). In specific, the testlet models tend to provide more accurate and precise parameter estimates than the independent item models (Bradlow, et al., 1999; DeMars, 2006; Jiao, et al., 2005; Jiao & Wang, 2008; Wainer, et al., 2000; Wainer, et al., 2007; Wang, et al., 2002; Wang & Wilson, 2005a, 2005b). The modest testlet length will have minimal effect on precision of estimates if LID is ignored in the testlets (Bradlow, et al., 1999, Wang, et al, 2002). In addition, the information correction ratio is a testlet-specific statistic that depends on the length of each testlet, so the adjusted SEE is weighted by the length of each testlet nonlinearly. Therefore, the balance of testlet length in the test also counts in this investigation in addition to LID and the test length.

The performance of the information correction method needs to be investigated in the contexts of 1-PL, 2-PL and 3-PL response theory models respectively, because each of them has a different representation of the information function. For 1-PL model, since each item has the same discrimination value, the distributions of the information function are equal. For 2-PL model, each item has a different slope, and distributions of information are different as well. The maximum amount of information provided by an item increases as the item discrimination increases. In the 3-PL model, With the presence of guessing parameters, all other things being equal, the amount of information an item provides decreases as the amount of guessing increases. In addition, compared to item discriminations and difficulties, guessing parameters are hard to estimate. These differences among the three types of models may lead to distinctive performances of the information correction method.

Thus, three simulation factors are manipulated: (a) LID—the variance of the random testlet variables, specified at 0, 0.25, 1, representing zero, small and large testlet effect respectively; (b) testlet length—short and long testlets (i.e., a testlet consisting of fewer than 10 items is regarded as the short testlet, while a testlet consisting of more than 10 items is regarded as the long testlet); and (c) balance of testlet length across the test—balanced design (i.e., equal number of items in each testlet), intermediately unbalanced design (e.g., 4 items in one testlet and 6 items in another, or 8 items in one testlet and 12 items in another) and extremely unbalanced design (e.g., 2 items in one testlet and 8 items in another, or 4 items in one testlet and 16 items in another). The data are generated by 1-PL, 2-PL and 3-PL TRT models respectively and are calibrated using IRT or TRT models with the same number of item parameters. These three factors in the context of three models compose $3 \times 2 \times 3 \times 3 = 54$ conditions. The conditions are described and numbered in Table 4.1.

Table 4.1: Simulation Design

Level of balance	models	# items per testlet	Variance of testlet effect ($\sigma_{\gamma h(i)}^2$)		
			.01	.25	1
Balanced	1-PL	5	S1	S2	S3
		10	S4	S5	S6
	2-PL	5	S7	S8	S9
		10	S10	S11	S12
	3-PL	5	S13	S14	S15
		10	S16	S17	S18
medium unbalanced	1-PL	4,6	S19	S20	S21
		8,12	S22	S23	S24
	2-PL	4,6	S25	S26	S27
		8,12	S28	S29	S30
	3-PL	4,6	S31	S32	S33
		8,12	S34	S35	S36
Extreme unbalanced	1-PL	2,8	S37	S38	S39
		4,16	S40	S41	S42
	2-PL	2,8	S43	S44	S45
		4,16	S46	S47	S48
	3-PL	2,8	S49	S50	S51
		4,16	S52	S53	S54

4.1.2. Data Generation

For each condition, the test consists of 60 dichotomous items and the sample size of examinees is 500. The values of the parameters are generated from the distributions specified in Table 4.2. The true values of ability parameters follow a standard normal distribution, $\theta \sim N(0,1)$. This set of θ values is fixed across all conditions. The values of the testlet variable are randomly generated from a normal distribution with a mean of zero and a variance of $\sigma_{\gamma d(i)}^2$. The true values of the difficulty parameters are generated from a standard normal distribution β~N(0,1) and truncated within a range of [-1.5, 1.5]. For responses from the 2-PL and 3-PL models, the true values of the discrimination parameters are generated from a normal distribution α~N(0.8,0.2), within the range of [0.6, 1.4]. For responses from the 3-PL model, the true values of the guessing parameters follow a normal distribution, ω~N(0.14, 0.05), within the range of [0,0.25]. The marginal distributions of these parameters are chosen based on a typical form of the Scholastic Assessment Test (SAT) (Bradlow, et al., 1999).

With the above item and person parameter values available, the probability of correct response is calculated by using the 1-PL TRT model (Equation 4.2), the 2-PL TRT model (Equation 4.3), and the 3-PL TRT model (Equation 2.12) respectively.

$$P(y_{ij}=1)=\frac{\exp(\theta_i-\beta_j-\gamma_{id(j)})}{1+\exp(\theta_i-\beta_j-\gamma_{id(j)})}. \tag{4.2}$$

Where β_j is the item difficulty of the j_{th} item;

θ_i is the person proficiency of the i_{th} person;

$\gamma_{id(j)}$ parameterizes the random effect for person i on testlet d(j);

$P(y_{ij}=1)$ is the probability of correct response from person i on item j.

$$P(y_{ij}=1)=\frac{\exp(\alpha_j(\theta_i-\beta_j-\gamma_{id(j)}))}{1+\exp(\alpha_j(\theta_i-\beta_j-\gamma_{id(j)}))} \tag{4.3}$$

where α_j is the item discrimination for item j.

The item scores are simulated using the Bernoulli distribution function based on the probability of the correct response calculated from Equation 4.2, 4.3 or 2.12. Each condition is replicated for 50 times.

Table 4.2: Simulation specifications

Parameters	Distributions
α	~N(0.8,0.2)
β	~N(0,1)
ω	~N(0.14, 0.05)
θ	~N(0,1)
γ	Zero: =0 Small: ~N(0,0.25) Large: ~N(0,1)

4.1.3. Analysis

Calibration. In order to evaluate the performance of the information correction method in adjusting the measurement error from IRT, each dataset is calibrated by IRT and TRT models respectively. The models for calibration use the same number of item parameters as in the models for data generation. Among all the estimation methods, ML estimation fails unless the examinees who obtain perfect scores or zero scores are removed prior to estimation. In contrast, the Bayesian MCMC method provides a flexible and straightforward approach for calibration with either IRT or TRT models. Because the prior distribution represents the belief about the parameter and will pull the posterior estimate towards the prior mean, the incorporation of prior information will increase the meaningfulness and accuracy of the posterior estimates. The calibration with the MCMC method is implemented by using WinBUGS embedded in R.

At the early stage of this research, since this study is targeted at the SEEs of the primary ability parameters, the item parameter values are fixed in order to speed up the estimation procedure. The prior distributions of the ability parameters and random testlet variables are specified as follows,

$$\begin{aligned} &\theta_i \sim N(0,1) \\ &\gamma_{id(j)} \sim N(0,\sigma^2_{\gamma d(j)}). \end{aligned} \tag{4.4}$$

The variance of the testlet effect $\sigma_{\gamma d(j)}{}^2$ indicates the strength of LID for testlet d(j). To estimate the variance of the testlet effect, a hyperprior for $\sigma_{\gamma d(j)}{}^2$ was specified as an inverse gamma distribution with shape parameter g=1 and scale parameter ξ=1.

$$\sigma^2_{\gamma_{d(j)}} \sim Gamma^{-1}(1,1) \tag{4.5}$$

It is also assumed that the information on examinees is exchangeable, that is, prior to observing item responses, the analyst's belief about the ability of any one examinee is no different from that of any other examinee.

To expedite convergence, ML point estimates of proficiency parameters from the IRT calibration are used as initial values. Two chains of iterations are run for each dataset. Convergence for a dataset of 60 items and 500 examinees usually occur within 1000 iterations (Bradlow, et al., 1999; Wainer, et al., 2000). To ensure that convergence would be achieved before a certain number of iterations, two chains of iterations are run first on a sample dataset generated using the same simulation specifications in each condition. Several convergence diagnostic criteria are available in WINBUGS: the dynamic trace lines, history plots, auto-correlation lines, Gelman-Rubin convergence statistics, and quantile plots. The definitions of these diagnostic criteria as well as their diagnostic graphs will be illustrated in the real data example.

The mean of the posterior distribution is regarded as the optimal estimate of the proficiency parameter; and the standard deviation of the posterior distribution is taken as the standard error of the proficiency estimate. Upon convergence, as one criterion ascertaining sufficient iterations have been run to best represent the posterior distribution, the MC errors should be no more than approximately 5% of the standard deviations of the posterior distributions.

The number of burn-in cycles and the sufficient number of iterations for the estimation of the posterior distributions depend on the complexity of the model and the sample size. For example, for calibration using the 2-PL TRT model, 4500 cycles are run for each chain and the first 1000 are discarded as burn-in cycles. The calibration of one dataset composed of 60 items nested in 10 testlets and 500 examinees is completed within 20 minutes on a desktop with a 1.8 Ghz Central Processor Unit; while IRT calibration usually takes no more than 10 minutes.

Following estimation, the point estimates and the estimated SEEs of the ability parameters from IRT and TRT models are compared. Parameter recovery from IRT and TRT model calibrations are compared and evaluated in terms of bias (Equation 4.6), absolute bias (Equation 4.7), empirical SEE (Equation 4.10), mean theoretical SEE (Equation 4.9) and root mean squared error (RMSE) (Equation 4.8) of the ability estimates averaged across all replications. The absolute bias is used in case the negative and positive biases set off. The equations for these statistics are shown as follows,

$$Bias(\hat{\theta}) = \frac{\sum_{r=1}^{R}(\hat{\theta}_r - \theta)}{R} \tag{4.6}$$

$$abs\ Bias(\hat{\theta}) = \frac{\sum_{r=1}^{R}|\hat{\theta}_r - \theta|}{R} \tag{4.7}$$

$$RMSE(\hat{\theta}) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\theta}_r - \theta\right)^2} \tag{4.8}$$

$$SEE(\hat{\theta}) = I^{-1/2}(\hat{\theta}) \tag{4.9}$$

$$SEE_{empirical}(\hat{\theta}) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\theta}_r - \frac{\sum_{t=1}^{R}\hat{\theta}_t}{R}\right)^2} \ . \qquad (4.10)$$

It is expected that when the variance of $\gamma_{id(j)}$ increases, that is, when the level of model misspecification increases, the IRT model will provide more biased estimates of proficiency. It is also hypothesized that the TRT model will yield less biased estimates of proficiency when there is substantial LID in testlets. As for precision, it is expected that when the variance of $\gamma_{id(j)}$ increases, the theoretical standard errors of proficiency estimates should increase if the model accounts for the testlet effect variable. The reason is that when the level of LID increases, the amount of overlapping information will also increase, which leads to less information in estimation. We might expect to see that the standard errors of proficiency estimates from the IRT calibration remain unchanged despite the increase in LID, because the IRT model does not account for this testlet effect variable, but when the testlet effect is considered, the level of underestimation as compared against SEE from the TRT calibration will increase, because the amount of information from the test decreases.

Information correction procedure. The variance components are derived from the generalizability analysis based on Table 3.1-3.3 that were described in the chapter of the theoretical framework. The information correction ratio (t_d) is calculated based on Equation 3.8. The square root of the inverse of this information correction ratio ($t_d^{-1/2}$) is applied to adjust the SEE of ability parameter calibrated with the unidimensional IRT models (i.e., 1-PL, 2-PL, and 3-PL IRT models). For each condition, these adjusted conditional SEEs are plotted against the ability scale and compared with SEEs from the IRT models as well as SEEs from the corresponding TRT models. The increase rates in SEE as a result of the information correction $\frac{SEE_{IRT-t_d} - SEE_{IRT}}{SEE_{IRT}}$ or as a result of the TRT modeling $\frac{SEE_{TRT} - SEE_{IRT}}{SEE_{IRT}}$ with the IRT SEE as the baseline are computed. SEID (Equation 4.1) is the dependent variable used to evaluate the effect of adjustment by the information correction method. A SEID value close to zero indicates sufficient adjustment in error variance and hence a good performance of the information correction method. To determine the effects of each manipulated factor on the performance of the information correction method, descriptive and inferential statistics (analyses of variance, ANOVA) are presented to determine whether the observed differences across simulation conditions in the dependent variable are of statistical significance.

4.2. A Real Data Example

4.2.1. Data

This is a large-scale test that was administered to 827 examinees in Grade 3 to assess their reading skills. The test consists of 40 multiple choice items nested in 9 testlets. Each testlet is associated with a reading passage. Since one of the testlets only contains two items, which presents difficulty in producing an accurate estimate of the random testlet effect, those two items were deleted. A brief summary of the testlet structure is shown in Table 4.3.

Table 4.3: the structure of reading comprehension test

testlet	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Number of items	6	3	3	6	6	5	3	6

4.2.2. Research questions

1. What are the characteristics of the test in terms of LID and dimensionality?
2. Which type of the response theory model fits the response data best?
3. How are the estimates from the IRT model compared with the estimates from the TRT model with the same number of item parameters by using Bayesian estimation through the MCMC procedure?
4. How are SEEs of proficiency estimates from the IRT model compared with SEEs adjusted by the information correction method, and SEEs from the TRT model?

4.2.3. Analysis

CI assessment. The local independence assumption of the IRT models is evaluated using Yen’s (1984) Q_3 statistics, which is calculated from the correlation of the residuals of an item pair based on IRT models. For test forms that exhibit no or minor LID, the unidimensional IRT models are more parsimonious models and might produce more accurate estimates than their TRT counterparts. Thus, it is necessary to know whether the test design and the item format conform to the characteristics of testlets and would allow the applications of testlet models and the information correction method. With respect to the real data, since we have no knowledge about the degree of LID in advance as we do for simulated data, we would need to perform LID tests and assess the magnitude of LID before proceeding with the application of testlet models and the information correction method. The distributional characteristics of the Q_3 statistics of each testlet are computed and compared with the expected value of Q_3 statistics. To understand which simulations are closest to the real data, the Q_3 statistics of one dataset in each condition in the 3-PL context are also estimated.

Factor analysis. To evaluate the unidimensionality assumption, the exploratory factor analysis is conducted on a tetrachoric correlation matrix of response variables on the test responses, since the test is composed of items with binary responses. The testlet model is essentially a special case of unidimensional IRT model, because the general factor will systematically affect examinees' performance in the tests, and the factors other than the primary factor can be regarded as nuisance factors limited within the testlet level. Thus, we expect one dominant factor as a result of the exploratory factor analysis.

The exploratory full-information item analysis is implemented in TESTFACT 4.0 (Wilson, Wood, & Gibbons, 1991), by using all the information in the data matrix through MMLE. First the smoothed tetrachoric correlation matrix is obtained, and then TESTFACT performs a principal factor analysis on the correlation matrix by using the minimum squared residuals (MINRES) method. Factors are extracted and factor loadings are calculated. In the initial solution generated by MINRES, the factors are orthogonal to each other, and can be subjected to varimax (factors being orthogonal) or promax (factors being oblique). Varimax rotation is chosen in this example. Determining the number of factors with the exploratory solution provided by TESTFACT involves examining the latent roots of the tetrachoric correlation matrix, the root mean square residual (RMSR) statistic for the matrix of residuals, chi-square difference statistics, and the number of substantial loadings for the factors

(Stone & Yeh, 2006). As suggested by many researchers (for example, Gorsuch, 1983), examination of scree plots is useful for determining the number of factors.

Model selection. While the testlet models yield more accurate results when fitting the data with certain magnitude of testlet effect, they would be over-fitting to the data with minimum amount of LID. Likewise, the 2-PL model would be parsimonious and have a better model fit compared with the 3-PL model if the pseudo guessing parameter values are not significantly different. The four types of models (i.e., 2-PL IRT, 2-PL TRT, 3-PL IRT and 3-PL TRT) are potential candidates for this test dataset, which all make sense if understood within their own conceptual framework. These four models are expected to lead to different solutions and interpretations, which need to be evaluated on the basis of model fits.

The information criteria, such as AIC, BIC, and CAIC, are especially useful in comparing models with a non-nested relationship. They are calculated using the MLEs of the model parameters. However, when the model parameters are estimated via methods other than maximum likelihood estimation, modified versions of these criteria are considered appropriate. Congdon (2003) suggested calculating AIC and BIC using the posterior means of the parameters in Bayesian modeling when the parameters were estimated via Markov Chain Monte Carlo (MCMC) sampling methods. The AIC and BIC described by Congdon were studied along with other model selection indices for mixture IRT models in Li, Cohen, Kim, and Cho (2006), and the results suggested that BIC performed the best in terms of correctness and consistency.

In this dissertation, the Deviance Information Criterion (DIC) is used to select the model with the best fit. DIC is a built-in function in WINBUGS 1.4 (Spiegelhalter et al., 2003) in which parameter calibration is implemented. Compared with other model fit indices of AIC or BIC, DIC is effective in complex hierarchical models where parameters may outnumber observations (Gelfand & Dey, 1994). DIC defines not only a measure of fit but also a measure of complexity. p_D is the complexity measure that is defined as the difference between the posterior mean of the deviance and the deviance at the posterior estimates of the parameters of interest (Spiegelhalter, Best, Carlin, & Van der Linden, 2002). The model with the minimum DIC is the one preferred.

Parameter calibration. Model estimation was implemented in WinBUGS through MCMC procedure. As suggested by the results of the previous simulation studies, the Bayesian procedure is relatively robust to different specifications of prior distributions so long as the parameters are well-identified and not too extreme (Gifford & Swaminathan, 1990; Swaminathan & Gifford, 1982, 1985, 1986). In that case, the prior distributions are specified based on the convention of a typical achievement test (Bradlow, et al., 1999;DeMars, 2006; Li, et al., 2005; Wainer, et al., 2000), and were given by

$$\begin{aligned}&\theta_i \sim N(0,1)\\&\alpha_j \sim \mathrm{logN}(0,0.25)\\&\beta_j \sim N(0,4)\\&\omega_j \sim beta(5,17)\\&\gamma_{id(j)} \sim N(0,\sigma^2_{\gamma d(j)})\end{aligned} \quad , \qquad (4.11)$$

where θ_i is the person proficiency of person i;α_j is the item discrimination on item j; β_j is the item difficulty of item j; and ω_j is the pseudo guessing parameter of item j. The hyper-prior distribution of the variance of testlet effect $\sigma^2_{\gamma_{d(j)}}$ is assumed to be an inverse gamma distribution with α and β parameters both set at 1, $\sigma^2_{\gamma_{d(j)}} \sim Gamma^{-1}(1,1)$. Two chains of item difficulty parameters with very divergent values (-2,-2,……,-2) and (2,2,……,2) are run for each model and the program is requested to generate the other starting values. The purpose for running two chains is to ensure the convergence of two chains when the estimates reach stationary.

Several convergence diagnostic criteria are available in WINBUGS: the dynamic trace lines, history, auto-correlation lines, Gelman-Rubin convergence statistics, and quantile plots. Dynamic trace lines, history and Gelman-Rubin convergence statistics are often used for the purpose of convergence diagnosis. Auto-correlation lines is not always a reliable indicator as estimation of some parameters can have reached stationary while their autocorrelations are still high. For 2-PL IRT model, the diagnostic graphs and statistics indicate that the two chains achieve convergence within the first 2000 iterations. To be conservative, the first 4000 burn-in cycles are discarded. When 3-PL IRT, 2-PL and 3-PL TRT are fit onto the data, the adaptive box is automatically checked, which suggests that WinBUGS is using a complex sampler such as a Metropolis sampler. On this circumstance, the default number of burn-in iterations is 4000. For 2-PL TRT model, the first 4000 cycles are burnt in. For 3-PL IRT and TRT models, the diagnostic indictors seem to suggest that convergence do not happen until 5000th iteration. In this case, the first 6000 iterations are discarded as burned-in cycles. The estimated standard errors and the point estimates of ability parameter are extracted from the statistics of the posterior distributions.

Information correction. The information correction ratios are estimated by following the steps shown in Table 3.1-3.3. The standard errors of the proficiency estimates from the selected IRT model are corrected by the estimated information correction ratios. The effect of the correction is evaluated by comparing the IRT SEEs, the IRT SEEs adjusted by the correction ratios and the TRT SEEs. To understand the possible differences and similarities between the real test analysis and the simulated data analysis, comparisons are made in terms of LID (indicated by Q_3), the testlet length, the balance of testlet length and the discrepancy between the adjusted IRT SEE and TRT SEE (indicated by SEID).

CHAPTER V: RESULTS

5.1. Simulation Study

5.1.1. Ability Parameter Recovery

The distributional characteristic statistics of the proficiency estimates are presented in Table 5.1. It contains the indicators of general accuracy (i.e., RMSE), bias (i.e., bias and absolute bias) and precision (i.e., empirical SEE and mean theoretical SEE) averaged across all examinees in each condition.

Bias and absolute bias evaluate estimates against their true parameters. The trend of change in bias across conditions is not obvious, possibly because the negative and positive biases across replications cancel out each other. The average absolute bias increases with the increase in LID. It seems to imply that proficiency estimates are more biased when the model is misspecified to a higher degree. Based on the bias statistics in Table 5.1, when the variance of the testlet effect is 0, that is, in the case of the test composed of independent items, the biases from the TRT calibration are generally higher than those from the IRT calibration, which indicates the overparameterization of TRT models on the independent item datasets. In contrast, in the conditions with high LID (i.e., $\sigma_{\gamma}^{2}=1$), the biases from the TRT calibration are slightly lower than those from the IRT calibration, which implies that TRT models fit the testlet data better and provide less biased estimates. Figure 5.1-5.6 show the absolute bias of IRT and TRT estimates from 1-PL, 2-PL and 3-PL contexts respectively. The absolute bias of IRT and TRT estimates are similar in pattern and they change in the same direction as the magnitude of LID. Long testlets seem to have higher biases than short testlets, especially in the conditions where the magnitude of LID is large, which is aligned with the results from the previous studies that short testlets tend to have moderate effects on estimation even though the independent item models are misspecified to the testlet dataset. The 3-PL models tend to result in larger biases than 2-PL models, which in turn result in larger biases than 1-PL models.

RMSE indicates the overall accuracy of proficiency estimates. The average RMSE increases as LID increases, which implies that the accuracy of proficiency estimates deteriorates as the level of model misspecification increases. This result is consistent with those from all the other simulation studies on testlet effects (Bradlow, et al., 1999; DeMars, 2006; Jiao, et al., 2005; Jiao & Wang, 2008; Wainer, et al., 2000; Wainer, et al., 2007; Wang, et al., 2002; Wang & Wilson, 2005a, 2005b). Figure 5.7-5.12 illustrate that RMSE follows the same pattern of change across conditions as the absolute bias, except that the magnitude of the differences in RMSE between the IRT and TRT calibrations are higher than those in biases.

The empirical SEE, which is computed as the standard deviation of the distribution of estimates across all replications, and the theoretical SEE which is the reciprocal of the square root of the information function, are two indices of the precision of the ability estimates. Figure 5.13-5.18 show that the empirical SEEs are not much different across conditions with the same number of item parameters (i.e., 1-PL, 2-PL and 3-PL respectively), except that they are slightly lower for conditions with higher LID. As hypothesized earlier, this is probably because IRT models do not account for LID among items within the testlet, and with less information available from the test, the empirical error variance tends to shrink towards the population variance. However, by using the testlet models, we expect that higher LID will lead to less information, which in turn will result in higher random errors of estimates. In addition, empirical SEEs from TRT calibration are slightly lower than those from IRT calibration for long testlets. These observations suggest the overestimation of precision in the cases of high LID or in the long testlet cases given other factors equal.

With respect to the mean theoretical SEE, Figure 5.19, 5.21 and 5.23 reveal that the mean theoretical SEEs from the IRT calibration are similar across conditions with the same number of item parameters, because IRT models do not account for LID among items within the testlet. However, Figure 5.20, 5.22 and 5.24 present that the mean theoretical SEEs from TRT calibration increase as LID goes up, and the mean theoretical SEEs for long testlets are higher than those for short testlets. This finding is in line with the hypothesis earlier. Namely, testlet models tend to provide more accurate parameter estimates than the independent item models; when the testlet effect variable is accounted for by the model, accuracy in proficiency estimates deteriorates when either the testlet length or LID increases. Mean theoretical SEE is only slightly higher for unbalanced tests than balanced tests in terms of testlet length, perhaps because the parameters are less easy to estimate when there are very few items in a testlet.

Table 5.1: Bias, Absolute Bias, RMSE, Empirical SEE and Mean SEE Averaged across Examinees

Condition	Bias		Absolute bias		RMSE		Empirical SEE		Mean SEE	
	IRT	TRT	IRT	TRT	IRT	TRT	IRT	TRT	IRT	TRT
S1	-0.0027	-0.0049	0.2264	0.2268	0.2815	0.2820	0.2682	0.2706	0.2866	0.3130
S2	0.0014	-0.0024	0.2514	0.2516	0.3082	0.3095	0.2616	0.2663	0.2845	0.3274
S3	0.0017	-0.0127	0.3109	0.3096	0.3670	0.3685	0.2523	0.2628	0.2846	0.3936
S4	-0.0006	-0.0027	0.2281	0.2285	0.2832	0.2834	0.2698	0.2687	0.2875	0.3191
S5	0.0051	0.0001	0.2821	0.2827	0.3412	0.3413	0.2642	0.2622	0.2851	0.3515
S6	0.0290	0.0164	0.3745	0.3725	0.4294	0.4257	0.2513	0.2445	0.2852	0.4628
S7	0.0042	0.0015	0.2610	0.2611	0.3234	0.3238	0.3027	0.3046	0.3245	0.3508
S8	0.0044	0.0025	0.2859	0.2867	0.3498	0.3502	0.2957	0.2928	0.3246	0.3665
S9	0.0097	0.0053	0.3468	0.3484	0.4108	0.4113	0.2852	0.2802	0.3227	0.4026
S10	0.0039	0.0029	0.2562	0.2558	0.3179	0.3172	0.2998	0.2974	0.3232	0.3549
S11	0.0002	-0.0016	0.2944	0.2937	0.3600	0.3586	0.2970	0.2930	0.3228	0.3817
S12	0.0213	0.0131	0.3843	0.3843	0.4478	0.4447	0.2864	0.2745	0.3227	0.4760
S13	-0.0013	-0.0041	0.2983	0.2990	0.3683	0.3692	0.3374	0.3392	0.3740	0.4018
S14	-0.0187	-0.0203	0.3156	0.3161	0.3879	0.3887	0.3339	0.3359	0.3747	0.4122
S15	0.0122	0.0110	0.3709	0.3704	0.4451	0.4451	0.3288	0.3321	0.3747	0.4594
S16	-0.0004	-0.0028	0.2948	0.2941	0.3652	0.3641	0.3392	0.3358	0.3741	0.4064
S17	0.0227	0.0221	0.3315	0.3297	0.4061	0.4032	0.3375	0.3315	0.3720	0.4252
S18	0.0148	0.0157	0.4190	0.4183	0.4914	0.4860	0.3265	0.3081	0.3762	0.5122
S19	0.0038	0.0005	0.2262	0.2267	0.2809	0.2819	0.2691	0.2717	0.2850	0.3116
S20	0.0015	-0.0033	0.2623	0.2616	0.3209	0.3210	0.2652	0.2694	0.2869	0.3293
S21	0.0204	0.0102	0.3220	0.3208	0.3785	0.3802	0.2537	0.2656	0.2849	0.3986
S22	0.0016	-0.0002	0.2270	0.2276	0.2817	0.2820	0.2671	0.2659	0.2868	0.3182

Table 5.1 (continued): Bias, Absolute Bias, RMSE, Empirical SEE and Mean SEE Averaged across Examinees

Condition	Bias		Absolute bias		RMSE		Empirical SEE		Mean SEE	
	IRT	TRT	IRT	TRT	IRT	TRT	IRT	TRT	IRT	TRT
S23	-0.0013	-0.0064	0.2804	0.2801	0.3392	0.3387	0.2683	0.2664	0.2879	0.3517
S24	0.0395	0.0260	0.3703	0.3639	0.4257	0.4182	0.2540	0.2482	0.2856	0.4665
S25	-0.0078	-0.0102	0.2581	0.2583	0.3203	0.3207	0.3023	0.3042	0.3232	0.3495
S26	0.0007	-0.0026	0.2842	0.2840	0.3493	0.3497	0.2999	0.3030	0.3243	0.3633
S27	0.0053	-0.0047	0.3364	0.3355	0.3990	0.4003	0.2828	0.2914	0.3225	0.4223
S28	0.0018	0.0007	0.2563	0.2566	0.3185	0.3185	0.3003	0.2980	0.3233	0.3553
S29	0.0043	0.0026	0.2959	0.2951	0.3612	0.3595	0.2963	0.2923	0.3237	0.3838
S30	0.0528	0.0416	0.3831	0.3780	0.4467	0.4385	0.2885	0.2765	0.3246	0.4778
S31	0.0093	0.0062	0.3027	0.3029	0.3735	0.3739	0.3385	0.3399	0.3773	0.4051
S32	0.0195	0.0168	0.3215	0.3215	0.3939	0.3943	0.3349	0.3369	0.3753	0.4140
S33	-0.0258	-0.0249	0.3767	0.3749	0.4497	0.4486	0.3278	0.3307	0.3743	0.4601
S34	-0.0003	-0.0022	0.2970	0.2970	0.3676	0.3672	0.3391	0.3355	0.3740	0.4065
S35	-0.0199	-0.0217	0.3384	0.3382	0.4128	0.4116	0.3357	0.3301	0.3746	0.4264
S36	0.0195	0.0162	0.4194	0.4163	0.4920	0.4842	0.3289	0.3097	0.3761	0.5163
S37	0.0052	0.0023	0.2261	0.2266	0.2802	0.2817	0.2699	0.2727	0.2874	0.3153
S38	0.0018	-0.0040	0.2607	0.2607	0.3185	0.3193	0.2652	0.2693	0.2868	0.3345
S39	0.0008	-0.0131	0.3245	0.3222	0.3807	0.3819	0.2513	0.2672	0.2847	0.4105
S40	0.0023	0.0002	0.2282	0.2286	0.2835	0.2837	0.2697	0.2680	0.2871	0.3219
S41	0.0062	0.0030	0.2894	0.2857	0.3485	0.3446	0.2652	0.2638	0.2869	0.3628
S42	0.0168	0.0044	0.3273	0.3230	0.3828	0.3820	0.2515	0.2666	0.2861	0.4806
S43	0.0067	0.0029	0.2586	0.2591	0.3204	0.3204	0.3009	0.3019	0.3216	0.3498

