# ABSTRACT

Title of Dissertation: DEVELOPING A COMMON SCALE FOR TESTLET MODEL PARAMETER ESTIMATES UNDER THE COMMON- ITEM NONEQUIVALENT GROUPS DESIGN

Dongyang Li, Doctor of Philosophy, 2009

Directed By: Professor Robert J. Mislevy
Department of Measurement, Statistics & Evaluation

An important advantage of item response theory (IRT) is increased flexibility for methods of test equating. Several methods of IRT scaling have been developed, but under the assumption of local independence of item responses, such as Haebara's linking procedure.  A recent development in IRT has been the introduction of Testlet Response Theory (TRT) models, in which local dependence among related sets of items is accounted for by the incorporation of "testlet effect" parameters in the model. This study extended Haebara's item characteristic curve scale linking method to the three-parameter logistic (3-PL) testlet model. Quadrature points and weights were used to approximate the estimated distribution of the testlet effect parameters so that the expected score of each item given $\theta$ can be computed and the scale linking parameters can be estimated. A simulation study was conducted to examine the performance of the proposed scale linking procedure by comparing it with the scale linking procedures that are based on the 3-PL IRT model and the graded response model. An operational data analysis was also performed to illustrate the application of the proposed scale linking method with real data.

# DEVELOPING A COMMON SCALE FOR TESTLET MODEL PARAMETER ESTIMATES UNDER THE COMMON- ITEM NONEQUIVALENT GROUPS DESIGN

By

Dongyang 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 Michel Wedel
Professor Robert W. Lissitz
Assistant professor Hong Jiao
Adjunct assistant professor Amy B. Hendrickson

# Acknowledgements

First and foremost, I wish to thank my advisor, Dr. Robert Mislevy, who with his immense knowledge guided me through this arduous journey. His patience and understanding made this dissertation possible.

I would also like to thank Dr. Robert Lissitz, Dr. Amy Hendrickson, Dr. Hong Jiao and Dr. Michel Wedel for being on my committee and providing insightful comments and advices. My thanks goes especially to Dr. Hendrickson, who got me interested in scaling and equating in the first place and inspired me to select this dissertation topic.

My thanks also go out to my colleagues at the Center for Applied Linguistics where I did my internship for the past one and a half years for their support.  I would especially like to thank Dr. David MacGregor and Dr. Dorry Kenyon for help clearing the way of using the WIDA data for my dissertation, and Dr. Carolyn Fidelman for allowing me to be flexible with my schedule as I was working on the dissertation.

I am also grateful to my parents Erxiao Li and Guiyun Huang for their unconditional and selfless support in my pursuit of the doctorate. I wish I could have done a better job fulfilling my filial duty.

Finally I wish to thank Qi Zhang for the constant encouragement along the way.

# Table of Contents

# List of Tables

# List of Figures

# Chapter 1  Introduction

The testlet is a popular type of item format that has been applied by a variety of tests and assessments. Different names have been used to describe the format. Some of the examples summarized by Haladyna (2004) include “interpretive exercises”, “scenarios”, “vignettes”, “item bundles”, “problem sets” and “super-items”. They all refer to the item format in which a stimulus is presented, followed by two or more questions that are related to the presented stimulus. Wainer & Kiely (1987) named this particular item format “testlet” and defined it as an aggregation of items on a single theme. In recent years, with advances in cognitive psychology and increasing emphasis on the improvement in the efficiency and validity of assessments, there is “a movement away from what is viewed as the atomistic nature of discrete multiple choice items toward the use of testlets that provides context.” (X. Wang, Bradlow, & Wainer, 2002, p. 125)

Researchers face special challenges when they try to link the scales of test forms that are composed of testlets. The unidimensional dichotomous IRT models such as the Rasch model, two-Parameter Logistic IRT (2-PL) model or three-Parameter Logistic IRT (3-PL) model may not work well with testlet-based tests. This is caused by a distinct characteristic of the testlet format: the items within a testlet usually demonstrate some degree of inter-dependence among themselves. This inter-dependence may occur because all the items within a testlet share a common stimulus such as a passage, a graph, a table or a diagram, etc. Another scenario that inter-dependence among items may occur within a testlet is that the items have to be solved in a step-wise fashion:  one item has to be solved before enough information can be gathered to solve the next item. These types of inter-dependence of items within a testlet are often considered a form of Local Item Dependence (LID) and are referred to as “testlet effect” in this study. Traditionally, researchers often ignore the testlet effect and treat the items as discrete and locally independent items. This may lead to biased model parameter estimates, underestimated standard error of measurement and inflated reliabilities. Such an approach to dealing with the testlet-based test forms may also cause inaccurate scaling and equating results.

In recent years, Testlet Response Theory (TRT) models (Bradlow, Wainer, & Wang, 1999; Wainer, Bradlow, & Du, 2000; Wainer, Bradlow, & Wang, 2007; Wainer & Wang, 2000) were developed to model the testlet effect. The models have generated a lot of interest among researchers and there have already been quite a few studies on the theories and applications of TRT models. However, few studies have been conducted on scale linking and test equating under the TRT framework. This limits more extensive applications of the TRT models. Most large scale testing programs need to perform some form of scale linking and equating procedures to put the item parameter estimates and test scores across test forms onto common scales. For some state assessment programs for basic skills, scale linking and equating are an integral part of test development and measurement so that horizontal and vertical comparisons of test scores over different test forms are possible. Without an applicable scale linking procedure that can be applied with the TRT models, it may not be justifiable to adopt such models in these testing programs.

The purpose of this dissertation is to develop a scale linking method for the TRT model parameter estimates when such models are used to calibrate testlet-based test forms. The dissertation follows this structure:

Chapter 2: Literature Review. The testlet format and the testlet effect are explained, followed by an introduction of the TRT theories and models. The scale linking is defined and several scale linking procedures such as Stocking & Lord method and Haebara method are described. It ends with a brief account of an extension of the Stocking & Lord method to the two parameter normal ogive TRT model proposed by Li, Bolt and Fu (2005) .

Chapter 3: Methodology and Simulation Study. The research questions are raised and the proposed TRT model scale linking method is explained in detail. A simulation study is performed to compare the performance of the proposed scale linking method with that of the 3-PL IRT model scale linking method and the graded response model scale linking method under different levels of the testlet effect. The results are presented and summarized.

Chapter 4: Real Data Analysis. The proposed TRT model scaling method is applied to link the scales of two test forms taken from the 2004-2005 ACCESS for ELLs® assessment to illustrate its application with operational data. The whole procedure, from testlet effect detection using the $Q_3$ index to the derivation of the scale linking parameter estimates, is explained.

Chapter 5: Conclusion and Discussion. This chapter summarizes the findings in the study and discusses their practical implications. Some caveats of the study are also presented and future research topics regarding this new scale linking method are proposed to address these issues.

# Chapter 2  Background and Literature Review

## *Testlet*

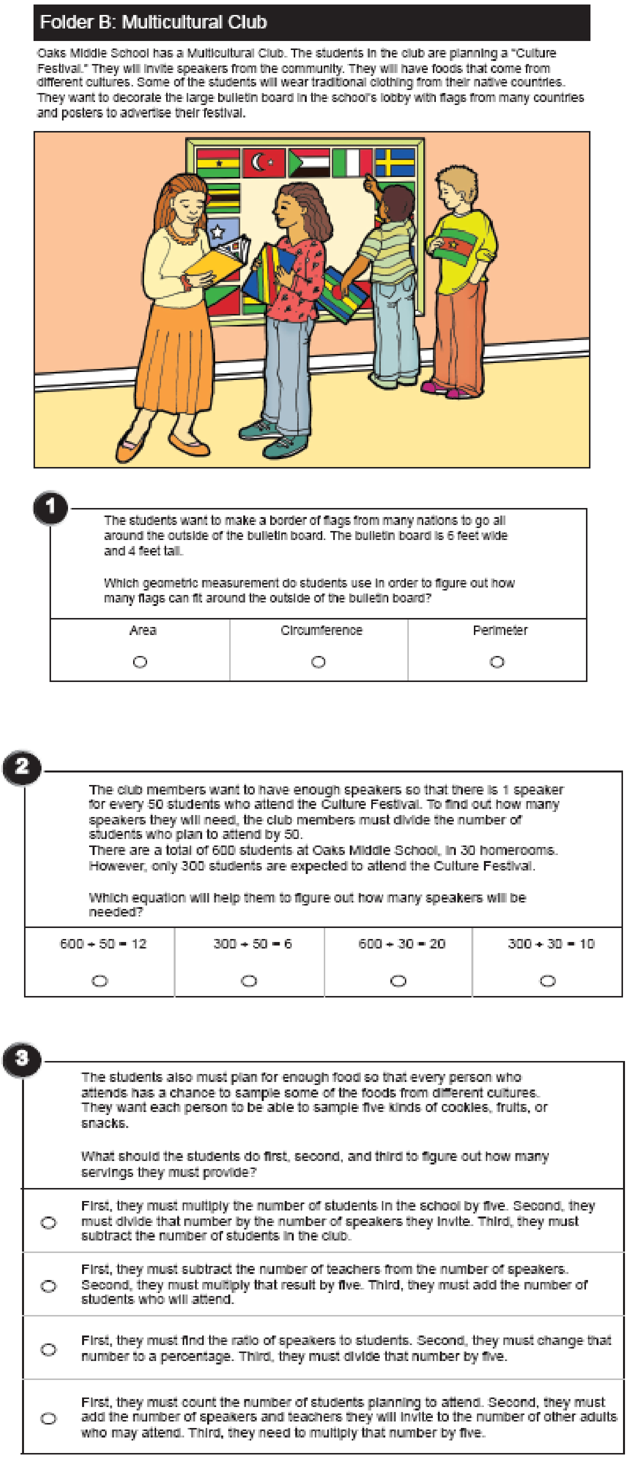
### *Testlet format*

Testlets are usually discussed as a special case of the use of multiple choice (MC) questions, and this will be the case addressed here. The MC items are considered “objective” items (Mehrens & Lehmann, 1984; Millman & Greene, 1989) since there is no raters’ bias involved when grading MC questions. Moreover, more MC questions can usually be administered than the constructed response (CR) questions given the same span of testing time. The objective scoring and greater numbers of test items usually lead to better reliabilities. Wainer and Thissen (1993) indicated that it is typically found that the reliability of the CR section is considerably less than that of a comparably timed MC section. Since more items can usually be covered using the MC items than using the CR items, the MC format sometimes enhances the validity of a test if the test targets a large content domain. One drawback of the MC format is that many researchers believe that MC is more suited for measuring recall level learning outcomes and is not optimal to elicit evidence about complex cognition (Bowman & Peng, 1972; Frederiksen, 1984; Morgenstern & Renner, 1984; Warren, 1979).

In 1989, Nickerson claimed that one of the new directions of assessments is that tests should assess thinking. He asserted that higher-order cognitive functioning should be a major goal in education and “the lack of adequate tools for assessing such functioning means that we are at a loss to judge the education enterprise as a whole" (Nickerson, 1989, p. 3). In recent years, cognitive psychology principles is increasingly prevalent in test design and item generation to improve construct validity (Embretson & Gorin, 2001) and studies have been conducted to incorporate cognitive information into test development (e.g. Mislevy, 1994; Mislevy, Steinberg, & Almond, 1999). These efforts greatly facilitate the assessment of higher order cognitive functioning. One example of such assessments is the classroom instruction and diagnostics assessment, by which higher cognitive activities such as test takers' problem solving schemes and their misconceptions are identified. The demand for assessing thinking has become so popular that the No Child Left Behind Act stipulates in (3)(C)(vi) that state assessments shall "involve multiple up-to-date measures of student academic achievement, including measures that assess higher-order thinking skills and understanding"(Congress, 2001). Since MC items are deemed to be more suitable for assessing low level cognitive activities, using stand alone MC items for such assessments sometimes may cause validity issues.

The testlet format can be considered as a bridge between the conventional MC format and the CR format. On the one hand, the testlet items retain the advantages of the MC format and can be efficiently administered and objectively scored. On the other hand, since a testlet incorporates several questions with the same stimulus, it reduces concerns about the atomistic nature of single independent small items (Wainer et al., 2000). Item writers have much more flexibility to test different aspects and stages of the cognitive activities using interrelated sets of items.

A sample testlet from the WIDA ACCESS® English language Assessment (WIDA, 2008) is presented in Figure 1. A paragraph describing "the planning of cultural festival" and a picture of several students standing in front of the bulletin board with national flags around it are presented in the stem of the testlet. The first item within the testlet asks the test takers to select a geometric measurement to decide on the number of flags that can be placed around the bulletin board. This targets 1) test takers' basic comprehension of the passage and the question with the help of the picture (with bulletin board and national flags on it) and 2) their recollection of the definitions and properties of the three geometric measurements. These are comparatively lower levels of cognition functioning. The second item is about finding the appropriate number of speakers. There is no graph to help test takers understand the question. Instead some irrelevant information is given (600 students and 30 rooms). To be able to answer the question, the test takers must 1) select information that is relevant to solve the problem, and 2) know how to apply the correct mathematics equation. The third question asks the test takers to decide on the number of snacks that need to be prepared based on the number of people attending the event. There is no mathematic equation in the response alternatives to visually aid test takers' comprehension of the question. To be able to answer this question, the test taker need to have a good understanding about how the number of attendees is decided and what equation is need to derive the number of snacks. This is a multiple step process that involves addition and multiplication. The three testlet items taps different levels and aspects of cognitive functioning, since the testlet format can carry with it more extensive and intensive context than the single-item MC format.



Folder B: Multicultural Club

Oaks Middle School has a Multicultural Club. The students in the club are planning a "Culture Festival." They will invite speakers from the community. They will have foods that come from different cultures. Some of the students will wear traditional clothing from their native countries. They want to decorate the large bulletin board in the school's lobby with flags from many countries and posters to advertise their festival.

1

The students want to make a border of flags from many nations to go all around the outside of the bulletin board. The bulletin board is 6 feet wide and 4 feet tall.

Which geometric measurement do students use in order to figure out how many flags can fit around the outside of the bulletin board?

| Area | Circumference | Perimeter |
|---|---|---|
| ○ | ○ | ○ |

2

The club members want to have enough speakers so that there is 1 speaker for every 50 students who attend the Culture Festival. To find out how many speakers they will need, the club members must divide the number of students who plan to attend by 50.
There are a total of 600 students at Oaks Middle School, in 30 homerooms. However, only 300 students are expected to attend the Culture Festival.

Which equation will help them to figure out how many speakers will be needed?

| 600 ÷ 50 = 12 | 300 ÷ 50 = 6 | 600 ÷ 30 = 20 | 300 ÷ 30 = 10 |
|---|---|---|---|
| ○ | ○ | ○ | ○ |

3

The students also must plan for enough food so that every person who attends has a chance to sample some of the foods from different cultures. They want each person to be able to sample five kinds of cookies, fruits, or snacks.

What should the students do first, second, and third to figure out how many servings they must provide?

○ First, they must multiply the number of students in the school by five. Second, they must divide that number by the number of speakers they invite. Third, they must subtract the number of students in the club.

○ First, they must subtract the number of teachers from the number of speakers. Second, they must multiply that result by five. Third, they must add the number of students who will attend.

○ First, they must find the ratio of speakers to students. Second, they must change that number to a percentage. Third, they must divide that number by five.

○ First, they must count the number of students planning to attend. Second, they must add the number of speakers and teachers they will invite to the number of other adults who may attend. Third, they need to multiply that number by five.

*Figure 1*. A sample testlet of Grade 6-8 Tier B Reading test from "ACCESS for ELLs Listening, Reading, Writing and Speaking Sample Items" (WIDA, 2008)

In the example, the third item also utilizes some of the information (about the “speakers”) in the second item. This is another feature that can be frequently observed among testlet items: the items not only share the common stimulus from the stem of the testlet, but they may also share information amongst themselves. Sometimes the shared information at the item level may be trivial like the one shown in the example, while in other occasions it may be crucial, and to answer an item correctly depends on drawing information from the previous item(s). Zenisky, Hambleton & Sireci (2001) pointed out that many real world tasks require solving related problems in stepwise fashion, thus including context-dependent testlet items in a test may improve construct validity.

***Local Item Dependence (LID)***

The responses to the items within a testlet usually exhibit higher degree of correlations among each other than the conventional MC items. Yen (1993) summarized two situations where the items within a testlet may be related:

> *“Passage dependence. If several items are attached to the same passage or setting, then LID can occur. This LID can be produced by a student's unusual level of interest or background knowledge about the passage or by the fact that information used to answer different items is interrelated in the passage.*
>
> *Item chaining. If items are organized in steps, then knowing the answer to one item increases the chances of a student's knowing the answer to the next one. While item chaining has long been an anathema to multiple-choice tests, it is*

*often seen as desirable in performance assessments because it models real life situations."* (Yen, 1993, p. 189)

In both situations, testlet items violate the local item independence assumption of the IRT models (Rosenbaum, 1988). IRT models assume that holding a person's latent trait constant, the probability function of the examinee's response pattern to a set of items is the product of the probabilities of his/her particular response to each of the items (Embretson & Yang, 2006). For example, for two items *i* and *j* in a test, the probability of answering both items correctly given the ability level $\theta$ is calculated as:

$$P(X_i = 1, X_j = 1 \mid \theta) = P(X_i = 1 \mid \theta) P(X_j = 1 \mid \theta) \qquad (2.1)$$

The estimation of IRT model parameters usually involves searching for constants that maximize the probability function of the joint response pattern. If the assumption of local item independence is violated, the constants derived through maximizing the product of the probabilities of each particular response to each item will generally not be the unbiased estimates that maximize the true probability function of the response pattern. As a result, when LID is present, the item and person parameter estimates using the IRT model may be inaccurate and biased. (Ackerman, 1987; Kingston & Dorans, 1984; Thissen, Steinberg, & Mooney, 1989; Wainer & Wang, 2000; Yen, 1980, 1993).

Items within a testlet may display stronger relationships among themselves and consequently these items may have higher correlations with the total score. Within the classical test theory framework, this is manifested as the stronger biserial correlations between the item scores and the total scores. Within the IRT framework, the testlet effect can produce higher item discriminations for the items that display LID (Masters, 1988).

The statistical dependence caused by the testlet format can also lead to the overestimation of test reliabilities if item-based methods are used to perform the estimation over tests composed of testlet. When estimating reliabilities of test scores using internal consistency indices such as Cronbach's *a* (Cronbach, 1951), the inter-correlations of the items within a testlet cause the score variances within testlets to be smaller than score variance between testlets, thus resulting in lower overall variance for the total scores. With the lower total score variance, the reliability statistics of the test scores are erroneously inflated. The positively biased estimates of reliability caused by the testlet format have been well studied by researchers. (Allen & Sudweeks, 2001; Feldt, 2002; Frisbie & Druva, 1986; Reese, 1999); Sireci, Thissen & Wainer (1991) examined the concept of the inflation of reliabilities within the IRT framework. Since the measurement precision is a function of the person parameter $\theta$ in IRT, there is no single overall reliability index for different test scores. Sireci et al. obtained the marginal reliability by integrating the measurement error variances of different proficiency levels over their distribution. Their study shows that failure to account for LID leads to overestimating the reliability of the test scores by as high as 10-15%. Crehan (1993) and Thissen et al. (1989) also detected inflated reliability estimates when context-dependent item sets are treated as stand-alone items by comparing results obtained using the 3-PL IRT models to those using the polytomous models. Lee (1999) found that the item-based estimation methods for the conditional standard error of measurement (SEM) would provide underestimates for tests composed of testlet.

Test information function (TIF) is a frequently used reliability statistics within the IRT framework. For item *i* given the trait parameter *θ*, the item information function $I_i$ is defined as

$$I_i(\theta) = \frac{1}{Se_i^2(\theta)} \tag{2.2}$$

where $Se_i^2(\theta)$ stands for the SEM squared for item *i* given *θ*. *TIF(θ)* is defined as the sum of the item information functions given *θ*:

$$TIF(\theta) = \sum_{i=1}^{n} I_i(\theta) \tag{2.3}$$

When the items within a testlet display LID, the summed value would be a biased and inflated estimate of TIF because the SEMs are underestimated (Yen, 1993).

Since LID caused by the testlet format impacts the estimation of the model parameters and test reliabilities/TIFs, various methods have been proposed to detect, mitigate or model the LID. The methods within the IRT framework can be categorized into three groups:

1) Identifying LID through comparing the observed response patterns and model-predicted response probabilities and observing the residual correlations or the Chi-Square statistics. This category includes the $Q_2$ index proposed by van den Wollenberg (1982); the $Q_3$ index proposed by Yen (1984) and the $X^2$ and $G^2$ index proposed by Chen & Thissen (1997).

A description of Yen's $Q_3$ index is provided here since it is applied in the real data analysis later in this study. $Q_3$ is a model based descriptive statistic which assesses the local item independence assumption of the unidimensional IRT models. First, for item *i*, a person *k*'s residual score $d_{ik}$ is defined:

$$d_{ik} = u_{ik} - \hat{P}_i(\hat{\theta}_k) \tag{2.4}$$

where $u_{ik}$ is the raw item score of person *k* on item *i* and $\hat{P}_i(\hat{\theta}_k)$ is the probability of person *k* answering item *i* correctly, which is derived using a specific unidimensional IRT model and its item and person parameter estimates. The correlation of the residual scores of item *i* and item *j* taken over examinees is the $Q_{3ij}$ statistic:

$$Q_{3ij} = r_{d_i d_j} \tag{2.5}$$

According to Yen(1984), when the tested IRT model is true, $d_{ik}$ and $d_{jk}$ should be distributed approximately as bivariate normal variables with a zero correlation since they are random error scores. Yen also noted that the $Q_3$ statistics may suffer from the half-whole contamination issue (Kingston & Dorans, 1982) since the observed item score is used to calculated the expected item score $\hat{\theta}$. As a result the $Q_3$ statistics tend to be slightly negative. The expected value of $Q_3$ when there is no LID is -1/(*n*-1), where *n* is the total number of items. The $Q_3$ statistics have been tested and applied by Yen (1993), Fennessy(1995), Chen & Thissen(1997) and Zenisky et al. (2001).

The $Q_3$ and $G^2$ indices can be used to detect the magnitude of the testlet effect when analyzing item properties. It is desirable to eliminate or replace items that display strong LID if the construct validity is not affected by such changes. However when the testlet format is used, the interrelated items within a testlet are often construct relevant and cannot simply be removed from the test. Consequently, test developers and psychometricians have to use models and methods to account for the testlet effect in such circumstances.

2) Mitigating the effect of LID using polytomous models instead of unidimensional dichotomous models for tests composed of testlets. This has been a frequently used approach (Bishop & Omar, 2002; Cook, Dodd, & Fitzpatrick, 1999; Lee, Kolen, Frisbie, & Ankenmann, 1998; Thissen et al., 1989) based on the notion that testlet is "a subset of items in a test form that is treated as a measurement unit in test construction, administration, and/or scoring." (Lee, Brennan, & Frisbie, 2000, p.10). By treating testlets instead of items as the primary scoring unit, the local item independence assumption of the IRT can be upheld, as claimed by Rosenbaum (1988) "that given the loss of local independence within testlets, local independence can still prevail between testlets." However, using testlet-based scoring method requires summing the individual item scores within each testlet. Information regarding the response patterns to items within a testlet is lost in the process. This can lead to the loss of measurement information about items as compared to discrete-item scoring (Zenisky et al., 2001). Moreover, the estimation of the latent traits can also be affected by collapsing item scores into testlet scores.

3) The third approach is to model the testlet effects. Several models accounting for LID effect caused by the testlet format have been proposed. Wang, Cheng & Wilson (2005) used a multidimensional item response model to detect specific forms of LID for items across tests connected by common stimuli. Andrich

(1985) proposed a “dispersion location model” which is a specialized form of the rating scales model and used the dispersion parameter to quantify the magnitude of the LID effect. Demars (2006) applied the bi-factor model to testlets by treating the testlet traits as the secondary trait. Among these proposed models, the testlet models based on the Test Response Theory emerged to be the most promising ones in treating LID caused by the testlet format.

### *Testlet Response Theory (TRT)*

Steinberg and Thissen (1996, p82) asserted that “IRT is not a theory; It should be called a collection of statistical models and methods for making sense out of data arising in the context of psychological measurement”. The same can be said about TRT: it is not a theory, but a family of statistical models that are used to analyze testlet-based tests data. TRT is explicitly described in “Testlet Theory and its Applications” (Wainer et al., 2007), where the authors presented various testlet models that they have developed over the years including the testlet models analog to the 2-PL IRT model (Bradlow et al., 1999), the 3-PL IRT model (Wainer et al., 2000); and the general model that can be fit to a mixture of 2-PL, 3-PL and polytomously scored items(X. Wang et al., 2002).

The TRT models differ from the IRT models in that they include the testlet parameters which capture LID within testlets. Taking the 3-PL IRT model for example:

$$p(y_{ij}=1)=c_j+(1-c_j)\mathrm{logit}^{-1}(\mathrm{t_{ij}}) \qquad (2.6)$$

where $p(y_{ij}=1)$ is the probability of person *i* answering item *j* correctly and $c_j$ is the guessing parameter (lower asymptote) for Item *j*. $t_{ij}$ is the latent linear score predictor, which can be extended to the following formula according to IRT:

$$t_{ij}=a_j(\theta_i - b_j) \qquad (2.7)$$

where $a_j$ is the discrimination parameter for Item *j*; $b_j$ is the difficulty parameter for Item *j* and $\theta_i$ is the latent trait parameter for Person *i*.

According to TRT, a new parameter that accounts for the testlet effect is added to the 3-PL model:

$$t_{ij}=\mathrm{a}_{\mathrm{j}}(\theta_{\mathrm{i}}-\mathrm{b}_{\mathrm{j}}-\gamma_{\mathrm{ig(j)}}) \qquad (2.8)$$

with $\gamma_{ig(j)}$ being the testlet effect of person *i* with item *j* that is nested within the testlet *g*. $\gamma_{ig(j)}$ is independent of the item parameters, the ability parameter $\theta$, and the testlet parameters $\gamma$ from other testlets (Bradlow et al., 1999). For a particular person, $\gamma_{ig(j)}$ parameter is specified to be the same for all the items that are nested within the same testlet. This would result in higher inter-item correlations for the expected item scores within testlets than the expected item scores between testlets. Therefore the testlet effect can be accounted for. The mean of the testlet parameters for a particular testlet across all examinees is usually set to 0 so that the scale of the parameters can be identified. The set of testlet parameters work their effect through the variance $\sigma^2_{\gamma(g)}$ . The degree of dependence among the items within a testlet depends on the value of the variance. The larger the variance, the larger the testlet effect is. If the variance is 0, the items are locally independent. By introducing the dependence effect parameter into IRT models, testlet models produce item and person parameter estimates without bias caused by the testlet effect.

The TRT model accounts for LID by including an additive term that affects the item difficulty. This is reasonable from a substantive perspective since the testlet effect is usually caused by examinees' background knowledge and understanding about the stimulus. Different levels of the knowledge may affect how difficult items within a testlet are for the examinees, so that for a given examinee, the items in a given testlet may tend to be a little easier or a little harder than other items, in relation to their relative difficulties for other examinees (Yen, 1980). From a technical perspective, this approach to modeling dependence by adding what amounts to another factor for just a small group of items is in fact identical to Spearman's two factor model of intelligence that he developed back in early 1900s (Spearman, 1904) and Holtzinger's bi-factor model (Holzinger & Swineford, 1937).

The testlet model can be embedded within a Bayesian framework that allows sharing of information across persons, items, and testlets (Wainer et al., 2000). Under the Bayesian hierarchical structure, $\lambda_{ij}$, the parameters of the likelihood function $p(y_{ij}|\lambda_{ij})$ are governed by a set of parameters $\Lambda$ through a set of prior distributions $\pi(\lambda_{ij}|\ \Lambda)$. The marginal posterior distribution can be given as (Wainer et al., 2007):

$$p(\lambda \mid Y) \propto \int p(Y \mid \lambda) p(\lambda \mid \Lambda) d\Lambda \qquad (2.9)$$

Markov Chain Monte Carlo (MCMC) (Geman & Geman, 1984) is often used to perform this integration through sampling from the posterior distributions. To obtain the posterior distributions, the MCMC algorithm goes through the following steps, as described by Wainer et al. (2007):

1. In the initial stage where the iteration number $t=0$, the starting values are given to the parameters $\lambda$ and $\Lambda$, denoted as $\lambda_0$ and $\Lambda_0$.
2. In the next iteration $t=t+1$, sample from the conditional distribution $p(\lambda_1|\Lambda_0, Y)$. Since $p(\lambda_1|\Lambda_0, Y)$ is proportional to the product of the likelihood function $p(Y|\lambda_1,\Lambda_0)$ and the prior $p(\lambda_1|\Lambda_0)$, and the prior and the posterior distributions are not conjugate, special methods such as the Metropolis-Hastings (Chib & Greenberg, 1995) algorithm can be used to implement sampling from the conditional distribution. According to this algorithm, a sample value $\theta^*$ is first obtained from a distribution that allows straightforward sampling such as a normal distribution $g(\theta),$ then the value $g(\theta^*)$ is compared with the height of the target density $f(\theta^*):$ if $f(\theta^*)$ is larger than $g(\theta^*),$ the new sampled value $\theta^*$ is accepted; if $f(\theta^*)$ is smaller than $g(\theta^*),$ the value $\theta^*$ is accepted with the probability $f(\theta^*)/g(\theta^*).$
3. Given the newly sampled value of $\lambda_1$, draw a sample from the conditional distribution $p(\Lambda_1|\lambda_1,Y).$
4. Repeat the previous steps. Multiple iterations are needed for the two stages of MCMC: 1) In the burn in stage, convergences need to be reached so that the sampling distributions become stationary; 2) In the sampling stage, values are

sampled from multiple draws to reach stable estimation of posterior distributions of the model parameters.

### *Scaling and Equating*

Test equating is "the process of deriving a function mapping score on an alternate form of a test onto the scale of the anchor form, such that after equating, any given scale score has the same meaning regardless of which test form was administered." (Haertel, 2004, p.1). Test equating is often performed for security reasons. It is common that test programs administer different test forms on different dates to minimize item exposure. While test developers strive to construct test forms that are similar in content and statistical characteristics using the test specifications as the guidelines, these test forms usually differ in their difficulties. It is necessary to equate these test forms so that the difference in the test difficulties can be accounted for.

When tests are developed and scored using IRT methods, test equating is usually performed within the IRT framework. Equating with IRT usually requires that the scales of the parameter estimates from different test forms be on the same IRT scale. This is due to the fact that the latent variable in many IRT models is unidentified up to a linear transformation, i,e.: if the latent trait parameters are linearly transformed, then a complementary linear transformation can be made to the item parameters so that the model produces exactly the same fitted probabilities (Hanson & Beguin, 2002). To solve this issue, constraints are generally imposed for model estimation. The prevalent practice is to set the scale of the latent trait/ability on the standard normal distribution *~N(0,1).* When model parameters are estimated for two test forms *X* and *Y* taken by two different groups of examinees, the trait parameters *θs* are scaled to have a mean of 0 and a standard deviation of 1 for both groups in the separate estimation processes, even though the two groups may be non-equivalent. Consequently, the two sets of parameter estimates for form *X* and form *Y* may be on different scales and it is necessary to transform them onto the same scale before equating can be performed. To accomplish this, the test forms need to 1) share a set of common items, or be taken 2) by a single group or 3) by random and equivalent groups of examinees. These three data collection designs are called common-item nonequivalent groups (CINEG) design, single group design and random groups design (Kolen & Brennan, 2004). This research focuses on the first option: CINEG. Under this design the parameters of the common items are estimated on different IRT scales due to the group difference in latent traits. A linear equation can be used to transform the two set of parameter estimates onto the same scale. For example, suppose Scale *I* and Scale *J* are linearly related for a 3-PL IRT model in that:

$$\theta_{Ji} = A\theta_{Ii} + B \qquad (2.10)$$

where *A* and *B* are the linear transformation coefficients, and $\theta_{Ji}$ and $\theta_{Ii}$ are person *i*'s latent trait *θ* on Scale *J* and Scale *I.* The transformation relations of the item *j*'s item parameters *a*, *b* and *c* between the two scales are:

$$a_{Ji} = a_{Ij} / A \qquad (2.11)$$

$$b_{Ji} = Ab_{Ij} + B \tag{2.12}$$

$$c_{Jj} = c_{Ij} \tag{2.13}$$

***Characteristic curve scaling methods***

Several methods have been developed to estimate *A* and *B* scale linking constants. Marco (1977) presented the Mean/Sigma method, which makes use of the means and standard deviations of the *b*-parameter estimates from the common items. Loyd and Hoover (1980) proposed the Mean/Mean method, which computes *A* and *B* linking constants using the means of *a* and *b* parameter estimates of the common items. Mislevy and Bock (1990) suggested using the means of the *b* parameters and the geometric means of the *a* parameters. Kolen and Brennan (2004) pointed out that one potential issue with these moment methods is that different combinations of the *a, b* and *c* item parameters can produce almost identical item characteristics curves over the latent trait range and the two methods can be overly influenced by the difference between one of the item parameter estimates, even though the item characteristic curves for the items on the two estimations are very similar. To solve this issue, Haebara (1980) and Stocking and Lord (1983) developed two scale transformation methods which search for *A* and *B* constants that minimize the differences between the estimated item characteristic curves or test characteristic curves over the common items. The characteristic curve methods take into account all item parameter estimates.

The Haebara function is defined as the sum of the squared difference of the estimated probability functions for all common items for an ability level. For the 3-PL model parameter estimates, the function is:

$$Hdiff(\theta_i) = \sum_{j:V} [p_{ij}(\theta_{Ji}; \hat{a}_{Jj}, \hat{b}_{Jj}, \hat{c}_{Jj}) - p_{ij}(\theta_{Ji}; \frac{\hat{a}_{Ij}}{A}, \hat{b}_{Ij} + B, \hat{c}_{Ij})]^2 \qquad (2.14)$$

where $\hat{a}, \hat{b},$ and $\hat{c}$ are the estimated values of the *a*, *b* and *c* parameters. *j:V* are the set of common items. The function *Hcrit* is then defined by either summing up *Hdiff(θ<sub>i</sub>)* over all examinees using a point estimate for each examinee as shown below:

$$Hcrit = \sum Hdiff(\theta_i) \qquad (2.15)$$

or integrating over $\theta$ with respect to a known or estimated density. The scale transformation constants can be estimated by finding *A* and *B* values that minimize the criterion *Hcrit*.

The Stocking & Lord method is similar to the Haebara method except that it aims to search for the scale linking constants that minimize the difference between the estimated test characteristic curves of the base test form and the new test form after the scale transformation. First, given $\theta_i$, the estimated number-correct test scores, i.e., true scores on the base form $\tau$ and the rescaled true scores on the new form $\tau^*$ can be estimated:

$$\hat{\tau}(\theta_i) = \sum_{j:V} p_{ij}(\theta_{Ji}; \hat{a}_{Jj}, \hat{b}_{Jj}, \hat{c}_{Jj}) \qquad (2.16)$$

$$\hat{\tau}^*(\theta_i) = \sum_{j:V} p_{ij}(\theta_{Ji}; \frac{\hat{a}_{Ij}}{A}, \hat{b}_{Ij} + B, \hat{c}_{Ij}) \tag{2.17}$$

The Stocking & Lord function is defined as the squared differences of the estimated true scores, for a given $\theta_i$:

$$SLdiff(\theta_i) = (\hat{\tau}(\theta_i) - \hat{\tau}^*(\theta_i))^2 \tag{2.18}$$

The function *SLcrit* is then defined by summing up *SLdiff(*$\theta_i$*)* over examinees and the scale transformation constants *A* and *B* can be estimated by finding the values that minimize the criterion *SLcrit*.

Past studies have shown that the characteristic curve methods perform better than the Mean/Mean and Mean/Sigma methods (Baker & Al-Karni, 1991; Hanson & Beguin, 2002; Way & Tang, 1991). The characteristic curve methods can also be used with polytomous IRT models. Baker (1992) extended the two methods to Samejima's (1969) graded response model and Hatorri (1998) applied them in Muraki's (1992) generalized partial credit model.

***Li et al.'s scale linking method for a testlet model***

Li, Bolt and Fu (2005) proposed a method of computing the linking coefficients for the two parameter normal ogive (2PNO) testlet model using an extension of the Stocking & Lord Method. The model specifies that the probability that an examinee *j* answers item *i* correctly as:

$$p_{ij} = \Phi(a_i(\theta_j - b_i - \gamma_{jd})) \tag{2.19}$$

where $\Phi$ is the standard normal cumulative distribution function and $\gamma_{jd}$ is the random testlet effect parameter for person *j* on testlet *d*. To perform scale linking, Li et al. adopted the reparameterization proposed by Glas, Wainer, & Bradlow (2000) and changed the probit $a_i(\theta_j-b_i-\gamma_{jd})$ to $a_i(\xi_j-b_i)$ where $\xi_j= \theta_j-\gamma_{jd}$. The probability of answering item *i* correctly conditional on $\theta$ can be obtained by integrating the testlet parameters out:

$$P(y_{di}=1|\theta;\sigma_{\xi_d})=\int P(y_{di}=1|\xi_d)h(\xi_d|\theta;\sigma_{\xi_d})d\xi_d \qquad (2.20)$$

where $\sigma_{\xi_d}$ is the standard deviation of $\xi_{jd}$, which is equal to the standard deviation of $\gamma_{d.}$ This parameter is assumed to be a known value, using the estimate obtained in the model estimation step. Given the above probability function, the true score for all the common items of the base test $\tau$ and of the transformed new test $\tau^*$ can be derived and the Stocking & Lord linking method can be implemented by minimizing the *SLcrit* function.

It should be noted that the method proposed by Li et al. allow that when "examinees from two populations respond to the same testlet, it may be that not only their $\theta$ distributions differ but also their $\gamma_{jd}$ distributions." (Li et al. 2005, p.343) Based on this proposition, the authors believe that the means of $\gamma_{jd}$ of the testlet effect parameters of the new test form need not be 0 after scale transformation. Their method accommodates this shift in the means by adding another constant $\mu_{\gamma d}$ in their calculation. By including the shifted means of the testlet effect parameters in scale linking, Li et al. in effect modifies the testlet model by adding a set of dimension parameters to it. This parameter accounts for Li et al.'s assumption that testlets can affect different populations differently: that the true score of examinees is systematically altered by a set of testlet related dimensions.

However, according to TRT, The interdependence of the testlet items results in the random interaction between the testlets and the examinees and the variances of the testlet effect parameters over the examinees are used to quantify the magnitude of this interaction. The researchers are generally not concerned with the means of the testlet parameters, which are customarily set to 0 to make the scale of the model identifiable in the estimation process. If a testlet affects different populations differently due to the inter-dependence of the items, it should be demonstrated in the difference in the variances of the testlet parameter distribution. Li et al.'s testlet related dimensions can be regarded as nuisance dimensions that originate from the content of each testlet.  This is different from the LID effect caused by items sharing the same stimulus. Note that this extension is beyond the scope of this research, as much remains to be learned about the standard situation in which scaling shifts are assumed common across testlets.

The characteristic curve method employed by Li et al. is an extension of the Stocking & Lord scale linking method. There have been few studies that compare the performances of the Stocking & Lord method and the Haebara method. Way and Tang (1991) found that methods based on the two criteria *Hcrit* and *SLcrit* produced similar results for dichotomous IRT models. Li and Yin(2008) compared several procedures for polytomous IRT model equating and found that using the Stocking & Lord method or the Haebara method doesn't have a significant impact on the final equating results. However, Kolen and Brennan (2004) suggested that the Haebara method may be theoretically superior to the Stocking & Lord method because *Hdiff*$(\theta_i)$ can be 0 only if the item characteristic curves are identical at $\theta_i$ whereas *SLdiff*$(\theta_i)$ can be 0 even if the item characteristic curves differ. Thus, the Haebara method can be viewed as being more stringent than the Stocking & Lord method (Kolen & Brennan, 2004).

Another caveat of Li et al.'s study is that they applied the characteristic curve scale linking method to the 2PNO testlet model. While this model has similar statistical characteristics as the 2-PL testlet model, it is not as popularly used as the logistic function- based testlet models.

To sum up, it would be of interest to devise such a scale linking procedure that 1) takes into account of the testlet effect using the logistic function-based testlet model and 2) is an extension of the Haebara item characteristic curve scale linking method.

# Chapter 3 Methodology and Simulation Study

## *Research Questions*

The objective of this dissertation is to propose a new scale linking method within the TRT framework: specifically, the Haebara item characteristic curve scale linking method is extended to the 3-PL testlet model. The study attempts to answer the following research questions:

1. How well does the proposed scale linking method recover the true linking relations for test forms composed of testlets?
2. Does the proposed 3-PL testlet model scale linking method perform better than the scale linking methods using the traditional dichotomous and polytomous IRT models when they are applied to testlet-based tests?

## *Methodology*

With the CINEG design, the scale transformation is based on the theory that if a model fits the data, a simultaneous linear transformation of the model parameters will result in the same probability function. Within the TRT framework, the model parameters that need to be rescaled include the testlet parameters as well as the item and person parameters. Suppose Scales *I* and Scale *J* are linearly related for a 3-PL testlet model, the scale transformation relations for the parameters are:

$$\theta_{Ji} = A\theta_{Ii} + B \tag{3.1}$$

$$a_{Jj} = a_{Ij} / A \tag{3.2}$$

$$b_{Jj} = Ab_{Ij} + B \tag{3.3}$$

$$c_{Jj} = c_{Ij} \tag{3.4}$$

$$\gamma_{Jig(j)} = A\gamma_{Iig(j)} \tag{3.5}$$

To prove such linear transformation is valid, when the parameters are on Scale *J*, the 3-PL testlet model can be written as:

$$p(y_{ij} = 1) = c_{Jj} + (1 - c_{Jj})\frac{\exp[a_{Jj}(\theta_{Ji} - b_{Jj} - \gamma_{Jig(j)})]}{1 + \exp[a_{Jj}(\theta_{Ji} - b_{Jj} - \gamma_{Jig(j)})]} \tag{3.6}$$

Replace $\theta_{Ji}$, $a_{Ji}$, $b_{Ji}$, $c_{Ji}$ and $\gamma_{Jig}$ with the expressions from (3.1) to (3.5)

$$= c_{Ij} + (1 - c_{Ij})\frac{\exp[\frac{a_{Ij}}{A}[(A\theta_{Ii} + B) - (Ab_{Ij} + B) - A\gamma_{Iig(j)})]}{1 + \exp[\frac{a_{Ij}}{A}[(A\theta_{Ii} + B) - (Ab_{Ij} + B) - A\gamma_{Iig(j)})]} \tag{3.7}$$

$$= c_{Ij} + (1 - c_{Ij})\frac{\exp[a_{Ij}(\theta_{Ii} - b_{Ij} - \gamma_{Iig(j)})]}{1 + \exp[a_{Ij}(\theta_{Ii} - b_{Ij} - \gamma_{Iig(j)})]} \tag{3.8}$$

where (3.8) is the same as formula (3.6) except that the parameters are on Scale *I* now. This shows that the linear scale transformation of the parameters using constants *A* and *B* results in the same probability functions.

The scale linking method based on the Haebara approach is proposed to search for *A* and *B* constants. First, a function called *Hdiff*($\theta_i$) is defined which computes the sum of the squared difference between the estimated true scores for each item for the ability level $\theta_i$:

$$Hdiff(\theta_i) = \sum_{j:V} [p_{ij}(\theta_{Ji}; \hat{a}_{Jj}, \hat{b}_{Jj}, \hat{c}_{Jj}; \gamma_{Jg(j)}) - p_{ij}(\theta_{Ji}; \frac{\hat{a}_{Ij}}{A}, \hat{b}_{Ij} + B, \hat{c}_{Ij}; A\gamma_{Ig(j)})]^2 \quad (3.9)$$

Now the issue is how to calculate the probability function $p_{ij}$ for each item given $\theta_i$ in formula (3.9). This is straightforward in the case of the 3-PL IRT model by using the estimated item parameter values. Note however that (3.9) includes, in addition to the standard 3-PL item parameter estimates, values of the testlet parameters $\gamma$. These values are not known in practice, so a practical procedure will need to find an approximation that deals with the testlet parameters as well. In the testlet model, $\gamma_{Jg(j)}$ and $\gamma_{Ig(j)}$ are a vector of values that are normally (and independently) distributed with the mean of 0 and standard deviation of $\sigma(\gamma_{Jg(j)})$ and $\sigma(\gamma_{Ig(j)})$ respectively. Consequently, the person parameters are vectors instead of single values. This greatly complicates the process of computing the *Hdiff*($\theta_i$) function.

Since the testlet parameter distribution is considered to be continuous, the probability of answering item *j* within testlet *g* correctly given $\theta_i$ can be obtained by:

$$
\begin{aligned}
&P_{ij}(\theta_{Ji};\hat{a}_{Jj},\hat{b}_{Jj},\hat{c}_{Jj};\boldsymbol{\gamma}_{\boldsymbol{Jg(j)}}) \\
&=\int\{\hat{c}_{Jj}+(1-\hat{c}_{Jj})\frac{\exp[\hat{a}_{Jj}(\theta_{Ji}-\hat{b}_{Jj}-\gamma_{Jig(j)})]}{1+\exp[a_{Jj}(\theta_{Ji}-\hat{b}_{Jj}-\gamma_{Jig(j)})]}\}\psi(\gamma_{Jig(j)})d(\gamma_{Jig(j)})
\end{aligned}
\tag{3.10}
$$

where $\psi(\gamma_{Jig})$ is the estimated distribution of $\boldsymbol{\gamma}_{\boldsymbol{Jig}}$. It is appropriate to obtain the expected item score by taking the integral over the $\boldsymbol{\gamma}_{\boldsymbol{Jig}}$ distribution because the testlet model assumes that the $\boldsymbol{\gamma}_{\boldsymbol{Jig}}$ parameter is independent of the ability parameter $\theta_i$. Otherwise, the expected item score would have to be obtained by integrating the logistic function over the $\boldsymbol{\gamma}_{\boldsymbol{Jig}}$ distribution conditional on the $\theta$ distribution.

Since $\boldsymbol{\gamma}_{\boldsymbol{Jig}}$ is specified to be drawn from a normal distribution with a mean of 0 and a standard deviation of $\sigma(\boldsymbol{\gamma}_{\boldsymbol{Jg}})$, and $\sigma(\boldsymbol{\gamma}_{\boldsymbol{Jg}})$ can be estimated, we can approximate the continuous distribution with a discrete distribution on a finite number of equally spaced quadrature points to compute the integral so that:

$$
\begin{aligned}
&P_{ij}(\theta_{Ji};\hat{a}_{Jj},\hat{b}_{Jj},\hat{c}_{Jj},\boldsymbol{\gamma}_{\boldsymbol{Jg(j)}}) \\
&\cong\sum_{k}((\hat{c}_{Jj}+(1-\hat{c}_{Jj})\frac{\exp[\hat{a}_{Jj}(\theta_{Ji}-\hat{b}_{Jj}-p_{k(\gamma_{Jg})})]}{1+\exp[\hat{a}_{Jj}(\theta_{Ji}-\hat{b}_{Jj}-p_{k(\gamma_{Jg})})]})W_{k(\gamma_{Jg})})
\end{aligned}
\tag{3.11}
$$

where $p_{k(\gamma_{Jg})}$ is the *kth* quadrature point and $W_{k(\gamma_{Jg})}$ is the corresponding weight. Similarly,

$$P_{ij}(\theta_{Ji};\frac{\hat{a}_{Ij}}{A},\hat{b}_{Ij}+B,\hat{c}_{Ij},A\boldsymbol{\gamma}_{\boldsymbol{Ig(j)}})$$

$$\cong \sum_k (\hat{c}_{Ij}+(1-\hat{c}_{Ij})\frac{\exp[\frac{\hat{a}_{Ij}}{A}(\theta_{Ji}-(A\hat{b}_{Ij}+B)-Ap_{k(\gamma_{Ig})})]}{1+\exp[\frac{\hat{a}_{Ij}}{A}(\theta_{Ji}-(A\hat{b}_{Ij}+B)-Ap_{k(\gamma_{Ig})})]})(W_{k(\gamma_{Ig})}) \qquad (3.12)$$

*Hdiff(θ$_i$)* as shown in formula (3.9) is a summation function performed over all the common items. The function *Hcrit* is then defined by adding up *Hdiff(θ)* for all examinees that have taken the base test form, again using point estimates of $\theta$ for each examinee in the summation.

$$Hcrit = \sum Hdiff(\theta_i) \qquad (3.13)$$

The scale transformation constants *A* and *B* can be estimated by finding the values that minimize the criterion *Hcrit*.