Table 5.1 (continued): Bias, Absolute Bias, RMSE, Empirical SEE and Mean SEE Averaged across Examinees

Condition	Bias		Absolute bias		RMSE		Empirical SEE		Mean SEE	
	IRT	TRT	IRT	TRT	IRT	TRT	IRT	TRT	IRT	TRT
S44	-0.0020	-0.0064	0.2958	0.2953	0.3613	0.3611	0.2987	0.3004	0.3244	0.3694
S45	0.0042	-0.0016	0.3508	0.3478	0.4148	0.4138	0.2870	0.2950	0.3227	0.4430
S46	0.0106	0.0097	0.2565	0.2573	0.3191	0.3193	0.3008	0.2985	0.3218	0.3568
S47	0.0016	-0.0017	0.3139	0.3117	0.3794	0.3761	0.2951	0.2901	0.3254	0.3971
S48	0.0271	0.0206	0.4138	0.3996	0.4767	0.4613	0.2864	0.2796	0.3222	0.4995
S49	0.0039	-0.0005	0.2975	0.2979	0.3688	0.3694	0.3408	0.3410	0.3744	0.4043
S50	0.0164	0.0128	0.3235	0.3230	0.3968	0.3963	0.3360	0.3362	0.3740	0.4185
S51	0.0022	0.0021	0.3809	0.3781	0.4541	0.4518	0.3273	0.3292	0.3740	0.4771
S52	0.0067	0.0048	0.2979	0.2982	0.3692	0.3690	0.3405	0.3360	0.3743	0.4098
S53	-0.0002	-0.0010	0.3423	0.3404	0.4175	0.4138	0.3393	0.3318	0.3748	0.4374
S54	-0.0285	-0.0246	0.4370	0.4187	0.5111	0.4885	0.3331	0.3169	0.3761	0.5296