### ***Simulation Design and Analysis***

The simulation study was performed to evaluate the effectiveness of the proposed linking method under the testlet model. It is compared against the scaling methods using the simpler 3-PL IRT model and the graded response model (GRM) (Samejima, 1969) to study if it performs better in recovering the true linking relationship between the two sets of parameters estimated on different scales.

***Data simulation***

Two test forms with common items were created for each dataset. Binary scores *(0, 1)* for two 30-item test forms (the base form and the new form) were simulated. There were 6 testlets in each test form and each testlet consisted of 5 items. 1000 subjects were simulated for each test form. The 3-PL testlet model was used to generate datasets with testlet characteristics. The generating distributions for the model parameters are presented in Table 1:

Table 1

*Simulation Specifications: Parameter Generating Distributions and Simulation Conditions*

| Parameters | Distributions | |
|---|---|---|
| | Base Form | New Form* |
| a | $\sim LN(-0.3, 0.35^2)$ | $\sim LN(-0.3, 0.35^2)$ |
| b | $\sim N(0,1^2)$ | $\sim N(0,1^2)$ |
| c | $\sim N(0.2, 0.05^2)$ | $\sim N(0.2, 0.05^2)$ |
| θ | $\sim N(0,1^2)$ | $\sim N(0.5,1.5^2)$ |
| $\gamma_g$ | Condition 1: =0 | Condition 1: =0 |
| | Condition 2: $\sim N(0,1^2)$ | Condition 2: $\sim N(0,1^2)$ |
| | Condition 3: $\sim N(0, \sqrt{2}^2)$ | Condition 3: $\sim N(0, \sqrt{2}^2)$ |

Note * The parameter distributions only apply to the first 3 testlets (15 items) for the new form. The last 15 items are the common items and have the same parameter values as those of the base form.

1. Item parameters: For the base form, the difficulty parameter *b*'s were created using the standard normal distribution *N(0,1);* the discrimination parameter *a*'s were created using the lognormal distribution *LN(-0.3, $0.35^2$)* and the guessing

parameter *c'*s were created using the normal distribution $N(0.2, 0.05^2)$ with the lower limit set at 0. 30 sets of item parameters were generated. For the new form, the item parameters of the first 15 items were generated from the same distributions as those for the base form. The last 15 items were specified to be the common items and they had exactly the same item parameter values as the last 15 items of the base form.

2. Person parameters: The $\theta$ parameters were specified to follow the normal distribution $N(0, 1)$ for the base form and $N(0.5, 1.5^2)$ for the new form. This reflected the non-equivalent nature of the examinee groups.

3. Testlet parameters: The testlet effect parameters were generated using normal distributions $N(0, var_{\gamma(g)})$. The degree of the testlet effect was determined by the variances of the testlet parameter values: $var_{\gamma(g)}$. In this study, three conditions of different degrees of testlet effects were simulated: 1-no testlet effect ($var_{\gamma(g)}=0$); 2-moderate testlet effect ($var_{\gamma(g)}=1$) and 3-strong testlet effect ($var_{\gamma(g)}=2$). These testlet effect conditions were similar to those simulation conditions specified in Bradlow et al (1999), which specified $var_{\gamma(g)}$ to be 0, 0.5, 1 and 2.

With the parameter values ready, the probability of getting each item right was calculated using the 3-PL testlet model. The item scores in the form ($0, 1$) were simulated using the Bernoulli distribution function which was dependent on this calculated probability. For each of the three conditions, 50 samples were generated.

***Model calibration***

After the datasets were simulated, the 3-PL IRT model, GRM and the 3-PL testlet model were fitted respectively to the data.

The 3-PL IRT model was used because its only difference from the 3-PL testlet model is that it doesn't account for the testlet effect. By including the results of scale linking procedure based on the 3-PL IRT model as a benchmark in the study, the improvement (if only) in the scale linking performance due to using the testlet model can be studied. The 3-PL IRT model is of the form:

$$p_{ij} = c_j + (1 - c_j)\frac{\exp[a_j(\theta_i - b_j)]}{1 + \exp[a_j(\theta_i - b_j)]} \qquad (3.14)$$

where $P_{ij}$ is the probability of correctly answering item $j$ for person $i$;; $a_j$ is the item slope parameter, $b_j$ is the item difficulty parameter and $c_j$ is the lower asymptote guessing parameter. The computer program BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 2005) was used to estimate the model parameters.

As discussed in the previous chapter, some researchers have employed polytomous IRT models in treating the testlet based test forms to account for the testlet effect. Therefore, it makes sense to include polytomous IRT model-based scale linking procedure in this study so that its performance can be compared with that of the proposed scale linking procedure based on the testlet model. Two polytomous models that are popularly used for such purposes are Samejima's (1969) GRM and Muraki's (1992) generalized partial credit model (GPCM). Past research has shown that the two models provide highly similar results when used to analyze items with multiple-category responses (Maydeu-Olivares, Drasgow, & Mead, 1994; Tang & Eignor, 1997; Thissen, Billeaud, McLeod, & Nelson, 1997). The GRM was used in this study. The GRM is a form of difference model (Thissen & Steinberg, 1986). It first models the “cumulative response functions,” which refers to the cumulated probability of scoring at or above a certain category. Note that this is different from the cumulative distribution function in its commonly used definition: the probability of receiving a certain outcome or a lower one. Next, the category response functions, or the probability of scoring at a specific category, are derived through calculating the differences between the cumulative functions of successive responses. The cumulative response category function is of the following form:

$$\begin{aligned} p_{ijk}^{*}(\theta_i) &= \frac{\exp[a_j(\theta_i - b_{jk})]}{1+\exp[a_j(\theta_i - b_{jk})]}, \qquad && \mathit{if} \quad 1<k<K; \\ &= 1 && \mathit{if} \quad k=1 \end{aligned} \tag{3.15}$$

where $P*_{ijk}(\theta_i)$ is the cumulative probability of scoring at or above category $k$ for person $i$; category $k=1, 2,..., K;$ $a_j$ is the item slope or step parameter and $b_{jk}$ is the item location (difficulty) parameter. For a specific item, the higher the category, the larger the difficulty parameter value for that category. Once the cumulative probability function is estimated, the category response function then can be calculated via:

$$\begin{aligned} p_{jk}(\theta_i) &= p_{jk}^{*}(\theta_i) - p_{j,k+1}^{*}(\theta_i), \quad && \mathit{if}\ 1<=k<K; \\ &= p_{jk}^{*}(\theta_i) && \mathit{if}\ k=K \end{aligned} \tag{3.16}$$

PARSCALE (Muraki & Bock, 1997) was used to perform the GRM estimation.

The WinBUGS program (Spiegelhalter, Thomas, Best, & Lunn, 2003) was used to implement the MCMC method for the 3-PL testlet model estimation. The following prior distributions were specified for the item parameters: the normal distribution *N(0,* $2^2$*)* for the difficulty parameter *b*; the lognormal distribution *LN(0,* $0.5^2$*)* for the discrimination parameter *a* and the beta distribution *Beta(5, 17)* for the guessing parameter *c*. As Patz and Junker (1999) pointed out in their study on applying MCMC to IRT models, it is common to use these types of prior distributions for the item parameters in the Bayesian estimation of the 3-PL IRT model. They used the same prior distributions for the *a* and *b* parameters and a quite similar beta prior distribution for the *c* parameter in their estimation of the 3-PL IRT model. Many simulation studies have been carried out to study the sensitivity of 3-PL item parameters to prior specifications (e.g., Harwell & Janosky, 1991), leading to the conclusion that the specifications noted above are sufficient to provide finite and stable estimates in the kinds of data normally seen in educational testing, without overwhelming response data. Since the 3-PL testlet model is an extension of the 3-PL IRT model, these prior distributions for the item parameters are adopted for the 3-PL testlet model estimation in this study.

All six prior distributions for the precision (the reciprocal function of the variance) of the testlets were set to be gamma distribution *Gamma(0.5, 1),* based on previous research by Bradlow, Wainer, and Wang (1999). The prior distribution of $\theta$ was specified to be standard normal distribution *N(0,1)* to set the scales of the person and item parameters.

Sinharay (2003) indicated that the number of iterations required to ensure convergence may be quite large for testlet models. In order to obtain the stable posterior distributions of the model parameters, it is necessary to ascertain the number of iterations that are needed before convergence can be achieved. Two chains of iterations were run first on a sample dataset generated using the above simulation specifications to check convergence.  While the initial values for the other parameters were randomly generated by WinBUGS, the starting values of the *b* parameters of the 30 items were all specified to be -1 for the first chain and 1 for the second chain so that the MCMC processes start from different spaces for the two chains. 25000 iterations were run in the test and the results were observed.

Since the Metropolis sampling method was used, WinBUGS required an adaptive phase of 4000 iterations before model estimation could be performed. The traces of all *a, b* and *c* item parameters, the first 10 $\theta$ parameters and the variances of the testlet parameters estimates for the 6 testlets were monitored. The traces of some randomly selected item parameter estimates are presented in Figure 2. They are: *a* parameters for Items 1 and 6, *b* parameters for Items 3 and 13 and *c* parameters for Items 11 and 20. As we can see, the two chains of the item parameter estimates converge very well after the initial 4000 iterations. Figure 3 presents the traces for the variances of the testlet parameter estimates of the 6 testlets. The two chains also converge well after 4000 iterations.

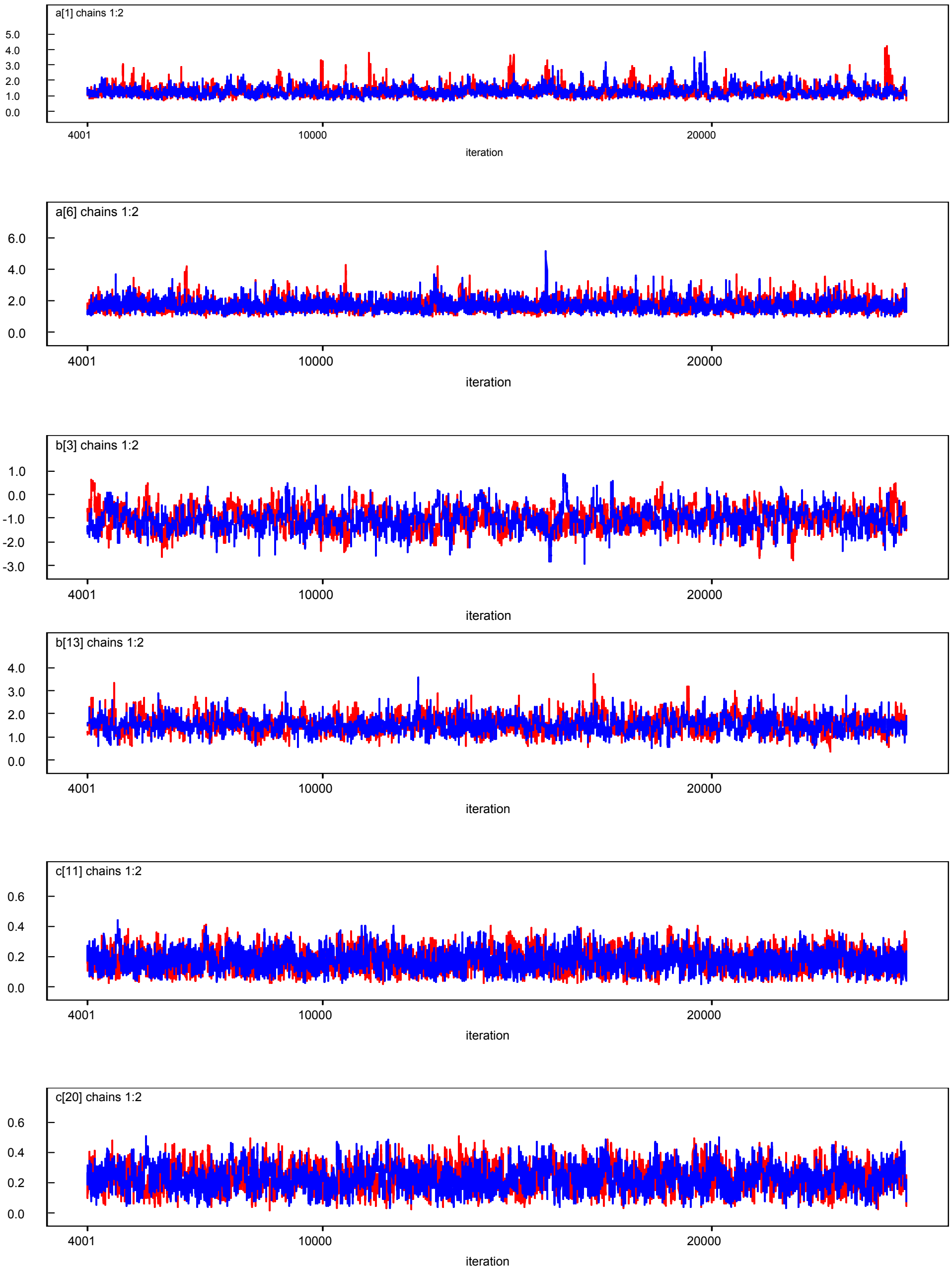


*Figure 2.* WinBUGS history output for some item parameters: $a_1$, $a_6$, $b_3$, $b_{13}$, $c_{11}$, $c_{20}$

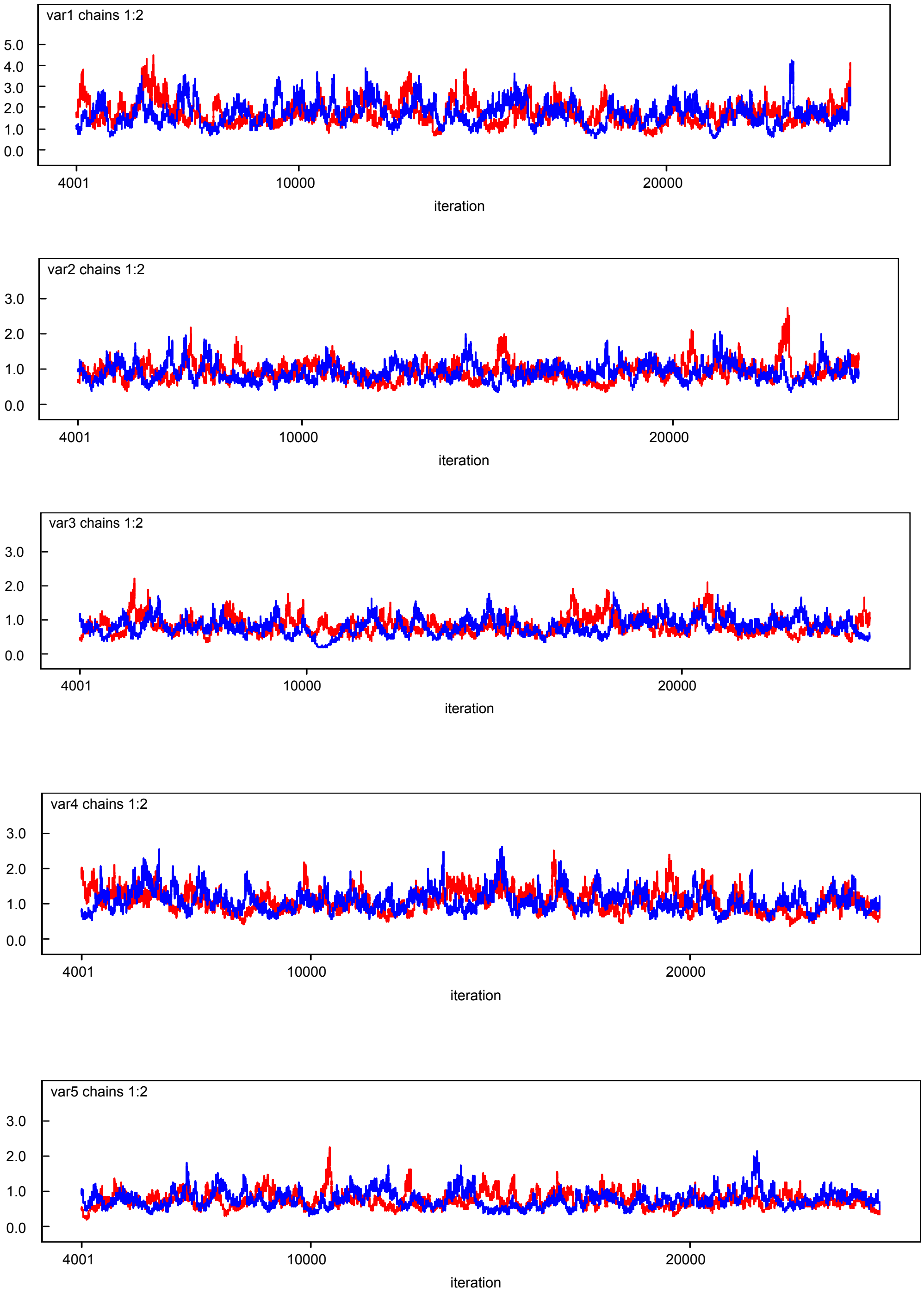


*Figure 3.* WinBUGS history output for variances of the testlet parameters

The WinBUGS also calculates the Gelman-Rubin (1992) index. According to the WinBUGS manual (Spiegelhalter et al., 2003, p.27), 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 R (= pooled / within) is red”. Since the variances of the testlet parameters may take more iterations to converge, the Gelman-Rubin plots for the variances of the testlet parameters are presented in Figure 4. The plots show that the blue and green curves overlap and the red curve hovers around 1 after about 6000 iterations for the variances of the testlet parameters. Judging from the history plots and the Gelman-Rubin plots, the convergence is reached after 6000 iterations.

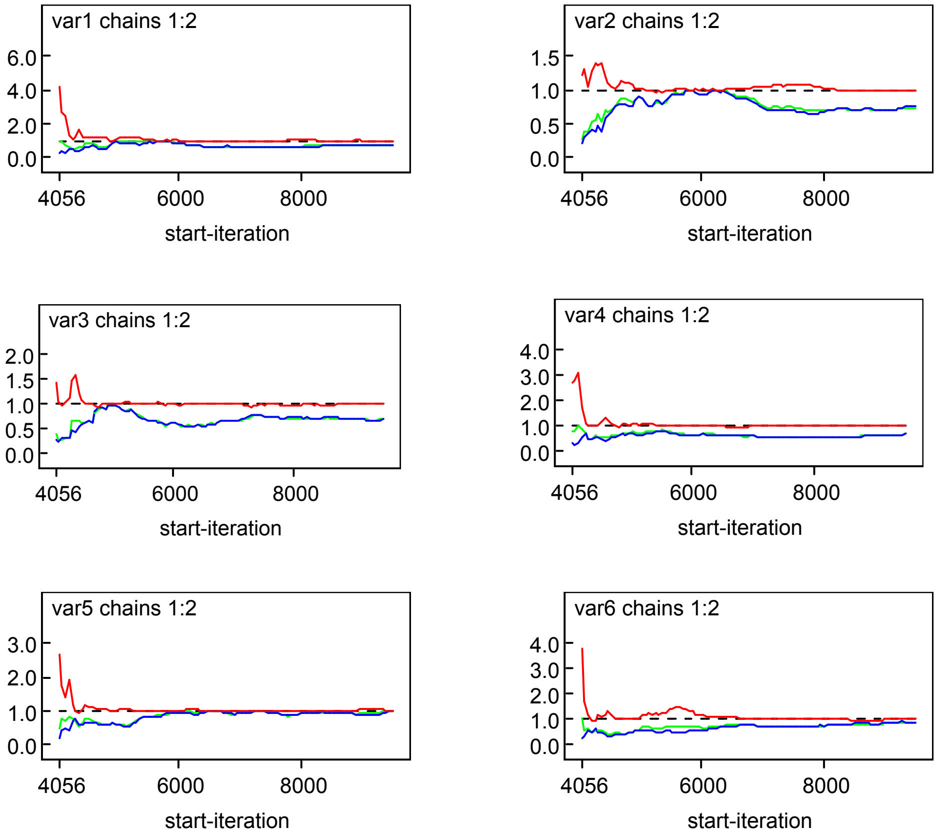


*Figure 4.* WinBUGS Gelman-Rubin plots for variances of the testlet parameters

Another convergence study was conducted specifically for the data simulated under Condition 1 since there may be a convergence issue for the testlet effect parameters when there is no testlet effect. The two chains with the same initial values as specified in the previous convergence study for the *b* parameters were run. The Gelman-Rubin plots are shown in Figure 5 and the history plots for the six variances of testlet effect parameters are shown in Figure 6. The two figures indicate that the two chains converge quite well after 6000 iterations.