Figure 5.1: Mean Absolute Bias from 1-PL IRT

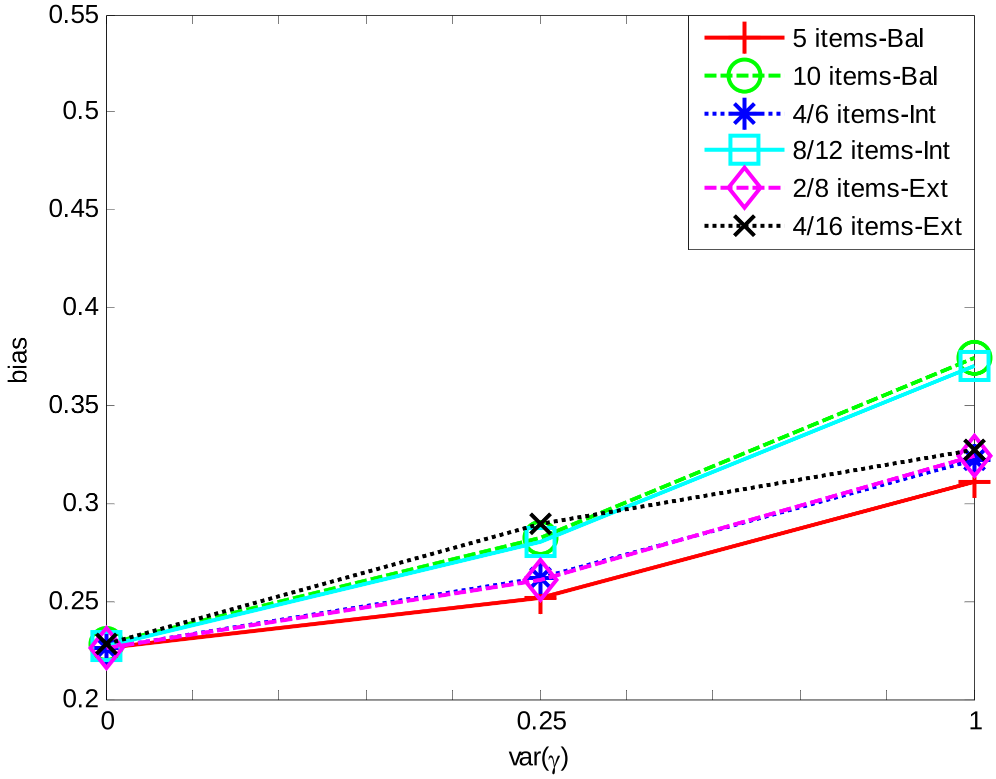


Figure 5.2: Mean Absolute Bias from 1-PL TRT

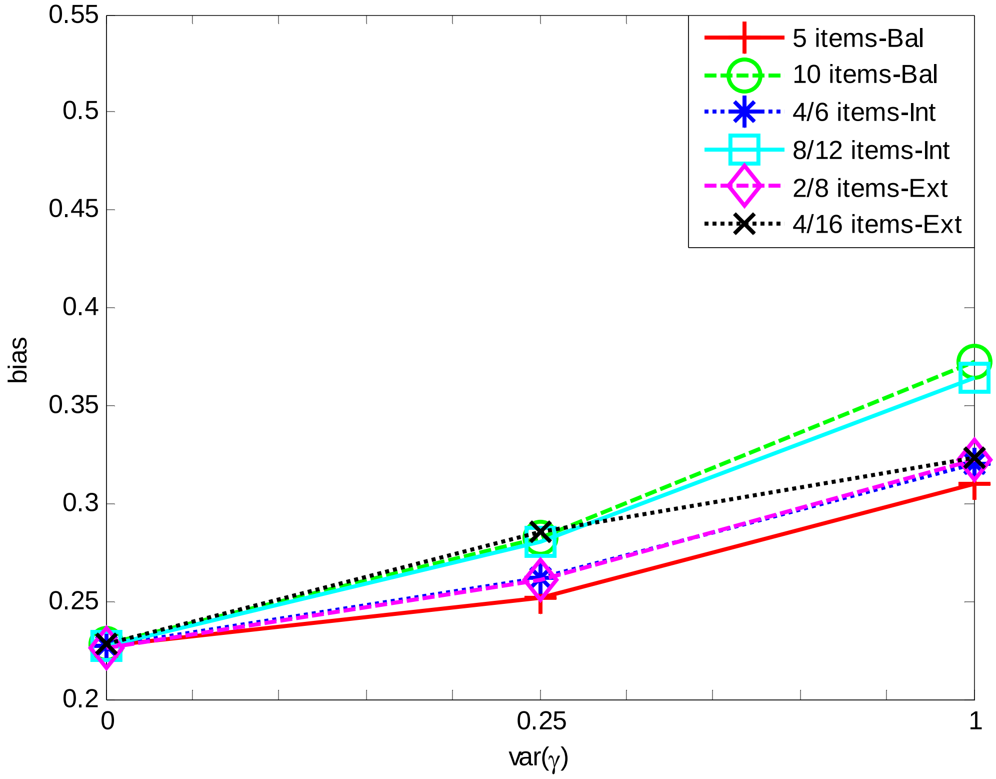


Figure 5.3: Mean Absolute Bias from 2-PL IRT

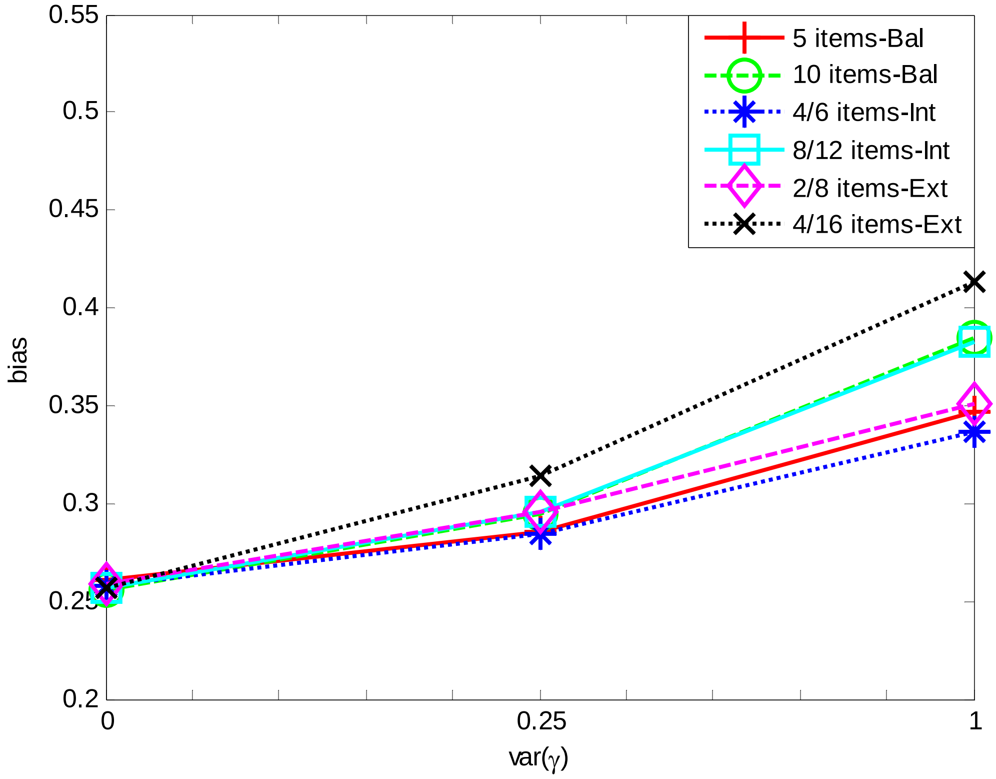


Figure 5.4: Mean Absolute Bias from 2-PL TRT

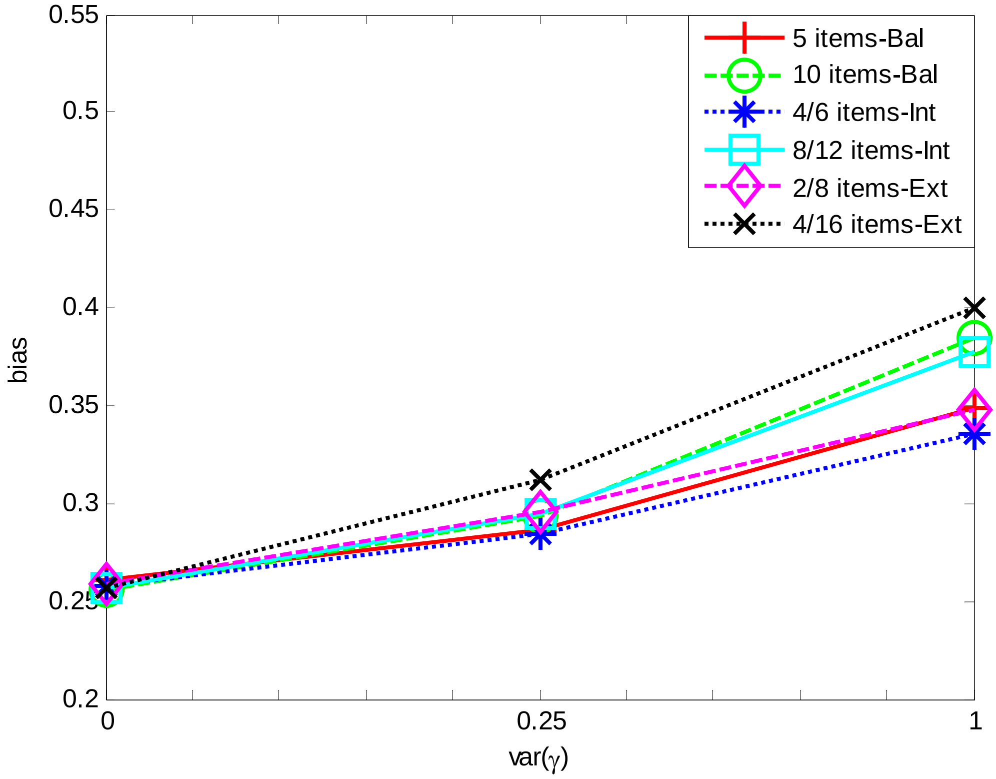


Figure 5.5: Mean Absolute Bias from 3-PL IRT

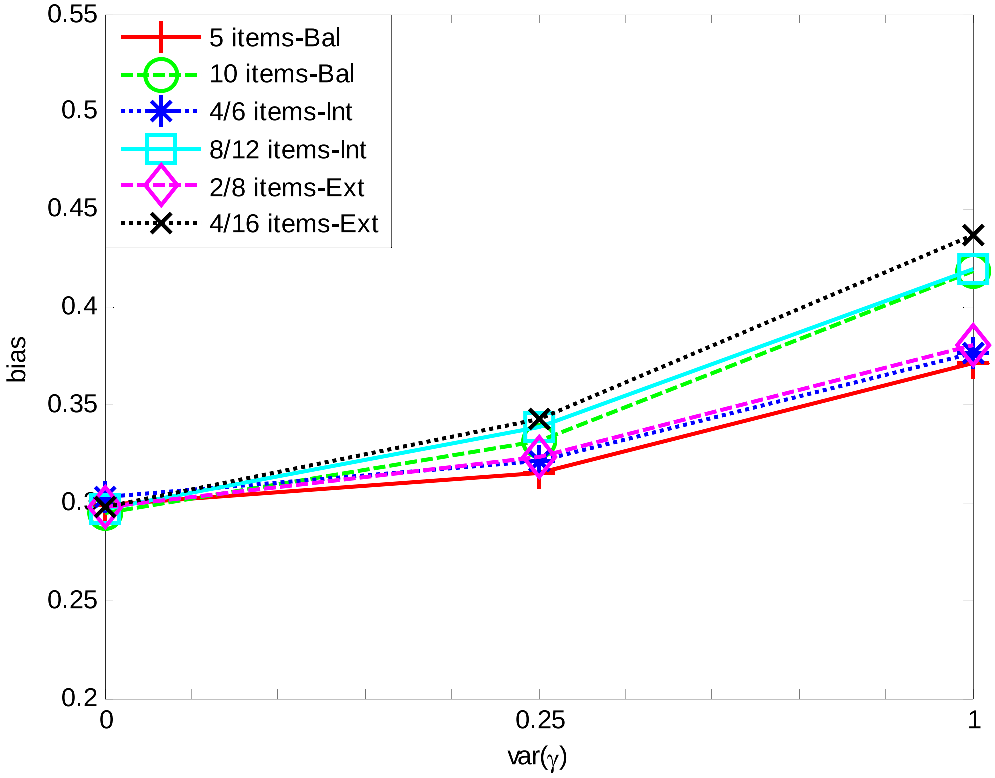


Figure 5.6: Mean Absolute Bias from 3-PL TRT

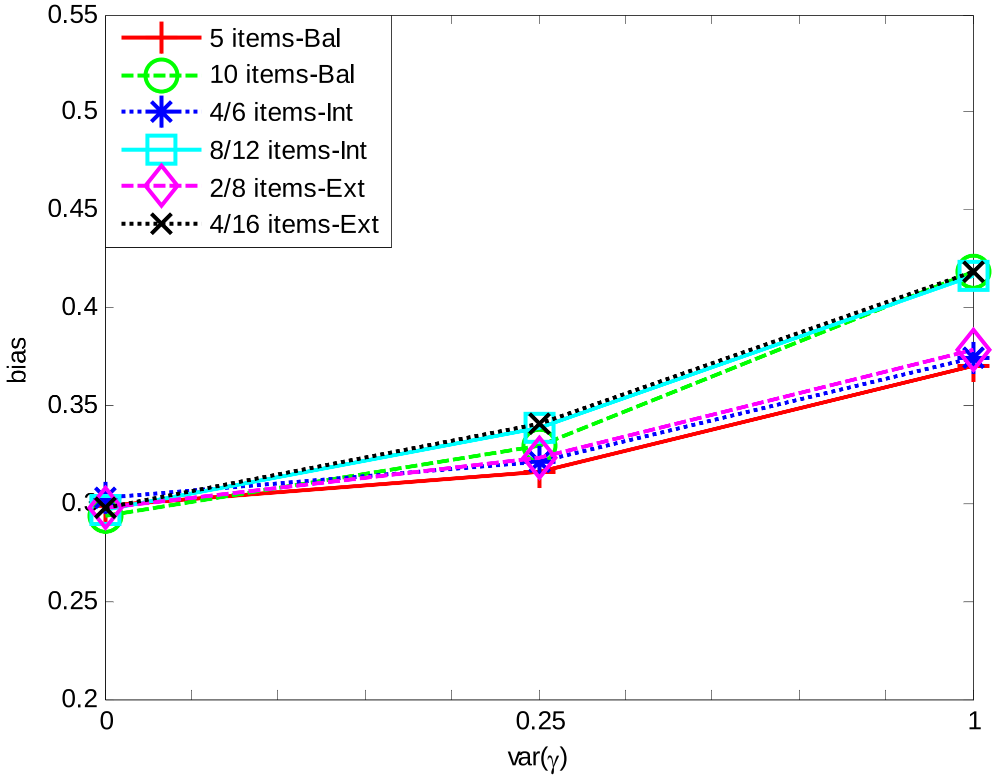


Figure 5.7: Mean RMSE from 1-PL IRT

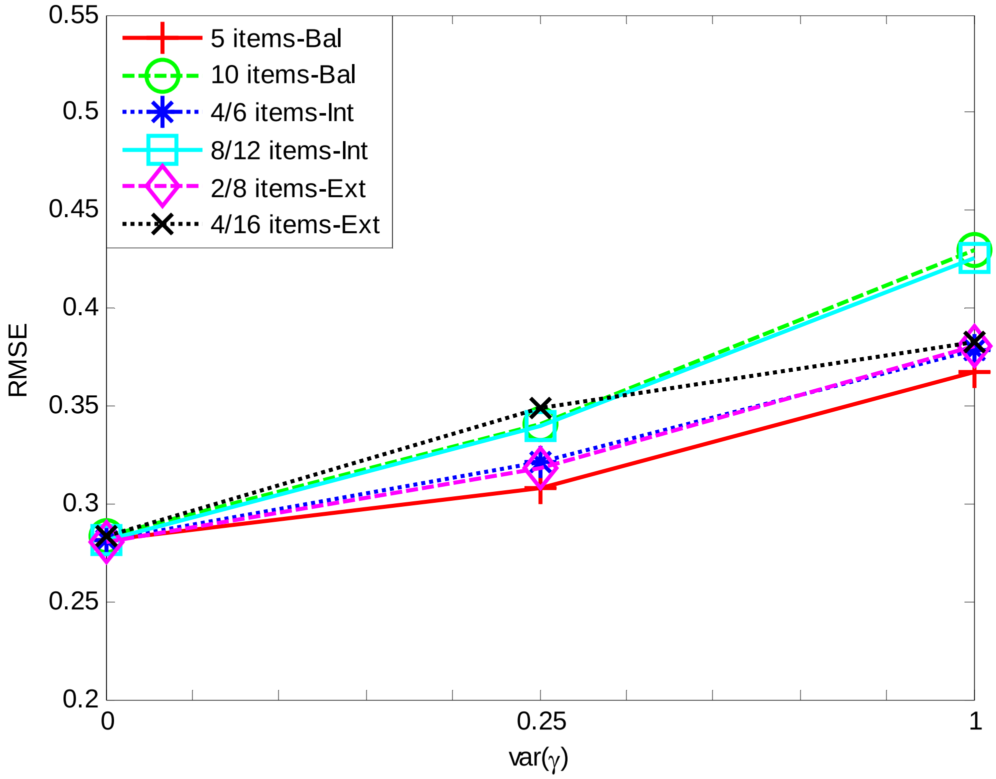


Figure 5.8: Mean RMSE from 1-PL TRT

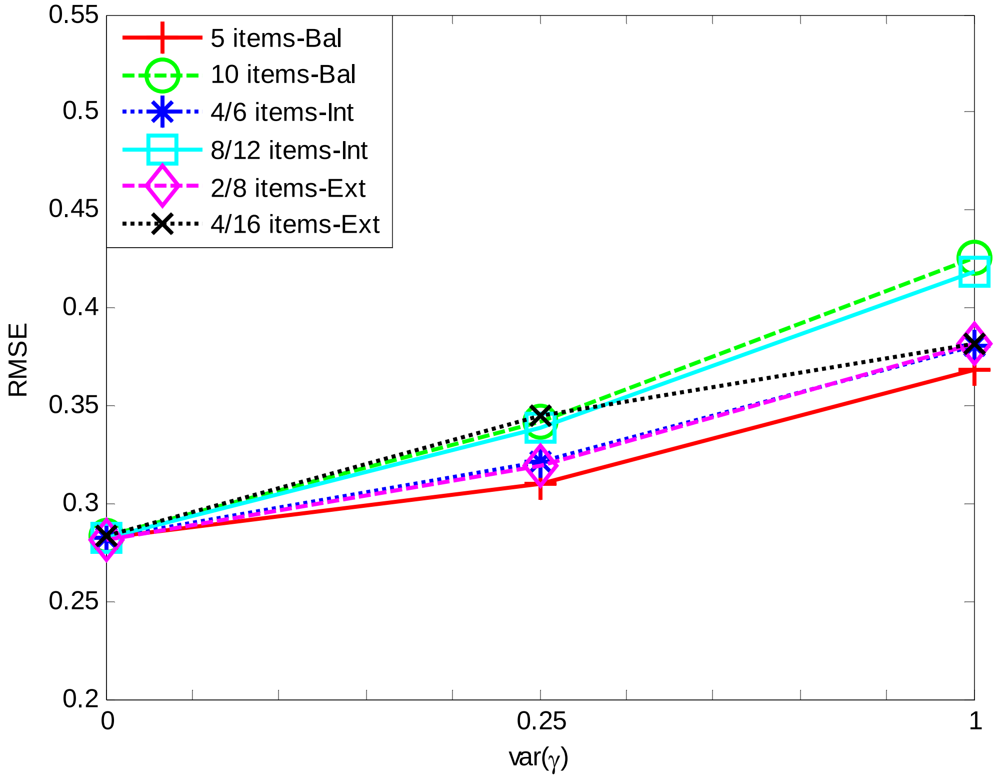


Figure 5.9: Mean RMSE from 2-PL IRT

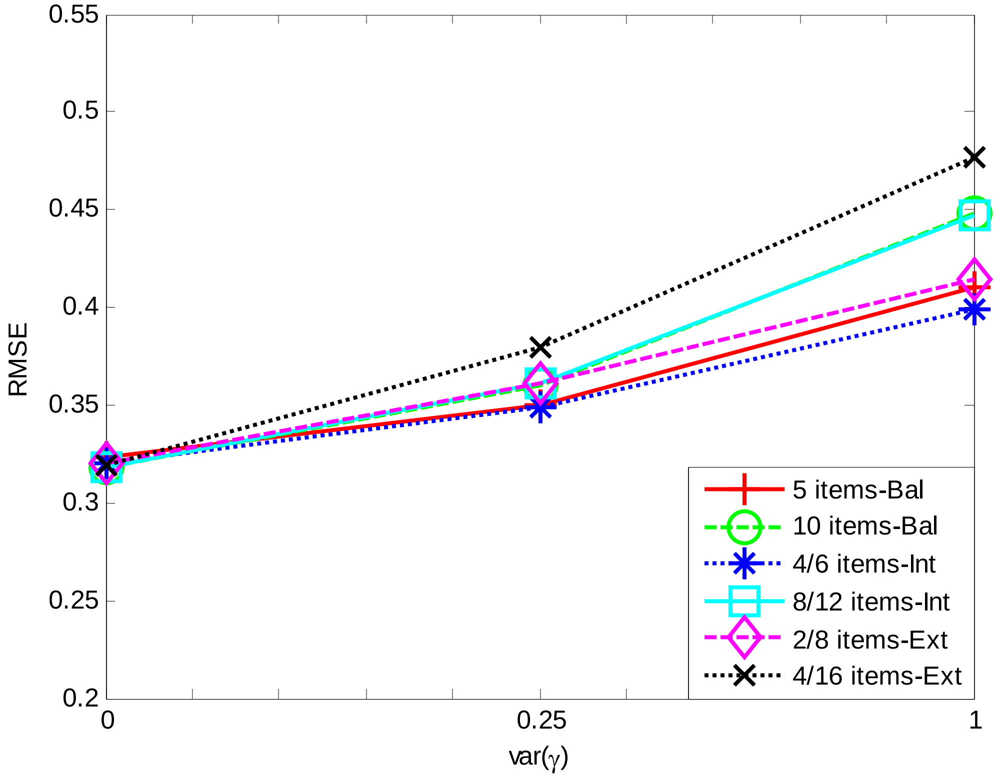


Figure 5.10: Mean RMSE from 2-PL TRT

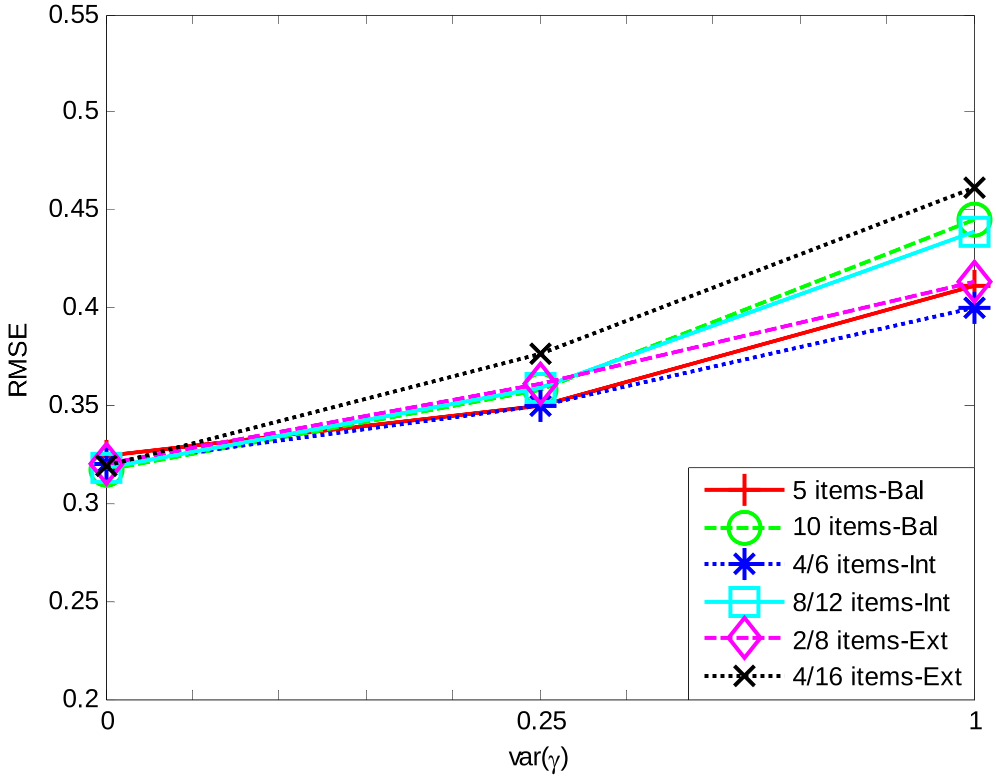


Figure 5.11: Mean RMSE from 3-PL IRT

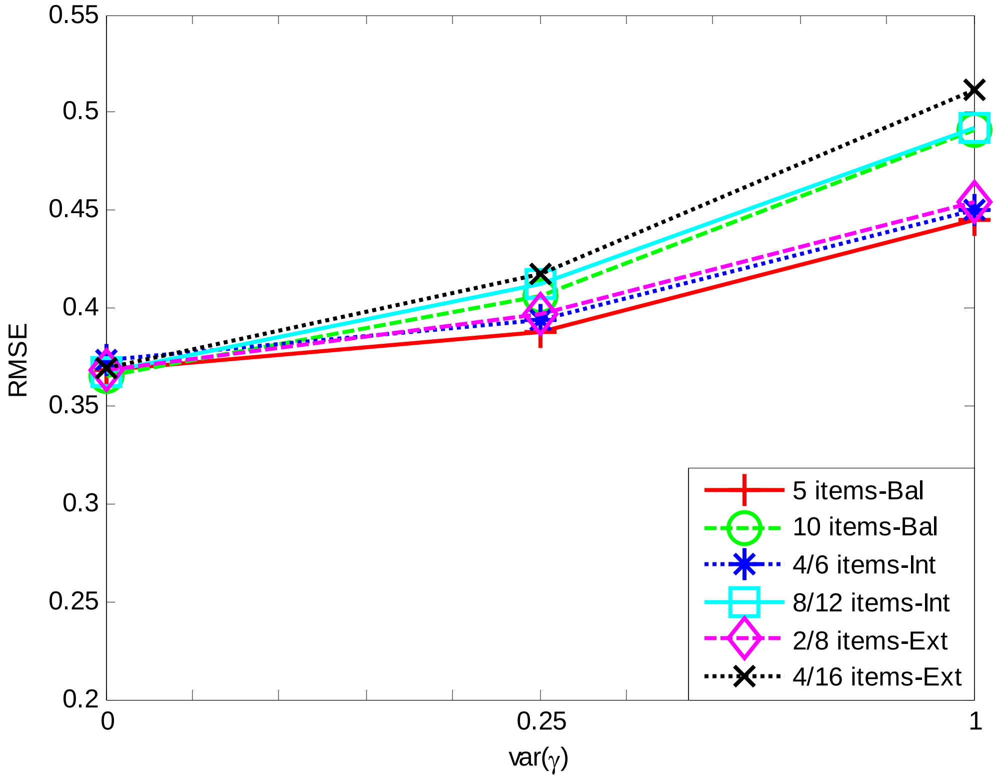


Figure 5.12: Mean RMSE from 3-PL TRT

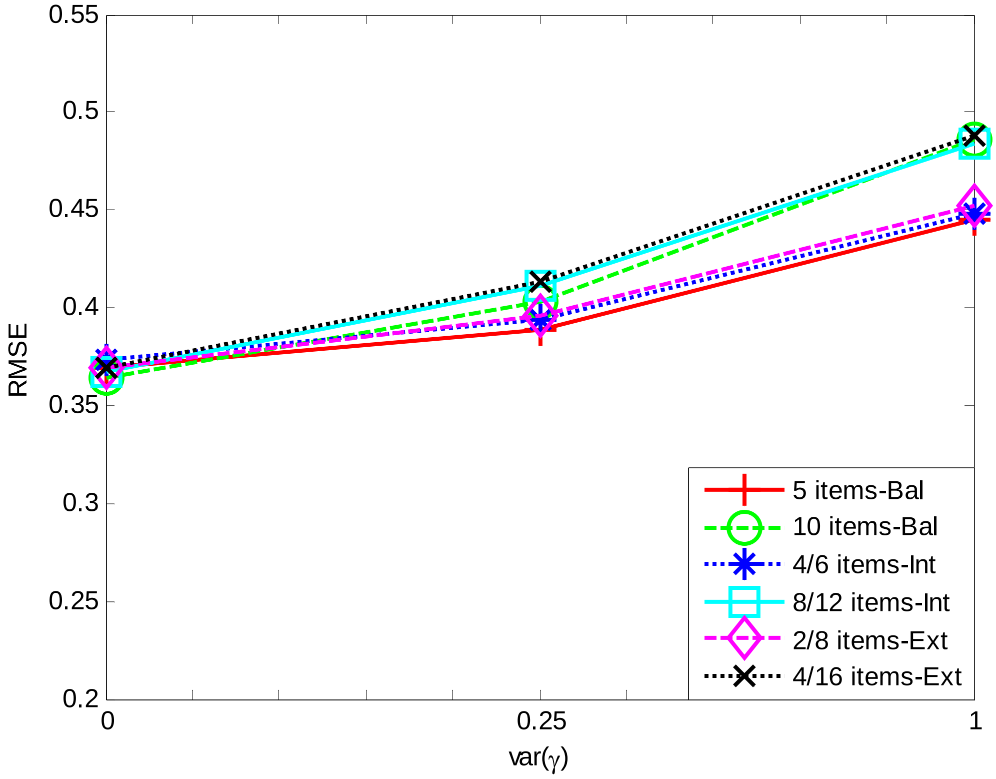


Figure 5.13: Mean Empirical SEE from 1-PL IRT

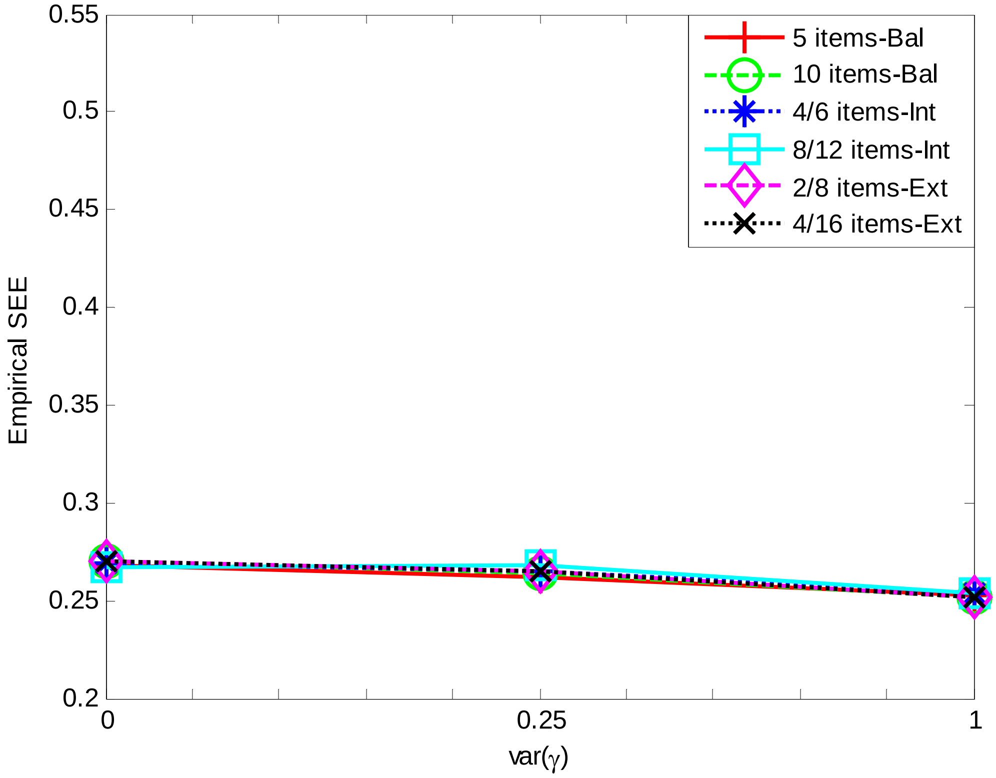


Figure 5.14: Mean Empirical SEE from 1-PL TRT

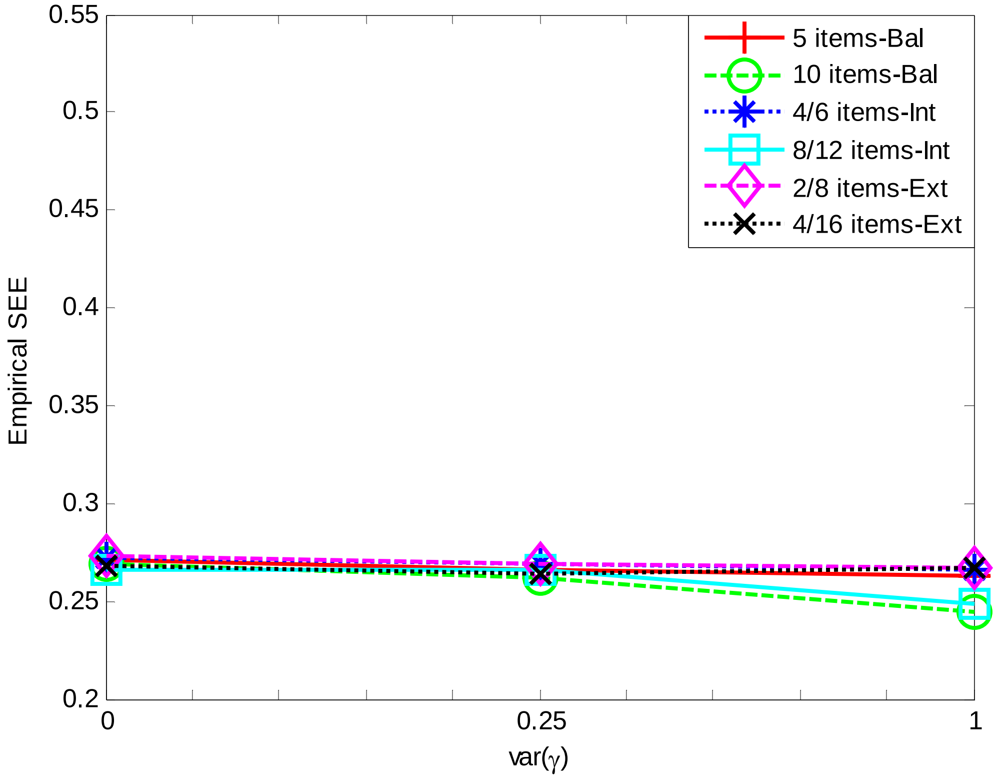


Figure 5.15: Mean Empirical SEE from 2-PL IRT

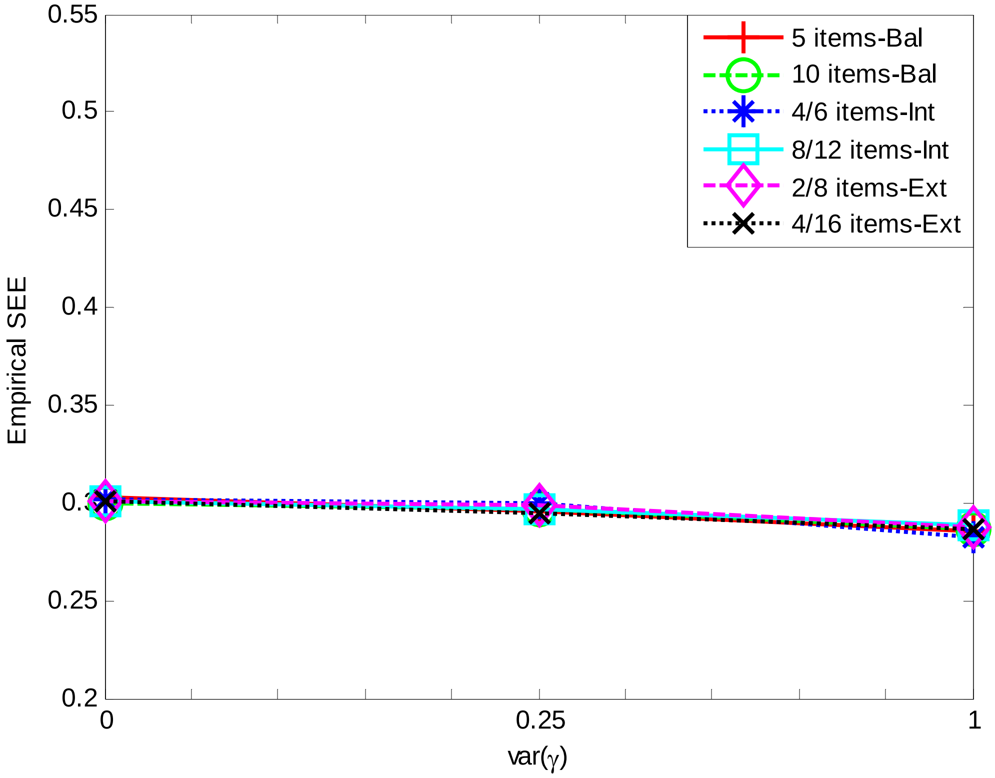


Figure 5.16: Mean Empirical SEE from 2-PL TRT

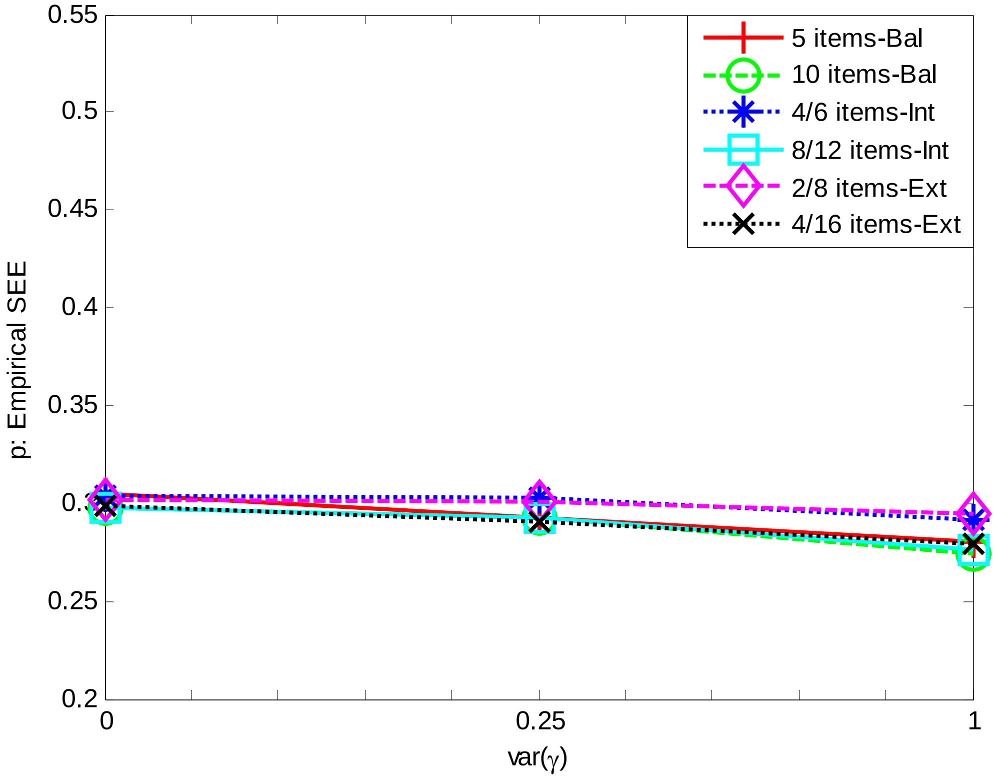


Figure 5.17: Mean Empirical SEE from 3-PL IRT

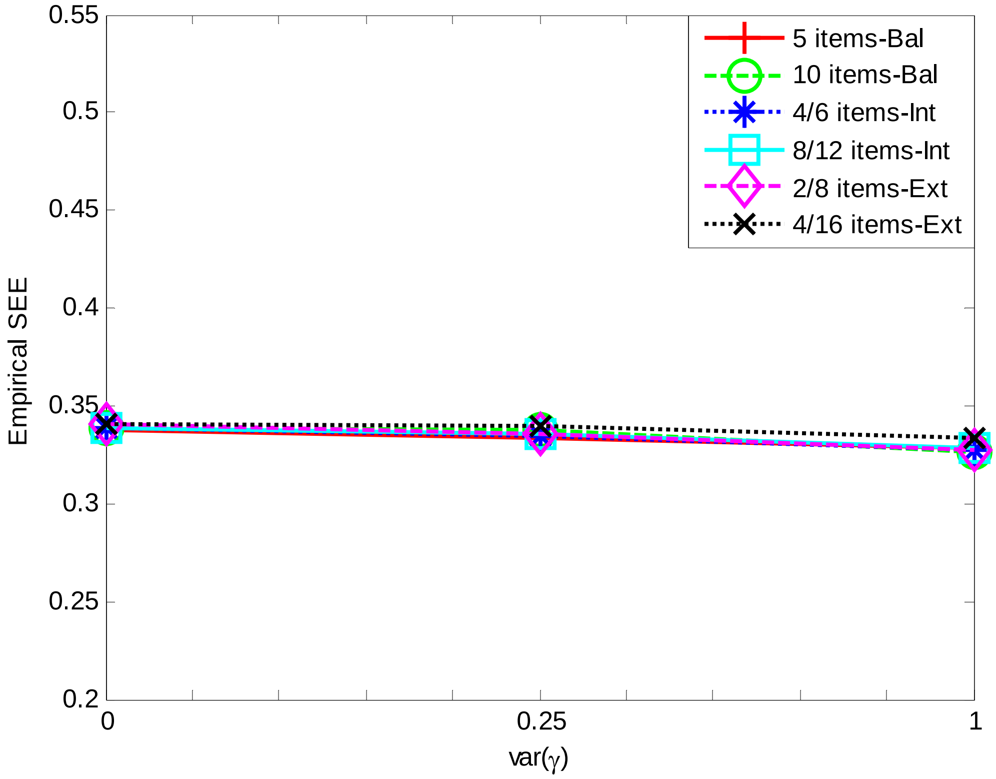


Figure 5.18: Mean Empirical SEE from 3-PL TRT

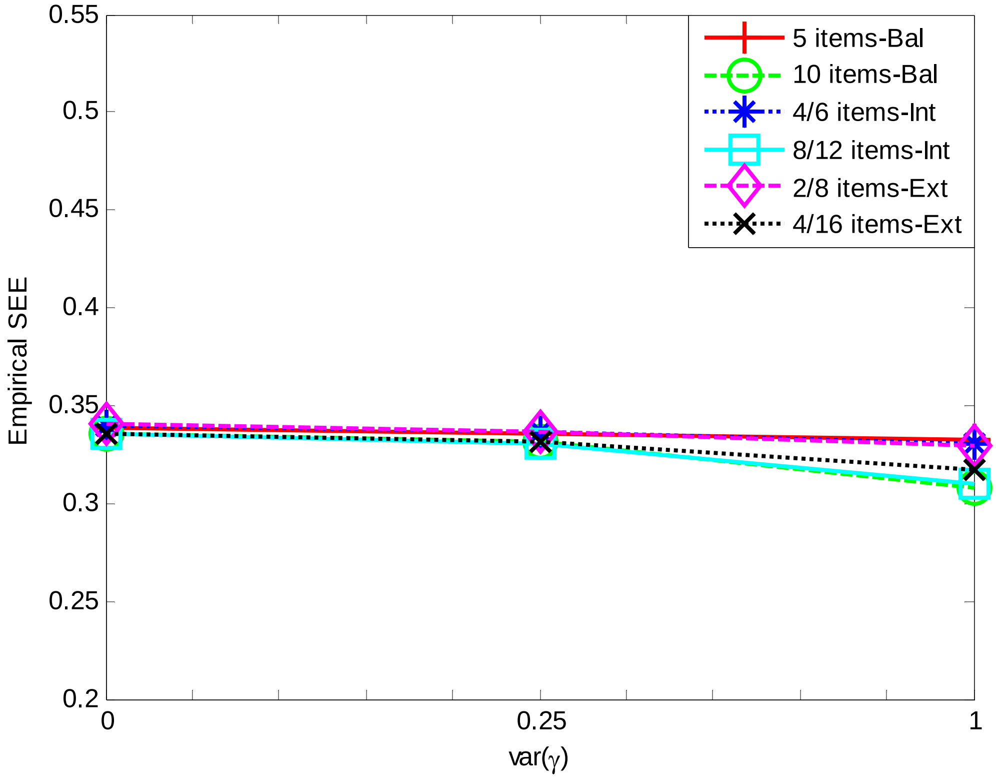


Figure 5.19: Mean Theoretical SEE from 1-PL IRT

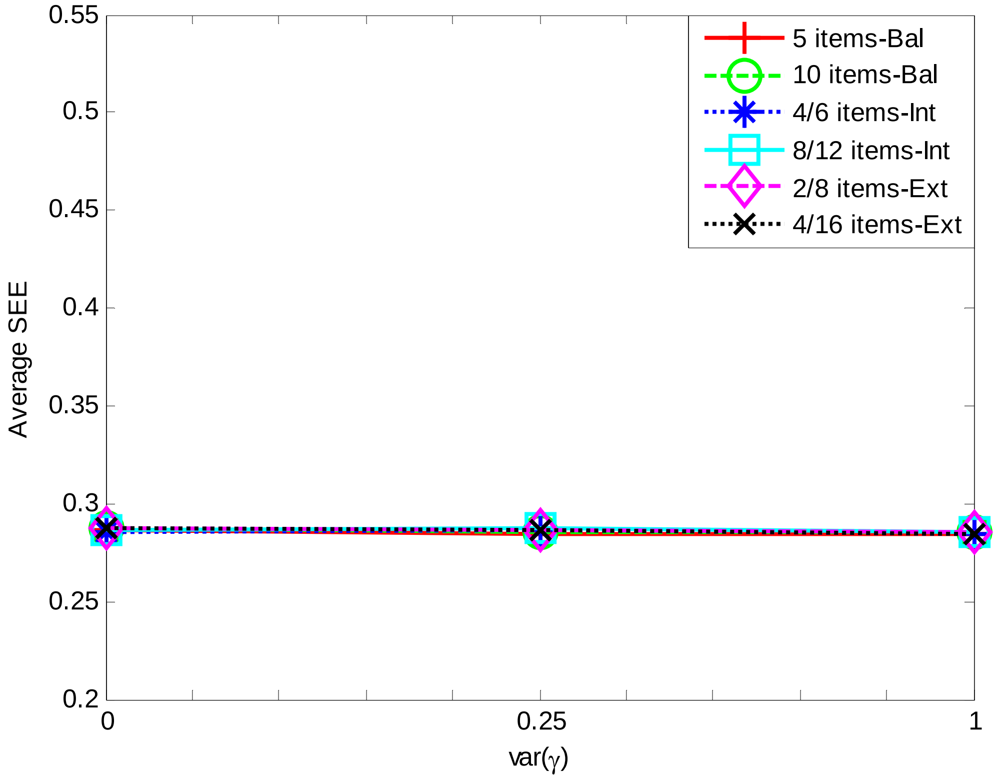


Figure 5.20: Mean Theoretical SEE from 1-PL TRT

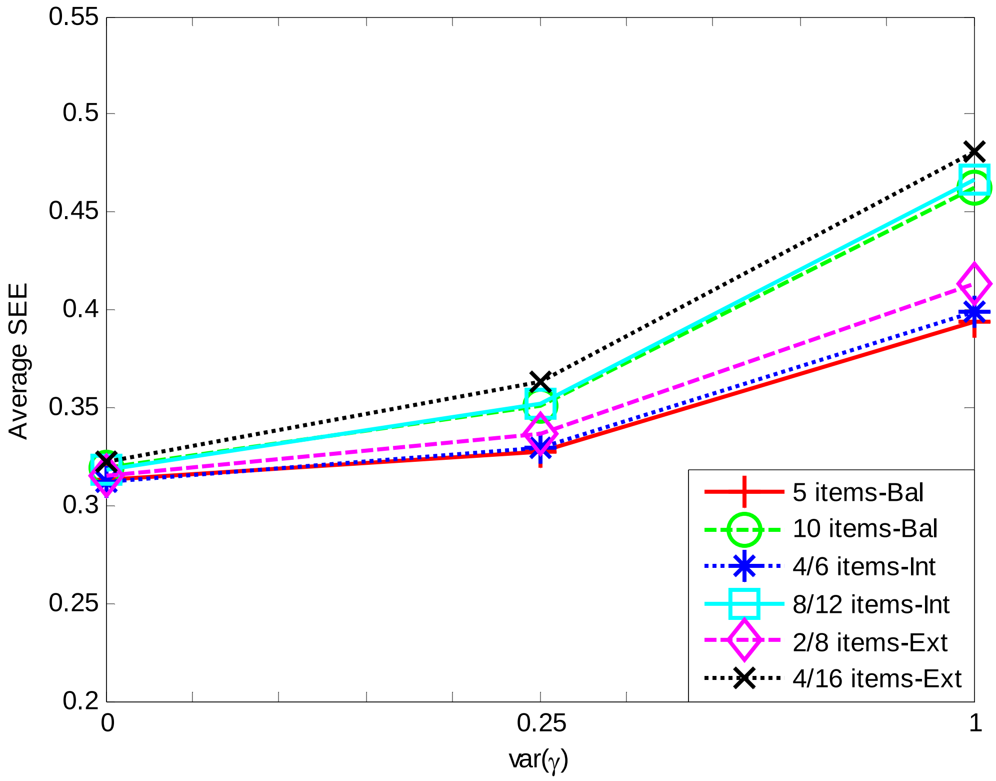


Figure 5.21: Mean Theoretical SEE from 2-PL IRT

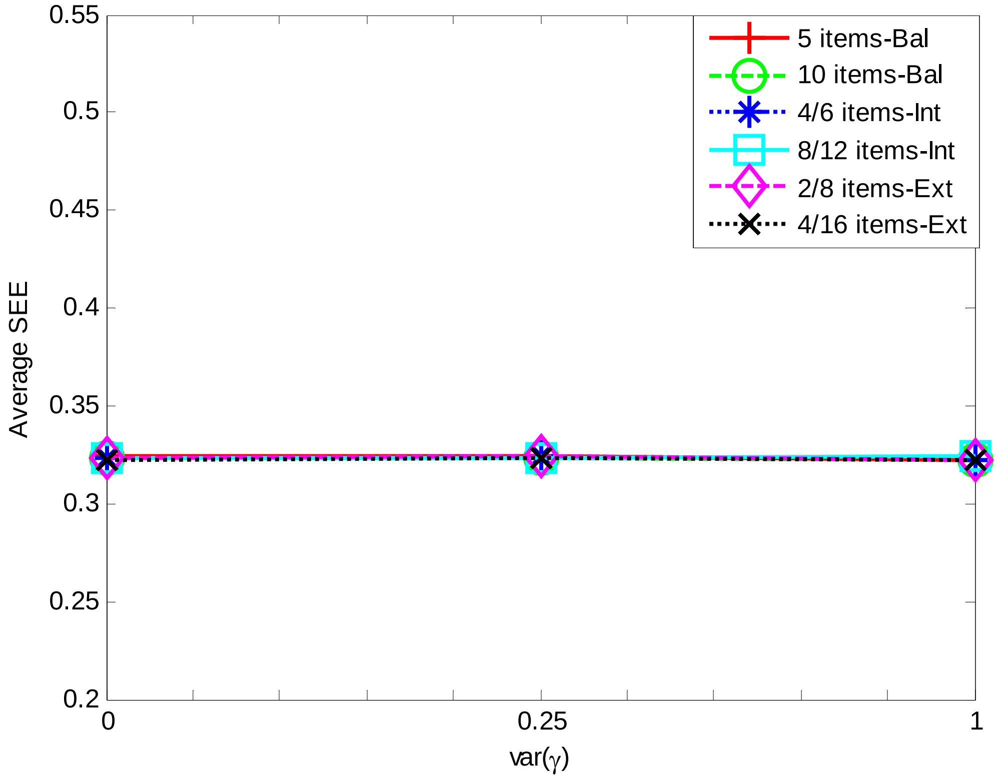


Figure 5.22: Mean Theoretical SEE from 2-PL TRT

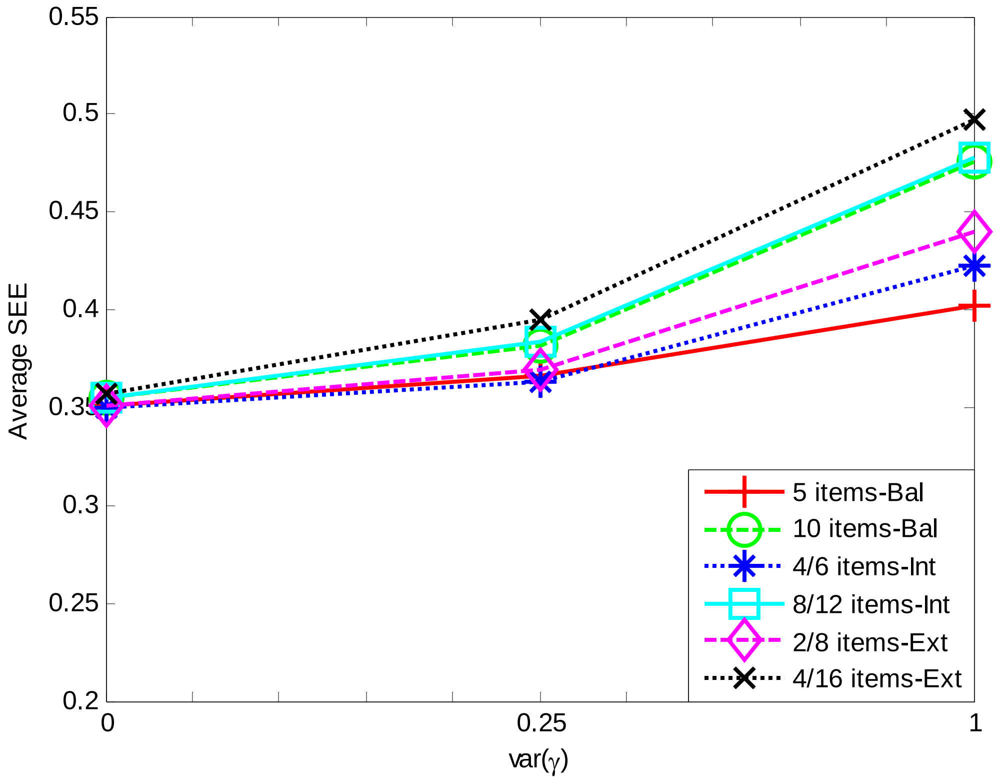


Figure 5.23: Mean Theoretical SEE from 3-PL IRT

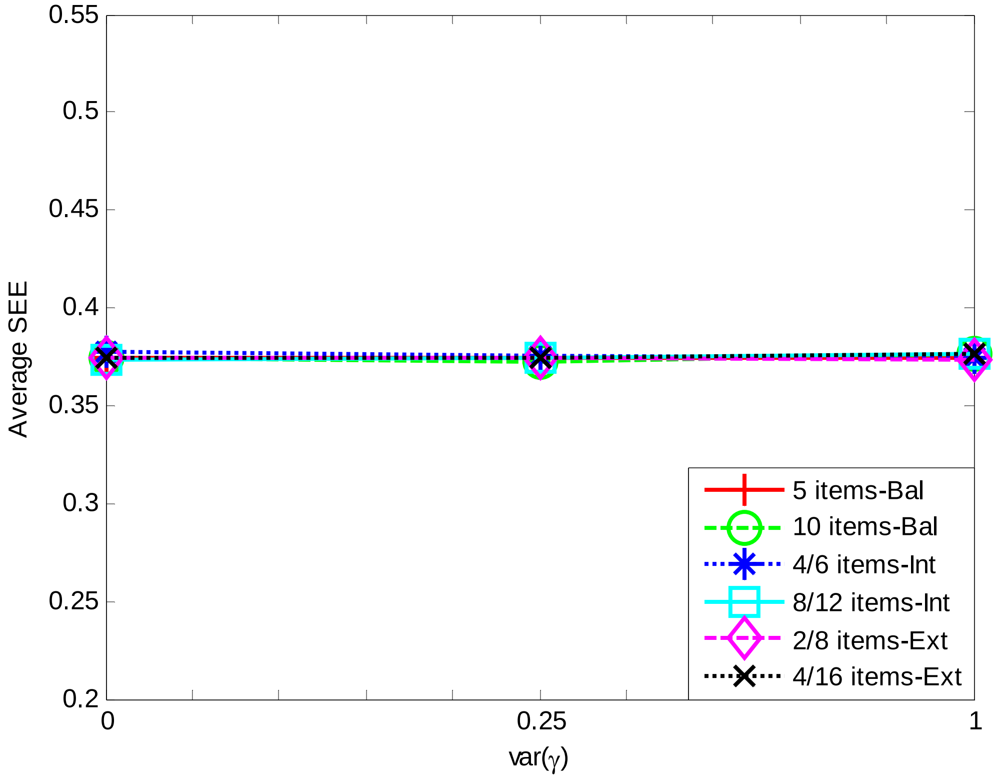
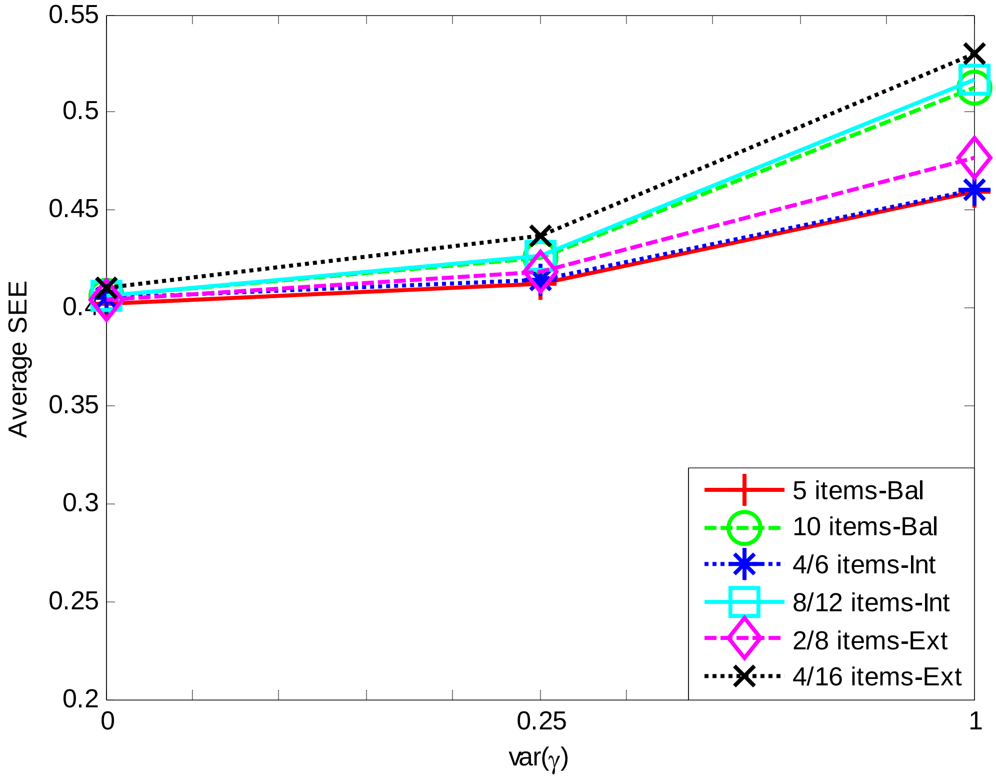


Figure 5.24: Mean Theoretical SEE from 3-PL TRT



5.1.2. Information Correction

In this section, I compare the conditional SEE from IRT models (SEE_{IRT}) and the conditional SEE adjusted by the information correction terms (SEE_{IRT-t_d}) with the conditional SEE from TRT models (SEE_{TRT}). Figure 5.25 illustrates the results of the comparison. Given the same values on other factors, SEE_{TRT} increases as LID goes up, so does SEE_{IRT-t_d}. The discrepancy between SEE_{IRT} and SEE_{TRT} becomes larger when LID increases or the testlet length decreases. However, it seems that across all conditions SEE_{IRT} can always be adjusted to the value that is close to SEE_{TRT} by using the information correction ratio, which suggests that the information correction method is effective for this purpose.