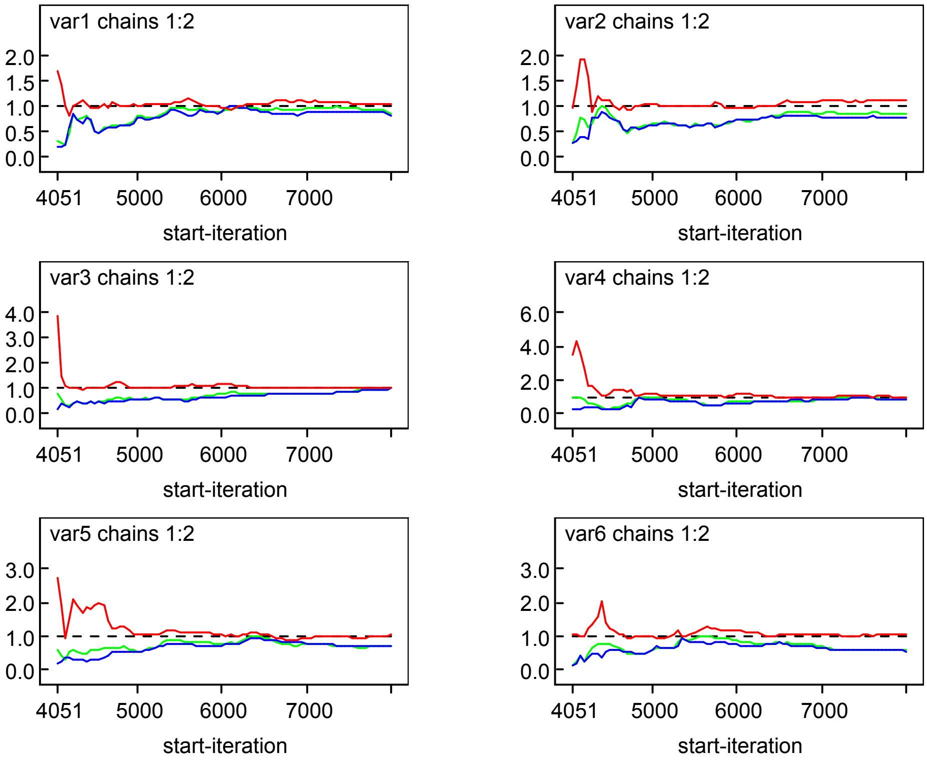


*Figure 5.* WinBUGS Gelman-Rubin plots for variances of the testlet parameters when there is no testlet effect

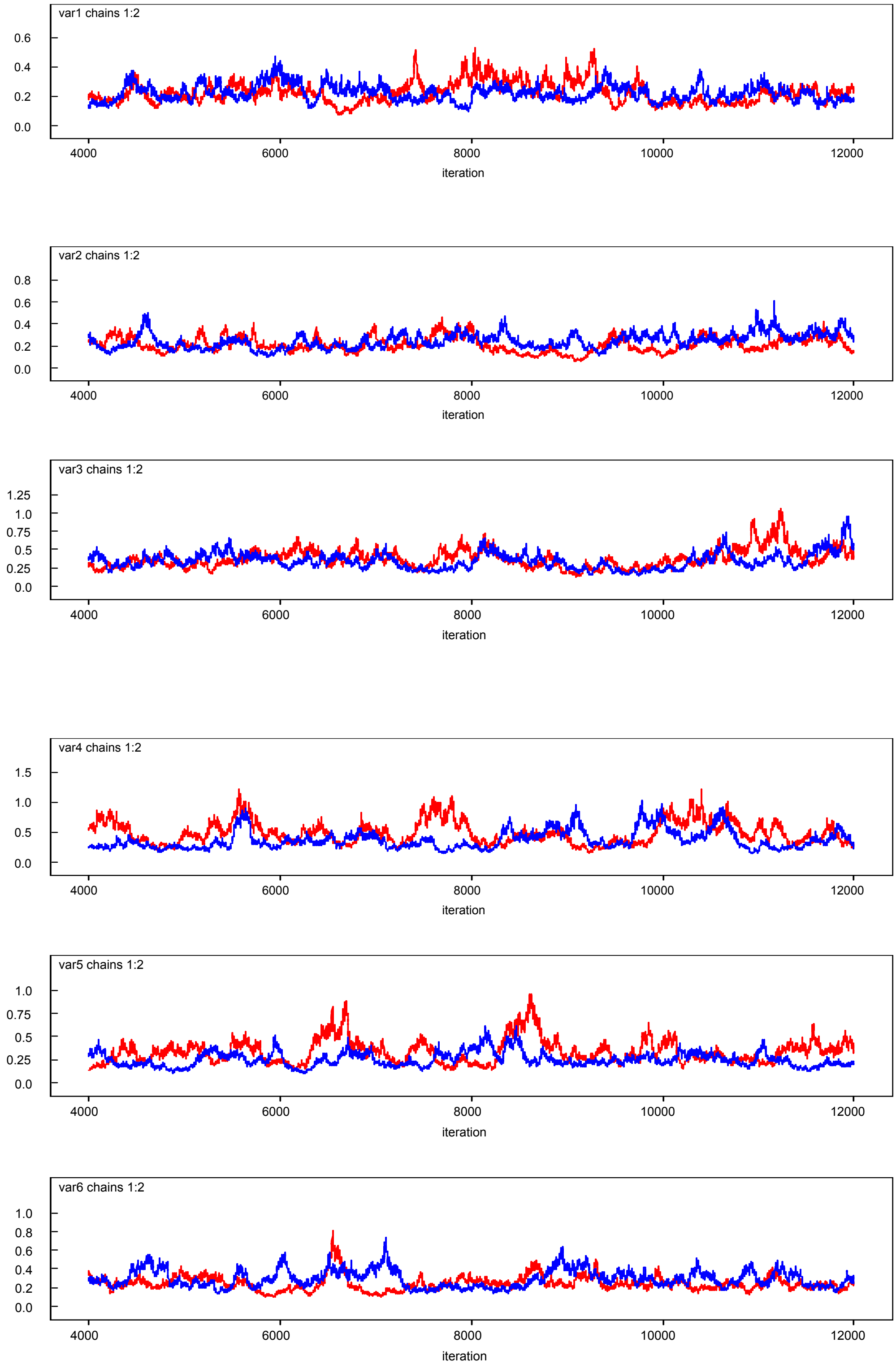


*Figure 6.* WinBUGS history output for variances of the testlet parameters when there is no testlet effect

In the actual model estimation, WinBUGS was programmed to run 12000 iterations and discarded the first 6000 iterations. Therefore a total of 6000 samples were used to estimate the model parameters. This is more conservative than several other TRT model estimation studies. For example, Bradlow et al. (1999) ran only 2000 iterations and used the last 1000 iterations for the 2-PL testlet model estimation.

***Scale linking***

To allow consistent comparisons, the Haebara item characteristic curve methods were used to perform the scale transformations for the 3-PL model and the GRM. The programs ST (Hanson & Zeng, 2004) and POLYST (Kim & Kolen, 2003) were used to perform scale linking for the 3-PL model-estimated parameters and GRM-estimated parameters respectively.

In the simulation study, each test form was taken by 1000 examinees and each test form had 15 common items embedded in 3 testlets. 20 evenly distributed quadrature points in the range of [-3, 3] were used to estimate the probability functions given $\theta_i$. The quadrature weights were calculated using the SAS PROBNORM function, a practice that has been applied by Xiao (1999). The criterion *Hcrit* in formula (3.13) was expanded:

$$
\begin{aligned}
Hcrit &= \sum Hdiff(\theta_i) \\
&= \sum_{N=1}^{1000}\sum_{j=1}^{15} \{(\sum_{20}((\hat{c}_{Ij}+(1-\hat{c}_{Ij})\frac{\exp(\frac{\hat{a}_{Ij}}{A}(\theta_{Ji}-(A\hat{b}_{Ij}+B)-Ap_{k(\gamma_{Ig})}))}{1+\exp(\frac{\hat{a}_{Ij}}{A}(\theta_{Ji}-(A\hat{b}_{Ij}+B)-Ap_{k(\gamma_{Ig})})]})W_{k(\gamma_{Ig})}) \\
&\quad -\sum_{20}((\hat{c}_{Jj}+(1-\hat{c}_{Jj})\frac{\exp(\hat{a}_{Jj}(\theta_{Ji}-\hat{b}_{Jj}-p_{k(\gamma_{Jg})}))}{1+\exp(\hat{a}_{Jj}(\theta_{Ji}-\hat{b}_{Jj}-p_{k(\gamma_{Jg})})]})W_{k(\gamma_{Jg})}))^{2}\ \}
\end{aligned}
\tag{3.17}
$$

Since this is a nonlinear minimization problem, the Newton Raphson method can be used to search for the constants *A* and *B*. There is no off-the-shelf program available to implement the method for scaling the 3-PL testlet model. The PROC NLP procedure of SAS was used to perform the computation, with *Hcrit* specified to be the objective function. Appendix A provides the kernel of the SAS NLP procedure code used in this study.

One issue with non linear programming is that multiple local optima may exist in the optimization process. To check if the global optima can be reached using the proposed method, multiple runs using different starting values were performed on a randomly generated dataset under each of the three conditions. For *A* parameter, the starting values were selected by taking four equally spaced values within the range [0.2, 5]: 0.20, 1.80, 3.40 and 5.00. For *B* parameter, the starting values were selected by taking four equally spaced values within the range [-5, 5]: -5.00, -1.67, 1.67 and 5.00. These two selected ranges were rather conservative and it was improbable for the true parameter values to be outside the ranges. The paired combinations of the two sets of values were adopted as the starting values. Altogether 16 (4 by 4) pairs of starting values were used under each of the three simulation conditions. As demonstrated in Table 2, under each of the three conditions, the same optimum results are reached using different starting values. This provides strong empirical support for the claim that there are no multiple local optima within the space where the true parameter values are likely to exist using the proposed scale linking method.

Table 2

*Estimation of the Scale Linking Parameters Using Multiple Starting Values For a Randomly Generated Dataset under Each of the Three Conditions*

| | | | Condition 1: Var(testlet)=0 | | | | Condition 2: Var(testlet)=1 | | | | Condition 3: Var(testlet)=2 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Trial | Starting Values | | Optimum Values | | Value of the Objective Function | | Optimum Values | | Value of the Objective Function | | Optimum Values | | Value of the Objective Function | |
| | A | B | A | B | Start | Final | A | B | Start | Final | A | B | Start | Final |
| 1 | 0.20 | -5.00 | 1.47 | 0.50 | 3849.95 | 11.02 | 1.58 | 0.62 | 2569.85 | 12.29 | 1.61 | 0.64 | 2573.18 | 14.10 |
| 2 | 0.20 | -1.67 | 1.47 | 0.50 | 3156.26 | 11.02 | 1.58 | 0.62 | 2408.54 | 12.29 | 1.61 | 0.64 | 2430.99 | 14.10 |
| 3 | 0.20 | 1.67 | 1.47 | 0.50 | 1529.51 | 11.02 | 1.58 | 0.62 | 2466.58 | 12.29 | 1.61 | 0.64 | 2542.39 | 14.10 |
| 4 | 0.20 | 5.00 | 1.47 | 0.50 | 1839.36 | 11.02 | 1.58 | 0.62 | 2763.63 | 12.29 | 1.61 | 0.64 | 2803.34 | 14.10 |
| 5 | 1.80 | -5.00 | 1.47 | 0.50 | 2473.33 | 11.02 | 1.58 | 0.62 | 2020.94 | 12.29 | 1.61 | 0.64 | 1988.60 | 14.10 |
| 6 | 1.80 | -1.67 | 1.47 | 0.50 | 717.82 | 11.02 | 1.58 | 0.62 | 711.37 | 12.29 | 1.61 | 0.64 | 630.96 | 14.10 |
| 7 | 1.80 | 1.67 | 1.47 | 0.50 | 165.77 | 11.02 | 1.58 | 0.62 | 162.07 | 12.29 | 1.61 | 0.64 | 126.51 | 14.10 |
| 8 | 1.80 | 5.00 | 1.47 | 0.50 | 1196.69 | 11.02 | 1.58 | 0.62 | 1834.90 | 12.29 | 1.61 | 0.64 | 1527.27 | 14.10 |
| 9 | 3.40 | -5.00 | 1.47 | 0.50 | 1343.48 | 11.02 | 1.58 | 0.62 | 1229.71 | 12.29 | 1.61 | 0.64 | 1129.42 | 14.10 |
| 10 | 3.40 | -1.67 | 1.47 | 0.50 | 389.45 | 11.02 | 1.58 | 0.62 | 413.76 | 12.29 | 1.61 | 0.64 | 341.20 | 14.10 |
| 11 | 3.40 | 1.67 | 1.47 | 0.50 | 66.96 | 11.02 | 1.58 | 0.62 | 47.20 | 12.29 | 1.61 | 0.64 | 38.31 | 14.10 |
| 12 | 3.40 | 5.00 | 1.47 | 0.50 | 540.35 | 11.02 | 1.58 | 0.62 | 627.36 | 12.29 | 1.61 | 0.64 | 469.35 | 14.10 |
| 13 | 5.00 | -5.00 | 1.47 | 0.50 | 880.99 | 11.02 | 1.58 | 0.62 | 848.67 | 12.29 | 1.61 | 0.64 | 742.24 | 14.10 |
| 14 | 5.00 | -1.67 | 1.47 | 0.50 | 302.80 | 11.02 | 1.58 | 0.62 | 326.70 | 12.29 | 1.61 | 0.64 | 260.33 | 14.10 |
| 15 | 5.00 | 1.67 | 1.47 | 0.50 | 77.21 | 11.02 | 1.58 | 0.62 | 66.30 | 12.29 | 1.61 | 0.64 | 49.54 | 14.10 |
| 16 | 5.00 | 5.00 | 1.47 | 0.50 | 282.13 | 11.02 | 1.58 | 0.62 | 257.75 | 12.29 | 1.61 | 0.64 | 195.57 | 14.10 |

***Evaluation criteria***

1. Scale linking parameters

In this study, the two test forms were specified to have the same levels of difficulties. The person parameters were drawn from different normal distributions *N(0,1)* and *N(0.5, 1.5$^2$)*. Therefore the true linking parameters were *A*=1.5 and *B*=0.5. The estimated linking parameters using the item characteristic curve methods based on the three models can be compared to see which one better recovers the true linking parameters. The loss function Mean Squared Error (MSE) was used to indicate the discrepancy between the estimated values and the true values. The MSE for the linking parameter estimate $\hat{A}$ using the testlet model was defined as $MSE(\hat{A}_{testlet})$:

$$MSE(\hat{A}_{(testlet)}) = \frac{\sum_{n=1}^{N} (\hat{A}_{n(testlet)} - A)^2}{N} \tag{3.18}$$

where $\hat{A}_{n(testlet)}$ is the estimated linking parameter based on the testlet model for sample *n* and *N* is the total number of the samples, which is 50 in this simulation study. $MSE(\hat{A}_{testlet})$ can be further dissected into two parts: the variance of *A* parameter estimates $Var(\hat{A}_{testlet})$ and the bias of the *A* parameter estimates squared $Bias^2(\hat{A}_{testlet})$:

$$\begin{aligned} MSE(\hat{A}_{testlet}) &= Var(\hat{A}_{testlet}) + Bias^2(\hat{A}_{testlet}) \\ &= \frac{\sum_{n=1}^{N} (\hat{A}_{n(testlet)} - \overline{A}_{testlet})^2}{N} + (\overline{A}_{testlet} - A)^2 \end{aligned} \tag{3.19}$$

The MSE and its components were also computed for *B* estimates. The MSE and the bias for the linking parameter estimates using the scale transformation methods based on the three models can be compared. The smaller the MSE and the bias, the better the method is in recovering the true linking parameters.

<u>*2. Item and person parameters*</u>

After the linking parameters were obtained, the estimated parameter values of the new form were rescaled so that they were on the same scale as the estimated parameter values of the base form. The effectiveness of the scale linking methods using the three models can be evaluated by observing how well these methods recover the true parameter values using the loss functions Root Mean Squared Deviation (RMSD) and Mean Absolute Difference (MAD). So for a specific sample, the RMSD and MAD of the rescaled *θ* estimators using the testlet model is:

$$RMSD_{\hat{\theta}(testlet)} = \sqrt{\frac{\sum_{i=1}^{I}(\hat{\theta}_{i(testlet)} - \theta_i)^2}{I}} \tag{3.20}$$

$$MAD_{\hat{\theta}(testlet)} = \frac{\sum_{i=1}^{I}|\hat{\theta}_{i(testlet)} - \theta_i|}{I} \tag{3.21}$$

where $\theta_i$ is the true *θ* value for person *i* and *I* is the total number of the examines, which is 1000 in the simulation study. The two loss functions can also be computed for the rescaled 3-PL model and GRM *θ* estimates.

Similar formulas can be used to compute RMSD and MAD for the rescaled item parameter estimates using the testlet model and the 3-PL model. In this case, instead of summing over the examinees, the squared or absolute differences are summed over the items. Note that only the rescaled item parameter estimates- $\hat{a},\ \hat{b},\ \hat{c}$ - of the 3-PL model and the testlet model were compared in this study because the GRM has a different set of item parameters that cannot be compared easily with item parameters of the other two models.

As discussed in Chapter 2, one benefit of treating testlet-based tests with TRT models instead of the traditional unidimensional dichotomous IRT models is that the reliability statistics produced by TRT models appropriately account for the testlet effect. Therefore, besides comparing the point estimates for the item and person parameters in evaluating the performance of the three scale linking procedures, it is also useful to compare the TIFs of the person parameters generated by the three procedures. It is expected that the 3-PL model scale linking procedure should produce TIFs that are larger than those produced by the GRM and the testlet model based scale linking procedures. These values are positively biased because unidimensional IRT models ignore the testlet effect in its model specification. The TIF inflation ratios can be calculated for the GRM-estimated TIFs vs.the 3-PL model estimated-TIFs; and for the testlet model estimated-TIFs vs. the 3-PL model-estimated TIFs.

## *Results*

### *Scale linking parameters*

In the simulation design, the true linking parameters were set to be *A=1.5* and *B=0.5* for all three simulated conditions. For each condition, 50 samples were simulated and the linking parameters *A* and *B* were estimated using the three scale linking procedures. Appendix B presents the scale linking parameters estimated under the three conditions. The means of the linking parameter *A* estimates are presented in Table 3. The estimators were denoted as $\hat{A}_{\text{3PL}}$ and $\hat{B}_{\text{3PL}}$ for 3-PL model procedure-estimated values, $\hat{A}_{\text{GRM}}$ and $\hat{B}_{\text{GRM}}$ for GRM procedure-estimated values, and $\hat{A}_{\text{Testlet}}$ and $\hat{B}_{\text{Testlet}}$ for testlet model procedure-estimated values. When the variances of the testlet parameters are 0, the mean of $\hat{A}_{\text{3PL}}$ is 1.4321. This is closer to 1.5 than the mean $\hat{A}_{\text{GRM}}$: 1.4058, and the mean of $\hat{A}_{\text{Testlet}}$: 1.4231. However, the ANOVA analysis shows that the three values are not significantly different from each other since the p value is 0.2871, well above $\alpha$=0.05. The three values are still not significantly different from each when the variances of the testlet parameters are 1 (p value=0.2823). However, it can be observed that as the variances of the testlet parameters get larger, $\hat{A}_{\text{GRM}}$ and $\hat{A}_{\text{Testlet}}$ become closer to the true parameter value 1.5 as compared to $\hat{A}_{\text{3PL}}$. When the variances of the testlet parameters are 2, the mean of $\hat{A}_{\text{GRM}}$ 1.4543 and the mean of $\hat{A}_{\text{Testlet}}$ is 1.4305, as opposed to the mean of the $\hat{A}_{\text{3PL}}$: 1.3702. The differences are statistically significant according to the ANOVA analysis.

Table 3

*Linking Parameter A Estimates using 3-PL, GRM and Testlet Model Scale Linking Procedures*

| Level of Testlet Effect | Estimators | | | | ANOVA | | p of the Tukey Test | | |
|---|---|---|---|---|---|---|---|---|---|
| | Statistic | $\hat{A}_{3PL}$ | $\hat{A}_{GRM}$ | $\hat{A}_{Testlet}$ | F | p | $\hat{A}_{3PL}$ vs. $\hat{A}_{GRM}$ | $\hat{A}_{3PL}$ vs. $\hat{A}_{Testlet}$ | $\hat{A}_{GRM}$ vs. $\hat{A}_{Testlet}$ |
| Var=0 | Mean | 1.4321 | 1.4058 | 1.4231 | 1.2587 | 0.2871 | 0.2656 | 0.8530 | 0.5629 |
| | SE | 0.0118 | 0.0122 | 0.0119 | | | | | |
| Var=1 | Mean | 1.4122 | 1.4425 | 1.4293 | 1.2758 | 0.2823 | 0.2518 | 0.6422 | 0.7670 |
| | SE | 0.0143 | 0.0133 | 0.0127 | | | | | |
| Var=2 | Mean | 1.3702 | 1.4543 | 1.4305 | 5.5700* | 0.0047 | 0.0042* | 0.0559 | 0.6312 |
| | SE | 0.0175 | 0.0183 | 0.0192 | | | | | |

Note: The true parameter value *A*=1.5.

*Difference is significant at α=0.05 threshold.

Table 4 presents the summary statistics for the linking parameter estimates $\hat{B}$ computed using the three scale linking procedures. The $\hat{B}$ estimates exhibit similar trend as $\hat{A}$ estimates: when there is no variance for the testlet parameters, the $\hat{B}$ produced by the three procedures are similar, with the means being 0.4898 for $\hat{B}_{3PL}$, 0.4843 for $\hat{B}_{GRM}$, and 0.5092 for $\hat{B}_{Testlet}$. All of these values are very close to the true parameter value 0.5 and the ANOVA analysis shows that the differences of these values are not significantly different at $\alpha$=0.05. When the variances of the testlet parameters are 1, the mean of $\hat{B}_{GRM}$ 0.5269 and the mean of the $\hat{B}_{Testlet}$ 0.5256 are similar. The two values are significantly different from the mean of $\hat{B}_{3PL}$ 0.4812. However, all three values are similarly close to the true parameter value 0.5. When the variances of the testlet parameters are 2, the GRM and the testlet model procedures produce better *B* parameter estimates than the 3-PL model: the mean of $\hat{B}_{GRM}$ is 0.5038 and the mean of $\hat{B}_{Testlet}$ is 0.4994, as opposed to the mean of $\hat{B}_{3PL}$: 0.4355. The post hoc Tukey multiple comparison test shows that both the mean of $\hat{B}_{GRM}$ and the mean of $\hat{B}_{Testlet}$ are significantly different from the mean of $\hat{B}_{3PL.}$

Tables 3 and 4 results demonstrate that the three procedures produce scale linking parameter estimates that are similarly close to the true parameter values when there is no or mild testlet effect. However, when strong testlet effects exist, the testlet model and the GRM procedures produce linking parameter estimates that are closer to the true parameter values than the 3-PL IRT model scale linking procedure.

Table 4

*Linking Parameter B Estimates using 3-PL, GRM and Testlet Model Scale Linking Procedures*

| Level of Testlet Effect | Estimators | | | | ANOVA | | p of the Tukey Test | | |
|---|---|---|---|---|---|---|---|---|---|
| | Statistic | $\hat{B}_{\text{3PL}}$ | $\hat{B}_{\text{GRM}}$ | $\hat{B}_{\text{Testlet}}$ | F | p | $\hat{B}_{\text{3PL}}$ vs $\hat{B}_{\text{GRM}}$ | $\hat{B}_{\text{3PL}}$ vs. $\hat{B}_{\text{Testlet}}$ | $\hat{B}_{\text{GRM}}$ vs. $\hat{B}_{\text{Testlet}}$ |
| Var=0 | Mean | 0.4898 | 0.4843 | 0.5092 | 0.9419 | 0.3922 | 0.9550 | 0.5671 | 0.3938 |
| | SE | 0.0121 | 0.0155 | 0.0126 | | | | | |
| Var=1 | Mean | 0.4812 | 0.5269 | 0.5256 | 6.2849* | 0.0024 | 0.0062* | 0.0082* | 0.9954 |
| | SE | 0.0102 | 0.0107 | 0.0102 | | | | | |
| Var=2 | Mean | 0.4355 | 0.5038 | 0.4994 | 10.8037* | 0.0000 | 0.0002* | 0.0005* | 0.9602 |
| | SE | 0.0108 | 0.0122 | 0.0118 | | | | | |

Note: The true parameter value *B*=0.5.

*Difference is significant at α=0.05 threshold.

Table 5 presents the Mean Squared Error (MSE), the error variance and the bias of the linking parameter *A* estimates. The MSE is the sum of the error variance and the bias squared, and smaller MSE values indicate better estimation performance. When variances of the testlet parameters are 0, $\hat{A}_{\text{3PL}}$ has the smallest MSE: 0.0114, followed by MSE of $\hat{A}_{\text{Testlet}}$: 0.0128. $\hat{A}_{\text{GRM}}$ has the largest MSE 0.0162. As the variances of the testlet parameters get larger, $\hat{A}_{\text{GRM}}$ and $\hat{A}_{\text{Testlet}}$ display smaller MSEs as compared to the $\hat{A}_{\text{3PL}}$. When variances of the testlet parameters are 1, $\hat{A}_{\text{3PL}}$ has the largest MSE: 0.0177 and the MSEs of $\hat{A}_{\text{GRM}}$ and $\hat{A}_{\text{Testlet}}$ are 0.0120 and 0.0129 respectively. When variances of the testlet parameters are 2, $\hat{A}_{\text{3PL}}$ has the largest MSE: 0.0319 and the MSEs of $\hat{A}_{\text{GRM}}$ and $\hat{A}_{\text{Testlet}}$ are 0.0185 and 0.0229 respectively.