When LID is zero, SEE_{IRT-t_d} seems to be overlapping with SEE_{IRT}, and SEE_{TRT} is higher than SEE_{IRT-t_d} across the ability scale. When LID is small, SEE_{TRT} is still higher than SEE_{IRT-t_d} conditional on θ values between -2 and 2, but their discrepancy is smaller compared with LID at zero. SEE_{IRT-t_d} conditional on extreme values on θ scale tends to be higher than SEE_{TRT}, which indicates overcorrection of random variances. When LID is large, the level of the discrepancy between SEE_{TRT} and SEE_{IRT-t_d} is similar with the discrepancy when LID is small, but the overcorrection on extreme θ values seems to be more salient, which suggests that the information correction presents satisfactory performance for the conditions with substantial testlet effects but overcorrection may occur to SEE conditional on extreme ability values.

Comparing condition 19-36 where the testlet lengths are intermediately unbalanced, condition 37-54 where the testlet length are extremely unbalanced with condition 1-18 where testlet lengths are equal within the test, the unbalanced conditions seem to result in better correction performances than the balanced conditions. In unbalanced conditions, SEE_{IRT-t_d} almost overlaps with SEE_{TRT} conditional on values in the middle part of the ability scale, but SEE_{TRT} is higher than SEE_{IRT-t_d} for that part of ability values in balanced conditions given equal values on other factors.

Comparing condition 1-6, 19-24 and 37-42 from the 1-PL context with condition 7-12, 25-30 and 43-48 that are from the 2-PL context, the discrepancies between SEE_{IRT-t_d} and SEE_{TRT} conditional on ability values in the middle seem to be smaller for 2-PL context than those for the 1-PL context, which implies better performance of the information correction method in the 2-PL context with this part of examinees, but the level of overcorrection is higher for the examinees with extreme ability values in the 2-PL context than the 1-PL context. In condition 13-18, 31-36 and 49-54 where the simulated datasets are generated and calibrated with 3-PL models, SEE_{IRT-t_d} almost overlaps with SEE_{TRT} in the middle part of the ability scale, but overcorrection on extreme ability values seems to be more serious. These observations from the conditional standard error plots suggest that the information correction method has a better performance for 3-PL models than for 2-PL models, which in turn has better performance for 2-PL models than for 1-PL models.

Figure 5.26-5.28 compare the means of the dependent variable in each condition—standard error increase discrepancy (SEID) (see Equation 4.1) , the discrepancy between the percent of increase in SEE as a result of adjustment and the percent of increase in SEE as a result of TRT modeling based on the SEE from the misspecified IRT models. A SEID value close to zero indicates sufficient adjustment in error variance and hence a good performance of the information correction method. The mean SEID (i.e., the SEID statistics averaged across all examinees) represents the effect of information correction in general for a particular condition. However, undercorrection for some SEEs and overcorrection for other SEEs may cancel out each other and result in a low mean SEID value as if all SEEs are appropriately corrected.

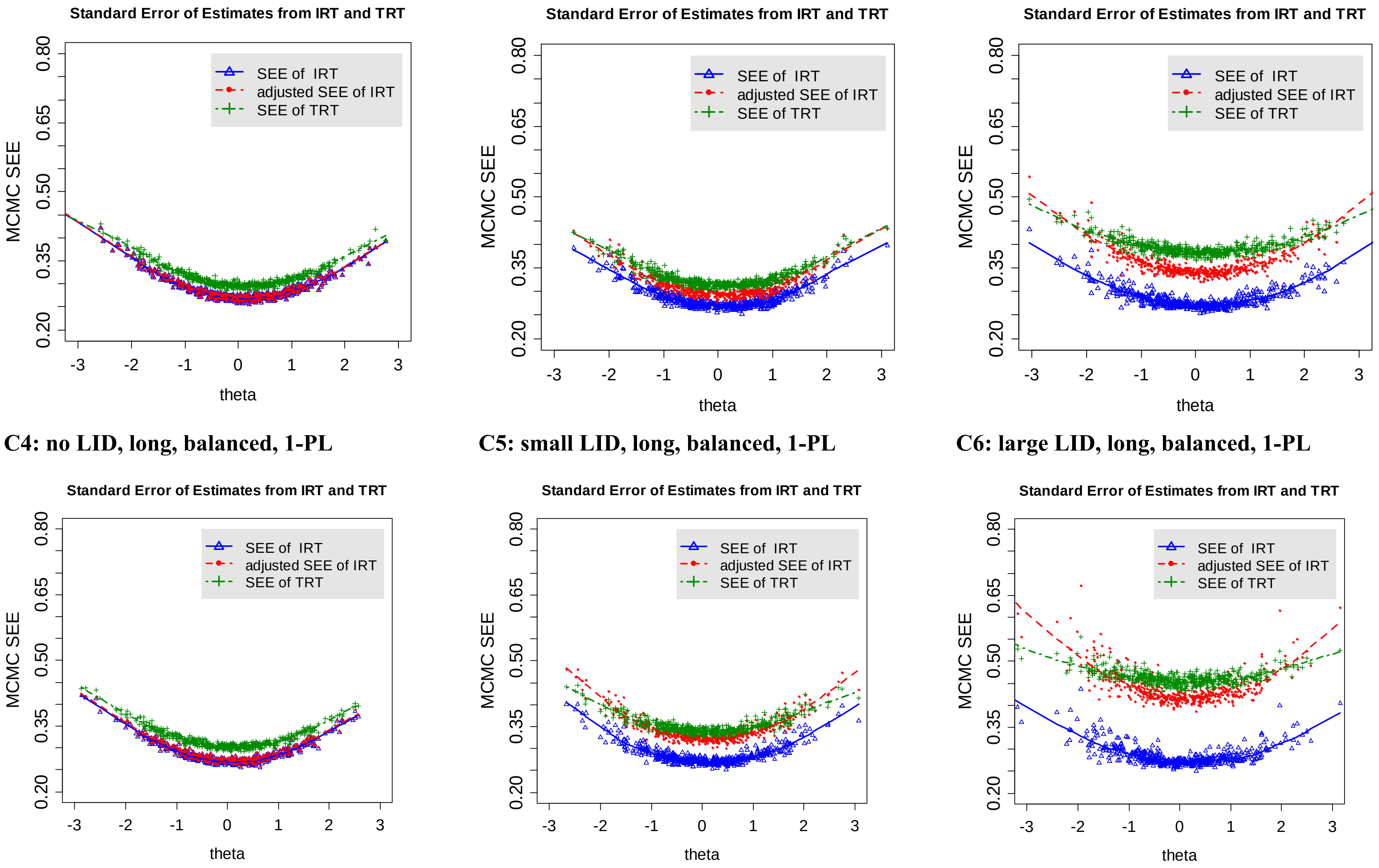
Figure 5.26 shows that in the 1-PL context, the mean SEID appears to be low and close to zero when LID is moderate ($\sigma_\gamma^2 = 0.25$), but comparatively high when LID is zero ($\sigma_\gamma^2 = 0$) or large ($\sigma_\gamma^2 = 1$). This implies that on the whole the information correction method might perform best in the conditions with moderate LID. Figure 5.27 also presents best correction performance with moderate LID conditions in the context of 2-PL models where the mean SEID is closer to zero for large LID than moderate LID given other factors equal. Similar patterns in regard to LID are shown in the context of 3-PL models (Figure 5.28). Based on the conditional SEE presented in Figure 5.25, the information correction method does not necessarily show better adjustment for moderate LID conditions than their large LID counterparts. Mean SEID values that are more deviated from zero on conditions with large LID may be attributed to overcorrection of SEE_{IRT} on extreme θ values.

With only a few exceptions, given equal values on other factors, the mean SEID statistics are closer to zero for conditions with extremely unbalanced design than those in the conditions with intermediately unbalanced design, which in turn are closer to zero than those in the conditions with balanced design. This result seems to imply that the adjustment effect improves as the degree of the unbalance of the testlet length increases, which is consistent with what has been observed from the conditional SEE plots. With a few exceptions in 2-PL and 3-PL contexts, the long-testlet test tends to result in mean SEID closer to zero than the short-testlet test given values on other factors fixed, which seems to suggest a better overall adjustment effect for long-testlet tests. The conditional SEE plots (Figure 5.25) show the same pattern of discrepancies in regard to the length of testlets.

The ANOVA results in Table 5.2-5.4 indicate that the three factors being manipulated (e.g., LID, testlet length and the balance of the testlet length) and their interactions in this study account for more than 99% of the total variance in the dependent variable. Based on the p values of the F tests, all of the three factors and their interactions are statistically significant, indicating significant effects on the adjustment of random errors using the information correction method.

In the context of 1-PL models (Table 5.2), balance of the testlet length explains 32.7% of the total variance; LID accounts for 20.2% of the total variance; and the interaction between these two terms explains 34.2% of the total variance. In the context of 2-PL models (Table 5.3), level of balance in testlet length accounts 35.2% of the variance in the dependent variable; LID accounts for 9.1% of the total variance and the interaction between these two 42.8% of the total variance. In contrast, in the context of 3-PL models (Table 5.4), the balance of the testlet length explains 4.8% of the total variance, the testlet length accounts for 3.8% of the variance, but the interaction between these two accounts for 52.3% of the total variance, and the interaction between the balance and the testlet length explains 28.5% of the variance.

Figure 5.25: SEE from IRT before and after adjustment, and SEE from TRT in 1-PL, 2-PL, 3-PL contexts respectively



C1: no LID, short, balanced, 1-PL

C2: small LID, short, balanced, 1-PL

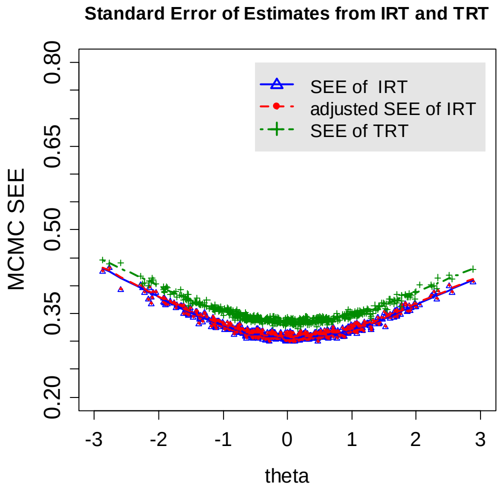
C3:large LID, short, balanced, 1-PL

C4: no LID, long, balanced, 1-PL

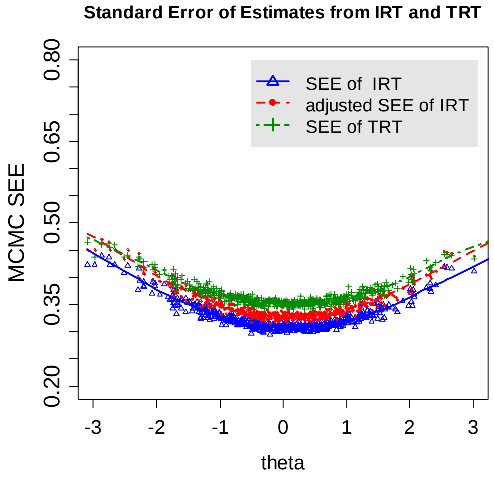
C5: small LID, long, balanced, 1-PL

C6: large LID, long, balanced, 1-PL

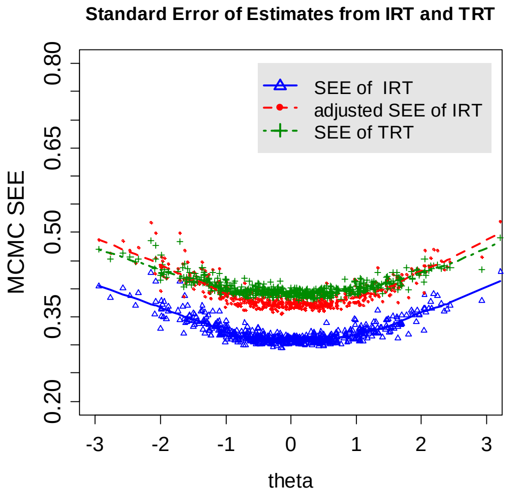
C7: no LID, short, balanced, 2-PL



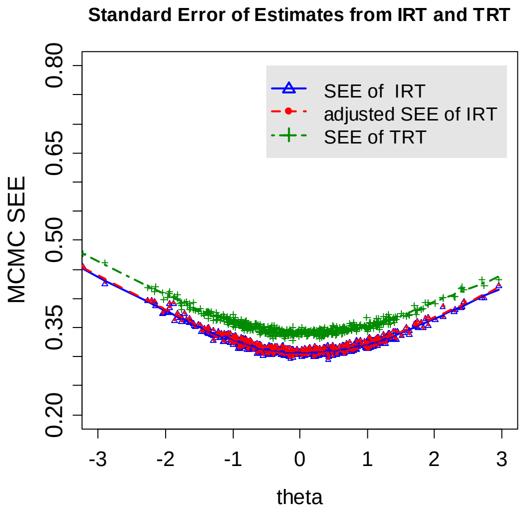
C8: small LID, short, balanced, 2-PL



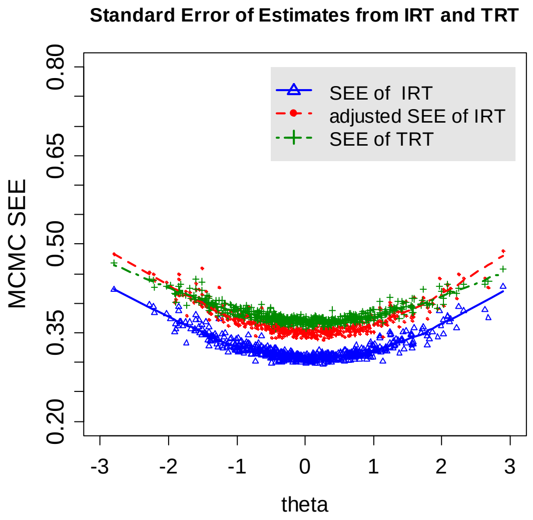
C9: large LID, short, balanced, 2-PL



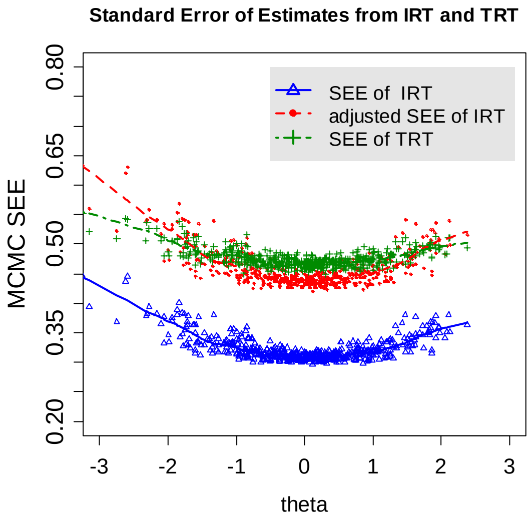
C10: no LID, long, balanced, 2-PL



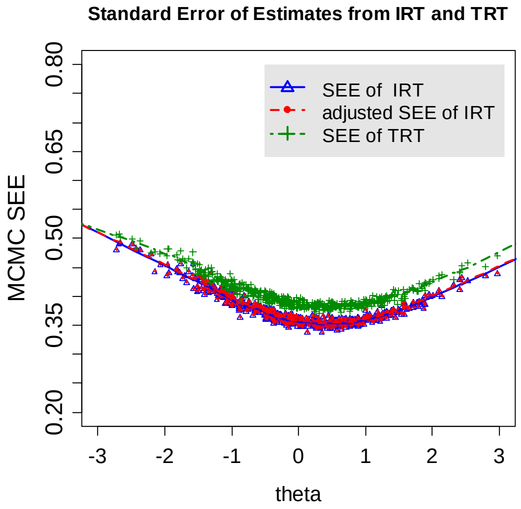
C11: small LID, long, balanced, 2-PL



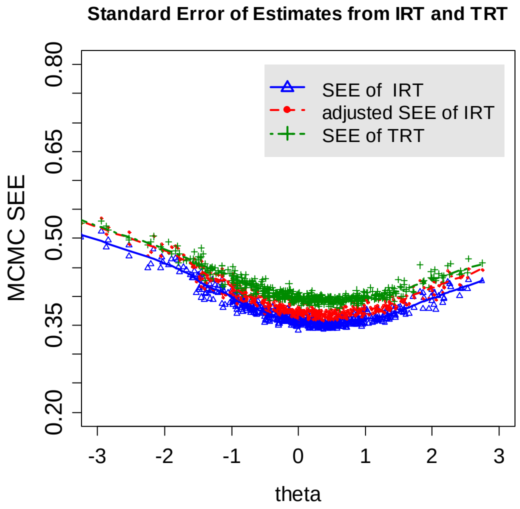
C12: large LID, long, balanced, 2-PL



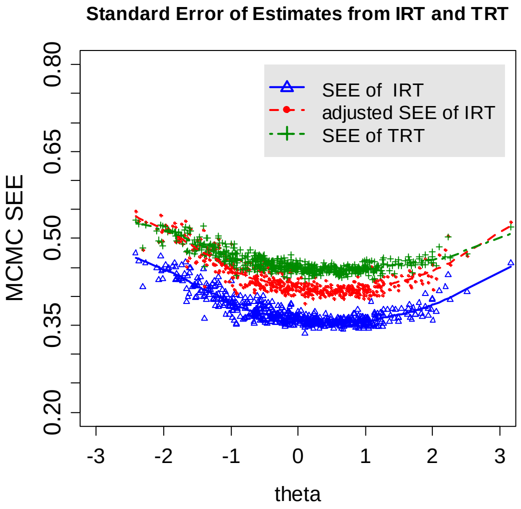
C13: no LID, short, balanced, 3-PL



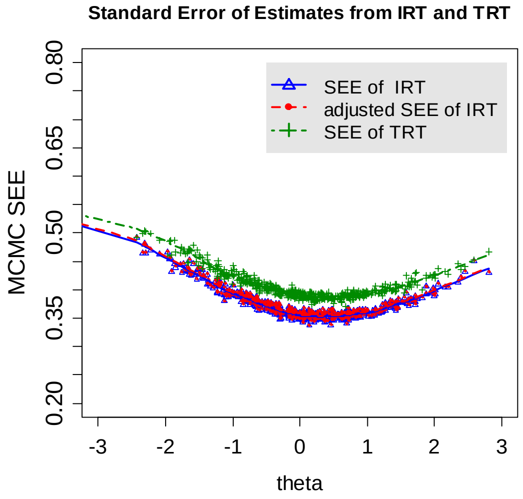
C14: small LID, short, balanced, 3-PL



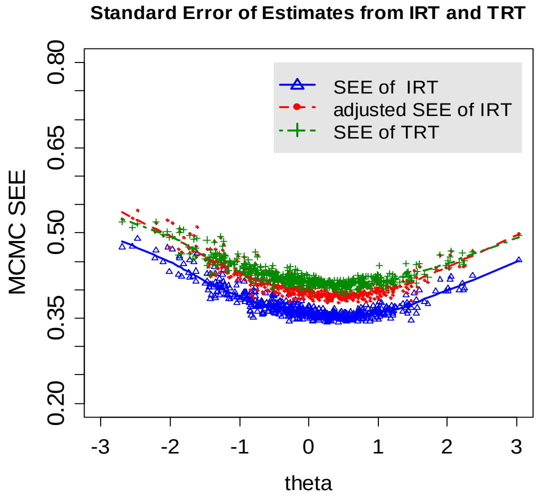
C15: small LID, short, balanced, 3-PL



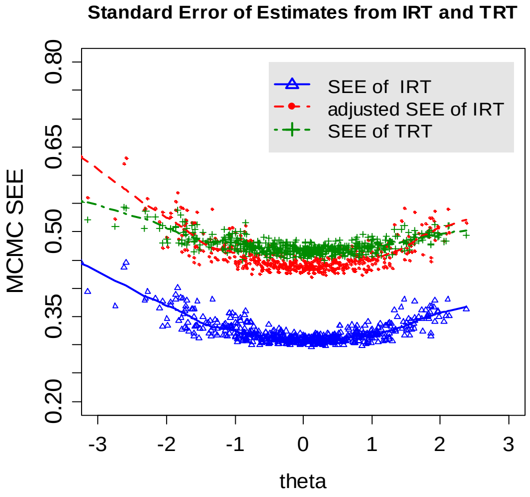
C16: no LID, long, balanced, 3-PL



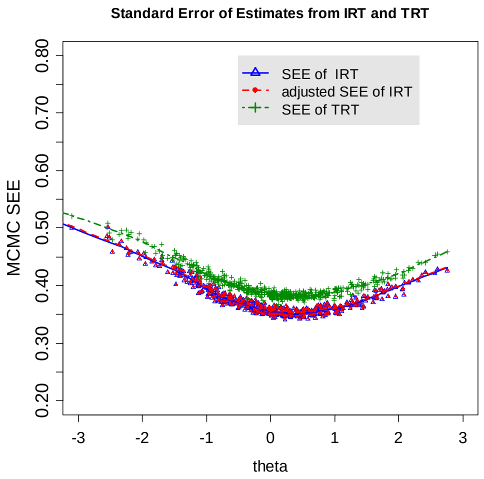
C17: small LID, long, balanced, 3-PL



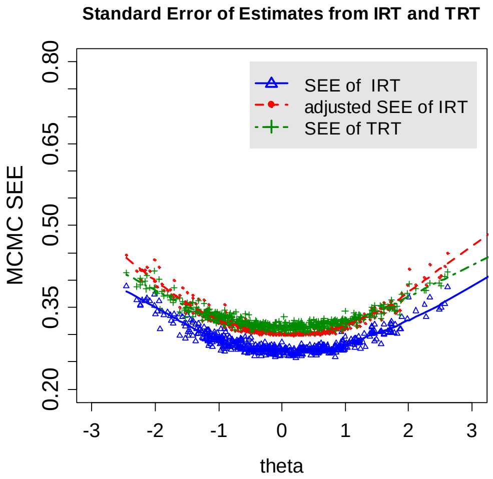
C18: large LID, long, balanced, 3-PL



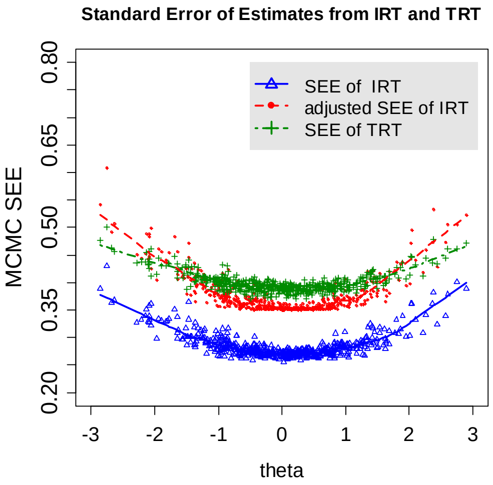
C19: no LID, short, medium unbalanced, 1-PL



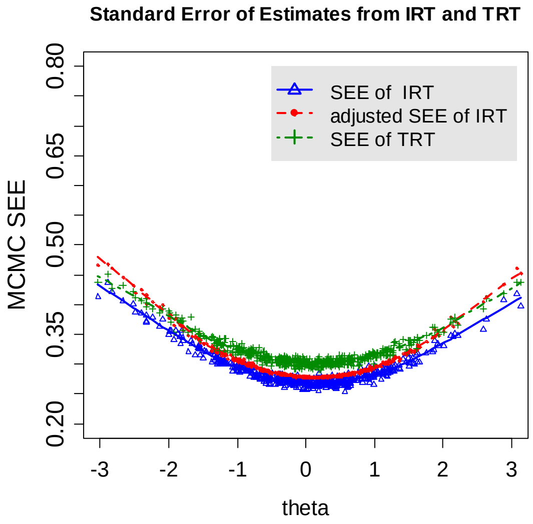
C20: small LID,short,medium unbalanced, 1-PL



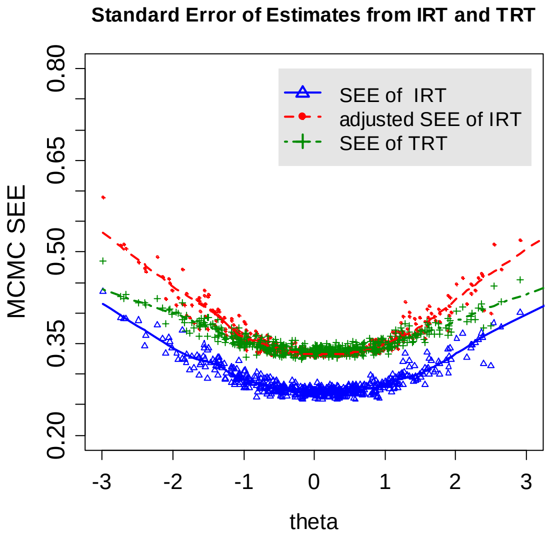
C21:large LID,short,medium unbalanced, 1-PL



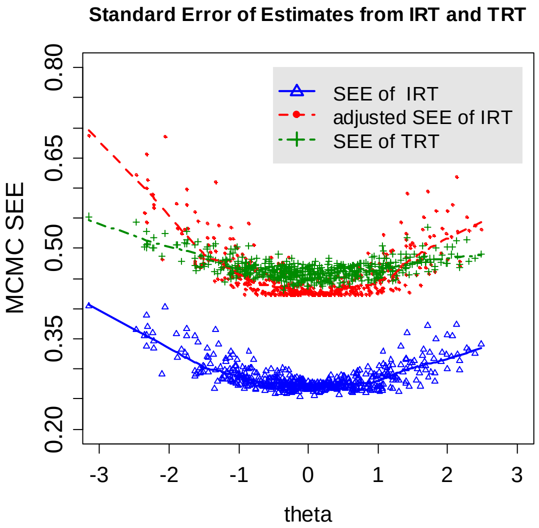
C22: no LID, long, medium unbalanced, 1-PL



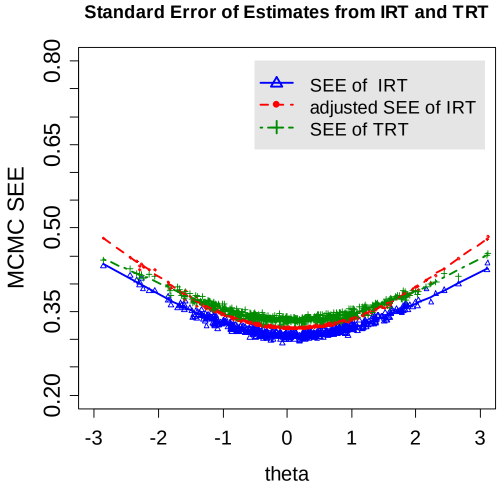
C23:small LID,long,medium unbalanced,1-PL



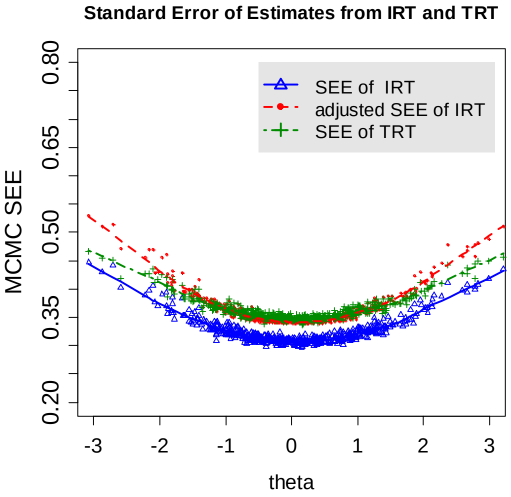
C24:large LID,long, medium unbalanced,1-PL



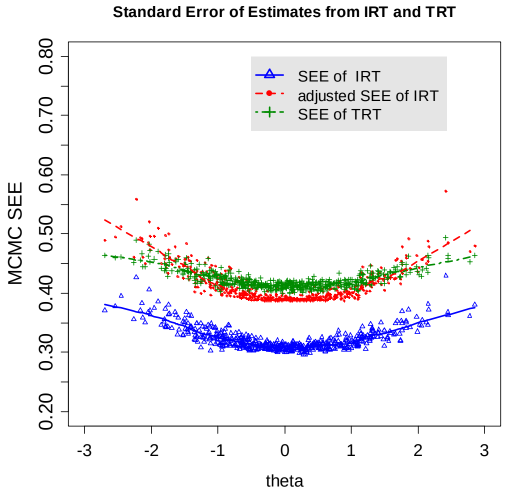
C25: no LID, short, medium unbalanced, 2-PL



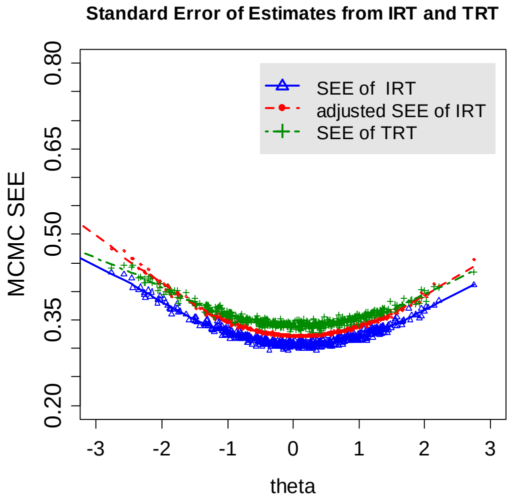
C26: small LID,short,medium unbalanced,2-PL



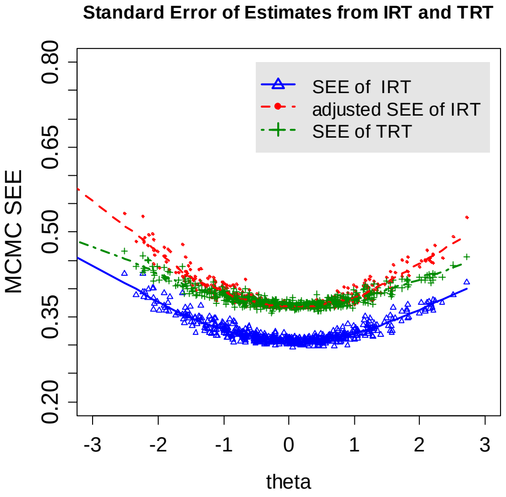
C27:largeLID,short,medium unbalanced, 2-PL



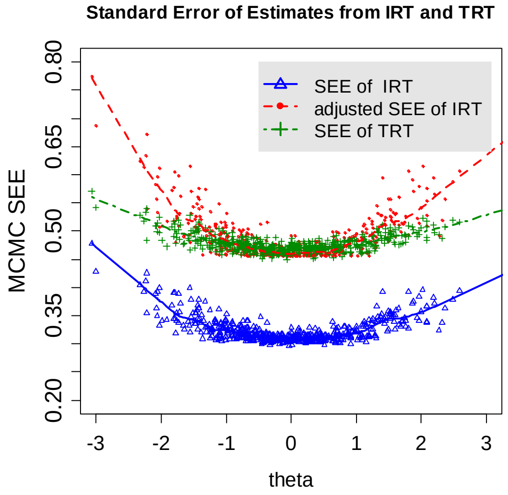
C28: no LID, long, medium unbalanced, 2-PL



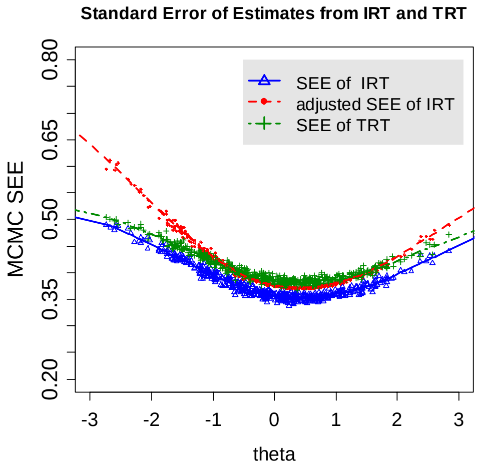
C29:small LID,long, medium unbalanced, 2-PL



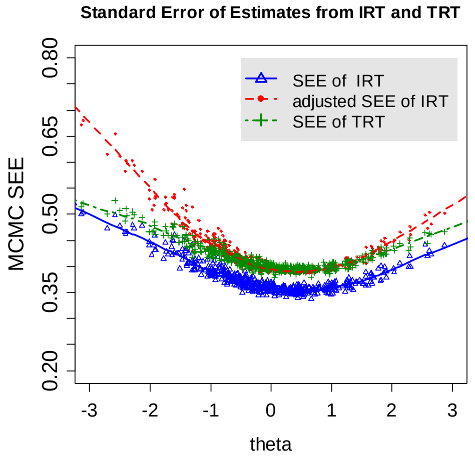
C30:large LID, long,medium unbalanced, 2-PL



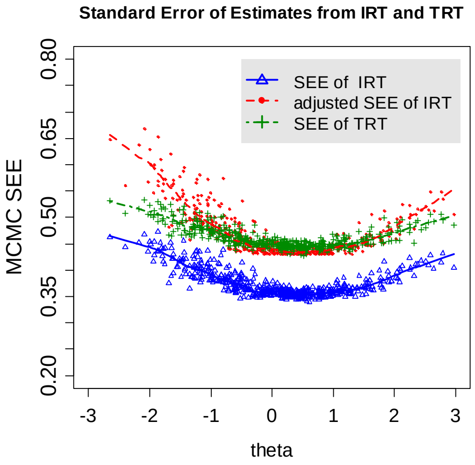
C31: no LID, short, medium unbalanced, 3-PL



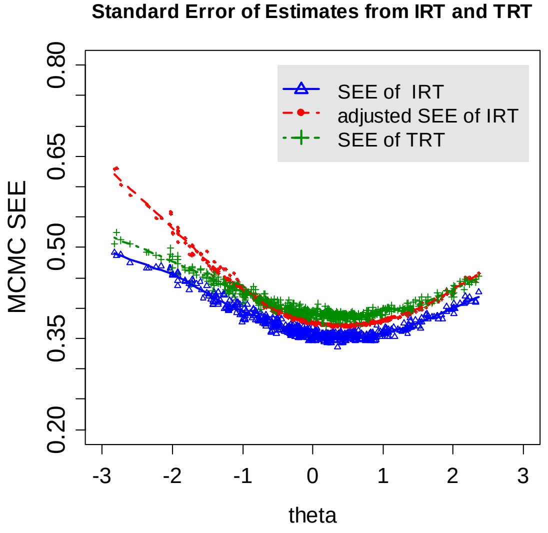
C32:small LID,short,medium unbalanced,3-PL



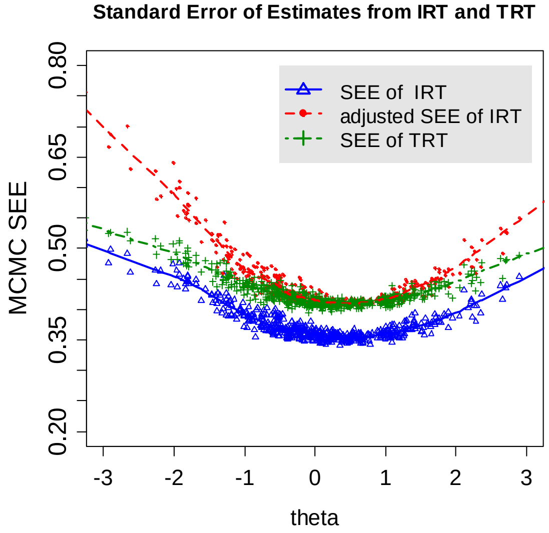
C33:large LID,short,medium unbalanced, 3-PL



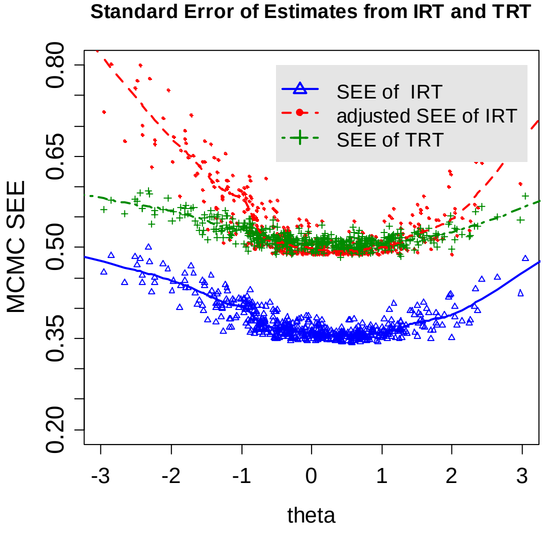
C34: no LID, long, medium unbalanced, 3-PL



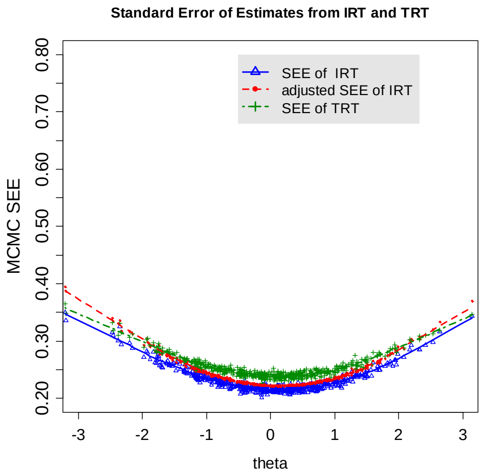
C35:small LID,long, medium unbalanced, 3-PL



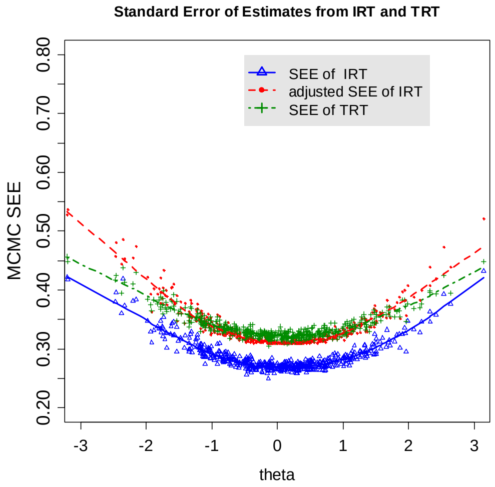
C36:large LID,long, medium unbalanced, 3-PL



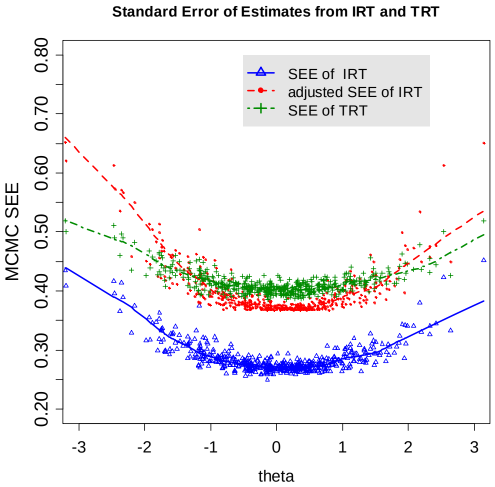
C37: no LID, short, extreme unbalanced, 1-PL



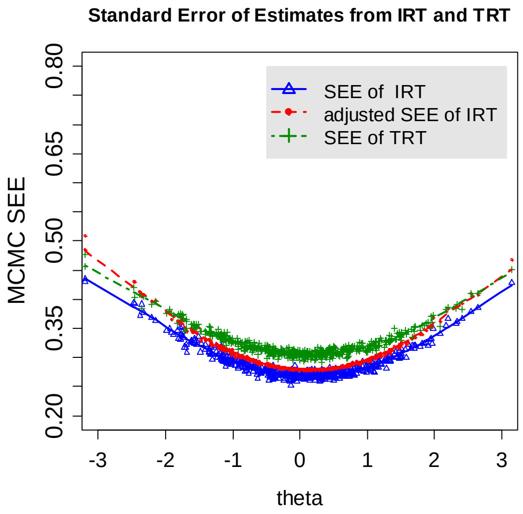
C38:small LID,short,extreme unbalanced, 1-PL



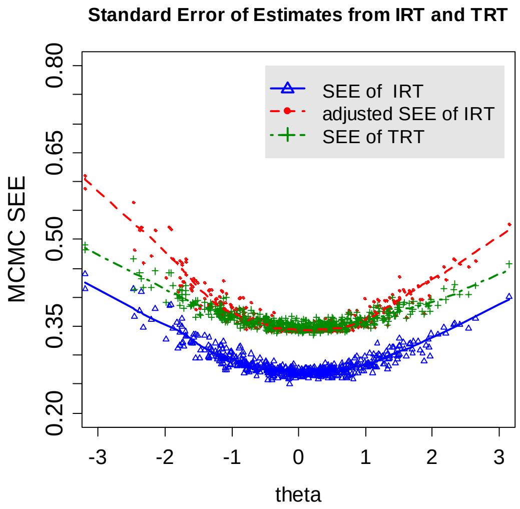
C39:large LID,short,extreme unbalanced,1-PL



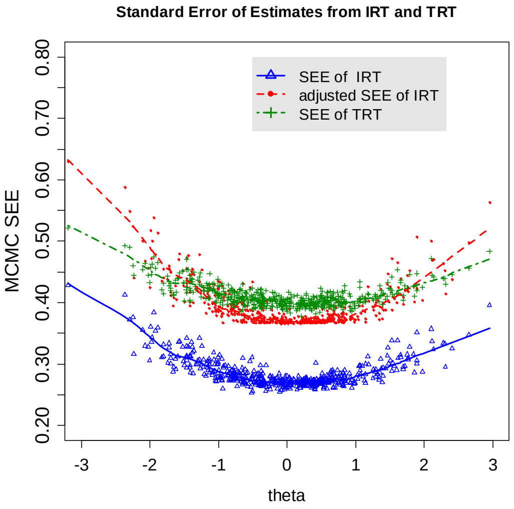
C40: no LID, long, extreme unbalanced, 1-PL



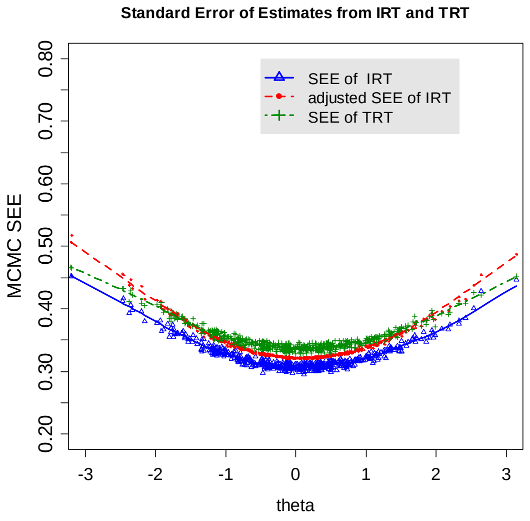
C41:small LID,long, extreme unbalanced,1-PL



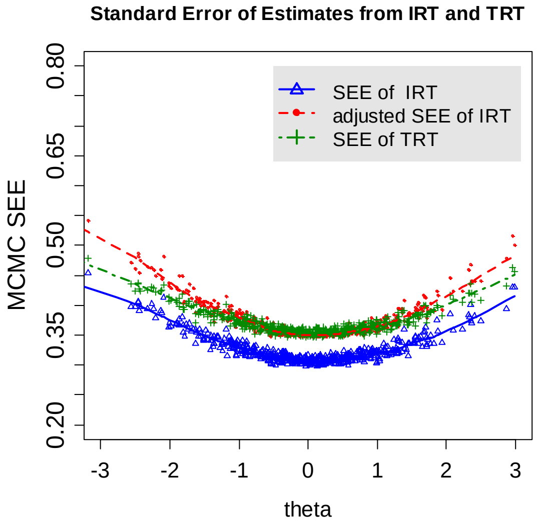
C42:large LID,long, extreme unbalanced, 1-PL



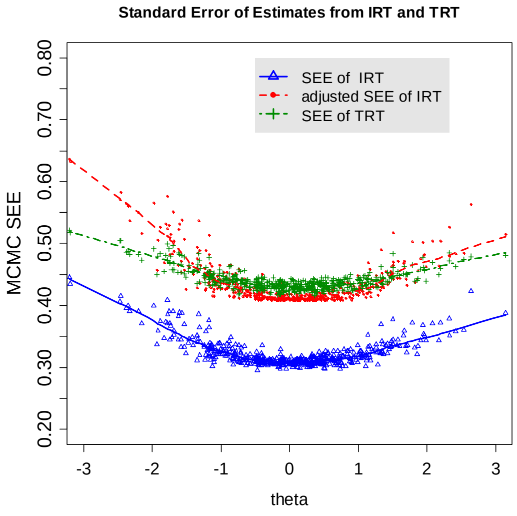
C43: no LID, short, extreme unbalanced, 2-PL



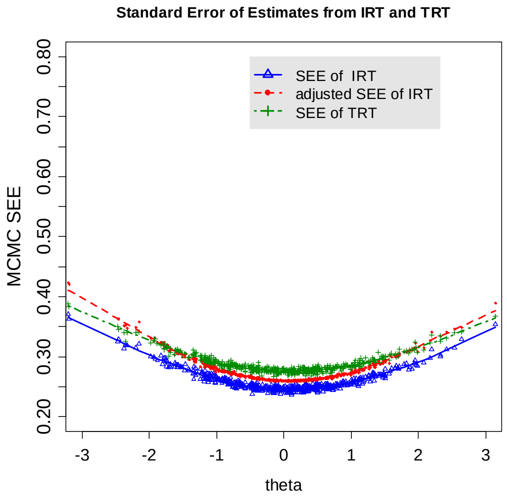
C44:small LID, short, extreme unbalanced, 2-PL



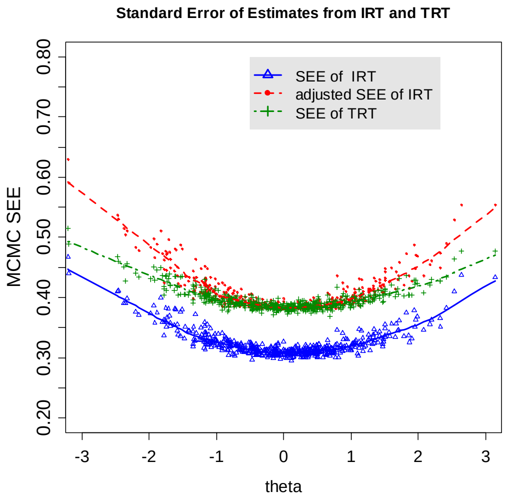
C45:largeLID,short,extreme unbalanced,2-PL



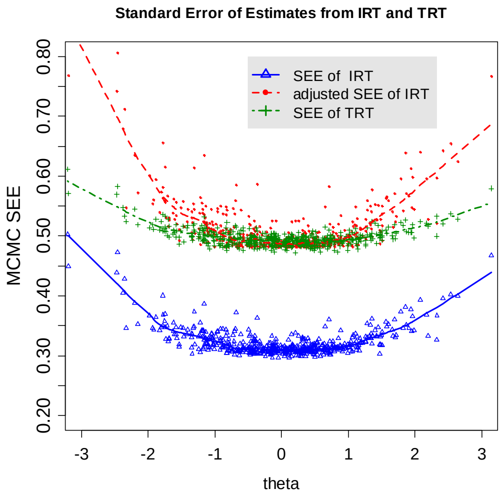
C46: no LID, long, extreme unbalanced, 2-PL



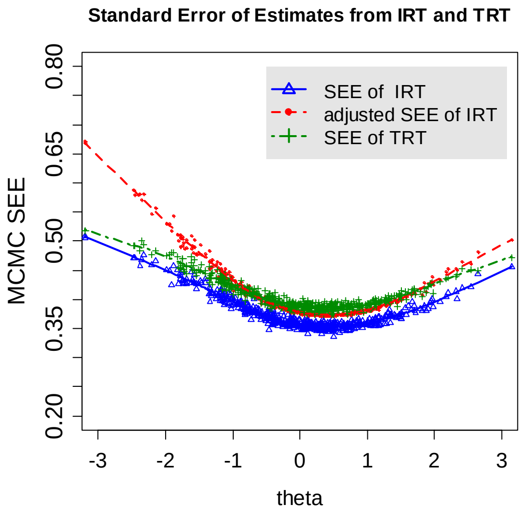
C47: small LID, long, extreme unbalanced, 2-PL



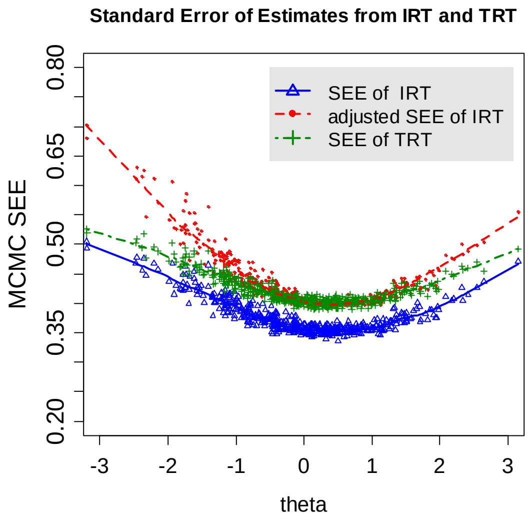
C48:large LID,long,extreme unbalanced, 2-PL



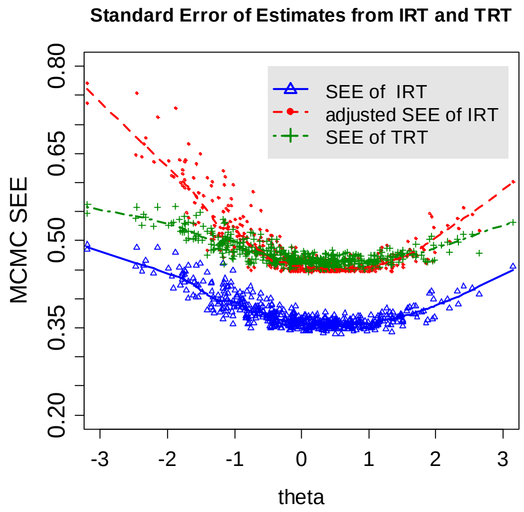
C49: no LID, short, extreme unbalanced, 3-PL



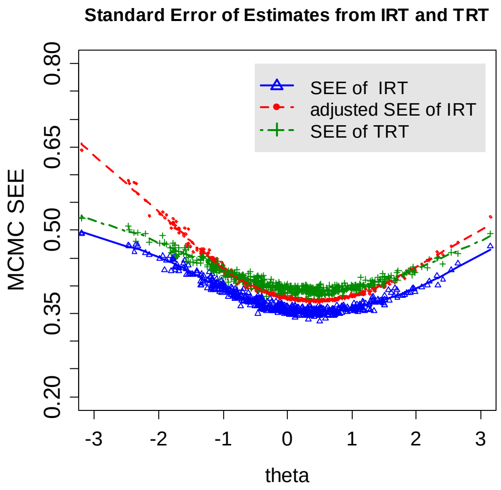
C50:small LID,short, extreme unbalanced, 3-PL



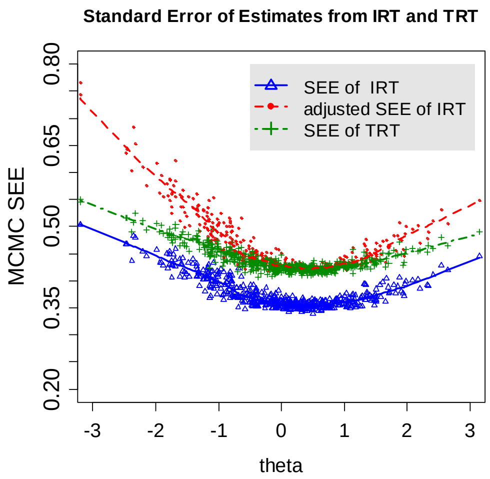
C51:large LID,short,extreme unbalanced,3-PL