Table 5

*MSE, Error Variance and Bias of Linking Parameter A Estimates using 3-PL, GRM and Testlet Model Scale Linking Procedures*

| Condition | Estimator | Bias | Bias Squared | Error Variance | MSE* |
|---|---|---|---|---|---|
| | $\hat{A}_{\text{3PL}}$ | -0.0679 | 0.0046 | 0.0068 | 0.0114 |
| Var=0 | $\hat{A}_{\text{GRM}}$ | -0.0942 | 0.0089 | 0.0073 | 0.0162 |
| | $\hat{A}_{\text{Testlet}}$ | -0.0769 | 0.0059 | 0.0069 | 0.0128 |
| | $\hat{A}_{\text{3PL}}$ | -0.0878 | 0.0077 | 0.0100 | 0.0177 |
| Var=1 | $\hat{A}_{\text{GRM}}$ | -0.0575 | 0.0033 | 0.0086 | 0.0120 |
| | $\hat{A}_{\text{Testlet}}$ | -0.0707 | 0.0050 | 0.0079 | 0.0129 |
| | $\hat{A}_{\text{3PL}}$ | -0.1298 | 0.0168 | 0.0150 | 0.0319 |
| Var=2 | $\hat{A}_{\text{GRM}}$ | -0.0457 | 0.0021 | 0.0164 | 0.0185 |
| | $\hat{A}_{\text{Testlet}}$ | -0.0695 | 0.0048 | 0.0181 | 0.0229 |

* MSE=Bias Squared + Error Variance

Table 6 shows the Mean Squared Error (MSE), the error variance and the bias of the linking parameter *B* estimates. When there is no variance in the testlet parameters, the MSE of $\hat{B}_{\text{3PL}}$ 0.0073 and that of $\hat{B}_{\text{Testlet}}$ 0.0079 are similar. Both are smaller than the MSE of $\hat{B}_{\text{GRM}}$: 0.0120. When variances of the testlet parameters are 1, $\hat{B}_{\text{3PL}}$ and $\hat{B}_{\text{Testlet}}$ still have similar MSEs: 0.0055 and 0.0057 respectively and the MSE of $\hat{B}_{\text{GRM}}$ is 0.0063. As the variances of the testlet parameters get even larger, $\hat{B}_{\text{GRM}}$ and $\hat{B}_{\text{Testlet}}$ display smaller MSEs as compared to the $\hat{B}_{\text{3PL}}$.When the variances of the testlet parameters are 2, the MSE of $\hat{B}_{\text{Testlet}}$ is 0.0069, followed by that of $\hat{B}_{\text{GRM}}$: 0.0073 and $\hat{B}_{\text{3PL}}$: 0.0098.

Table 6

*MSE, Error Variance and Bias of Linking Parameter B Estimates using 3-PL, GRM and Testlet Model Scale Linking Procedures*

| Condition | Estimator | Bias | Bias Squared | Error Variance | MSE* |
|---|---|---|---|---|---|
| Var=0 | $\hat{B}_{\text{3PL}}$ | -0.0102 | 0.0001 | 0.0072 | 0.0073 |
| | $\hat{B}_{\text{GRM}}$ | -0.0157 | 0.0002 | 0.0118 | 0.0120 |
| | $\hat{B}_{\text{Testlet}}$ | 0.0092 | 0.0001 | 0.0078 | 0.0079 |
| Var=1 | $\hat{B}_{\text{3PL}}$ | -0.0188 | 0.0004 | 0.0051 | 0.0055 |
| | $\hat{B}_{\text{GRM}}$ | 0.0269 | 0.0007 | 0.0056 | 0.0063 |
| | $\hat{B}_{\text{Testlet}}$ | 0.0256 | 0.0007 | 0.0051 | 0.0057 |
| Var=2 | $\hat{B}_{\text{3PL}}$ | -0.0645 | 0.0042 | 0.0057 | 0.0098 |
| | $\hat{B}_{\text{GRM}}$ | 0.0038 | 0.0000 | 0.0073 | 0.0073 |
| | $\hat{B}_{\text{Testlet}}$ | -0.0006 | 0.0000 | 0.0069 | 0.0069 |

* MSE=Bias Squared + Error Variance

The error variances in Table 5 and Table 6 indicate how efficient the three procedures are in estimating the scale linking parameters. Under Condition 1 when the variances of the testlet effects are 0, the error variance is for 0.0068 for $\hat{A}_{3PL}$, for 0.0073 $\hat{A}_{GRM}$ and for 0.0069 for $\hat{A}_{Testlet.}$ The error variance is for 0.0072 for $\hat{B}_{3PL}$, for 0.0118 $\hat{B}_{GRM}$ and for 0.0078 for $\hat{B}_{Testlet.}$ The 3-PL model procedure is the most efficient method in estimating the scale linking parameters. However, the testlet model procedure has very similar error variance values as the 3-PL model procedure, indicating that it is almost as efficient as the 3-PL model in estimating the scale linking parameters. The 3-PL model is the correct and the most parsimonious model when there is no testlet effect in the test forms. Using a scale linking method based on more complex model usually leads to less efficient estimation of the scale linking parameters. In this case, the loss in efficiency by using the more complex testlet model is very small—about 2-percent for the *A* parameter and 8-percent for the *B* parameter. Therefore, while a penalty is paid for using the testlet model which is larger than the 3-PL model, the cost of inefficiency of using the testlet model when the 3-PL model is correct is minimal.

It is of interest to investigate the bias of the linking parameter estimates using the three models and study which of the models produce less biased, more accurate estimators when the testlet effects differ. Figure 7 shows the absolute values of biases of $\hat{A}_{3PL}$, $\hat{A}_{GRM}$, and $\hat{A}_{Testlet}$. When the variances of the testlet parameters are 0, $\hat{A}_{3PL}$ has the smallest bias. When the variances get larger, the biases of $\hat{A}_{GRM}$, and $\hat{A}_{Testlet}$ get smaller while the bias of $\hat{A}_{3PL}$ gets larger. As the variances of the testlet parameters reach 2, $\hat{A}_{\text{3PL}}$ displays much larger bias than $\hat{A}_{\text{GRM}}$, and $\hat{A}_{\text{Testlet}}$. Figure 8 shows the absolute values of bias of $\hat{B}_{\text{3PL}}$, $\hat{B}_{\text{GRM}}$, and $\hat{B}_{\text{Testlet.}}$ The three estimates have similar levels of bias when variances of the testlet parameters are 0 and 1, but when the variances are 2, $\hat{B}_{\text{3PL}}$ has much larger bias than that of $\hat{B}_{\text{GRM}}$ and $\hat{B}_{\text{Testlet}}$.

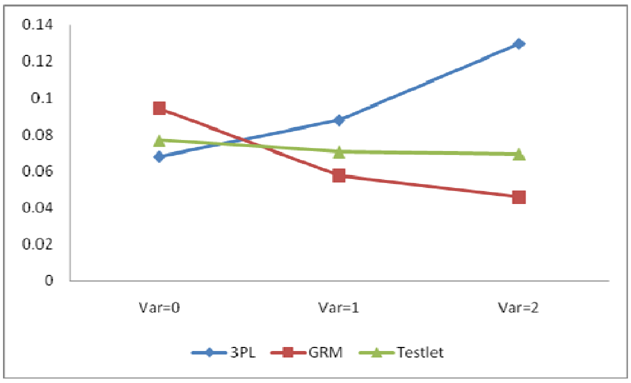


*Figure 7.* Absolute values of bias of the linking parameter *A* estimates using 3-PL, GRM and Testlet model scale linking procedures

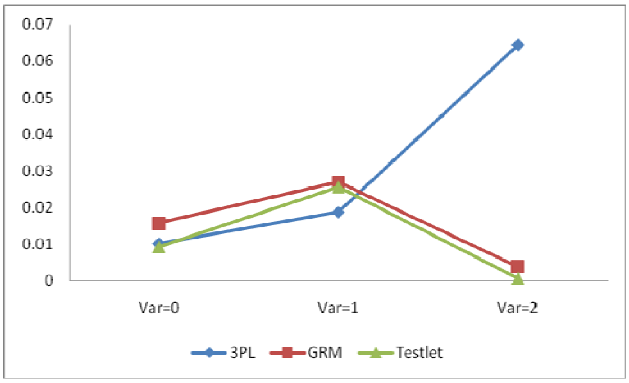


*Figure 8.* Absolute values of bias of the linking parameter *B* estimates using 3-PL, GRM and Testlet model scale linking procedures

The results indicate that when the variances of the testlet parameters are 0 and there is no testlet effect, the scale linking parameter estimates produced by the three procedures are similar. The linking parameter estimates of the 3-PL model procedure has similar MSE and bias as those of the GRM and the testlet model, the differences of the means of the estimated linking parameter values using the three procedures are not significantly different. When the variances of the testlet parameters are 1 and there is some degree of testlet effect, the linking parameter $A$ estimated using the GRM and the testlet model procedures is a little better than that of the 3-PL model procedure, with lower MSE and smaller bias, however, the differences of the means of $\hat{A}_{\text{3PL}}$, $\hat{A}_{\text{GRM}}$, and $\hat{A}_{\text{Testlet}}$ are not statistically significant. The means of $\hat{B}_{\text{3PL}}$, $\hat{B}_{\text{GRM}}$, and $\hat{B}_{\text{Testlet}}$ are also similarly close to the true parameter value 0.5 and these estimates share similar MSEs and biases. When the variances of the testlet parameters are 2 and the testlet effects are large, the GRM procedure and the testlet model procedure-estimated linking parameters perform much better than those estimated using the 3-PL model procedure: the means of linking parameters estimated using the GRM and the testlet model procedures are much closer to the true parameter values than those estimated using the 3-PL model procedure. The GRM procedure and the testlet model procedure-estimated linking parameters also have much smaller MSEs and biases than the 3-PL model procedure estimated values. It is evident that as the variances of the testlet parameters get larger, the testlet model procedure and the GRM model procedure perform better in estimating the scale linking parameters than the 3-PL model procedure.

It should also be noted that under each of the three simulation conditions, the GRM procedure and the testlet model procedure perform quite similarly in estimating the scale linking parameters. The differences of the means of the linking parameter estimates are not significantly different and they have similar biases and MSEs.

***Person parameters***

After scale linking parameters were estimated, the $\theta$ parameter estimates of the examinees taking the new forms were transformed onto the scale of the $\theta$ parameter estimates of the examinees taking the base form using Formula (3.1). Appendix C provides detailed information about the computed evaluation criteria for person parameter $\theta$ estimates. Specifically, correlations between $\theta$ estimates and true $\theta$ values, the RMSD and MAD loss functions of $\theta$ estimates and mean TIF estimates for each sample under each of the three conditions are included in the appendix.

Table 7 presents the correlations between the estimated θ values using the three models and the true θ values for examinees taking the new form. Higher correlation indicates better estimation performance for the model. The three models produce very similar correlations under each of the three testlet effect conditions. For example, when the variances of the testlet parameters are 0, the mean correlation of the 3-PL model-estimated person parameter and the true person parameter $r(\hat{\theta}_{3PL}, \theta)$=0.9039, the mean correlation of the GRM-estimated person parameter and the true person parameter $r(\hat{\theta}_{GRM}, \theta)$=08934, and the mean correlation of the testlet model-estimated person parameter and the true person parameter $r(\hat{\theta}_{Testlet}, \theta)$=0.9037. However, A Post Hoc Tukey multiple comparison shows that while there is no significant difference for $r(\hat{\theta}_{3PL}, \theta)$ vs. $r(\hat{\theta}_{Testlet}, \theta)$, the differences for $r(\hat{\theta}_{3PL}, \theta)$ vs. $r(\hat{\theta}_{GRM}, \theta)$, and $r(\hat{\theta}_{GRM}, \theta)$ vs. $r(\hat{\theta}_{Testlet}, \theta)$ are statistically significant at $a$=0.05 threshold under all three simulation conditions.

Table 7

*Correlations between the True Person Parameters and the Estimated Person Parameters of Examinees Taking the New Form*

| Level of Testlet Effect | | Correlation | | | ANOVA | | p of the Tukey Test | | |
|---|---|---|---|---|---|---|---|---|---|
| | Statistic | $r(\hat{\theta}_{3PL}, \theta)$ | $r(\hat{\theta}_{GRM}, \theta)$ | $r(\hat{\theta}_{Testlet}, \theta)$ | *F* | *p* | $r(\hat{\theta}_{3PL}, \theta)$ *vs.* $r(\hat{\theta}_{GRM}, \theta)$ | $r(\hat{\theta}_{3PL}, \theta)$ *vs.* $r(\hat{\theta}_{Testlet}, \theta)$ | $r(\hat{\theta}_{GRM}, \theta)$ *vs.* $r(\hat{\theta}_{Testlet}, \theta)$ |
| Var=0 | Mean | 0.9039 | 0.8934 | 0.9037 | 27.3572* | 0.0000 | 0.0000* | 0.9895 | 0.0000* |
| | SE | 0.0011 | 0.0012 | 0.0012 | | | | | |
| Var=1 | Mean | 0.8693 | 0.8617 | 0.8699 | 13.6275* | 0.0000 | 0.0000* | 0.9445 | 0.0000* |
| | SE | 0.0012 | 0.0013 | 0.0012 | | | | | |
| Var=2 | Mean | 0.8366 | 0.8305 | 0.8385 | 6.6923* | 0.0017 | 0.0217* | 0.7002 | 0.0018* |
| | SE | 0.0016 | 0.0017 | 0.0016 | | | | | |

*Difference is statistically significant at *a=0.05*

Table 8 shows mean RMSD and MAD of the rescaled $\theta$ estimates for the three procedures. Smaller mean RMSD and MAD indicate better performance in parameter estimation and scaling. Figure 9 and Figure 10 present mean MAD and RMSD of the rescaled $\theta$ estimates in graphic form. These figures show that the MAD and RMSD of the rescaled person parameter estimates of the 3-PL model procedure and the testlet model procedure are similar as the two curves almost overlap each other and they are both smaller than MAD and RMSD of the person parameter estimates of the GRM procedure under all three simulation conditions.

Table 8

*MAD and RMSD of the Rescaled Person θ Parameter Estimators for the Examinees Taking the New Form*

| Level of Testlet Effect | | MAD | | | RMSD | | |
|---|---|---|---|---|---|---|---|
| | | $\hat{\theta}_{3PL}$ | $\hat{\theta}_{GRM}$ | $\hat{\theta}_{Testlet}$ | $\hat{\theta}_{3PL}$ | $\hat{\theta}_{GRM}$ | $\hat{\theta}_{Testlet}$ |
| Var=0 | Mean | 0.5062 | 0.5332 | 0.5068 | 0.6457 | 0.6793 | 0.6463 |
| | SE | 0.0034 | 0.0033 | 0.0033 | 0.0044 | 0.0044 | 0.0043 |
| Var=1 | Mean | 0.5932 | 0.6083 | 0.5911 | 0.7502 | 0.7700 | 0.7480 |
| | SE | 0.0024 | 0.0024 | 0.0024 | 0.0030 | 0.0032 | 0.0031 |
| Var=2 | Mean | 0.6604 | 0.6701 | 0.6567 | 0.8307 | 0.8433 | 0.8266 |
| | SE | 0.0032 | 0.0033 | 0.0031 | 0.0037 | 0.0038 | 0.0036 |

Table 7 and Table 8 indicate that the 3-PL model procedure and the testlet model procedure produce comparable rescaled person estimates and both perform better than the GRM procedure. This is reasonable since the testlet model is based on the 3-PL model and the only difference between the two is that the testlet model contains a set of testlet parameters. On the other hand, the GRM model is a different model and less information is utilized in its estimation of the person parameters. Therefore the person parameter estimates can be different from those of the 3-PL model and the testlet model. However, Table 7 and Table 8 also indicate that while MAD and RMSD of the $\theta$ estimates for the GRM procedure are different from those of the $\theta$ estimates for the other two models, the differences are not very large. The three procedures perform similarly in person parameter estimation and scaling.

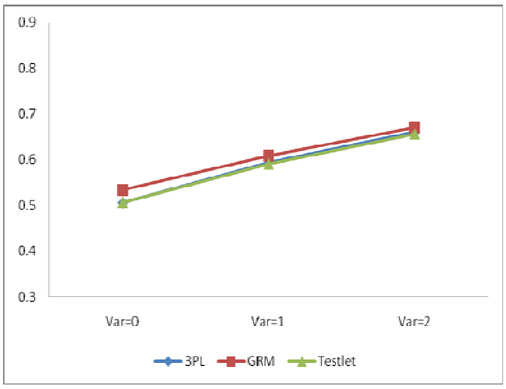


*Figure 9*. Mean MAD of the rescaled person parameters estimates for examinees taking the new form

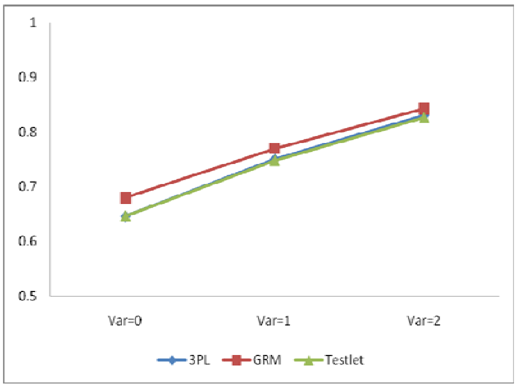


*Figure 10.* Mean RMSD of the rescaled person parameters estimates for examinees taking the new form

Each sample contained 1000 simulated subjects and the test information function (TIF) was estimated for each examinee during the model estimation process. Three TIFs were computed for each person as three models were fit to the data. Since the 3-PL model doesn't take into account LID caused by the testlet format, the TIF estimated by the 3-PL model is usually inflated as compared to the TIF estimated using the GRM and the testlet model. Table 9 presents the mean TIF values and the mean TIF ratios under each simulation condition and Figure 11 is the graphic representation of the mean TIF information. From the information, the following can be learned:

Table 9

*Mean TIF and Mean TIF Ratio of the Examinees Taking the New Form*

| | | Mean TIF | | | Mean TIF Ratio | |
|---|---|---|---|---|---|---|
| | | $TIF_{3PL}$ | $TIF_{GRM}$ | $TIF_{Testlet}$ | $\frac{TIF_{3PL}}{TIF_{GRM}}$ | $\frac{TIF_{3PL}}{TIF_{Testlet}}$ |
| Var=0 | Mean | 5.8497 | 5.2898 | 4.9762 | 1.1162 | 1.1692 |
| | SD | 0.4943 | 0.4386 | 0.3798 | 0.0678 | 0.0103 |
| Var=1 | Mean | 5.2984 | 4.0046 | 3.9150 | 1.3347 | 1.3463 |
| | SD | 0.3553 | 0.2361 | 0.2145 | 0.0389 | 0.0434 |
| Var=2 | Mean | 4.9471 | 3.2689 | 3.2035 | 1.5266 | 1.5378 |
| | SD | 0.3238 | 0.1871 | 0.1531 | 0.0431 | 0.0553 |

1) The TIF values estimated by the 3-PL model are consistently higher than those estimated by the GRM and the testlet model. The differences are small when

the testlet effect is nonexistent and gets increasingly large as the testlet effect becomes stronger. When the variances of the testlet parameters are 0, the average of the mean TIFs for each sample is 5.8497 for the 3-PL model, 5.2898 for the GRM and 4.9762 for the testlet model. The 3-PL model produces has higher mean TIF values than the GRM and the testlet model, but the differences are not very large as the mean $\frac{TIF_{3PL}}{TIF_{GRM}}$ is 1.1162 and the mean $\frac{TIF_{3PL}}{TIF_{Testlet}}$ is 1.1692. As the variances of the testlet parameters get larger, the gaps between the mean TIF values produced by the 3-PL model vs. the GRM and the testlet model get larger: when the variances of the testlet parameters are 1, the mean $\frac{TIF_{3PL}}{TIF_{GRM}}$ is 1.3347 and the mean $\frac{TIF_{3PL}}{TIF_{Testlet}}$ is 1.3463 and when the variances of the testlet parameters are 2, the mean $\frac{TIF_{3PL}}{TIF_{GRM}}$ is 1.5266 and the mean $\frac{TIF_{3PL}}{TIF_{Testlet}}$ is 1.5378.

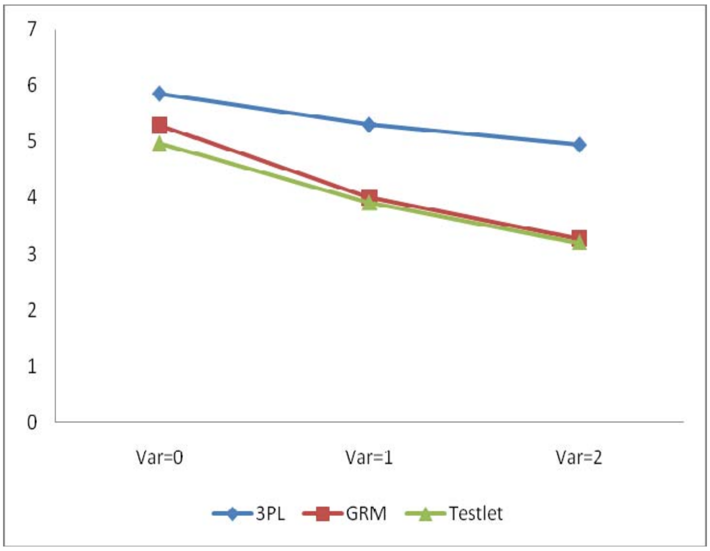


*Figure 11.* Mean TIF of the examinees taking the new form

2) All of the estimated mean TIF curves of the three models display a downward trend as the variances of the testlet parameters increase. When the testlet effects are large and the items within a testlet are locally dependent, the information that a particular item provides about the person's ability $\theta$ may not be unique. Holding other factors constant, this information redundancy may lead to less accurate and stable person parameter estimation. Since the TIF is negatively related to the standard error of measurement for the person parameter estimates, larger testlet effects eventually result in smaller TIF. This is true regardless which model is used to perform parameter estimation and scale linking.

3) The TIFs estimated by the GRM and those by the testlet model are very close, especially when the testlet effects are large. When the variances of the testlet parameters are 1, the average of the mean TIFs of the GRM is 4.0046 and the average of the mean TIF of the testlet model is 3.9150. When the variances of the testlet parameters are 2, the average of the mean TIFs of the GRM is 3.2689 and the average of the mean TIF of the testlet model is 3.2035. The testlet model performs quite similarly as the GRM in capturing the loss of the test information due to the testlet effect in this simulation study.

As a result, while the TIFs estimated by the GRM and the testlet model are very similar, they are smaller than those estimated by the 3-PL model. In this simulation study, the 3-PL model is the most parsimonious model to measure TIFs when there is no testlet effect (the true testlet parameter variances are 0). The GRM and the testlet model would still attempt to measure and account for the testlet effects, which may be present in some samples due to sampling errors. Thus the GRM and the testlet model would sometimes slightly underestimate the TIF in the population under such situations. However, when the testlet effects exist and especially when they are strong, the testlet model and the GRM are superior to the 3-PL model in accurately estimating TIFs due to their abilities to account for the loss of information caused by LID.

As far as the person parameter estimation is concerned, researchers are often interested in two pieces of information: the point estimator $\theta$ and the TIF. The simulation study shows that the testlet model procedure produces similar $\theta$ estimates as the 3-PL model procedure. This is reasonable considering the similarities between the 3-PL IRT model and the 3-PL testlet model. The two models produce slightly better $\theta$ estimates than the GRM under the three simulated conditions in this simulation study because the GRM utilizes less information from the data than the 3-PL model and the testlet model. However, the testlet model procedure performs similarly as the GRM procedure in estimating TIF in the simulation study. Both perform better than the 3-PL model procedure which overestimates TIF when the testlet effects are evident in a test.