C52: no LID, long, extreme unbalanced, 3-PL



C53:small LID, long, extreme unbalanced, 3-PL



C54:large LID,long, extreme unbalanced, 3-PL

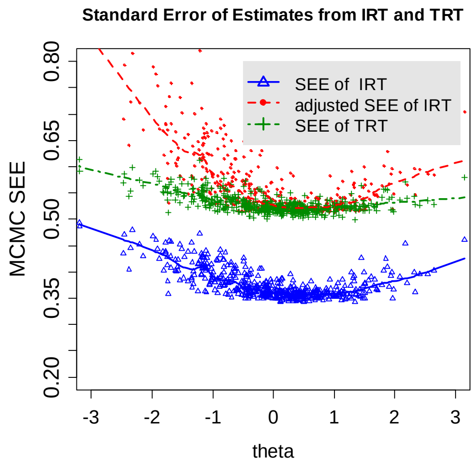


Figure 5.26: Standard Error Increase Discrepancy (SEID*) from 1-PL

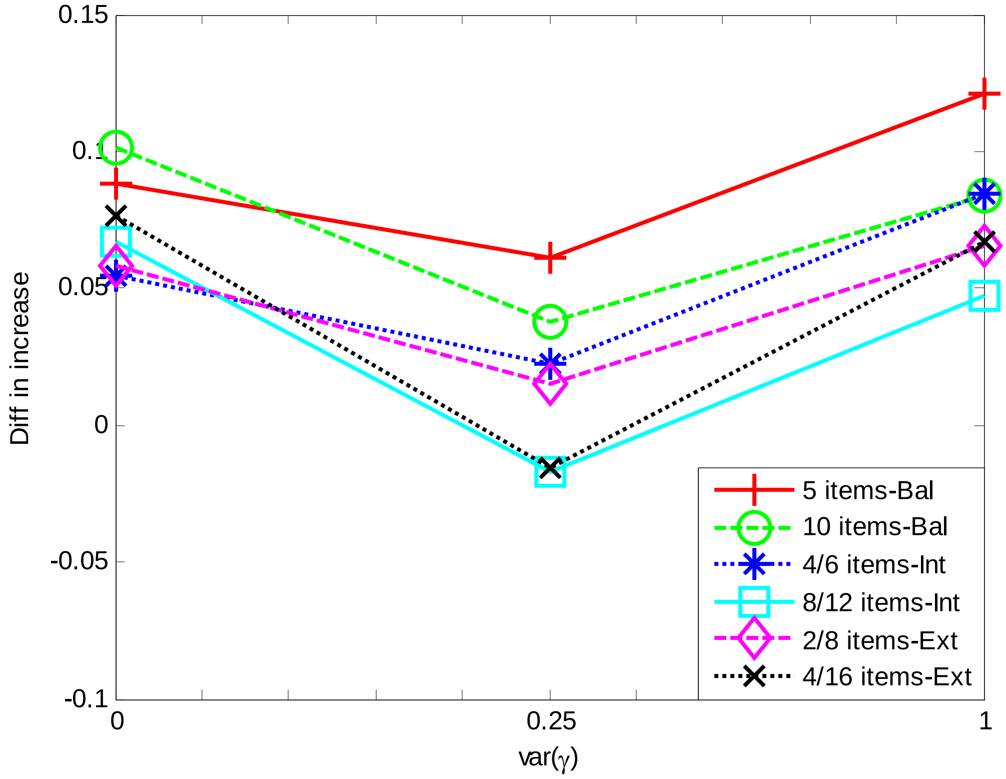


Figure 5.27: Standard Error Increase Discrepancy (SEID) from 2-PL

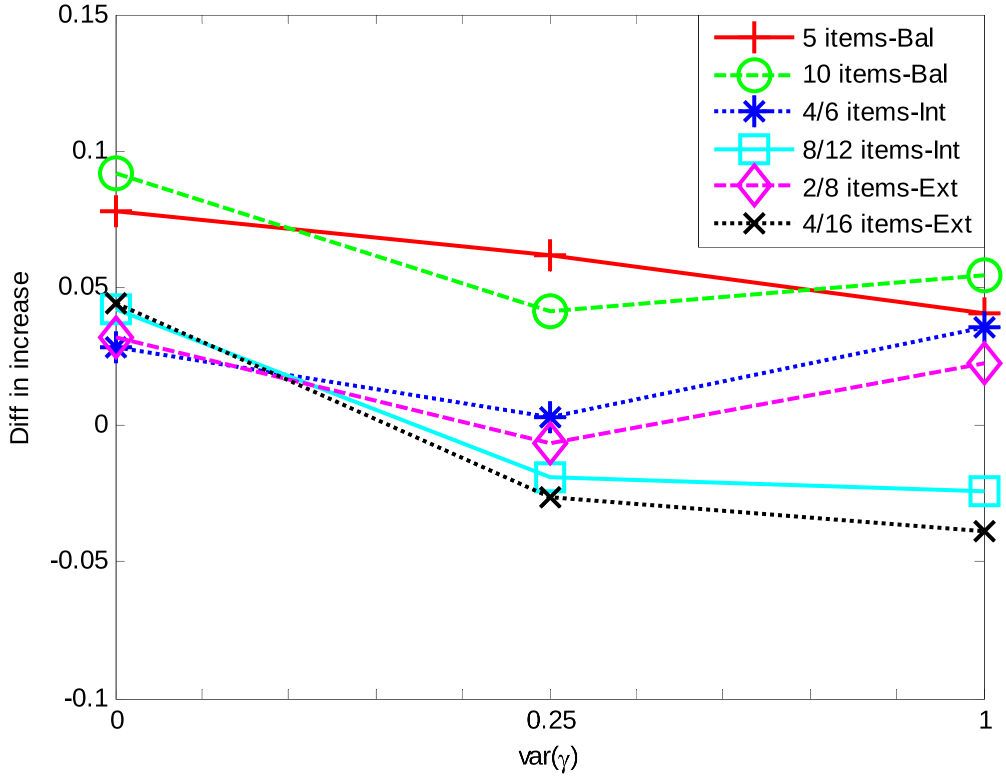
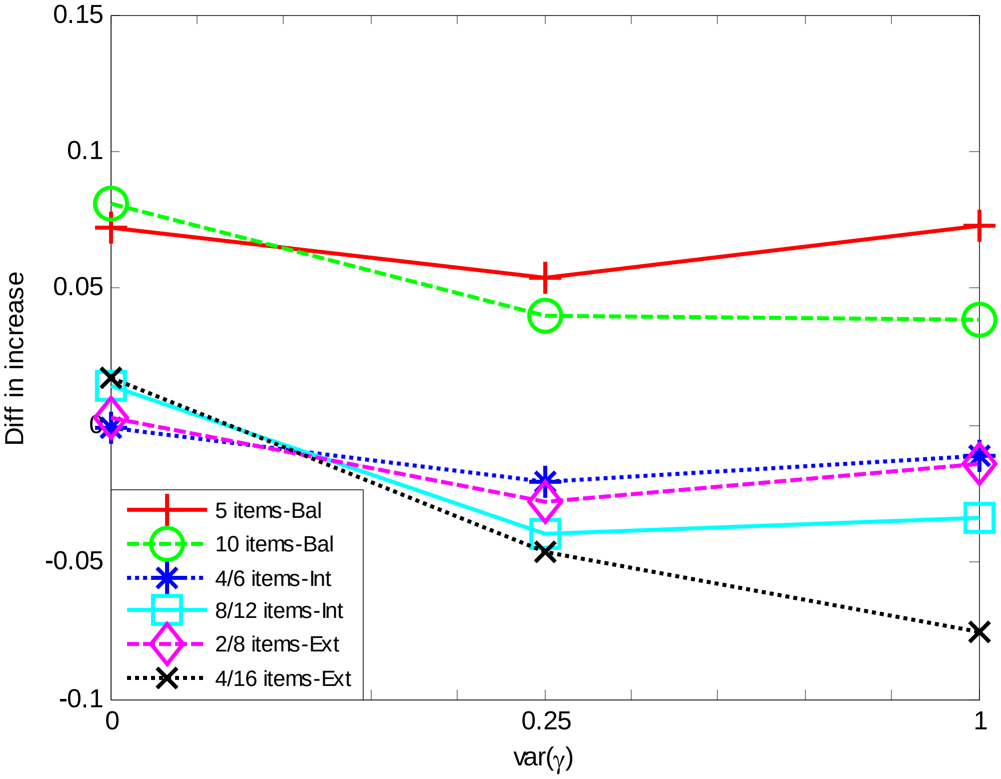


Figure 5.28: Standard Error Increase Discrepancy (SEID) from 3-PL



Note: The standard error increase discrepancy (SEID): the difference between the SEEs of proficiency from IRT models adjusted by information correction ratio compared against those from TRT models

$$\frac{SEE_{TRT} - SEE_{IRT-t_d}}{SEE_{IRT}}$$

Table 5.2: Tests of Between-subjects Effects on SEID in 1-PL Context

	Sum of Squares	df	Mean Square	F	Sig.	Eta Squared
Intercept	9.376	1	9.376	180682.566	.000	0.731
balance	4.192	2	2.096	40392.665	.000	0.327
length	.097	1	.097	1869.847	.000	0.008
LID	2.587	2	1.293	24925.110	.000	0.202
balance * length	.804	2	.402	7750.104	.000	0.063
balance * LID	4.384	4	1.096	21121.991	.000	0.342
length * LID	.049	2	.025	474.198	.000	0.004
balance * length * LID	.660	4	.165	3177.644	.000	0.051
Error	.046	882	5.19E-005			0.004
Total	22.195	900				
Corrected Total	12.819	899				1

Table 5.3: Tests of Between-subjects Effects on SEID in 2-PL Context

	Sum of Squares	df	Mean Square	F	Sig.	Eta Squared
Intercept	3.892	1	3.892	74479.218	.000	0.527
balance	2.602	2	1.301	24896.414	.000	0.352
length	.059	1	.059	1137.186	.000	0.008
LID	.673	2	.337	6440.084	.000	0.091
balance * length	.477	2	.239	4567.438	.000	0.065
balance * LID	3.163	4	.791	15128.766	.000	0.428
length * LID	.022	2	.011	207.791	.000	0.003
balance * length * LID	.340	4	.085	1625.238	.000	0.046
Error	.046	882	5.23E-005			0.006
Total	11.275	900				
Corrected Total	7.382	899				1

Table 5.4: Tests of Between-subjects Effects on SEID in 3-PL Context

	Sum of Squares	df	Mean Square	F	Sig.	Eta Squared
Intercept	2.829	1	2.829	53125.683	.000	0.482
balance	.283	2	.141	2655.039	.000	0.048
length	.222	1	.222	4176.241	.000	0.038
LID	.045	2	.023	425.653	.000	0.008
balance * length	1.671	2	.836	15692.662	.000	0.285
balance * LID	3.070	4	.768	14415.767	.000	0.523
length * LID	.239	2	.120	2248.230	.000	0.041
balance * length * LID	.289	4	.072	1358.801	.000	0.049
Error	.047	882	5.32E-005			0.008
Total	8.696	900				
Corrected Total	5.867	899				1

5.2. Real Data Analysis

5.2.1. Conditional Independence

The distributional statistics for Q_3 local item dependence measures are shown in table 5.5. Although Q_3 is a correlation between residuals of an item pair, Q_3 has a tendency to be slightly negative when the CI holds (Chen & Thissen, 1997; Yen, 1984, 1993). Yen (1993) demonstrated that the expected value of Q_3 is approximately -1/(n-1), and n denotes the number of test items. The expected value for Q_3 can be used as a criterion for comparing the overall level of local dependence of within-testlet item pairs. When CI holds, the average of Q_3 from within-testlet item pairs will be similar to the expected values of the Q_3. Table 5.5 shows that average within-testlet Q_3 statistics have more positive values compared to the expected values of Q_3. This suggests that CI is violated. Yen and Fitzpatrick (2006) suggested paying special attention to testlets with average Q3 values greater than 0.2. In the paper by Lee et al. (2001), the magnitude of the LID is evaluated by referring to the number of standard deviations between the observed and expected mean of Q3 using the standard deviation of the observed Q3 statistics within each testlet. They regarded one SD as appreciable magnitude of difference. Following Lee et al. (2001) approach the t-scores of the within-testlet Q3 statistics were calculated. By referring to the t-scores of the observed Q_3 statistics, the magnitude of the differences between the observed values and the expected value of Q_3 seem to be approximately one SD or even larger, except for testlet 4, 5 and 7 where these differences are moderate.

Table 5.5: Means, Standard Deviations and T-scores of Q3 Statistics within Testlets

Testlet	1	2	3	4	5	6	7	8
Mean	.0252	.0824	.0852	.0023	.0063	.0304	.0401	.1162
SD	.0508	.0491	.0903	.0501	.0395	.0357	.1120	.0440
t-score	1.0273	2.2260	1.2438	0.5847	0.8426	1.6078	0.5998	3.2549

Note: The expected value of Q3 is -1/(38-1)=-.0270

To understand which simulation condition is closest to the response matrix of the real test, the Q_3 statistics of one simulated dataset in each condition in the 3-PL context are estimated. Their means, standard deviations and the t-scores of the mean Q_3 within each testlet are presented as follows (Table 5.6). Q_3 statistics tend to increase as the variance of testlet effect variable specified in data generation becomes larger. By comparing both means and t-scores in Table 5.5 and Table 5.6, it is found that the Q_3 pattern of the real test data is somewhere between condition 32 and 33. Condition 32 and 33 share the features of short testlet length and intermediate level of unbalance in testlet length. The only difference between them is that in condition 32 the dataset is generated with a small variance of testlet effect while in condition 33 the dataset is generated with a large variance of testlet effect.

Table 5.6: Means, Standard Deviations and T-scores of Q3 within Testlets for 3-PL Conditions

condition	testlet	1	2	3	4	5	6	7	8	9	10	11	12
C13: Bal/S/ No LID	Mean	-0.02	-0.04	-0.02	-0.03	-0.02	-0.01	-0.03	0.00	-0.02	-0.02	-0.02	0.02
	SD	0.04	0.06	0.05	0.06	0.06	0.03	0.03	0.04	0.05	0.04	0.03	0.04
	t-score	-0.20	-0.33	-0.08	-0.28	-0.06	0.31	-0.48	0.52	-0.11	-0.01	-0.10	0.93
C14 Bal/S/ Small LID	Mean	0.02	0.02	0.02	0.03	0.01	0.01	0.00	0.00	0.01	-0.01	0.00	-0.01
	SD	0.04	0.04	0.05	0.05	0.03	0.03	0.07	0.06	0.03	0.06	0.04	0.07
	t-score	1.03	0.90	0.74	1.00	0.87	0.76	0.27	0.21	0.85	0.11	0.39	0.15
C15 Bal/S/ Large LID	Mean	0.05	0.08	0.07	0.08	0.06	0.08	0.08	0.07	0.08	0.06	0.09	0.05
	SD	0.03	0.03	0.07	0.05	0.03	0.04	0.04	0.03	0.04	0.04	0.06	0.03
	t-score	2.23	2.75	1.24	1.82	2.55	2.19	2.31	2.68	2.28	1.71	1.88	1.91
C16 Bal/L/ No LID	Mean	-0.01	-0.01	-0.03	-0.02	-0.02	0.01						
	SD	0.04	0.04	0.05	0.05	0.04	0.04						
	t-score	0.09	0.05	-0.23	-0.10	-0.10	0.54						
C17 Bal/L/ Small LID	Mean	0.00	0.01	0.01	0.00	0.02	0.01						
	SD	0.05	0.05	0.05	0.04	0.04	0.05						
	t-score	0.42	0.52	0.53	0.40	0.78	0.46						
C18 Bal/L/ Large LID	Mean	0.06	0.06	0.05	0.04	0.06	0.05						
	SD	0.06	0.04	0.05	0.05	0.04	0.04						
	t-score	1.48	1.76	1.39	1.29	1.83	1.59						
C31: M Unb/S/ No LID	Mean	-0.01	0.00	0.00	0.00	-0.01	-0.03	0.00	-0.02	-0.01	-0.01	-0.02	0.01
	SD	0.02	0.05	0.05	0.04	0.04	0.06	0.04	0.05	0.05	0.03	0.02	0.05
	t-score	0.22	0.40	0.22	0.32	0.16	-0.14	0.37	0.03	0.12	0.17	-0.08	0.55
C32: M Unb/S/ Small LID	Mean	0.02	0.01	-0.01	0.01	-0.01	0.02	0.00	-0.02	-0.01	0.00	0.01	-0.01
	SD	0.06	0.04	0.04	0.06	0.04	0.04	0.05	0.04	0.05	0.04	0.08	0.05
	t-score	0.63	0.79	0.15	0.43	0.08	0.74	0.36	-0.01	0.25	0.43	0.37	0.06
C33: M Unb/S/ Large LID	Mean	0.05	0.09	0.07	0.08	0.07	0.07	0.06	0.05	0.05	0.07	0.06	0.05
	SD	0.05	0.05	0.03	0.04	0.01	0.05	0.06	0.04	0.04	0.05	0.06	0.05
	t-score	1.33	2.07	2.76	2.71	6.00	1.86	1.32	1.68	1.65	1.53	1.35	1.37
C34: M Unb/L/ no LID	Mean	-0.02	-0.01	-0.01	-0.02	-0.01	-0.01						
	SD	0.04	0.04	0.04	0.04	0.05	0.05						
	t-score	-0.10	0.14	0.14	0.02	0.20	0.13						

Table 5.6 (continued):

	testlet	1	2	3	4	5	6	7	8	9	10	11	12
C35: M Unb/L/ Small LID	Mean	0.01	0.01	0.03	0.02	0.00	0.01						
	SD	0.04	0.04	0.05	0.04	0.04	0.05						
	t-score	0.66	0.62	0.92	0.78	0.32	0.43						
C36: M Unb/L/ Large LID	Mean	0.08	0.05	0.07	0.04	0.06	0.06						
	SD	0.04	0.05	0.04	0.05	0.05	0.04						
	t-score	2.24	1.51	1.99	1.33	1.52	1.78						
C49: E Unb/S/ No LID	Mean	-0.05	-0.02	0.01	-0.02	0.03	-0.01	-0.06	-0.02	-0.01	-0.02	0.02	0.00
	SD	NA	0.04	NA	0.04	NA	0.04	NA	0.05	NA	0.04	NA	0.04
	t-score	NA	-0.11	NA	-0.04	NA	0.11	NA	-0.02	NA	0.00	NA	0.39
C50: E Unb/S/ Small LID	Mean	0.12	0.00	0.11	0.03	-0.01	0.01	0.05	0.00	0.06	0.00	0.00	0.02
	SD	NA	0.05	NA	0.05	NA	0.05	NA	0.05	NA	0.03	NA	0.05
	t-score	NA	0.31	NA	0.90	NA	0.68	NA	0.38	NA	0.63	NA	0.76
C51: E Unb/S/ Large LID	Mean	0.16	0.08	0.04	0.04	0.13	0.05	0.16	0.05	0.11	0.06	0.04	0.06
	SD	NA	0.06	NA	0.05	NA	0.05	NA	0.03	NA	0.05	NA	0.04
	t-score	NA	1.72	NA	1.25	NA	1.22	NA	1.96	NA	1.64	NA	1.88
C52: E Unb/L/ no LID	Mean	-0.02	-0.02	-0.02	-0.01	-0.03	-0.02						
	SD	0.03	0.05	0.04	0.04	0.03	0.04						
	t-score	-0.12	0.03	-0.06	0.19	-0.49	0.00						
C53: E Unb/L/ Small LID	Mean	0.02	0.00	-0.02	0.00	0.03	0.01						
	SD	0.03	0.05	0.03	0.05	0.02	0.04						
	t-score	1.23	0.27	-0.21	0.31	2.19	0.52						
C54: E Unb/L/ Large LID	Mean	0.10	0.04	0.10	0.04	0.10	0.05						
	SD	0.02	0.04	0.04	0.04	0.08	0.05						
	t-score	6.55	1.17	3.01	1.35	1.42	1.50						

Note: the expected value of Q3 is -1/(60-1)= -0.01695

S: short testlet;

L: long testlet

Bal: Balanced testlet length

M Unb: medium unbalanced testlet length

E Unb: extremely unbalanced testlet length

5.2.2. Unidimensionality

Table 5.7 presents the eigenvalues of the largest components as a result of the principle component analysis implemented in TESTFACT. The analysis yields 9 components with eigenvalues larger than 1. The first component accounts for over 40% of the overall variance in the dependent variable. The largest eigenvalue is about 6 times as large as the second largest eigenvalue. All these obviously suggest that one factor is dominant in this dataset. When we further look at the scree plots of the principle component analysis (Figure 5.29), the eigenvalue of the first component is significantly larger (12.873) than those of all the other components, which provides more evidence that one dimension is sufficient to account for the variance in this dataset. From the table of factor loadings, almost all items load highly on the first factor compared with loadings on other factors, which also agrees that this test is for the most part unidimensional. The second factor, which has an eigenvalue that is much smaller in size and only slightly larger than the remaining eigenvalues, shows a tendency to be related to item position: Items earlier in the test tend to have higher positive loadings, while items toward the end of the test tend to have higher negative loadings, although for all items but one, smaller in absolute value than their loadings on the dominant first factor. This pattern may suggest speededness in the test, but it does not distinguish the testlet structure.

Table 5.7: Eigenvalues of the Principle Components

Component	Eigenvalues	% of variance	Cumulative %
1	12.873	40.267	40.267
2	2.180	6.111	46.378
3	1.292	3.298	49.676
4	1.268	2.864	52.540
5	1.224	1.929	54.469
6	1.175	1.359	55.828
7	1.108	1.086	56.914
8	1.049	0.671	57.585
9	1.008	0.471	58.055
10	0.944	0.448	58.504

Figure 5.29: Scree Plot from Principle Component Analysis on Tetrachoric Correlation Matrix

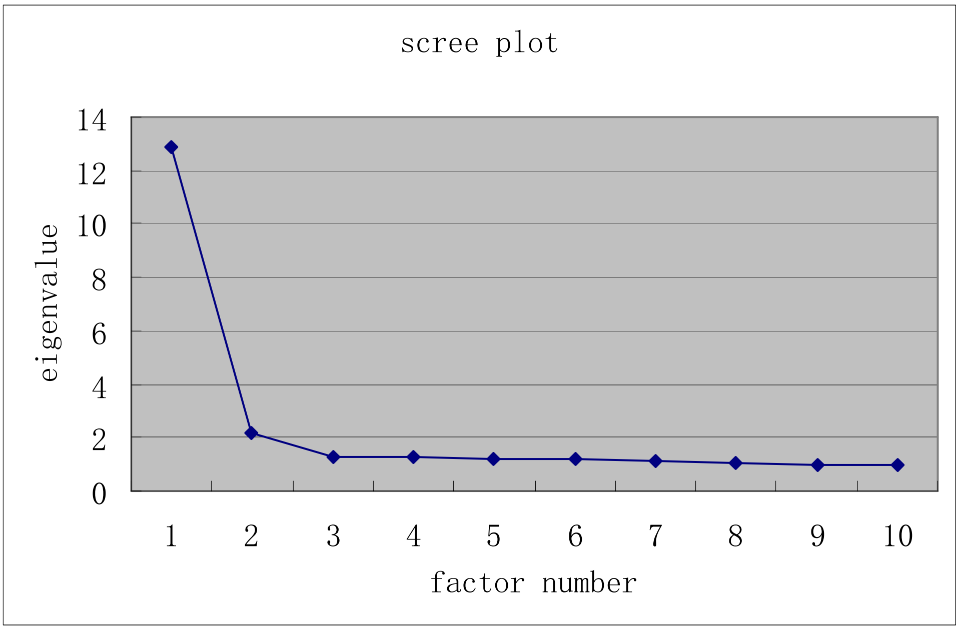


Table 5.8: Factor Loadings on the First 10 Principle Components

Item #	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7	Factor 8	Factor 9	Factor 10
1	0.65	0.17	-0.07	-0.04	-0.09	0.01	0.06	0.03	-0.01	-0.21
2	0.69	0.39	0.11	0.10	-0.22	0.09	0.06	-0.09	-0.07	0.14
3	0.53	0.05	0.01	0.09	-0.10	-0.01	0.10	0.06	0.09	-0.12
4	0.77	0.29	-0.02	-0.05	-0.13	-0.14	0.02	-0.08	0.02	-0.09
5	0.54	0.12	-0.09	-0.20	0.12	-0.20	-0.02	0.03	0.16	-0.05
6	0.36	0.23	-0.47	-0.47	-0.30	-0.07	-0.16	-0.15	0.07	0.15
7	0.34	0.32	0.04	-0.23	0.28	0.26	-0.30	0.07	-0.06	-0.21
8	0.43	0.22	0.10	0.04	-0.22	0.19	-0.13	-0.03	0.06	-0.08
9	0.64	0.23	0.08	-0.04	0.09	0.09	-0.05	0.14	0.05	-0.14
10	0.67	0.15	-0.10	-0.10	0.22	0.00	0.09	0.12	0.03	0.18
11	0.50	0.21	0.31	-0.15	0.10	0.08	0.11	-0.23	-0.05	0.07
12	0.64	0.24	0.20	-0.18	0.31	-0.10	0.19	0.00	0.10	0.05
13	0.45	0.06	0.08	0.05	-0.06	0.03	-0.07	0.29	0.13	0.04
14	0.53	0.12	-0.05	0.00	-0.05	-0.08	0.05	0.31	-0.04	0.01
15	0.57	0.00	-0.28	0.03	0.01	-0.06	0.03	0.14	-0.19	-0.05
16	0.53	0.11	-0.06	0.03	0.13	0.04	0.08	-0.10	-0.20	-0.07
17	0.72	0.08	0.05	0.05	-0.19	-0.02	0.03	0.03	0.03	-0.04
18	0.61	0.03	-0.10	0.08	0.10	-0.10	0.06	0.10	-0.21	0.01
19	0.63	0.07	-0.05	0.18	0.00	0.05	-0.04	-0.15	-0.01	-0.06
20	0.75	-0.02	-0.09	0.14	-0.05	0.14	0.14	-0.03	0.07	-0.10
21	0.56	-0.06	-0.01	0.12	-0.07	-0.01	0.07	-0.05	-0.04	-0.10
22	0.72	0.06	0.02	0.13	-0.07	0.03	0.17	-0.04	-0.05	0.09
23	0.54	0.08	0.08	0.07	0.03	0.07	0.03	-0.25	0.05	0.21
24	0.59	-0.03	0.07	0.14	-0.12	0.03	-0.08	-0.06	-0.08	0.01
25	0.53	-0.03	0.07	0.18	-0.07	0.00	-0.30	0.15	0.05	0.24
26	0.30	-0.14	-0.43	0.26	0.26	0.22	-0.04	-0.21	0.27	-0.06
27	0.36	-0.03	0.12	-0.05	0.17	-0.05	-0.26	-0.17	-0.13	0.01
28	0.61	-0.16	0.05	0.14	0.04	-0.10	-0.19	-0.04	0.06	0.07
29	0.66	0.00	-0.02	0.12	0.17	-0.12	-0.03	0.14	0.21	0.17
30	0.61	-0.28	0.04	-0.02	0.03	-0.46	-0.01	-0.14	0.00	-0.02
31	0.52	-0.25	-0.07	0.06	0.05	-0.22	-0.17	-0.07	-0.12	-0.17
32	0.38	-0.10	-0.21	0.05	0.13	0.17	-0.01	0.07	-0.34	0.21

Table 5.8 (Continued):

Item #	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7	Factor 8	Factor 9	Factor 10
33	0.72	-0.39	0.00	-0.05	-0.09	0.06	0.05	0.03	0.05	-0.08
34	0.57	-0.41	0.12	-0.20	-0.01	0.08	0.03	0.07	-0.05	-0.02
35	0.59	-0.40	0.15	-0.16	-0.11	0.21	-0.18	0.02	0.02	0.03
36	0.16	-0.30	-0.13	-0.23	0.01	0.22	0.23	0.02	0.01	0.05
37	0.61	-0.40	0.07	-0.09	0.03	0.10	0.06	0.02	0.10	0.00
38	0.60	-0.35	0.08	-0.15	-0.06	-0.02	0.06	-0.07	-0.05	0.00

5.2.3. Model Selection

A minimum value of DIC indicates a parsimonious model with good model fit. As a result of MCMC estimation, it turns out that 3-PL TRT model is preferred as it has the smallest DIC among the four models (Table 5.9). On the one hand, it provides further evidence that LID in this test is significant enough to be addressed and needs to be accounted for by the testlet model. On the other hand, it is necessary to model the pseudo guessing parameters, as their values vary significantly from one item to another. Therefore, 3-PL IRT model is selected to calibrate the parameters. The estimated standard errors of ability estimates are adjusted by the ratio of variances from the response matrix and compared with the standard errors of ability estimates calibrated through the 3-PL TRT model.

Table 5.9: Deviance Information Criterion Values for four Models

	IRT	TRT
2-PL	33640.5	32960.5
3-PL	33514.9	32921.0

5.2.4. Parameter Calibration

After the 3-PL IRT and 3-PL TRT models are selected for calibration, the next step is to check convergence in MCMC estimation. It is done by examining whether the simulated Markov chain converges to a stationary distribution. A random subset of parameters is selected for this purpose. A large number of iterations are usually required to ensure the convergence and stable estimates for the complicated models such as 3-PL IRT and TRT models (Sinharay, 2003). Three approaches are generally used in assessing convergence. The first approach is to examine the history plot, which shows the full history of the sample values for the parameter being monitored. Second, we can look at trace plots of the sample values versus iteration to see when the simulation appears to have stabilized. If the chains starting from divergent initial values in the trace plot or history plot appear to be overlapping one another, we have evidence to believe that convergence has taken place. Figure 5.30 and Figure 5.33 demonstrate the history plots for a subset of parameters from 3-PL IRT and TRT calibrations, respectively. For each parameter in the case of TRT calibration, two chains are mixing well and have converged to a stabilized distribution before 5000 iterations are completed. Figure 5.31 and 5.34 show the trace plots for the same parameters. We observe the trend of convergence again. The third diagnostic approach is the Gelman-Rubin index. For the Gelman-Rubin plots, the width of the central 80% interval of the pooled runs is green, the average width of the 80% intervals within the individual runs is blue, and their ratio (pooled /within) is red (Brooks & Gelman, 1998). The Gelman-Rubin plots for the variances of the testlet parameters are presented in Figure 5.32 and 5.35. The blue and green curves overlap, and the red curve hovers around 1 after about 10000 iterations for the variances of the testlet parameters. Based on the history plots, trace plots and the Gelman-Rubin plots, the convergence is achieved after 6000 iterations for 3-PL IRT calibration and 10000 iterations for 3-PL TRT calibration .