***Item parameters***

After scale linking parameters were estimated, the item parameter estimates of the new forms were transformed onto the scale of the base forms using Formulae (3.2) and (3.3). The performance of the testlet model scale linking procedure can also be evaluated by examining how well it recovers the true item parameters as compared to the 3-PL model scale linking procedure. The comparison was done on the rescaled discrimination parameter estimates $\hat{a}$, difficulty parameter estimates $\hat{b}$, and guessing parameter estimates $\hat{c}$ for the 3-PL model procedure and the testlet model procedure in the study. The GRM procedure was excluded from the item parameter estimation and scaling comparison since it has category related step and difficulty parameters instead of item paramters. Appendix D provides detailed information about the computed evaluation criteria for item parameter estimates on the base forms and the new forms. Specifically, correlations between item parameter estimates and true item parameter values and the RMSD and MAD loss functions of the item parameter estimates are included in this appendix.

For each sample, correlations between the estimated item parameters and the true item parameters can be computed. Table 9 presents the summary statistics of the correlations for the new test forms. According to the table, the mean correlation for the item parameters estimated using the testlet model procedure and the 3-PL model procedure are very similar under the condition when the variances of the testlet parameters are 0. However, when the testlet effects get larger, the mean correlation for the item parameters estimated using the testlet model procedure becomes increasingly larger than those for the item parameters estimated using the 3-PL model procedure. This is especially true for *a* discrimination parameter estimates. When the variances of the testlet parameter are 0, mean $r(\hat{a}_{3PL}, a)$ is 0.9075, and mean $r(\hat{a}_{testlet}, a)$ is 0.9078. They are very close. When the variances of the testlet parameter are 1, mean $r(\hat{a}_{3PL}, a)$ is 0.8585, and mean $r(\hat{a}_{testlet}, a)$ is 0.8822. When the variances of the testlet parameter are 2, mean $r(\hat{a}_{3PL}, a)$ is 0.8117 and mean $r(\hat{a}_{testlet}, a)$ is 0.8599. When the testlet effects get larger, the testlet model scale linking procedure produces discrimination parameter estimates that are better correlated with the true discrimination parameter values than the discrimination parameter estimates of the 3-PL model procedure. Table 10 also shows that both models are best at estimating *b* parameters, with the mean correlations ranging from 0.9387 to 0.9583 between these two models, followed by *a* parameter estimation, with mean correlations ranging from 0.8117 to 0.9075. Both models are not good at estimating *c* parameters, with correlations ranging from 0.3228 to 0.3491 for the 3-PL model and 0.3305 to 0.4105 for the testlet model.

Table 10

*Correlations between the Estimated Item Parameters and the True Item Parameters for the New Form*

| | | 3PL | | | Testlet | | |
|---|---|---|---|---|---|---|---|
| | | $r(\hat{a}_{3PL}, a)$ | $r_l(\hat{b}_{3PL}, b)$ | $r(\hat{c}_{3PL}, c)$ | $r(\hat{a}_{testlet}, a)$ | $r(\hat{b}_{testlet}, b)$ | $r(\hat{c}_{testlet}, c)$ |
| var=0 | Mean | 0.9075 | 0.9525 | 0.3228 | 0.9078 | 0.9583 | 0.3305 |
| | SD | 0.0492 | 0.0183 | 0.1479 | 0.0501 | 0.0150 | 0.1420 |
| var=1 | Mean | 0.8585 | 0.9512 | 0.3371 | 0.8822 | 0.9591 | 0.3673 |
| | SD | 0.0577 | 0.0187 | 0.1676 | 0.0495 | 0.0172 | 0.1519 |
| var=2 | Mean | 0.8117 | 0.9387 | 0.3491 | 0.8599 | 0.9480 | 0.4105 |
| | SD | 0.0756 | 0.0324 | 0.1796 | 0.0766 | 0.0236 | 0.1922 |

Table 11 presents summary statistics of MAD and RMSD for the rescaled item parameter estimates of the new test form. Figure 12 and Figure 13 are the graphic representation of the mean MAD and mean RMSD respectively. These statistics confirm the finding in Table 10 that the tesltet model procedure performs consistently better than the 3-PL model procedure in estimating the item parameters *a*, *b* and *c*. Figures 12 and 13 also show that as the testlet effect increases, the MAD and the RMSD statistics of the item parameter estimates increase for both the 3-PL model and the testlet model. The testlet effect has an impact on the accuracy of the item parameter estimation and scale linking for both procedures.

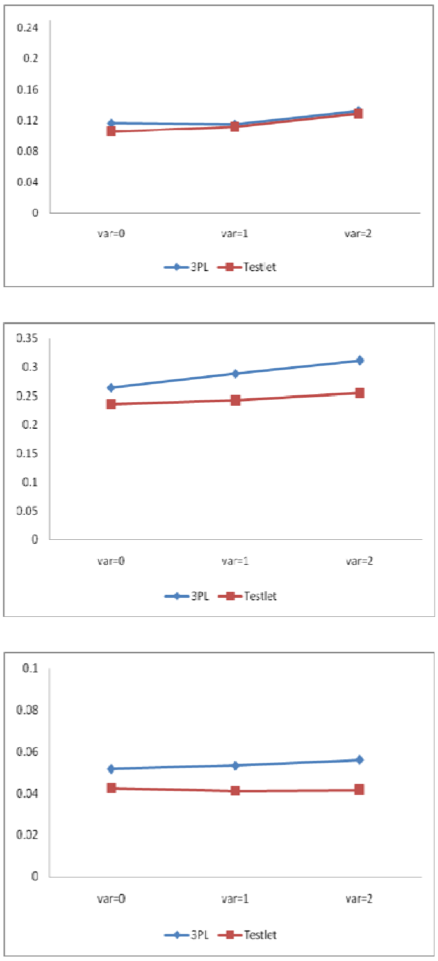


*Figure 12.* Mean MAD of item parameter estimates (from top to bottom: $\hat{a}$ , $\hat{b}$ , $\hat{c}$ )

Table 11

*MAD and RMSD of the Rescaled Item Parameter Estimates of the New Form*

| | | MAD | | | | | | RMSD | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 3PL | | | Testlet | | | 3PL | | | Testlet | | |
| | | $\hat{a}$ | $\hat{b}$ | $\hat{c}$ | $\hat{a}$ | $\hat{b}$ | $\hat{c}$ | $\hat{a}$ | $\hat{b}$ | $\hat{c}$ | $\hat{a}$ | $\hat{b}$ | $\hat{c}$ |
| Var=0 | Mean | 0.1162 | 0.2639 | 0.0517 | 0.1057 | 0.2355 | 0.0424 | 0.1467 | 0.3449 | 0.0649 | 0.1393 | 0.3067 | 0.0531 |
| | SD | 0.0304 | 0.0435 | 0.0077 | 0.0222 | 0.0373 | 0.0056 | 0.0360 | 0.0554 | 0.0083 | 0.0293 | 0.0512 | 0.0065 |
| Var=1 | Mean | 0.1148 | 0.2886 | 0.0533 | 0.1118 | 0.2425 | 0.0413 | 0.1501 | 0.3689 | 0.0672 | 0.1482 | 0.3156 | 0.0519 |
| | SD | 0.0219 | 0.0530 | 0.0072 | 0.0217 | 0.0385 | 0.0049 | 0.0293 | 0.0713 | 0.0087 | 0.0321 | 0.0591 | 0.0062 |
| Var=2 | Mean | 0.1318 | 0.3117 | 0.0562 | 0.1288 | 0.2554 | 0.0417 | 0.1732 | 0.3978 | 0.0697 | 0.1688 | 0.3329 | 0.0518 |
| | SD | 0.0268 | 0.0455 | 0.0070 | 0.0308 | 0.0405 | 0.0060 | 0.0413 | 0.0644 | 0.0084 | 0.0422 | 0.0616 | 0.0074 |

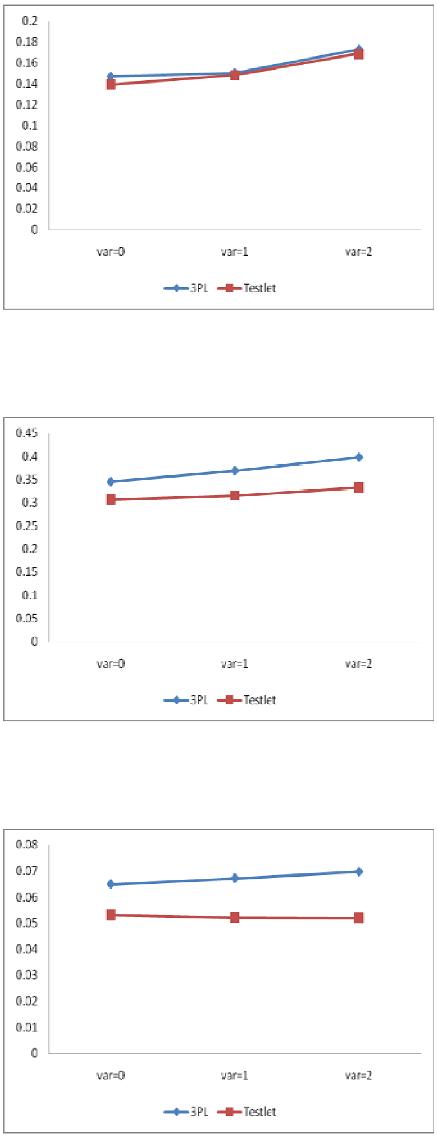


*Figure 13.* Mean RMSD of item parameter estimates (from top to bottom: $\hat{a}$ , $\hat{b}$ , $\hat{c}$ )

# Chapter 4 Real Data Analysis

To illustrate the application of the proposed testlet model scale linking procedure with real data, the operational data from 2004-2005 ACCESS for ELLs® assessment developed by the Center for Applied Linguistics and World-Class Instructional Design and Assessment (WIDA) Consortium was used.

## *ACCESS for ELLs® Assessment*

ACCESS for ELLs® stands for Assessing Comprehension and Communication in English State-to-State for English Language Learners (ELLs). It is an English language proficiency assessment given annually to students in kindergarten through Grade 12 who have been identified as ELLs. "The results of this test are used to monitor student progress in acquiring English for the academic environment, to plan support for continuing English language development, and to satisfy legal requirements for assessment and accountability." (WIDA, 2008, p.5). The assessment is aligned with the WIDA English Language Proficiency (ELP) standards for ELLs, which state expectations for student performance at six levels (1-entering; 2-beginning; 3-developing; 4-expanding; 5-bridging and 6-reaching) of the language development continuum. A set of model performance indicators (MPI) are used to illustrate the ELP standards in different content areas. The standards are further divided into five grade clusters: PreK-K, 1-2, 3-5, 6-8 and 9-12 and four domains: Reading, Listening, Writing and Speaking. Moreover, to make the tests scores reliable and appropriate for as many individuals as possible, the test items are presented in three tiers-A, B and C- for each grade level cluster. The student's teacher makes the decision as to which tier to place the student based on the information they have about his/her English proficiency level. Figure 14 shows how the different tiers map to the English proficiency levels. Part of the adjacent tiers overlap which allows each tier to measure a common proficiency scale. "You can think of ACCESS for ELLs® as one enormous test divided into multiple parts, each designed for students within a particular grade level cluster and range of proficiency levels." (WIDA, 2008, p. 8) This design allows the delivery of test results that can be linked to a common scale across grades and tiers.

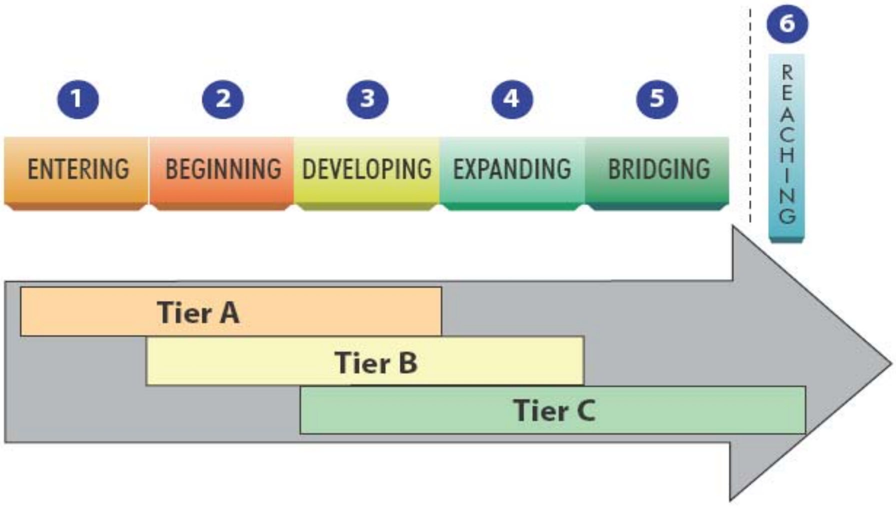


*Figure 14.* The proficiency levels of WIDA ELP standards: figure taken from "WIDA Annual Technical Report for ACCESS for ELLs English Language Proficiency Test, Series 102, 2006-2007 Administration" (MacGreger, Louguit, Ryu, Li, & Kenyon, 2008)

The test design makes ACCESS for ELLs a good candidate for the application of the testlet model scale linking method. Test takers of different groups and tier levels are put onto the same scale using the common items in the overlapping part of the adjacent tiers. This is consistent with the CINEG design under which the characteristic curve scale linking method can be applied. Furthermore, items on the ACCESS ELLs are all multiple choice questions arranged in "thematic folders", which are collections of "test items at consecutive proficiency levels organized along a common content topic" (WIDA, 2008). These folders are de facto testlets. Readers can refer to Figure 1 in Chapter 2, which is one sample folder taken from the Tier B form of the grade cluster 6-8 Reading test, for an example of the testlet. Each folder/testlet is assigned to a tiered test form for a certain grade level. For instance, a folder for tier C typically contains items with difficulty levels that correspond to Level 3, 4, 5 and 6 Model Performance Indicators. With the testlet format being used for all items on ACCESS for ELLs, the testlet model can be an option to calibrate the tests.

The ACCESS for ELLs® contains a comprehensive set of test forms that target different grade clusters, English proficiency levels and content domains. 300,000 ELL students over 15 states took the test in 2004-2005. Two Reading test forms that are adjacent to each other in tier levels: the Tier B form and the Tier C form of grade cluster 3-5 were selected for the real data analysis. Student data collected from one state[1] was used. 1663 ELL students took the Tier B form and 1418 ELL students took the Tier C form. Each of these two forms contains 30 items

[1] The name of the state is not disclosed in this study due to the agreement with WIDA consortium for using the data.

embedded in 8 folders. Altogether the two forms share five common folders, which contain 19 items. The common folders are positioned differently in the two test forms. (Appendix E shows the item structure and the common folders of the two test forms.) The items on the two test forms were rearranged so that the common folders were identically positioned in the two test forms. The new item structures for the two forms are shown in Table 12. After the adjustment, the last five folders, Folders 4 to 8 of the two forms are the common folders. They are highlighted in Table 12.

### *Yen's $Q_3$ Analysis of Testlet Effects and Factor Analysis*

While it is possible to use the proposed testlet model scale linking method for this test because the test design and the item format allow such an application, the testlet model may not be the most parsimonious model for this particular set of tests. As demonstrated in the simulation study, the scale linking method based on unidimensional dichotomous IRT model such as the 3-PL model might be a better choice for testlet-based test forms that exhibit no or minor local item dependence effects. With real data, since we have no knowledge about the extent of the testlet effect beforehand as we do with simulated data, it is a good practice to perform some form of LID tests to check how strong the testlet effects are for the items within each test before we proceed with model calibration.

Table 12

*Rearranged Item and Folder Numbers for Tier B Form and Tier C Form*

| | | Tier B Form | | Tier C Form | |
|---|---|---|---|---|---|
| New Folder Number | New Item Number | Original Folder Number | Original Item Number | Original Folder Number | Original Item Number |
| Folder 1 | **1** | Folder 1 | **1** | Folder 1 | **1** |
| | **2** | | **2** | | **2** |
| | **3** | | **3** | | **3** |
| Folder 2 | **4** | Folder 3 | **9** | Folder 2 | **4** |
| | **5** | | **10** | | **5** |
| | **6** | | **11** | | **6** |
| | **7** | | **12** | | **7** |
| | **8** | | **13** | | **8** |
| Folder 3 | **9** | Folder 8 | **28** | Folder 6 | **21** |
| | **10** | | **29** | | **22** |
| | **11** | | **30** | | **23** |
| Folder 4 | **12** | Folder 2 | **4** | Folder 3 | **9** |
| | **13** | | **5** | | **10** |
| | **14** | | **6** | | **11** |
| | **15** | | **7** | | **12** |
| | **16** | | **8** | | **13** |
| Folder 5 | **17** | Folder 4 | **14** | Folder 4 | **14** |
| | **18** | | **15** | | **15** |
| | **19** | | **16** | | **16** |
| Folder 6 | **20** | Folder 5 | **17** | Folder 7 | **24** |
| | **21** | | **18** | | **25** |
| | **22** | | **19** | | **26** |
| Folder 7 | **23** | Folder 6 | **20** | Folder 5 | **17** |
| | **24** | | **21** | | **18** |
| | **25** | | **22** | | **19** |
| | **26** | | **23** | | **20** |
| Folder 8 | **27** | Folder 7 | **24** | Folder 8 | **27** |
| | **28** | | **25** | | **28** |
| | **29** | | **26** | | **29** |
| | **30** | | **27** | | **30** |

Chen & Thissen(1997) found that while both $Q_3$ and $G^2$detect LID with some power, $Q_3$ outperforms $G^2$ for the most part. Therefore, Yen's $Q_3$ analysis was adopted in the study. The $Q_3$ statistics were computed for each pair of items within each folder for both forms. The mean and standard deviation of the $Q_3$ statistics for each folder are displayed in Table 13.

Table 13

*$Q_3$ for each Folder in the Two Forms*

| | | Folder 1 | Folder 2 | Folder 3 | Folder 4 | Folder 5 | Folder 6 | Folder 7 | Folder 8 |
|---|---|---|---|---|---|---|---|---|---|
| Tier B Form | Mean | 0.0412 | -0.0060 | -0.0121 | -0.0194 | 0.0089 | 0.0059 | -0.0028 | 0.0621 |
| | SD | 0.0441 | 0.0493 | 0.0600 | 0.0346 | 0.0381 | 0.0332 | 0.0364 | 0.0369 |
| Tier C Form | Mean | 0.0439 | -0.0094 | 0.0124 | 0.0183 | -0.0077 | 0.0248 | 0.0270 | 0.0524 |
| | SD | 0.0596 | 0.0408 | 0.0548 | 0.0634 | 0.0410 | 0.0269 | 0.0391 | 0.0735 |

*The expected value of $Q_3$ is *-1/(30-1)=-0.035*

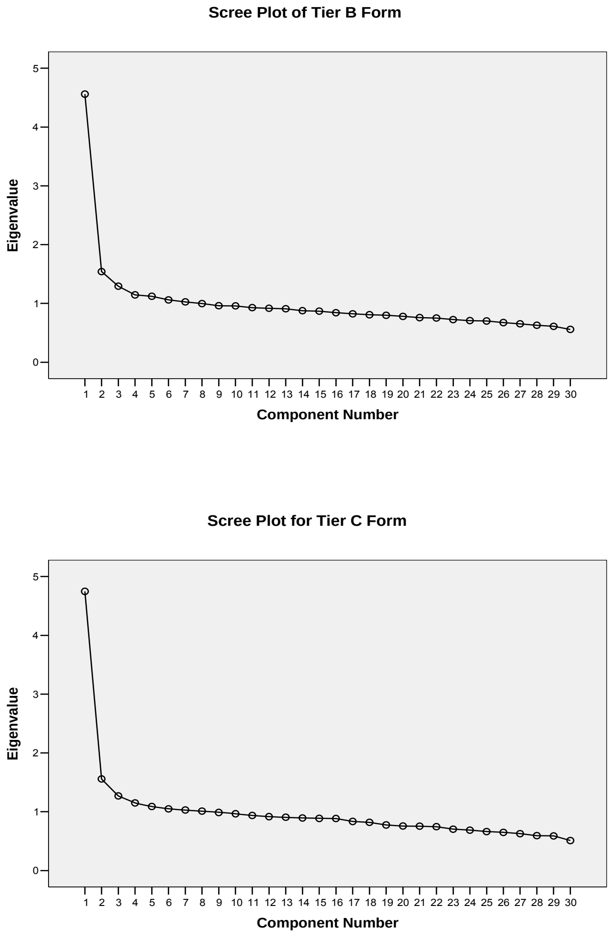
As demonstrated in the table, the mean $Q_3$ statistics are very small for all folders in both forms. The local item independence assumption of the 3-PL model doesn't appear to be violated, through the lens of the $Q_3$ statistic.

The factor analysis was also performed to study the unidimensionality assumption of the IRT models. The testlet model can still be considered as a special form of unidimensional IRT model since the LID effects it accounts for are limited within the testlet level. A factor analysis may reveal if any nuisance factors systematically affect examinees' performance on the tests. Table 14 shows the eigenvalues of the 10 largest components according to the principle component analysis. The analysis yielded 7 components with eigenvalues larger than 1 for the Tier B form and 8 components for the Tier C form. For both forms, the first component accounts for over 15% of the variance; and the eigenvalue of the first component is about three times as large as the eigenvalue of the second component while the differences of the remaining successive eigenvalues are very small. This indicates that there is one dominant dimension in the two test forms. Figure 15 shows the scree plots of the factor analysis for the two test forms. The plots show a clear "L" shape with the turning point located at the second dimension. The unidimensional model can be applied in this case.

Table 14

*Eigenvalues of the Components in the Two Forms*

| Component | Eigenvalues (Tier B Form) | | | Eigenvalues (Tier C Form) | | |
|---|---|---|---|---|---|---|
| | Total | % of Variance | Cumulative % | Total | % of Variance | Cumulative % |
| 1 | 4.561 | 15.203 | 15.203 | 4.748 | 15.827 | 15.827 |
| 2 | 1.542 | 5.14 | 20.343 | 1.557 | 5.192 | 21.018 |
| 3 | 1.293 | 4.311 | 24.654 | 1.269 | 4.231 | 25.249 |
| 4 | 1.146 | 3.819 | 28.473 | 1.151 | 3.836 | 29.085 |
| 5 | 1.122 | 3.739 | 32.212 | 1.089 | 3.628 | 32.713 |
| 6 | 1.061 | 3.536 | 35.748 | 1.049 | 3.497 | 36.21 |
| 7 | 1.026 | 3.421 | 39.169 | 1.03 | 3.433 | 39.644 |
| 8 | 0.998 | 3.328 | 42.497 | 1.012 | 3.373 | 43.017 |
| 9 | 0.961 | 3.204 | 45.701 | 0.987 | 3.292 | 46.309 |
| 10 | 0.959 | 3.195 | 48.896 | 0.966 | 3.219 | 49.528 |



Scree Plot of Tier B Form

Scree Plot for Tier C Form

*Figure 15.* Scree plots produced by the principle component analysis

## *Model Estimation*

BILOG-MG, PARSCALE and WinBUGS were used to fit the 3-PL model, the GRM and the testlet model respectively. The prior distributions specified in the simulation study were employed here for the estimation of the testlet model parameters. 12000 iterations were run and iterations 6001 to 12000 were used to estimate the model parameters. Table 15 and Table 16 present the 3-PL model and the testlet model-estimated item parameters for the Tier B form and the Tier C form respectively.

Table 15

*Tier B Form Item Parameter Estimates*

| | 3-PL Model | | | Testlet Model | | |
|---|---|---|---|---|---|---|
| | $\hat{a}$ | $\hat{b}$ | $\hat{c}$ | $\hat{a}$ | $\hat{b}$ | $\hat{c}$ |
| Item 1 | 0.655 | -2.865 | 0.270 | 0.591 | -3.425 | 0.254 |
| Item 2 | 1.054 | -1.358 | 0.159 | 1.222 | -1.495 | 0.147 |
| Item 3 | 0.633 | -0.320 | 0.174 | 0.639 | -0.457 | 0.148 |
| Item 4 | 1.269 | -2.484 | 0.254 | 1.267 | -2.690 | 0.228 |
| Item 5 | 1.742 | 1.699 | 0.200 | 1.570 | 2.028 | 0.203 |
| Item 6 | 2.661 | 1.200 | 0.333 | 2.491 | 1.323 | 0.325 |
| Item 7 | 2.861 | 0.881 | 0.247 | 2.842 | 0.884 | 0.223 |
| Item 8 | 1.784 | 0.885 | 0.150 | 1.704 | 0.926 | 0.137 |
| Item 9 | 1.559 | -1.105 | 0.219 | 1.747 | -1.251 | 0.178 |
| Item 10 | 1.157 | 1.526 | 0.143 | 1.273 | 1.665 | 0.152 |
| Item 11 | 2.568 | 1.641 | 0.281 | 2.563 | 1.780 | 0.262 |
| Item 12 | 1.950 | 0.946 | 0.149 | 1.894 | 1.014 | 0.132 |
| Item 13 | 2.197 | 1.024 | 0.442 | 2.586 | 1.081 | 0.432 |
| Item 14 | 2.436 | 0.173 | 0.178 | 2.388 | 0.074 | 0.129 |
| Item 15 | 2.203 | 0.593 | 0.194 | 2.140 | 0.637 | 0.180 |
| Item 16 | 1.635 | 0.363 | 0.178 | 1.259 | 0.337 | 0.144 |
| Item 17 | 2.290 | -2.785 | 0.247 | 2.703 | -3.248 | 0.224 |
| Item 18 | 1.162 | -0.203 | 0.149 | 1.325 | -0.227 | 0.146 |
| Item 19 | 1.585 | 1.326 | 0.188 | 1.967 | 1.399 | 0.181 |
| Item 20 | 1.209 | 0.511 | 0.154 | 1.231 | 0.486 | 0.137 |
| Item 21 | 1.094 | 2.134 | 0.271 | 0.909 | 2.313 | 0.237 |
| Item 22 | 2.239 | 1.198 | 0.246 | 2.696 | 1.308 | 0.241 |
| Item 23 | 1.679 | -1.268 | 0.153 | 1.728 | -1.406 | 0.147 |
| Item 24 | 1.514 | -0.135 | 0.192 | 1.580 | -0.162 | 0.189 |
| Item 25 | 1.048 | 1.436 | 0.149 | 1.121 | 1.494 | 0.149 |
| Item 26 | 1.605 | 0.658 | 0.218 | 1.980 | 0.731 | 0.233 |
| Item 27 | 1.334 | -0.573 | 0.241 | 1.453 | -0.716 | 0.206 |
| Item 28 | 1.858 | 0.589 | 0.234 | 2.413 | 0.596 | 0.224 |
| Item 29 | 1.866 | 0.797 | 0.204 | 2.500 | 0.829 | 0.196 |
| Item 30 | 1.208 | 2.097 | 0.192 | 1.099 | 2.393 | 0.177 |
| | | | | | | |
| Mean | 1.669 | 0.286 | 0.214 | 1.763 | 0.274 | 0.199 |
| SD | 0.579 | 1.366 | 0.065 | 0.650 | 1.540 | 0.065 |

Table 16

*Tier C Form Item Parameter Estimates*

| | 3-PL Model | | | Testlet Model | | |
|---|---|---|---|---|---|---|
| | $\hat{a}$ | $\hat{b}$ | $\hat{c}$ | $\hat{a}$ | $\hat{b}$ | $\hat{c}$ |
| Item 1 | 0.995 | -2.219 | 0.220 | 0.984 | -2.509 | 0.210 |
| Item 2 | 1.140 | -2.351 | 0.226 | 1.220 | -2.579 | 0.209 |
| Item 3 | 1.202 | -0.757 | 0.178 | 1.390 | -0.790 | 0.188 |
| Item 4 | 1.745 | 0.053 | 0.175 | 1.610 | -0.001 | 0.153 |
| Item 5 | 2.015 | -0.046 | 0.193 | 2.318 | -0.073 | 0.184 |
| Item 6 | 1.682 | 0.384 | 0.293 | 1.684 | 0.353 | 0.272 |
| Item 7 | 0.793 | 1.169 | 0.185 | 0.740 | 1.275 | 0.179 |
| Item 8 | 1.693 | -2.220 | 0.232 | 1.690 | -2.461 | 0.203 |
| Item 9 | 0.711 | 2.487 | 0.163 | 0.740 | 2.762 | 0.158 |
| Item 10 | 1.087 | 2.179 | 0.195 | 1.238 | 2.425 | 0.184 |
| Item 11 | 1.402 | 2.904 | 0.202 | 1.524 | 3.572 | 0.200 |
| Item 12 | 1.637 | 0.474 | 0.156 | 1.758 | 0.510 | 0.153 |
| Item 13 | 2.298 | 0.177 | 0.445 | 2.480 | 0.086 | 0.412 |
| Item 14 | 1.747 | -0.523 | 0.253 | 1.711 | -0.730 | 0.193 |
| Item 15 | 2.510 | -0.108 | 0.267 | 2.353 | -0.270 | 0.208 |
| Item 16 | 1.203 | -0.378 | 0.230 | 0.954 | -0.591 | 0.195 |
| Item 17 | 2.130 | -3.761 | 0.301 | 2.920 | -3.890 | 0.226 |
| Item 18 | 0.954 | -1.160 | 0.222 | 0.925 | -1.337 | 0.195 |
| Item 19 | 1.184 | 0.574 | 0.229 | 1.287 | 0.605 | 0.229 |
| Item 20 | 1.028 | -0.104 | 0.314 | 0.926 | -0.416 | 0.233 |
| Item 21 | 1.179 | 1.545 | 0.343 | 0.972 | 1.507 | 0.284 |
| Item 22 | 1.588 | 0.278 | 0.212 | 1.754 | 0.172 | 0.168 |
| Item 23 | 1.919 | -2.342 | 0.217 | 1.972 | -2.654 | 0.211 |
| Item 24 | 1.521 | -1.292 | 0.198 | 1.620 | -1.419 | 0.194 |
| Item 25 | 0.913 | 0.726 | 0.216 | 0.963 | 0.701 | 0.203 |
| Item 26 | 1.261 | -0.748 | 0.148 | 1.379 | -0.753 | 0.168 |
| Item 27 | 1.452 | -1.616 | 0.276 | 1.487 | -1.924 | 0.241 |
| Item 28 | 1.587 | -0.702 | 0.197 | 1.846 | -0.800 | 0.194 |
| Item 29 | 1.938 | -0.309 | 0.134 | 3.354 | -0.341 | 0.135 |
| Item 30 | 1.518 | 1.066 | 0.204 | 1.863 | 1.262 | 0.217 |
| | | | | | | |
| Mean | 1.468 | -0.221 | 0.227 | 1.589 | -0.277 | 0.206 |
| SD | 0.450 | 1.519 | 0.064 | 0.629 | 1.687 | 0.051 |

The two tables show that the item parameter estimates of the two models are quite comparable. This is consistent with the findings from the simulation study. Figure 16 presents the scatter plots for the 3-PL model-estimated item parameters vs. the testlet model-estimated item parameters for the Tier B form and Figure 17 presents the scatter plots for the 3-PL model estimated item parameters vs. the testlet model estimated item parameters for the Tier C form. As we can see, the two sets of item parameters estimates are highly correlated, the correlations for Tier B form $r(\hat{a}_{3PL}, \hat{a}_{testlet}) = 0.931$, $r(\hat{b}_{3PL}, \hat{b}_{testlet}) = 0.999$, and $r(\hat{c}_{3PL}, \hat{c}_{testlet}) = 0.975$. The correlations for Tier C form $r(\hat{a}_{3PL}, \hat{a}_{testlet}) = 0.885$, $r(\hat{b}_{3PL}, \hat{b}_{testlet}) = 0.997$, and $r(\hat{c}_{3PL}, \hat{c}_{testlet}) = 0.926$. The high correlations between the 3-PL model estimated item parameters and the testlet model estimated item parameters have also been observed in the simulation study. They arise from the common properties of the two models with regard to the marginal relationship between item response and proficiency, as the testlet model only differs from the 3-PL model through its inclusion of the testlet parameters. Moreover, the weak testlet effects for the two test forms as indicated by the $Q_3$ analysis result in less impact on the estimation of *a, b,* and *c* item parameters.

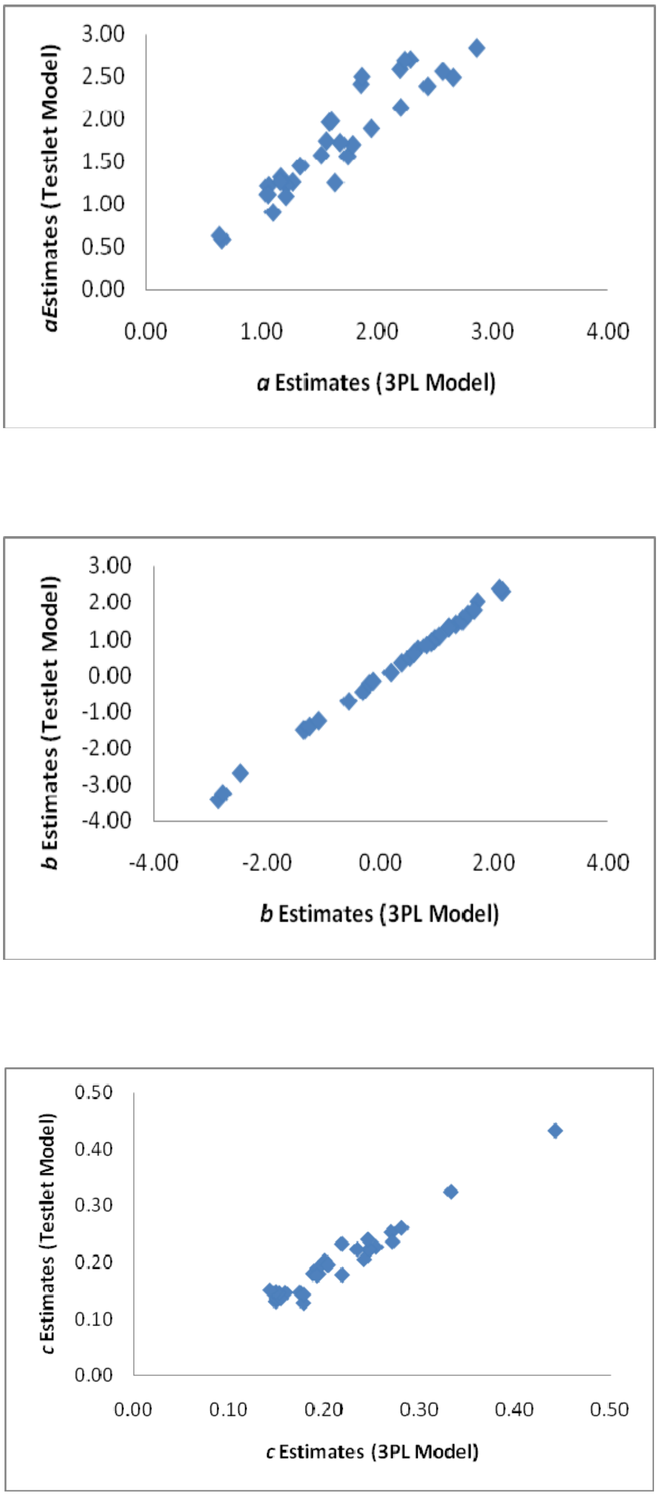
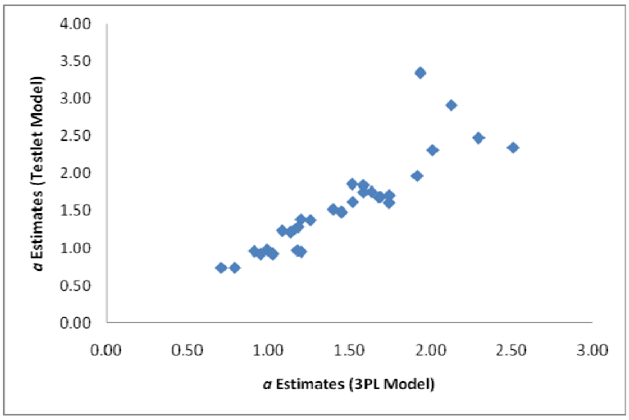
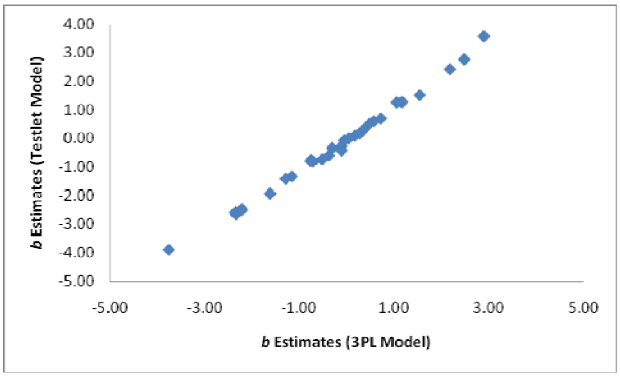


*Figure 16.* Tier B form item parameter estimates (3-PL model vs. testlet model)





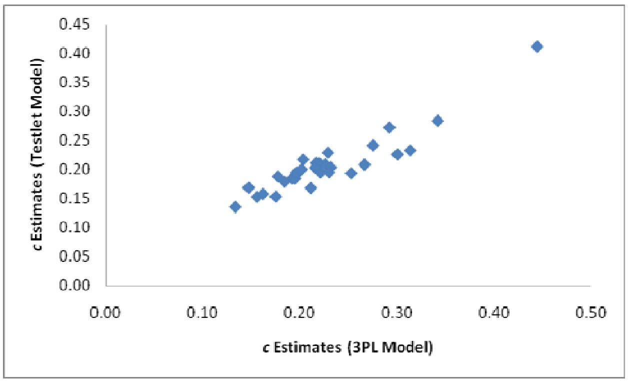


*Figure 17.* Tier C form item parameter estimates (3-PL model vs. testlet model)

Table 17 presents the variances of the testlet model estimated testlet parameters for the two test forms. With the exception of Folder 1 of tier B test form with a testlet parameter variance of 1.0200 and Folder 3 of Tier C test form with a variance of 1.0720, the folders in the two forms have small testlet parameter variances ranging from 0.2450 to 0.6861. This confirms the findings in the $Q_3$ test that the testlet effects in the two test forms are not very strong, in terms of this statistic.

Table 17

*The Testlet Model-Estimated Variances of the Testlet Parameters*

| | | Folder 1 | Folder 2 | Folder 3 | Folder 4 | Folder 5 | Folder 6 | Folder 7 | Folder 8 |
|---|---|---|---|---|---|---|---|---|---|
| Tier B Form | Estimate | 1.0200 | 0.2450 | 0.4118 | 0.4070 | 0.5877 | 0.4023 | 0.3496 | 0.5603 |
| | SE | 0.3177 | 0.0495 | 0.1129 | 0.0663 | 0.1568 | 0.1177 | 0.0747 | 0.1038 |
| Tier C Form | Estimate | 0.6861 | 0.2662 | 1.0720 | 0.4372 | 0.3753 | 0.4689 | 0.4400 | 0.6820 |
| | SE | 0.1707 | 0.0445 | 0.3584 | 0.0707 | 0.1297 | 0.1030 | 0.1090 | 0.1077 |

For $\theta$ parameters, the three models produced highly correlated parameter estimates. For Tier B form: $r(\hat{\theta}_{3PL}, \hat{\theta}_{testlet}) = 0.996$ and $r(\hat{\theta}_{GRM}, \hat{\theta}_{testlet}) = 0.966.$ For Tier C form: $r(\hat{\theta}_{3PL}, \hat{\theta}_{testlet}) = 0.996$ and $r(\hat{\theta}_{GRM}, \hat{\theta}_{testlet}) = 0.983.$ However the testlet model and the GRM estimated θ parameters have lower TIFs than the 3-PL model since the latter model ignores the bias caused by the testlet effect. As shown in Table 18, on average, the 3-PL model estimated TIFs are over 30% higher than those estimated by the GRM and the testlet model. The TIF inflation ratio is substantial considering the mild testlet effects displayed in the test forms as determined by the $Q_3$ statistics.

*Table 18*

*TIF and TIF Ratios Estimated by the Three Models*

| | | TIF | | | TIF Ratio | |
|---|---|---|---|---|---|---|
| | | $TIF_{3PL}$ | $TIF_{GRM}$ | $TIF_{Testlet}$ | $\frac{TIF_{3PL}}{TIF_{GRM}}$ | $\frac{TIF_{3PL}}{TIF_{Testlet}}$ |
| Tier B Form | Mean | 6.114 | 4.463 | 4.447 | 1.384 | 1.341 |
| | SD | 2.184 | 0.503 | 1.008 | 0.514 | 0.217 |
| Tier C Form | Mean | 6.077 | 4.551 | 4.354 | 1.346 | 1.387 |
| | SD | 1.276 | 0.645 | 0.655 | 0.277 | 0.153 |

## *Scale Linking*

Since the separate model estimation was performed on the two test forms, the two forms were estimated on different scales. The scale linking procedures were performed to put the scale of the parameter estimates of the Tier C form onto that of the Tier B form.

Before scale linking was performed, the two sets of item parameter estimates for the common items using the 3-PL model were compared and plotted in Figure 18.The two sets of item parameter estimates for the common items using the 3-PL testlet model were compared and plotted in Figure 19. The purpose of these comparisons is to see if the points indicating the two sets of item parameter estimates are well behaved and do not deviate greatly from the line that best fit the scattered plot. If outliers exist, it may pose a threat to the stable estimation of the scale linking parameter.

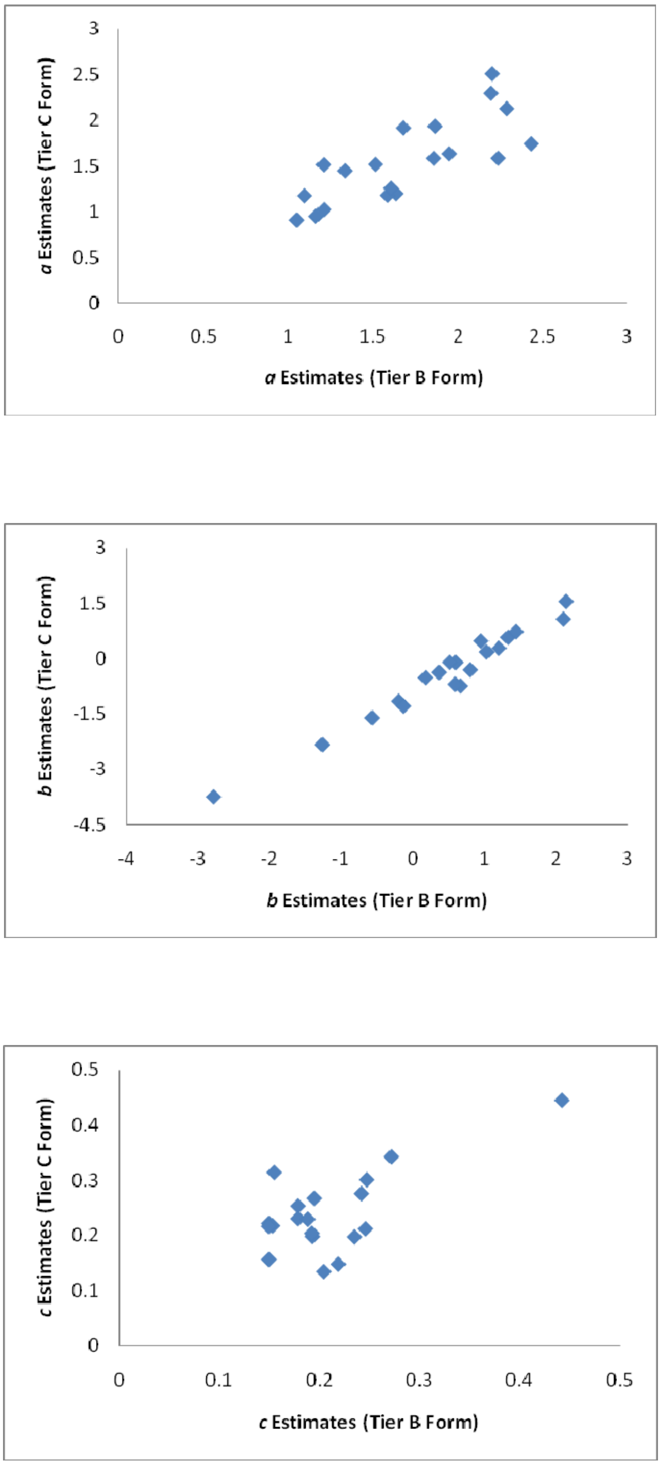


*Figure 18.* 3-PL model item parameter estimates (Tier B form vs. Tier C form)

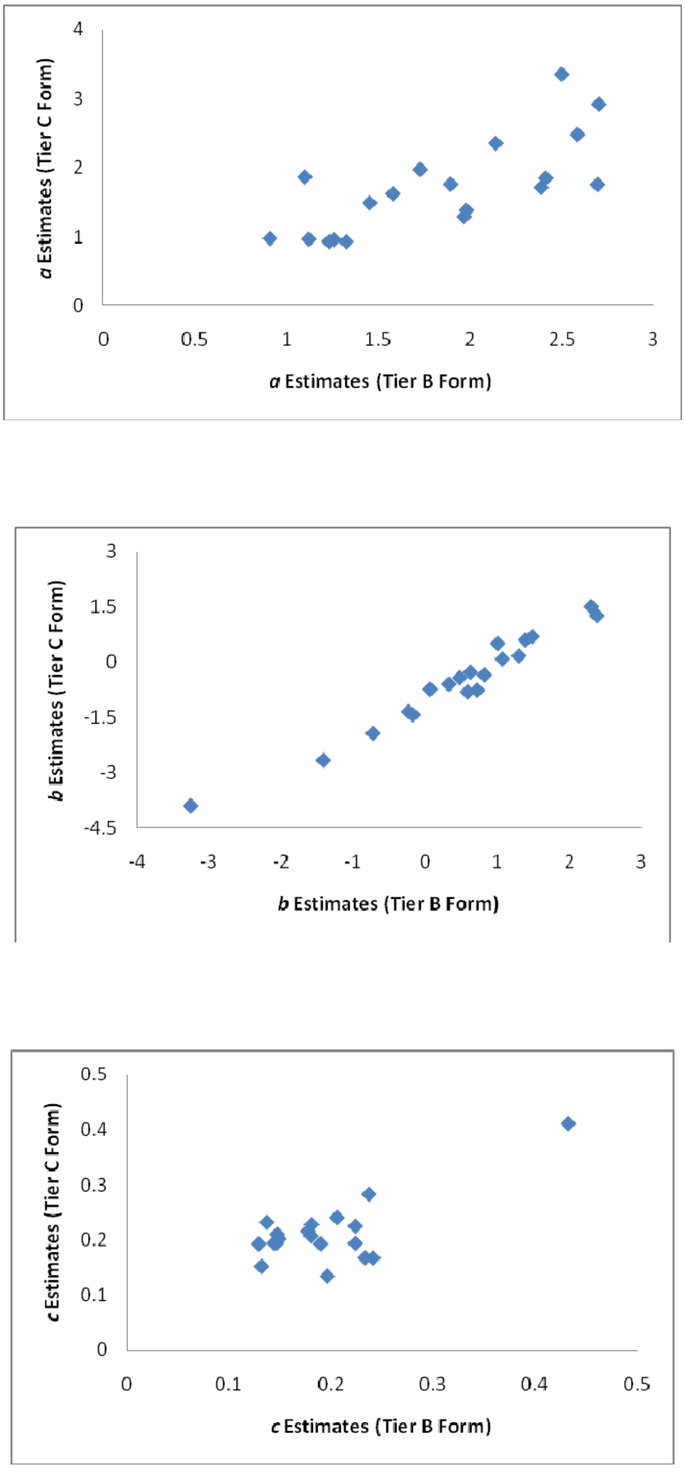


*Figure 19.* Testlet model item parameter estimates (Tier B form vs. Tier C form)

As Figure 18 shows, the points formed by the two sets of *b* parameter values estimated by the 3-PL model are very well aligned and almost form a straight line. The points on the scatter plots of the *a* parameter and *c* parameter estimates are more spread-out than the *b* parameter estimates. This is expected since the estimation of *a* and *c* parameters are usually less stable than that of *b* parameters. There are no off-diagonal outliers indicating discordance in the estimated item parameter values using the two forms for all the common items. The same conclusion can also be drawn for the testlet model-estimated item parameter values, as demonstrated in Figure 19. All common items can be included in the scale linking process.