Figure 5.30: WigBUGS history plots for parameters $a_3, b_{10}, c_{14}, \theta_{26}, \sigma^2_{\gamma 1}$ from 3-PL TRT

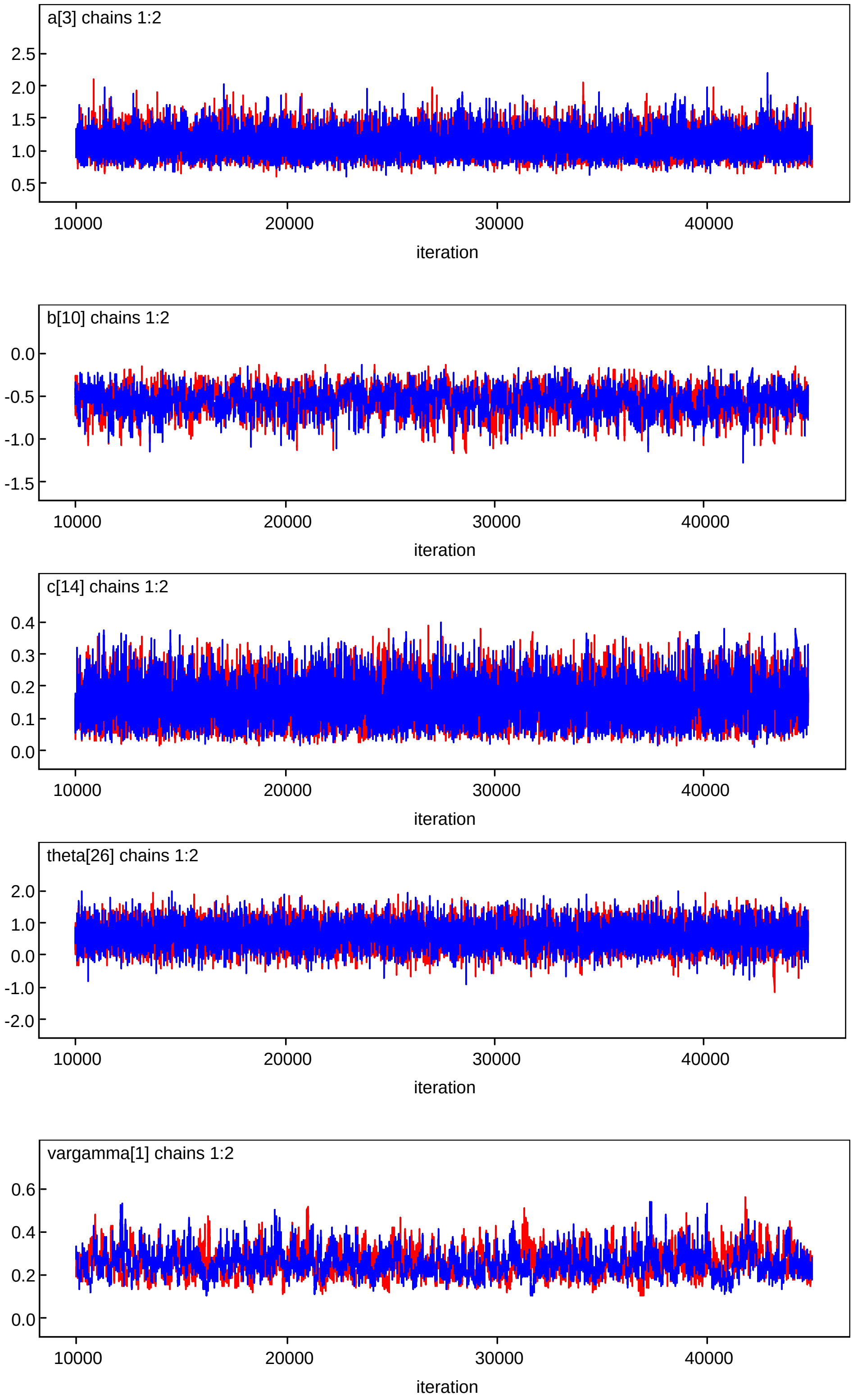


Figure 5.31: WigBUGS Traceplots for parameters $a_3, b_{10}, c_{14}, \theta_{26}, \sigma^2_{\gamma 1}$ from 3-PL TRT

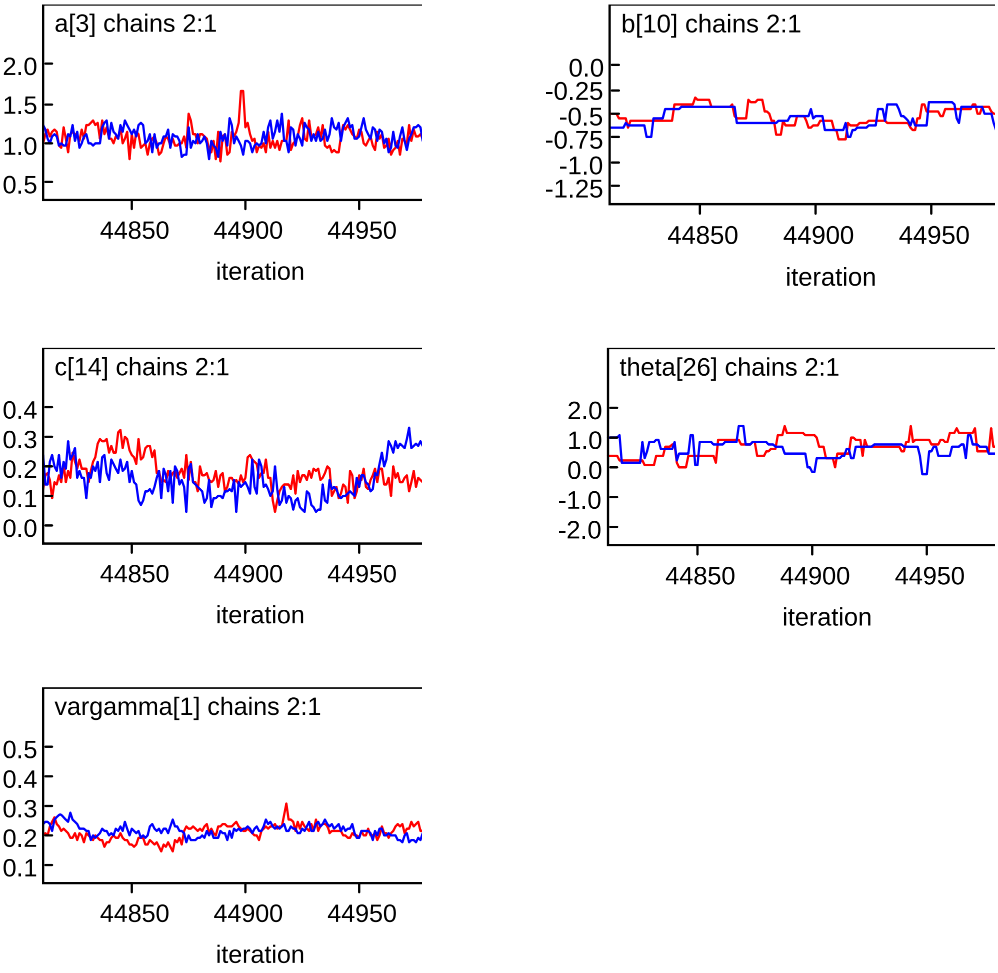


Figure 5.32: WigBUGS Gelman-Rubin plots for parameters $a_3, b_{10}, c_{14}, \theta_{26}, \sigma^2_{\gamma1}$ from 3-PL TRT

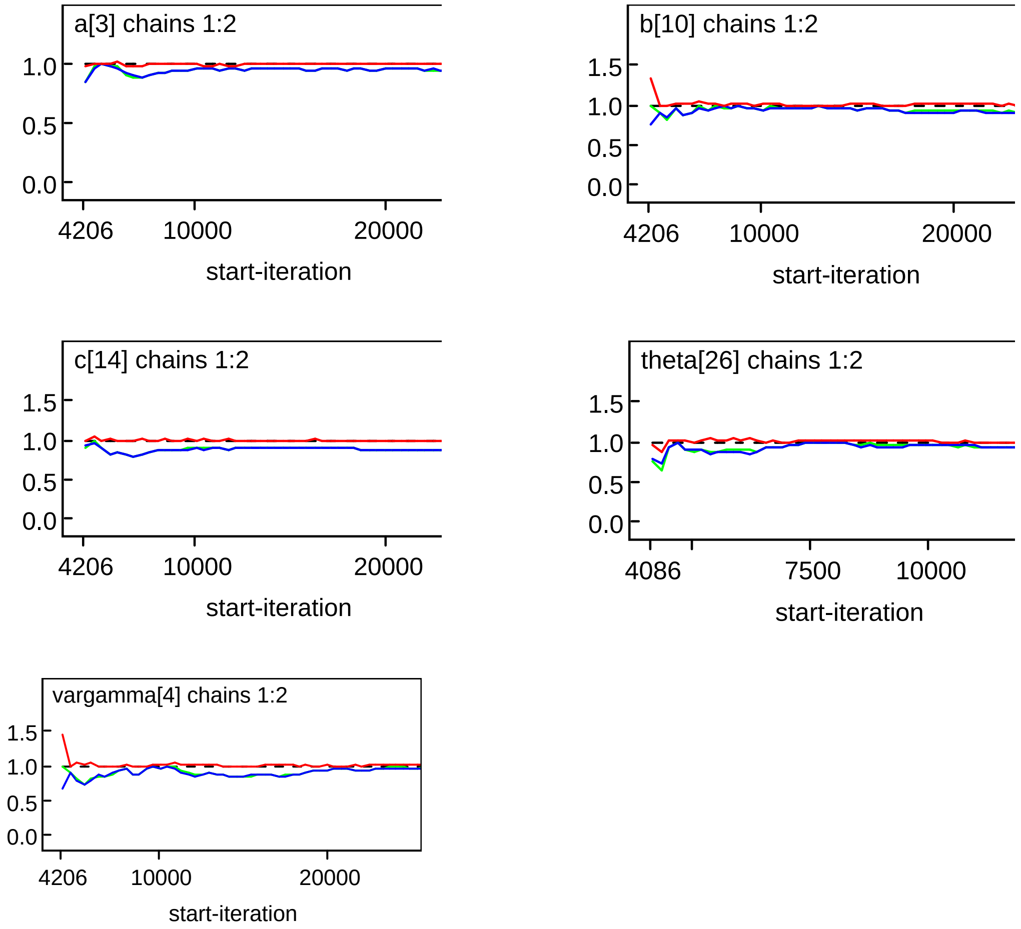


Figure 5.33: WigBUGS history plots for parameters $a_5, b_{37}, c_4, \theta_{18}$ from 3-PL IRT

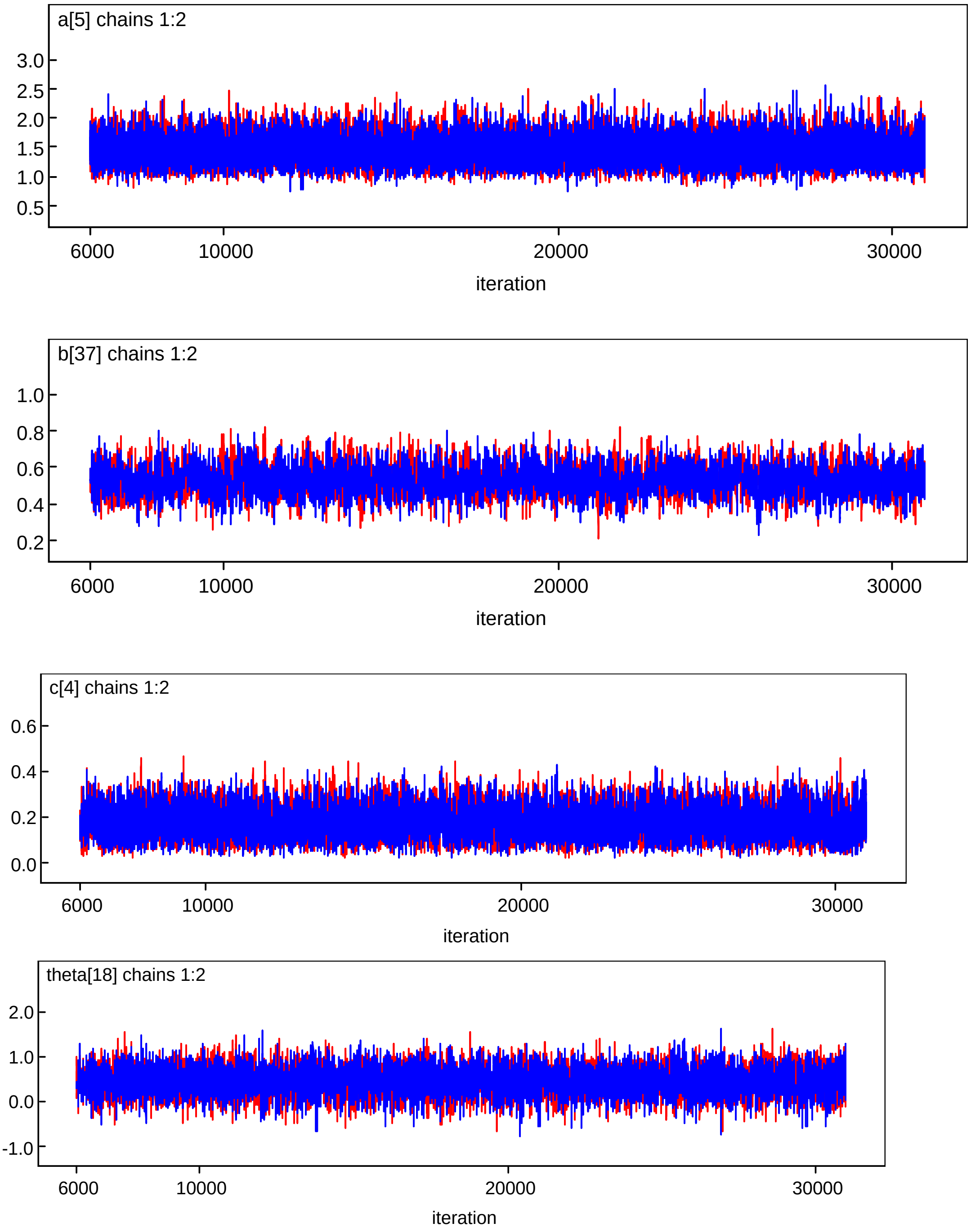


Figure 5.34: WigBUGS trace plots for parameters $\mathbf{a_5, b_{37}, c_4, \theta_{18}}$ from 3-PL IRT

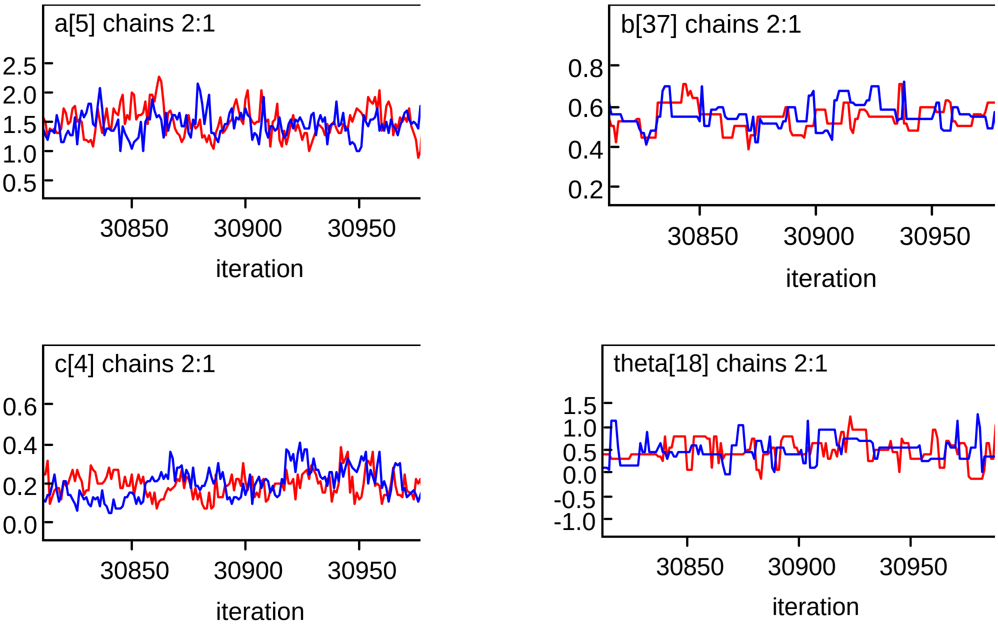
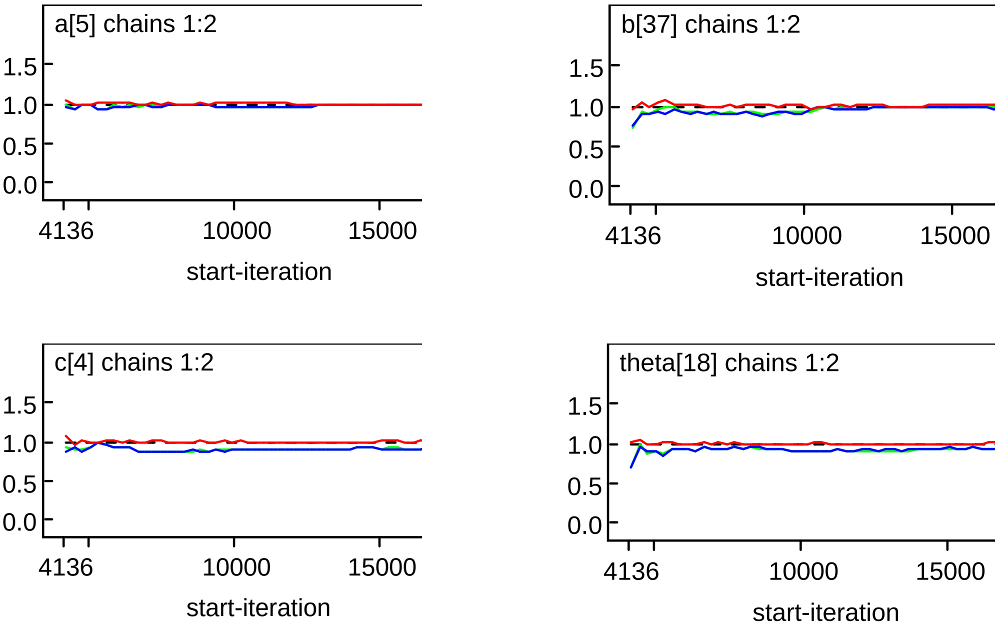


Figure 5.35: WigBUGS Gelman-Rubin plots for parameters $\mathbf{a_5, b_{37}, c_4, \theta_{18}}$ from 3-PL IRT



For the 3-PL IRT model, I ran 50,000 iterations in the numerical implementation. The first 6,000 iterations are discarded as burn-in cycles, so the parameters are estimated from the posterior distributions based on the 6,001st to the 50,000th iterations. For 3-PL TRT model estimation, the posterior distributions are estimated based on the 10,001st to the 70,000th iterations. The means of the Bayesian posterior distributions are used for item parameters in the calculation.

Table 5.10 shows the means of the posterior distributions of item parameter estimates and their standard deviations from 3-PL IRT and 3-PL TRT models. Table 5.11 provides the correlations of parameter estimates and the correlations of standard errors of estimates between two calibration models. Figure 5.35 represents the scatter plots for the 3-PL IRT model estimates against the 3-PL TRT model estimates. Both the figure and the correlation statistics show that the two sets of item parameter estimates are highly correlated.

Table 5.10: Means and Standard Deviations of Posterior Distributions of the Item Parameters Estimates

	α				β				ω			
	3-PL TRT		3-PL IRT		3-PL TRT		3-PL IRT		3-PL TRT		3-PL IRT	
item #	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE
1	1.739	0.259	1.638	0.208	-0.573	0.143	-0.549	0.142	0.184	0.055	0.182	0.057
2	1.804	0.228	1.771	0.188	-1.189	0.139	-1.125	0.123	0.151	0.057	0.139	0.053
3	1.098	0.158	1.122	0.141	-0.190	0.182	-0.207	0.162	0.160	0.055	0.148	0.052
4	2.844	0.465	2.504	0.329	-1.353	0.120	-1.252	0.122	0.167	0.059	0.180	0.062
5	1.543	0.270	1.443	0.217	0.625	0.117	0.611	0.119	0.155	0.036	0.158	0.037
6	1.008	0.246	0.889	0.212	0.569	0.262	0.592	0.290	0.248	0.066	0.248	0.071
7	0.762	0.130	0.680	0.110	-1.757	0.428	-1.844	0.435	0.251	0.091	0.238	0.087
8	0.842	0.134	0.835	0.113	-0.724	0.281	-0.694	0.245	0.188	0.069	0.181	0.065
9	2.444	0.604	1.542	0.178	-0.219	0.112	-0.215	0.116	0.135	0.042	0.135	0.044
10	2.929	0.794	1.883	0.277	-0.559	0.145	-0.598	0.158	0.259	0.058	0.231	0.066
11	1.027	0.142	1.005	0.124	-0.868	0.219	-0.825	0.207	0.174	0.064	0.170	0.063
12	1.809	0.268	1.468	0.166	-0.776	0.139	-0.754	0.140	0.151	0.053	0.152	0.055
13	0.833	0.111	0.879	0.113	-1.036	0.255	-0.971	0.239	0.179	0.067	0.177	0.066
14	1.104	0.143	1.064	0.127	-0.463	0.179	-0.464	0.170	0.155	0.056	0.151	0.054
15	1.819	0.381	1.559	0.252	0.382	0.130	0.335	0.131	0.189	0.046	0.173	0.046
16	1.264	0.209	1.218	0.180	0.173	0.168	0.141	0.160	0.181	0.053	0.172	0.052
17	2.083	0.294	2.098	0.255	-0.267	0.104	-0.224	0.096	0.143	0.042	0.153	0.041
18	1.534	0.217	1.361	0.160	-0.001	0.128	-0.040	0.122	0.142	0.045	0.127	0.042
19	1.769	0.273	1.693	0.235	-0.413	0.152	-0.381	0.148	0.209	0.058	0.214	0.058
20	3.143	0.544	2.570	0.304	0.164	0.071	0.153	0.070	0.130	0.027	0.127	0.028
21	1.627	0.297	1.487	0.225	0.362	0.136	0.318	0.129	0.189	0.046	0.177	0.045
22	2.008	0.263	2.020	0.251	-0.746	0.126	-0.664	0.123	0.160	0.054	0.176	0.055
23	1.104	0.150	1.122	0.141	-0.162	0.175	-0.168	0.160	0.156	0.053	0.151	0.051
24	1.230	0.150	1.282	0.149	-0.298	0.148	-0.274	0.140	0.137	0.049	0.137	0.049
25	1.164	0.173	1.059	0.135	-0.178	0.176	-0.199	0.169	0.159	0.054	0.150	0.052
26	1.650	0.498	1.545	0.405	2.161	0.245	2.072	0.255	0.130	0.021	0.134	0.022
27	0.802	0.192	0.827	0.174	0.754	0.296	0.689	0.263	0.204	0.066	0.201	0.063
28	1.643	0.273	1.546	0.227	-0.485	0.165	-0.383	0.167	0.192	0.059	0.222	0.063
29	1.712	0.241	1.502	0.166	-0.170	0.113	-0.161	0.110	0.121	0.040	0.121	0.040

30	1.697	0.321	1.368	0.182	-0.682	0.174	-0.610	0.179	0.178	0.061	0.197	0.065
31	1.466	0.278	1.293	0.208	0.651	0.126	0.711	0.131	0.108	0.035	0.137	0.039
32	1.230	0.404	1.077	0.248	1.718	0.240	1.445	0.198	0.184	0.041	0.155	0.042
33	3.304	0.684	3.736	0.649	0.259	0.084	0.341	0.066	0.127	0.028	0.180	0.027
34	2.101	0.379	2.706	0.482	0.531	0.106	0.598	0.081	0.162	0.032	0.222	0.029
35	1.712	0.268	2.114	0.360	0.166	0.130	0.348	0.105	0.154	0.042	0.233	0.040
36	1.062	0.388	1.321	0.444	3.130	0.475	2.691	0.490	0.154	0.028	0.168	0.025
37	2.448	0.454	2.732	0.420	0.504	0.094	0.530	0.075	0.149	0.029	0.191	0.027
38	1.794	0.295	2.153	0.342	0.156	0.040	0.359	0.098	0.156	0.040	0.205	0.037
Mean	1.662	0.305	1.582	0.239	-0.021	0.172	-0.018	0.167	0.168	0.049	0.174	0.049
SD	0.641	0.157	0.633	0.117	0.918	0.087	0.855	0.088	0.034	0.014	0.034	0.014

Item difficulty (b) estimates obtained through IRT are almost perfectly correlated with those from TRT (r=0.995). Item discrimination (a) estimates are highly correlated (r=0.882). Agreements on person proficiencies (θ) are also high (r=0.994). In contrast, there is less agreement for guessing estimates (r=0.756). These results are consistent with those from Wainer et al. (2000). Namely, it is relatively easy to obtain accurate estimates on difficulty parameters, but it is less easy to obtain accurate estimates on discrimination parameters, and guessing parameters are hard to estimate. It seems that the correlations of errors for item difficulty (b), guessing (c) and proficiency (θ) estimates are considerably high. Correlation between errors of item discrimination is lower (0.680). Comparison between the magnitudes of errors shows that errors resulting from TRT are significantly greater than those from IRT for ability parameters ($t_{(826)}$= 58.66, p<0.001), discrimination parameter ($t_{(37)}$= 3.44, p=0.001) and difficulty parameter ($t_{(37)}$= 3.05, p=0.004), but there is no significant difference in errors of guessing parameter ($t_{(37)}$= 0.33, p=0.743). The differences between the estimates from the IRT and TRT models indicate their impact on parameter estimation.

Table 5.11: Correlations of Means and Correlations of Standard Deviations of Posterior Distributions between 3-PL IRT and 3-PL TRT

	α	β	ω	θ
Means	.882	.995	.756	.994
Standard deviations	.680	.980	.982	.935

Figure 5.35: Scatter Plots of Parameter Estimates 3-PL IRT vs. 3-PL TRT

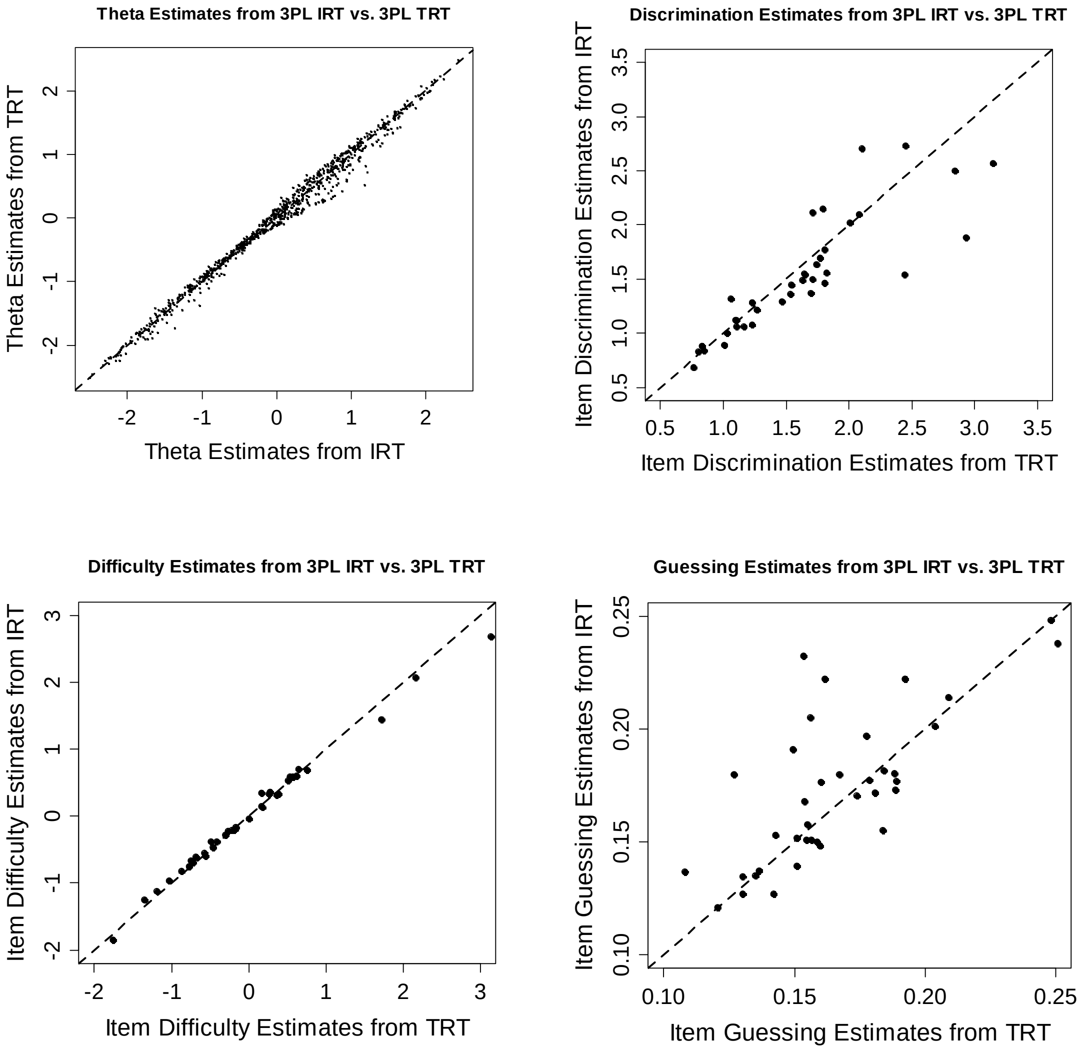


Table 5.12 represents the variances of the testlet effect variable in each testlet. The magnitudes of the testlet effect range from small to moderate. For example, testlet 1, 4, 5 have small variances of testlet effect variable. There were substantial effects for testlet 2, 7 and 8. All estimates of the variances of testlet effects have acceptable standard errors. By comparing the estimates of σ_{γ}^{2} with the mean of Q_3, we may notice that high σ_{γ}^{2} statistics are associated with high Q_3, while low σ_{γ}^{2} tend to be associated with low Q_3.

Table 5.12: Means and Standard Deviations of Posterior Distributions of Testlet Effect Variances and Correction Ratios

testlet	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Number of items	6	3	3	6	6	5	3	6
$\sigma_{\gamma d(j)}{}^{2}$	.25	.71	.48	.19	.19	.38	.65	.61
SE	.06	.19	.12	.05	.04	.09	.19	.11
Mean of Q_3	.0252	.0824	.0852	.0023	.0063	.0304	.0401	.1162
Ratio*	1.1633	1.0683	1.0683	1.1633	1.1633	1.1325	1.0683	1.1633

Note: Ratio is the G-theory correction ratio

5.2.5. Information Correction

Based on the analysis above, this test is characterized by short testlets and LID that ranges from moderate to large. Testlet lengths are unbalanced to an intermediate extent across the test. 3-PL IRT and TRT models are used for calibration. Both the test characteristics and the Q_3 pattern are close to those of simulation condition 32 or 33. By referring to Figure 5.28, for the condition with moderate or large LID ($\alpha_{\gamma}^{2} = .25$ or $\alpha_{\gamma}^{2} = 1$) and 4 or 6 items in each testlet, the mean SEID statistic values are close to zero. It implies that on average IRT SEE can be adjusted to be very close to TRT SEE. By referring to condition 32 and 33 of Figure 5.25, the adjusted IRT SEE and TRT SEE are almost overlapping when θ values range from -1.5 to 1.5 on the ability scale, while IRT SEE seems to have been overadjusted as compared against the TRT SEE for extreme θ values on the ability scale. Thus, it is speculated that IRT SEE could be corrected to a level that is close to TRT SEE in the test of this example.

The partition of variances is shown as follows in table 5.13. The correction ratios are listed in table 5.12. The correction ratios that are specific to each testlet are calculated as a function of the testlet length. The conditional standard errors of proficiency estimates from 3-PL-IRT, 3-PL-TRT, as well as those from 3-PL IRT adjusted by the correction ratios are plotted in Figure 5.36.

It is noticed that IRT SEEs have been increased as a result of adjustment. The mean SEID is -0.0459 which indicates that on the average, the adjusted IRT SEE is 5% higher than the targeted TRT SEE based on IRT SEE. It suggests a satisfactory adjustment effect in general compared with the mean SEID values in the simulation study. However, by referring to the conditional SEE plots (Figure 5.36), the IRT SEEs conditional on the TRT ability estimates between -1 and 1 seem to be under-adjusted compared with TRT SEE, but the magnitude of this underadjustment is very small and no more than 0.05. In contrast, the IRT SEEs conditional on TRT ability estimates beyond either -2 or 2 seem to be overadjusted but the two extreme ends of the ability scale include less than 2% of the examinees in this test. By looking at the SEID statistics conditional on the ability estimates (Figure 5.37), the majority of the SEID values range between 0 to 0.3 for ability estimates between -1.5 to 1.5.

In condition 33 of the simulation study, the IRT SEEs conditional on θ values between -1 and 1 have been adjusted to be overlapping to TRT SEEs; while in this real data example, the IRT SEEs conditional on this part of ability scale are a little bit lower than the TRT SEEs. One possible explanation for this difference is that all testlets in the simulation study are generated to have the same magnitude of the variance of testlet effect, but LID in this real example analysis is unequal across testlets. It is possible that the correction ratio should not only be a function of the testlet length but also be made dependent on the error variances specific to each testlet.

Table 5.13: Mean Squares, Variance of the Random Variance Components from G-theory Analysis

Source of variation	df	MS	σ^2
Examinees	826	1.7395	0.0396
Testlets x Examinees	6616	0.2282	0.0128
(Items: Testlets) x Examinees	24810	0.1682	0.1682

Note: $r = \sum_d \frac{k_d^2}{n}$ =5.1579 $t = \frac{n-r}{m-1}$ =4.6917

Figure 5.36: Standard Errors of Ability Estimates

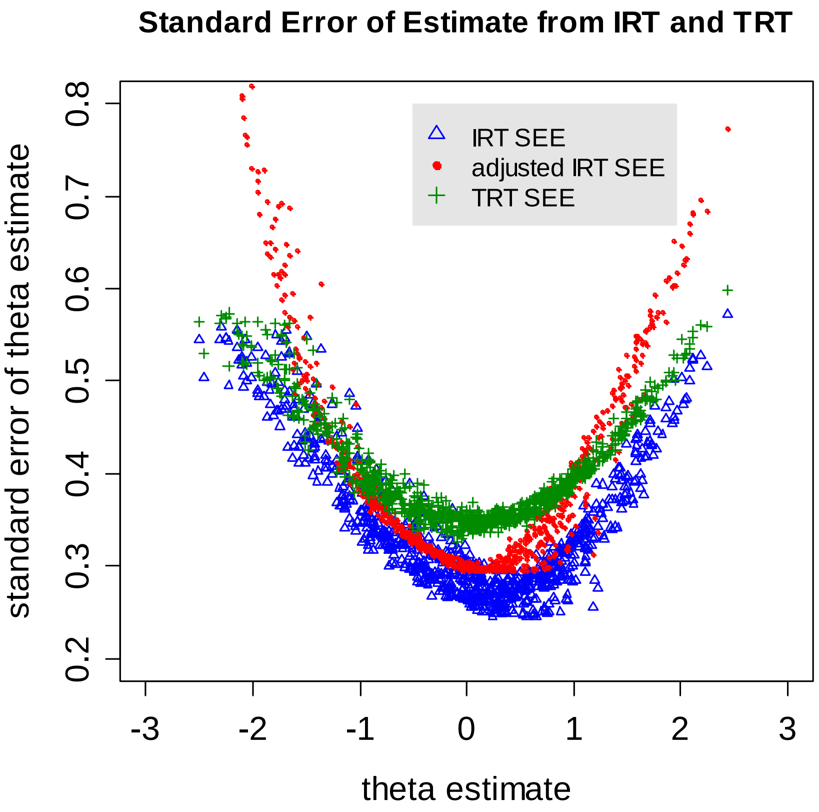
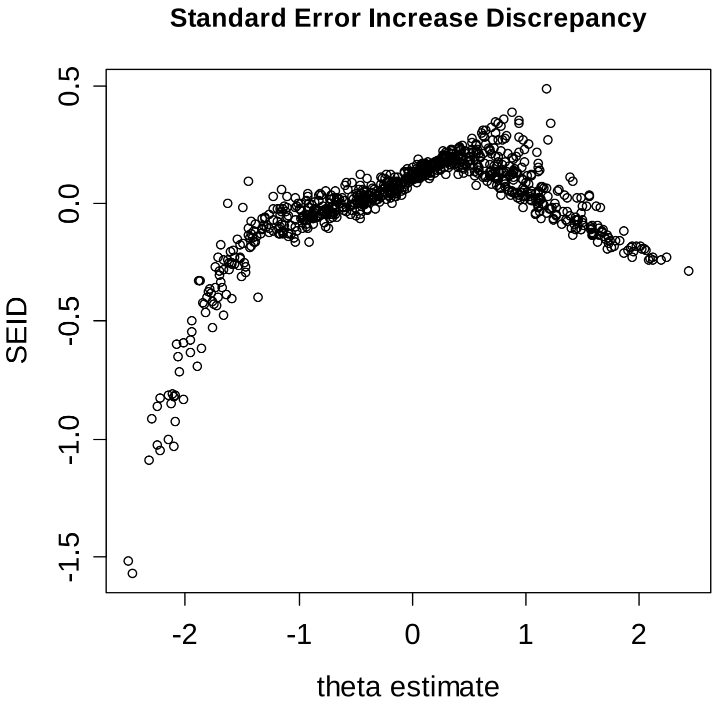


Figure 5.37: Standard Error Increase Discrepancy Conditional on TRT Ability Estimates



Chapter VI: CONCLUSIONS AND DISCUSSION

Previous research has repeatedly shown that the measurement error for the examinee proficiency parameter is often underestimated when a unidimensional conditional-independence IRT model is specified for a testlet dataset. The information correction method makes it possible to adjust the underestimated measurement error from the IRT models to a more appropriate level by using a design effect ratio of error variances derived from the generalizability analysis. The precision of proficiency estimates is crucial for proficiency classification and CAT scoring when the scores have consequences for the individual examinees.

The work conducted in this research extends the information correction of multiple ratings proposed by Bock, Brennan and Muraki (2002) and demonstrates how GT and IRT could be implemented sequentially to obtain more accurate precision estimates in a testlet-based test. The simulation study is designed to examine the performance of the information correction method in the 1-PL, 2-PL and 3-PL IRT contexts with the varying magnitude of LID, testlet length and balance of testlet length. Through a real data analysis, the measurement errors yielded by the information correction method are compared to those from the TRT model.

6.1. Summary of Findings

LID in a testlet dataset will lead to overlapping information, which in turn will result in higher error variances of the proficiency estimates if LID is appropriately accounted for by the testlet model. In comparison, the mean of the item scores of an examinee provided by GT is a linear unbiased estimator of the ability of that individual in the item-score, rather than the IRT proficiency, metric, and the error variance is composed of random error variances of the facets and their interactions including the variance of the person-testlet interaction term. The variance of the ability parameter from the IRT model corresponds to the random error variance of estimates from an independent item design in GT, while the variance of the primary ability parameter from the TRT model corresponds to the random error variance of estimates from a testlet design in GT. Therefore, the error variance ratio from GT models of the independent item design to the testlet design could be used to correct the error variance of proficiency estimates from IRT. That is, the correction ratio is easier to calculate in the item score metric using GT, and this ratio is applied in the IRT metric. As TRT provides relatively more accurate estimates when it is specified to a testlet dataset, comparison could be made between the expected TRT error variances and the adjusted IRT error variances.

In the simulation study, the variance of testlet effect variable, the testlet length and the balance of testlet length are manipulated to evaluate the performance of the information correction method. The data are generated based on 1-PL, 2-PL and 3-PL TRT models respectively, and are calibrated by their own true models as well as the IRT models with the same number of item parameters using the Bayesian MCMC method. In regard to the ability parameter recovery from IRT versus TRT in each simulation condition, it is shown that in the case of independent item test ($\sigma_{\gamma}^{2} = 0$), the biases from the TRT calibration are generally higher than those from the IRT calibration, which indicates the overparameterization of TRT models. However, in the responses simulated with significant LID, the biases from the TRT calibration are lower than those from the IRT calibration, which indicates that TRT might have a better model fit than IRT in this situation. It seems that the increase in LID will lead to the increase in biases for both IRT and TRT models. Long testlets seem to have higher biases than short testlets, especially in the conditions where the magnitude of LID is large. RMSE demonstrates a similar pattern as in bias. IRT provides more accurate estimates for tests with no LID, while TRT offers more accurate estimates for tests with LID; accuracy decreases with the increase in LID and long-testlet tests tend to result in less accurate estimates than short-testlet tests. The empirical SEEs are not much different between IRT and TRT calibrations. However, the mean theoretical SEEs from TRT calibration increases as LID goes up. In addition, the short-testlet test seems to result in lower mean theoretical SEE than the long-testlet test.

To evaluate the performance of the information correction method, a criterion variable—standard error increase discrepancy (SEID) — is chosen to quantify the discrepancy between the percent of increase in the adjusted IRT SEE and the percent of increase in TRT SEE with IRT SEE as the baseline. A SEID value close to zero indicates sufficient adjustment in standard error and hence a good performance of the information correction method. According to the conditional SEE plots in each condition, the discrepancy between SEE_{IRT} and SEE_{TRT} becomes larger when LID increases or the testlet length decreases. However, SEE_{IRT} can always be adjusted to the value that is close to SEE_{TRT} by using the information correction ratio, which suggests that the information correction method is effective for this purpose.

The conditional standard error plots further show that when LID is zero, SEE_{IRT-t_d} is lower than the targeted SEE_{TRT} across the ability scale, but when LID is substantial, SEE_{IRT-t_d} is very close to SEE_{TRT} conditional on θ values in the middle part of the scale. SEE_{IRT-t_d} conditional on extreme values on θ scale tends to be higher than SEE_{TRT}, which suggests that the information correction presents satisfactory performance for the conditions with substantial testlet effects but overcorrection may occur to SEE conditional on extreme ability values. The unbalanced conditions seem to result in better correction than the balanced conditions. 3-PL models have better information correction performance than 2-PL models, which in turn result in better correction than 1-PL models. However, it is not clear whether the effect of the type of the IRT models on the performance of the information correction is generalizable, because each dataset in this simulation study is calibrated by using the same model with which that dataset is generated. Unless we have a simulation design where an IRT model is specified to a dataset generated with a different IRT model, we will not be able to know whether the results about the IRT models are systematic.

The mean SEID (i.e., the SEID statistics averaged across all examinees) represents the effect of information correction in general for a test dataset. The mean SEID appears to be close to zero when LID is moderate ($\sigma_{\gamma}^{2} = 0.25$), but comparatively deviated from zero when LID is zero ($\sigma_{\gamma}^{2} = 0$) or large ($\sigma_{\gamma}^{2} = 1$). This implies that the information correction method might perform best in the conditions with moderate LID. However, the conditional SEE plots do not show better adjustment for moderate LID conditions than their large LID counterparts. Mean SEID values that are more deviated from zero on conditions with large LID may be attributed to overcorrection of SEE_{IRT} on extreme θ values. With only a few exceptions, the mean SEID statistic also suggests that the adjustment effect improves as the degree of the unbalance of the testlet length increases, which is consistent with what has been observed from the conditional SEE plots. Based on both mean SEID and conditional SEE plots, the long-testlet test tends to result in a better overall adjustment than the short-testlet test.

The ANOVA results indicate that the three factors being manipulated (e.g., LID, testlet length and the balance of the testlet length) and their interactions in this study account for more than 99% of the total variance in the dependent variable. Based on the p-values of the F tests, all of the three factors and their interactions are statistically significant, which implies significant effects on the adjustment of random errors using the information correction method. Balance of the testlet length, LID and the interaction between these two terms explain a large proportion of the total variance.

It is shown in the real data analysis that the average within-testlet Q_3 statistics has more positive values compared to the expected values of Q_3, suggesting that CI is violated in this test. The SD of the observed Q_3 statistics indicates that for the five out of the eight testlets the magnitude of the differences between the observed values and the expected value of Q_3 are approximately one SD or even larger. By comparing both means and t-scores of the Q_3 statistics between the real test example and the simulated datasets, it is found that the Q_3 pattern of the real test data is somewhere between those of condition 32 and 33.

Through the exploratory factor analysis, the first component accounts for over 40% of the overall variance in the dependent variable. The largest eigenvalue is about 6 times as large as the second largest. According to the scree plot, the eigenvalue of the first component is significantly larger than those of all the other components. From the table of factor loadings, almost all items have higher loadings on the first factor than they do on other factors. All these suggest that one factor is dominant in this dataset. As for the model fitting, 3-PL TRT model is preferred among the four (i.e., 2-PL IRT, 2-PL TRT, 3-PL IRT and 3-PL TRT) since it has the smallest DIC, which not only suggests that LID in this test is significant enough, but also justifies the need to model the pseudo guessing parameter. As to the result of parameter calibration using 3-PL IRT and 3-PL TRT, the correlation statistics shows that the two sets of item parameter estimates are highly correlated. The estimates of the variances of the testlet effect variable indicate that the magnitude of LID ranges from small to moderate values in this test.

The test of this real data example is characterized by short testlets, LID from moderate to large magnitude and testlet lengths that are unbalanced to an intermediate extent. 3-PL IRT and TRT models are used for calibration. These characteristics match those of simulation conditions where both the mean SEID and the conditional SEID statistics indicate the satisfactory adjustment of the error variances. In this real data example, the mean SEID that is close to zero suggests good adjustment effect in general. However, the IRT SEEs conditional on ability estimates between -1 and 1 seem to be under-adjusted compared with TRT SEE, but the magnitude of this underadjustment is very small. In contrast, the IRT SEEs conditional on ability estimates beyond either -2 or 2 seem to be overadjusted but this part of scale covers less than 2% of the examinees in this test. LID in this real example analysis is unequal across testlets. It is possible that the correction ratio should not only be a function of the testlet length but also be made dependent on the error variances specific to each testlet.

6.2. Implications for Testing Practices

Testlets have gained increasing popularity in recent years in that they can save time and cost in test development, and often require the integration of the knowledge and skills which cannot be represented in simple independent multiple-choice items. This research addresses the problem that the conditional-independent IRT models do not account for LID in testlets, which would lead to an underestimation of the measurement error, substantial on some occasions. This issue can be critical for scoring in high-stake tests or for classification in proficiency. In complex computerized simulations, the precision estimates will affect the IRT evidence accumulation process. Therefore, it is necessary to have a relatively accurate estimate of the measurement precision and quantify the local dependency of testlets.

Although some testlet models have demonstrated satisfactory performance in terms of model-data fit and parameter recovery, each of them has limitations. GT model has not been sufficiently developed to connect continuous latent values with discrete scores. Models of the IRT approach such as the bi-factor model, the multilevel model and the TRT models are complex and usually take long time to converge because of the ways they are currently estimated.

In this dissertation, it is shown how GT and IRT could be used sequentially to correct the measurement precision. The information correction method is efficient and straightforward as it is easy to derive the error variances of person parameters in either the testlet design or the independent item design from the GT analysis, as well as the precision estimates from IRT models. Given the corresponding relationship in error variance ratios between the generalizability models and response theory models, it should be reasonable to apply the information correction term to testing practices.

The simulation study provides evidence that the underestimated measurement errors from IRT calibration could be adjusted to the appropriate level through the information correction method despite the varying LID, testlet length, balance of testlet length and number of the item parameters in the model. The expected values of error variances from the TRT calibration can be assumed as the benchmark because TRT models account for LID and thus can produce more accurate estimates about the testlet datasets. Given the robustness of variance ratios, estimation of the information correction should be adequate for practical work.

In addition to demonstrating the adequate performance of the information correction method, results from ANOVA show the impact from the three factors (e.g., LID, testlet length and the balance of the testlet length) on information correction. All the three factors and their interactions present as statistically significant and account for more than 99% of the total variance in the dependent variable. The Balance of the testlet length, LID and the interaction between these two explain a large proportion of the total variance. In other words, the information correction method is more effective for tests with certain characteristics. The information correction method seems to perform better on long-testlet tests than short-testlet tests, better on tests with LID than the tests without LID, and better on tests with unbalanced testlet lengths than tests with balanced testlet length. The 3-PL model context seems to have more satisfactory adjustment results than the 2-PL context which in turn has better adjustment effect than the 1-PL context. Although the results of the significance tests rely upon the dependent variable that has been selected, this analysis has roughly depicted a picture about how the information correction method performs in each situation.

The real data example has provided more details about the information correction procedure. By comparing the real test and the simulated tests, it is shown how close the error variance from a real data example can be adjusted to the results we would expect. In addition, it allows an investigation into how the correction coefficients work on the measurement error of each examinee’s ability parameter when the calibration of item parameters is involved. It is noteworthy that diagnosis tests are necessary to detect LID and dimensionality so as to ensure if the correction procedure is applicable.

6.3. Limitations and Directions for the Future Research

Presented in this dissertation are the initial studies on the information correction methods, so it is beset with certain limitations. First, as shown in both the simulation study and the real data analysis, the standard errors of proficiency estimates given extreme proficiency values have been overcorrected in many conditions, while the standard errors of proficiency estimates given the proficiency values in the middle are undercorrected in some of the conditions. Because the correction ratio term is a function of the testlet length and is applied to all examinees across the ability scale, the standard errors that are already high conditional on the extreme ability values will be magnified with the multiplicative coefficient. In the future studies we may cut the ability scale into intervals and estimate correction ratio for each interval. Alternatively, we may also build a correction term as a function of not only the testlet length but also ability values for the future study.

In order to focus on the change in SEE in ability estimation, the estimation procedure was simplified by fixing all item parameter values in calibration. Previous research in the literature cited by in Chapter 2 demonstrated that under this condition, SEE for abilities is underestimated when the testlet structure is ignored, but bias is negligible. However, in the full hierarchical Bayesian framework, the estimation of the item parameters can affect the estimation of ability parameters. It will be of interest to examine the bias and the standard errors of the item parameter estimates as well as their influences on the performance of the information correction method. In addition, it is worthwhile to consider about correcting the measurement error of the item parameters if overestimation also happens to the precision of the item estimates. Therefore, we may want to estimate the item parameters in the follow-up simulation studies.

Another limitation of the simulation study is that in each condition, the correct model was used to estimate ability parameters. For example, the 3-PL TRT model was used to estimate the dataset generated with 3-PL TRT. By this means, we are able to find out whether the type of IRT models have systematic effect on the performance of the information correction method. A future topic for research can be the consequence of fitting misspecified TRT models to data.

LID is on a continuous scale rather than the categorical data points at 0, .25 and 1 specified in the simulation study. Conclusions can be made from the ability parameter recovery that IRT models fit the data with zero LID better than TRT models while TRT models present better model fits for datasets generated with $\sigma_{\gamma_d}^2 = .25$. However, it is not clear from the cases in this study which model is better for datasets generated with LID at any point between 0 and .25. Values between 0 and .25 along the LID scale would need to be examined to improve the decision-making about which model should be used for ability estimation from a testlet dataset.

The response datasets were generated with equal LID for each testlet in the simulation study, but in the real test LID often varies across the testlets. It seems that the correction results from the real data analysis are somewhat different from what was expected based on the simulation study. Thus, we may consider the simulation conditions of unequal LID in the future. The mixed test format that consists of both independent items and testlets is not discussed in either simulation or the real data example. However, the information correction term is applied at the testlet level rather than the test level so the approach is directly applicable to this situation. In addition, situations in which LID is zero are examined as part of the simulation, the results of which naturally apply to the independent items in the mixed format. The mixed format condition would be studied with an implementation similar to that of the "unequal LID across testlets" condition of this study.

The expected value of TRT random error was treated as the benchmark in this study to evaluate the adjusted standard error, because TRT models provide better parameter recovery than IRT models based on the simulation studies. However, it is also true that on some occasions when LID is zero in the response matrix, TRT models provide less accurate estimates than IRT. Therefore, it is inevitable that the estimation of TRT parameters is confounded with the correction results when the discrepancy between the TRT standard error and the adjusted standard error is regarded as a criterion variable. It is worth further investigation for a research design in which effects from the confounded variables could be cleared.

REFERENCES

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

Bock, R. D., Brennan, R. L., & Muraki, E. (2002). The information in multiple ratings. *Applied Psychological Measurement, 26*(4), 364-375.

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

Brennan, R. L. (1992). *Elements of generalizability theory (revised edition).* Iowa City, IA: ACT publications.

Brennan, R. L. (1997). A perspective on the history of generalizability theory. *Educational Researcher, 16*, 14-20.

Brennan, R.L. (2001). *Generalizability theory*. New York, NY: Springer.

Briggs, D., & Wilson, M. (2007). Generalizability in item response modeling. Journal of Educational Measurement, 44(2), 131-155.

Brooks, S.P. and Gelman A. (1998): Alternative Methods for Monitoring Convergence of Iterative Simulations. Journal of Computational and Graphical Statistics, 7, 434-455.

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

Congdon, P. (2003). *Applied Bayesian modeling.* New York: Wiley

Cornfield, J. 1951. Modern methods in the sampling of human populations. *American Journal of Public Health*, 41, 654-661.

Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational Measurement: Issues and Practice, 57*, 373-399.

Damien, P., Wakefield, J.C., & Walker, S.G. (1999). Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables. *Journal of the Royal Statistical Society, B Statistical Methodology*, 61, 331-344.

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

Douglas, J., Kim, H. R., Habing, B., & Gao, F. (1998). Investigating local dependence with conditional covariance functions. *Journal of Educational and Behavioral Statistics*, 23, 129-151.

Du, Z. (1998). Modeling conditional item dependencies with a three-parameter logistic testlet model (pp. 98): Columbia University.

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

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

Gelfand, A.E., & Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 398-409.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bifactor analysis. *Psychometrika, 57,* 423–436.

Gifford, J.A., & Swaminathan, H. (1990). Bias and the effect of priors in Bayesian estimation of parameters of item response models. *Applied Psychological*

Measurement, 14 (1), 33-43.

Gilks, W.R. (1992). Derivative-free adaptive rejection sampling for Gibbs sampling. In J.M. Bernardo, J.O. Berger, A.P. Dawid, & A.F.M. Smith (Eds.), *Bayesian statistics 4* (pp. 169-194). Oxford, UK: Clarendon Press.

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

Goldstein, H. (1995). *Multilevel statistical models* (2nd ed.). London, Edward Arnold.

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Habing, B., & Roussos, L. A. (2003). On the need for negative local item dependence. *Psychometrika*, 68, 435-451.

Hambleton, R. K., & Swaminathan, H. (1985). *Item Response Theory*. Boston: Kluwer.Nijhoff Publishing.

Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 54, 93-108.

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

Jiao, H., & Wang, S. (2008). Comparison of estimation methods of one-parameter testlet models. Paper presented at the 2008 National Council on Measurement in Education Annual Meeting in New York, New York.

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

Kingston, N.M., & Dorans, N.J. (1982). The feasibility of using item response theory as a psychometric model for the GRE Aptitude Test (ETS Research Report 82-12).

Princeton NJ: Educational Testing Service.

Koretz, D., Stecher, B., Klein, S., & McCaffrey, D. (1994). "The Vermont Portfolio Assessment Program: Findings and Implications." *Educational Measurement: Issues and Practice,* 13 (3), 5-16.

Lee, G., & Frisbie, D. A. (1999). Estimating reliability under a generalizability theory model for test scores composed of testlets. *Applied Measurement in Education, 12*(3), 237-255.

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*(4), 357-372.

Li, F., Cohen, A. S., Kim, S.-H., & Cho, S.-J. (in press). Model selection methods for mixture dichotomous IRT models. *Applied Psychological Measurement.*

Li, Y., Bolt, D. M., & Fu, J. (2005). A test characteristic curve linking method for the testlet model. *Applied Psychological Measurement, 29*(5), 340-356.

Li, F., & Rijmen, F. (2009). *A Vertical Linking Design for Periodic Assessments and Tests that Consist of Situated Tasks.* Paper presented at the 2008 National Council on Measurement in Education Annual Meeting in San Diego, California.

Lindley, D.V., & Smith, A.F.M. (1972). Bayes estimates for the linear model. Journal of the Royal Statistical Society, 34, (Series B), 1-41.

Lord, F.M. (1952). A theory of test scores. Psychometric Monographs, No. 7.

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

McDonald, R. P. (1994). Testing for approximated dimensionality. In D. Laveault, B.D. Zumbo, M.E. Gessaroli, & M.W. Boss (Eds.), *Modern theories in measurement: Problems and issues* (pp. 31-61). Ottawa, Canada: University of

Ottawa, Edumetrics Research Group.

Messick. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher,* 23(2), 13-23.

Mislevy (1986). Bayes modal estimation in item response models. *Psychometrika*, 51(2), 177-195.

Mislevy, R.J., Beaton, A., Kaplan, B.A., & Sheehan, K. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, 29(2), 133-161.

Novick, M.R., & Jackson, P.H. (1974). *Statistical methods for educational and psychological research*. New York: McGraw-Hill.

Patz, R.J., Junker, B.W., Johnson, M.S., & Mariano, L.T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics*, 27(4), 341-384

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

Reddy, S. K. (1992) Effects of ignoring correlated measurement error in structural equation models. *Educational and Psychological Measurement, 52,* 549-570.

Ritter, C., & Tanner, M.A. (1992). Facilitating the Gibbs sampler: The Gibbs stopper and the Griddy-Gibbs sampler. *Journal of the American Statistical Association*, 87, 861-868.

Rosenbaum, P. R. (1988). Items bundles. *Psychometrika, 53*(3), 349-359.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. (*Psychometrika Monograph No. 17*) Richmond, VA: Psychometric Society.

Sinharay, S. (2003). *Assessing convergence of the Markov Chain Monte Carlo Algorithms: A review* (ETS RR-03-07). Princeton, NJ: Educational Testing Service.

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

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society*, Series B, 64, 583-616.

Spiegelhalter, D., Thomas, A., & Best, N. (2003). *WingBUGS* (Version 1.4) [Computer program]. Cambridge, UK: MRC Biostatistics Unit, Institute of Public Health.

Sternberg, R. J. (1977). *Information processing and analogical reasoning: The componential analysis of human abilities*. Hillsdale, NJ: Erlbaum.

Stone, C. A., & Yeh, C-C. (2006). Assessing the dimensionality and factor structure of multiple-choice exams: An empirical comparison of methods using the Multistate Bar Examination. *Educational and Psychological Measurement, 66,* 193-214.

Swamininathan, H., & Gifford J.A. (1982). Bayesian estimation in the Rasch model. *Journal of Educational Statistics*, 7, 175-191.

Swamininathan, H., & Gifford J.A. (1985). Bayesian estimation in the two-parameter logistic model. *Psychometrika*, 50, 349-364.

Swamininathan, H., & Gifford J.A. (1986). Bayesian estimation in the three-parameter logistic model. *Psychometrika*, 51, 589-601.

Tanner, M.A., & Wong, W.H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82, 528-550.

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

Verhelst, N., & Verstralen, H. (2001). An IRT model for multiple raters. In A. Boomsma, M. van Duijin, & T. Snijders (Eds.), *Essays on item response theory* (pp.88-108). New York: Springer-Verlag

Wainer, H. (1995). Precision & differential item functioning on a testlet-based test: The 1991 law school admissions test as an example. *Applied Measrement in Education, 8*(2), 157-187.

Wainer, H., Bradlow, E. T., & Du, Z. (2000). Testlet response theory: An analog for the 3-PL model useful in testlet-based adaptive testing. In W. J. v. d. Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 245-270). Boston, MA: Kluwer-Nijhoff.

Wainer, H., Bradlow, E.T., & Wang, X. (2007). *Testlet response theory.*

Wainer, H., Brown, L.M., Bradlow, E.T., Wang, X., Skorupski, W.P., Boulet, J., & Mislevy, R.J. (2006). An application of testlet response theory in the scoring of a complex certification exam D.M. Williamson, R.J. Mislevy, & I.I. Bejar (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp169-200). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

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

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

Wang, X., Bradlow, E. T., & Wainer, H. (2002). A general bayesian model for testlets:

Theory and applications. *Applied Psychological Measurement, 26*(1), 109-128.

Wang, W., & Wilson, M. (2005a). Exploring local item dependence using a random-effects facet model. *Applied Psychological Measurement, 29*(4), 296-318.

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

Wilson, D. T., Wood, R., & Gibbons, R. (1991). *TESTFACT: Test scoring, item statistics, and item factor analysis.* Chicago: Scientific Software International.

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

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

Yen, W.M., & Fitzpatrick, A.R. (2006). Item response theory. In R.L. Brennan (Ed.), *Educational measurement* (4th ed.; pp111-154). Westport, CT: American Council on Education and Praeger Publishers.

Zeller, A. (1971). *An introduction to Bayesian inference in econometrics*, New York: John Wiley & Sons.