For comparison's purpose, the Haebara item characteristic curve linking were applied for the 3-PL model and the GRM model parameter estimates using ST and POLYST respectively, and the proposed scale linking procedure was performed for the testlet model parameter estimates using the PROC NLP procedure of the SAS program. Table 19 presents the estimated linking parameters for the procedures based on the three models. The linking parameter estimates of the 3-PL model and the GRM do not differ very much, with the GRM procedure producing slope and intercept parameters that are just slightly higher than those of the 3-PL procedure. This may be due to the fact that the two test forms do not display strong testlet effects that would impact the 3-PL model estimation. The testlet model scale linking procedure produces results that fall somewhere between the results of the 3-PL model procedure and the GRM model procedure: its slope estimate is similar as that of the 3-PL model and its intercept estimate is similar as that of the GRM. The linking parameters affect the shape of the distribution of the rescaled $\theta$ parameters of the Tier C test form. The mean of the rescaled $\theta$ parameter distribution of the testlet model procedure is 0.995. It shows that the student group taking the Tier C form has substantially higher English reading ability than the student group taking the Tier B form: the mean difference is almost 1 logit point. The rescaled $\theta$ distribution of the students taking the new form using the testlet model procedure has a standard deviation of 0.847, which is less than 1. This shows that the distribution of the reading abilities of the students taking the Tier C form is a little tighter than that of the students taking the Tier B form.

Table 19

*Scale Linking Parameter Estimates*

| **Method** | **Slope (*A*)** | **Intercept (*B*)** |
|---|---|---|
| 3-PL | 0.849 | 0.887 |
| GRM | 0.958 | 0.990 |
| Testlet | 0.847 | 0.995 |

After scale linking parameters were estimated, the item parameter estimates of the new forms were transformed onto the scale of the base forms using Formulae (3.2) and (3.3). The rescaled item parameter estimates using the 3-PL model procedure and the testlet model procedure are presented in Table 20.

Table 20

*Tier C Form Item Parameter Estimates after Scale Linking*

| | **3-PL Model** | | | **Testlet Model** | | |
|---|---|---|---|---|---|---|
| | $\hat{a}$ | $\hat{b}$ | $\hat{c}$ | $\hat{a}$ | $\hat{b}$ | $\hat{c}$ |
| Item 1 | 1.171 | -0.998 | 0.220 | 1.162 | -1.129 | 0.210 |
| Item 2 | 1.342 | -1.110 | 0.226 | 1.441 | -1.188 | 0.209 |
| Item 3 | 1.415 | 0.244 | 0.178 | 1.642 | 0.327 | 0.188 |
| Item 4 | 2.055 | 0.932 | 0.175 | 1.902 | 0.994 | 0.153 |
| Item 5 | 2.372 | 0.848 | 0.193 | 2.738 | 0.933 | 0.184 |
| Item 6 | 1.981 | 1.213 | 0.293 | 1.989 | 1.294 | 0.272 |
| Item 7 | 0.934 | 1.880 | 0.185 | 0.874 | 2.074 | 0.179 |
| Item 8 | 1.994 | -0.998 | 0.232 | 1.996 | -1.088 | 0.203 |
| Item 9 | 0.837 | 3.000 | 0.163 | 0.874 | 3.333 | 0.158 |
| Item 10 | 1.279 | 2.738 | 0.195 | 1.462 | 3.048 | 0.184 |
| Item 11 | 1.650 | 3.353 | 0.202 | 1.800 | 4.019 | 0.200 |
| Item 12 | 1.928 | 1.289 | 0.156 | 2.077 | 1.427 | 0.153 |
| Item 13 | 2.706 | 1.037 | 0.445 | 2.929 | 1.068 | 0.412 |
| Item 14 | 2.056 | 0.443 | 0.253 | 2.021 | 0.377 | 0.193 |
| Item 15 | 2.956 | 0.795 | 0.267 | 2.779 | 0.767 | 0.208 |
| Item 16 | 1.416 | 0.566 | 0.230 | 1.127 | 0.494 | 0.195 |
| Item 17 | 2.508 | -2.307 | 0.301 | 3.449 | -2.298 | 0.226 |
| Item 18 | 1.123 | -0.098 | 0.222 | 1.092 | -0.137 | 0.195 |
| Item 19 | 1.395 | 1.375 | 0.229 | 1.520 | 1.508 | 0.229 |
| Item 20 | 1.210 | 0.799 | 0.314 | 1.094 | 0.643 | 0.233 |
| Item 21 | 1.388 | 2.199 | 0.343 | 1.148 | 2.271 | 0.284 |
| Item 22 | 1.870 | 1.123 | 0.212 | 2.072 | 1.141 | 0.168 |
| Item 23 | 2.259 | -1.102 | 0.217 | 2.329 | -1.252 | 0.211 |
| Item 24 | 1.791 | -0.210 | 0.198 | 1.914 | -0.206 | 0.194 |
| Item 25 | 1.075 | 1.504 | 0.216 | 1.137 | 1.588 | 0.203 |
| Item 26 | 1.485 | 0.252 | 0.148 | 1.629 | 0.357 | 0.168 |
| Item 27 | 1.710 | -0.485 | 0.276 | 1.756 | -0.634 | 0.241 |
| Item 28 | 1.869 | 0.291 | 0.197 | 2.181 | 0.318 | 0.194 |
| Item 29 | 2.281 | 0.624 | 0.134 | 3.962 | 0.707 | 0.135 |
| Item 30 | 1.787 | 1.792 | 0.204 | 2.201 | 2.063 | 0.217 |
| Mean | 1.728 | 0.700 | 0.227 | 1.877 | 0.761 | 0.206 |
| SD | 0.530 | 1.290 | 0.064 | 0.743 | 1.428 | 0.051 |

# Chapter 5  Conclusion and Discussion

The TRT models are a comparatively new family of IRT models that have been employed by researchers to tackle the issue of LID effect caused by the testlet format. By including a set of person-testlet interaction parameters in addition to the usual item and person parameters, the TRT models are able to account for the testlet effects which have been ignored by the traditional unidimensional IRT models.

The inclusion of the testlet parameters in the TRT models complicates the scale linking process, especially when characteristic curve scale linking methods are used. This is due to the fact that the probability of answering an item correctly as computed using a TRT model must be calculated with a vector of person parameter values instead of a single value. One of these person parameters, $\theta$, is analogous to the ability parameter in a standard unidimensional IRT model. The testlet effect parameters for students, however, can be considered nuisance parameters in this situation and must be marginalized over in the course of estimating linking parameters. There have been very few studies on scale linking for TRT model parameter estimates. Li et al. (2005) extended Stocking & Lord's test characteristic curve scale linking method to the TRT models. In their scale linking procedure, they included a set of testlet related dimension parameters that shift when the scale is transformed. While this practice complicates the computation even further, the effect of adding nuisance dimension parameters on the performance of the scale linking procedure remains to be studied.

This study extended Haebara's item characteristic curve scale to TRT models. Quadature points and weights were used to approximate the estimated distribution of the testlet effect parameters so that the expected score of each item given $\theta$ can be computed. The Newton Raphson method was used to obtain $A$ and $B$ scale linking parameters that minimize the item characteristic curve differences. Nonlinear programming procedure NLP of the SAS program was applied to implement the method. The proposed procedure was performed for the 3-PL testlet model in this study. The 3-PL testlet model was selected because Rasch/1-PL and 2-PL testlet models can be treated as the nested model of the 3-PL testlet model. If the proposed scale linking procedure works for the 3-PL testlet model, it should also work for the 1-PL and 2-PL testlet models.

## ***Summary of Findings***

A simulation study and a real data study were conducted to compare the performance of the proposed 3-PL testlet model scale linking procedure with that of the 3-PL IRT procedure and the GRM procedure. The findings are summarized:

1) When there is no testlet effect in the test forms, the testlet model scale linking procedure still performs well. The 3-PL IRT model is the true model under this condition. Therefore the 3-PL model-based scale linking procedure should perform better than the testlet model-based scale linking procedure. This was

confirmed in this study: under the condition that the variances of the testlet effect parameters are 0, the 3-PL model scale linking procedure produced linking parameter estimates that had the lowest MSE among the three procedures. Its person parameter estimates had the largest correlation with the true person parameter values and the lowest loss functions MAD and RMSD on average. However, the testlet model procedure-produced scale linking parameter estimates were not significantly different from those of the 3-PL model procedure. The bias and the MSE of the testlet model procedure-produced linking constant estimates are similar to those of the 3-PL model procedure. The testlet model procedure even produced item parameter estimates that were better correlated with the true parameter values and had smaller mean MAD and RMSD than those of the 3-PL model procedure, although the differences were very minimal. This "better performance" of the testlet model procedure was caused by sampling error. It is understandable that the 3-PL testlet model scale linking method produced comparable results as the 3-PL IRT model scale linking method, since the 3-PL IRT model can be regarded as a restricted model nested within the 3-PL testlet model with its testlet parameters being 0s.

The 3-PL IRT model is the most parsimonious model when dealing with the test forms that do not display testlet effects, since it is the true model and it has less model parameters to estimate. Therefore, the 3-PL IRT model scale linking procedure should be the preferred scale linking procedure. However, the testlet model procedure also proved to be working well under such situations. Moreover, while the 3-PL testlet model is a less parsimonious model, the simulation study showed that the error variances of scale linking parameter estimates produced by the 3-PL testlet model scale linking procedure are not much larger than those produced by 3-PL IRT model scale linking procedure. The loss of efficiency in scale linking parameter estimation by using the 3-PL testlet model scale linking procedure when there is no testlet effect is trivial according to the simulation study.

However, it is also observed that when there are no testlet effects in the test forms, the testlet model procedure tends to underestimate the reliability statistics. In the simulation study, the testlet model procedure underestimated TIF by about 15% under Condition 1. While the simulation design specified that there was no variance in the testlet model parameters, some sampled test forms might still display minor testlet effects which were captured by the testlet model. This was reflected in the estimation of TIF. Although the underestimation of TIF might not be serious in this case, we should be aware of this downside of using the testlet model scale linking procedure when there is no testlet effect in the test forms.

2) When there are testlet effects and particularly when the testlet effects are strong, the testlet model scale linking procedure usually performs better than the 3-PL IRT model procedure. The testlet model procedure in such situations (when the variances of the testlet effect are 1 or 2 in the simulation study) produced scale linking estimates that were generally closest to the true parameter values. Its item parameter estimates and $\theta$ parameter estimates had the highest correlations with the true parameter values and the lowest mean MAD and RMSD. The superiority of the testlet model scale linking procedure over the 3-PL model scale linking procedure becomes more evident as the testlet effects become larger. In the simulation study, when the variances of the testlet effect were 1, the testlet model procedure performed

better than the 3-PL model procedure in almost all categories. However, the mean of the scale linking parameter estimate $\hat{B}$ of the 3-PL model was closer to the true *B* value than that of the testlet model procedure. Granted the scale linking parameter estimates $\hat{A}$ and $\hat{B}$ should be evaluated jointly than in isolation, but it still showed that the testlet model procedure may not always produce the best model or scale linking parameters when the testlet effects are not very strong. When the variances of the testlet parameters were 2, the simulation study showed that the testlet model procedure produced better results than the 3-PL model procedure in all categories of the evaluation criteria.

The testlet model procedure dominated the 3-PL model procedure in the scale linking performance when there were strong testlet effects in the simulation study. A practical implication is involved in this finding: when testlet effects are strong, the difference in the scale parameter estimates produced by the 3-PL model procedure and the testlet model procedure may impact examinees' rescaled θ values. In the simulation study, the mean value of the scale linking parameter *B* estimates is 0.44 for the 3-PL model procedure and 0.50 for the testlet model procedure. It is apparent that the latter procedure produced better *B* parameter estimates since the true value of *B* parameter is 0.5. When scale linking is performed, holding other factors constant, using *B* parameter estimates of the testlet model procedure would lead to an improvement of 0.06 logit point (0.50-0.44) over the 3-PL model procedure for the θ estimates. This difference would probably be considered trivial by test practitioners in most testing programs.

The mean value of the scale linking parameter *A* estimates is 1.43 for the testlet model procedure and 1.37 for the 3-PL model procedure under condition 3. This means that if the 3-PL model scale linking method is used, the standard deviation of the rescaled $\theta$ distribution would be underestimated by 4.20% as compared to the testlet model procedure for the examinees taking the new form due to the two models' differences in *B* parameter estimation. This may still seem to be a small number that doesn't warrant attention. However, in vertical scaling situations, when the common scale is obtained by separate calibration and chained linking design for test forms from multiple grade levels, the effect of underestimating the standard deviations of the rescaled $\theta$ parameters by using the 3-PL model procedure can multiply and become a serious issue. The scale shrinkage issue in vertical scaling has been discussed and debated by scholars (Camilli, Yamamoto, & Wang, 1993; Yen, 1985, 1986). The use of the traditional unidimensional dichotomous IRT models for test forms that display LID effects may be a possible cause of scale shrinkage according to this study.

3) The testlet model procedure produces better reliability statistics. One major criticism of using the unidimensional IRT models for tests that exhibit testlet effects is that they produce positively biased reliabilities because they do not consider LID among the items within the testlets. The testlet model corrects this issue since it accounts for the testlet effects. In the simulation study, the testlet model procedure produced TIFs that were smaller than those produced by the 3-PL model procedures. The discrepancies got larger as the testlet effects increased. The mean TIF inflation ratio rose from 1.12 (when variances of the testlet parameter=0) to 1.33 (when

variances of the testlet parameter=1) and finally to 1.53 (when variances of the testlet parameter=2) as the magnitude of the testlet effects get larger. The TIFs produces by the testlet model procedure are quite similar as the TIFs produced by the GRM procedure. The GRM, along with other polytomous IRT models have been employed by researchers to deal with the inflated reliability issue caused by the testlet effects. The study shows that the testlet model procedure performs quite similarly as the GRM procedure in this aspect.

An important practical implication is associated with the finding. When testlet effects are strong, the 3-PL model may substantially overestimate TIF. The TIF values indicate how stable the examinees' $\theta$ estimates are and overestimated TIF values would lead to unjustified confidence about the estimation of $\theta$ values. The large inflation of TIF by using the wrong model may have an especially negative effect in computerized adaptive tests (CAT) that often use the estimated TIF values to determine whether stable and reliable ability estimates have been reached and the test can be stopped.

4) The testlet model scale linking procedure has several advantages over the GRM scale linking procedure. While both the testlet model and the polytomous model do a good job estimating test reliabilities without bias caused by the testlet effects, the testlet model procedure is superior to the GRM procedure in two aspects as demonstrated in the simulation study:

First, the testlet model utilizes more information in its model estimation than the GRM. The GRM uses the testlet scores for the model estimation and these scores are obtained by summing over the item scores within each testlet. The specific response patterns to the items within the testlet are lost during the process. This results in less accurate parameter estimates for the GRM. For example in the simulation study, under each of the three simulation conditions, the $\theta$ estimates of the GRM had the lowest correlations with the true $\theta$ values and largest mean RMSD and MAD loss functions. The testlet procedure produced $\theta$ estimates that had the highest correlations with the true $\theta$ values and lowest RMSD and MAD statistics under Conditions 2 and 3 when there were testlet effects; and they were only marginally inferior to the $\theta$ estimates of the 3-PL model procedure under condition 1 when there was no testlet effect, but still better than those produced by the GRM procedure.

Secondly, The testlet model procedure can produce item parameter estimates that are consistent with the item parameter estimates of the traditional unidimensional IRT models. The testlet models are based on the unidimensional IRT models and they are identical to the corresponding IRT models except that they include the testlet effect parameters. The simulation study and the real data study demonstrated that the item parameters estimated by the testlet model were quite comparable with the item parameter estimates produced by the 3-PL model and the inclusion of the testlet parameters only made the estimation of the item parameters even more accurate. On the contrary, the polytomous models have a different set of model parameters. For example, the GRM has step parameters and difficulty parameters that are concerned with the response categories instead of individual items. Therefore, the polytomous model procedures cannot be used in situations where the estimation and scaling of item parameters are required. In the simulation study and the real data study, only the

3-PL model procedure and the testlet model procedure were included when comparing the procedures' performance in item parameter estimation and scaling.

5) It is a good practice to check for the magnitude of the testlet effect before proceeding with model fitting and scale linking. When working with testlet based test data, the magnitude of the testlet effects is unknown to researchers. By conducting a LID test using such indices as Yen's $Q_3$, researchers can make an informed decision about which model and scale linking procedure they should employ. If the $Q_3$ and other LID indices turn out to be large, a testlet model can be fitted to the data so that the variances of the testlet effect parameters can be estimated. Large variance values usually confirm the previous finding that the testlet effects are large and warrant attention. The application of the testlet model scale linking procedure can be justified in such cases.

However, the study also shows that the $Q_3$ analysis is not sensitive to testlet effects that are not very strong. As a result, readers should be aware that when the $Q_3$ values are low, it doesn't necessarily mean that there is no testlet effect. Mild to medium testlet effects may still exist in such cases, and, as the real-data study demonstrated, substantial effects on the test information function can result.

## *Caveats*

The simulation study demonstrated that the proposed procedure performs well in linking scales for testlet model parameter estimates. However, there are several caveats in the scale linking method and the simulation study that readers should be aware of.

1) The 3-PL testlet model used in the scale linking procedure assumes that testlet parameters for each testlet follow a normal distribution N(0, $var_{\gamma(g)}$). The magnitude of the testlet effects are determined by the variances of the testlet parameters. In the simulation study, the true testlet parameters were specified to be normally distributed and the normal distribution was also used to estimate the testlet parameters. The practice of assuming normal distributions for testlet parameters has almost been exclusively applied by researchers in their specification and estimation of TRT models (Bradlow, Wainer, & Wang, 1999; Wainer, Bradlow, & Du, 2000; Wainer, Bradlow, & Wang, 2007; Wainer & Wang, 2000). Readers should be aware that while this is a generally accepted practice, there is no guarantee that the true testlet parameters are normally distributed universally for different tests that target different content domains and examinees in real life. The discrepancy between the assumed testlet parameter distribution and the true testlet parameter distribution can lead to inaccuracies in model and scale linking parameter estimation. Therefore it is recommended to study the behavior of the testlet effects parameters and investigate the fitness of TRT models that employ different testlet parameter distributions. The proposed scale linking method can be adapted to accommodate different distributions for the testlet parameters through assigning quadrature points and weights that approximate the specific distributions.

2) The proposed testlet model scale linking method makes the assumption that all persons share the same testlet effect parameter distribution within a specific testlet.

For example, if the distribution of the testlet effect parameters over all examinees for a testlet is estimated to be normally distributed with a variance of 1.5, all persons are assigned the same set of quadratic points and weights to approximate the N(0, 1.5) distribution when computing the expected item score given $\theta$ in the proposed scale linking procedure. This assumption is made based on the belief that the testlet parameter bears no relationship with the person's latent trait $\theta$. If such correlations do exist, the proposed linking procedure can also be adapted to accommodate such situation using the following approach: after the testlet parameters are estimated, persons with similar *θs* can be grouped together and the variances of their testlet effect parameter distribution can be estimated. A specific set of quadrature points and weights that approximate that testlet parameter distribution can be assigned to the examinee group with the specific level of latent trait.

3) When calculating the cumulative differences between characteristic curves for Haebara and Stocking & Lord scale linking methods, there are different approaches in specifying the examinees used in the summations. As indicated in Formula (3.13), the *Hcrit* in the proposed scale linking procedure is derived by summing over the estimated traits of all the examinees who have taken the base test form. This summation approach was used by Stocking and Lord (1983). There are also other summation methods. For example, the summation can be made over equally spaced trait values (Baker & Al-Karni, 1991), or over the estimated traits of all examinees (Haebara, 1980). The integration function over the trait distribution can also be used instead of the summation function if the distribution can be estimated (Zeng & Kolen, 1994). As long as there are a large number of examinees who are

administered the base test form and they are well distributed and representative of the population, the summation approach used in this study should work fine. However, readers should be aware of the other options when the sample size is small.

4) The proposed method extends Haebara's scale linking method to the testlet model. The simulation study compared the scale linking procedure with the 3-PL model and GRM model based procedures. The biggest difference in these procedures is that they use different models. It can be inferred that as the testlet effects get stronger, it is natural that the testlet model scale linking procedure performs better since the testlet model is a better fit model under the situation. Another way of evaluating the performance of the proposed scale linking method is to compare it with other testlet model based scale linking method. For example, we can compare the proposed testlet model scale linking procedure with Li et al.'s (2005) procedure, which extends the Stocking & Lord test characteristic curve scale linking method to the testlet model, or we can compare the proposed procedure with the testlet model based concurrent calibration scale linking method. Since these scale linking methods are based on the same testlet model, such comparisons can reveal the differences in the procedures' performances that are due to the scale linking approaches instead of the models employed.

5) According the proposed scale linking method, in the process of estimating the scale linking parameters by finding the values that minimize the *Hcrit* function, the expected item scores are obtained using the estimated item parameters, which are the means of the posterior distributions of the item parameters since the estimation is performed using the Bayesian method. However this may not be the optimal practice

from a Bayesian perspective; it would be preferable to integrate the criterion over the distributions of the item parameters as well as the distribution of the testlet effect parameters. In the context of MCMC estimation, for example, the optimization to find linking parameters could be performed at every iteration using the current draws from every item parameter's and every examinee $\theta$ parameter's full conditional distribution. The full posterior distribution of the $A$ and $B$ parameters can thus be obtained across iterations and the posterior means would be scale linking parameter estimates using this approach. The posterior means may not necessarily be the same as the optimization results executed on the posterior mean of the parameter estimates.

At the time Haebara method was developed in the early 80s, most IRT models were estimated using the frequentist approach. Finding the linking functions by using the point estimates of the item parameters was consistent with best practices. The testlet model could not even be estimated at the time. The advances in Bayesian inferences and computer technology in recent years allow us to explore more complex models and improve upon the current practices in educational measurement. What this study has done is taking a step in the direction: generalize the Haebara method to the testlet models and integrate over the testlet effect parameters. Future research could address the possibility of integrating over all the item parameters, not just the testlet effect parameters.

6) The focus of the dissertation is to propose a new scale linking method under CINEG design for the testlet model. While its effectiveness in linking scales for test forms composed of testlets has been demonstrated via the comparison of the proposed method with the 3-PL model and GRM scale linking approaches, this

dissertation doesn't intend to be a comprehensive comparison study on the merits of using different model-based scale linking methods under different conditions. The range of simulation conditions in the study is limited. Further studies can be conducted on the effectiveness of the proposed method under different conditions. For example, Bradlow et al. (1999) asserted that with short testlet that have only 4-6 items, fitting each testlet item as if it was independent and ignoring the overestimation of the precision of the error of measurement can be deemed acceptable. In the simulation study, each testlet only had 5 items and the testlet model and the linking method based on it seemed to be working well. But it would also be of interest to see if the proposed method would work better with testlets that have more items. Other conditions, such as sample sizes, numbers of common testlets/items, and non-uniform levels of testlet effect for different testlets can also be simulated and analyzed.

### *Application of the Proposed Scale Linking Method*

As discussed in Chapter 2, the testlet format is better at eliciting evidence about high-order cognitive functioning than the stand alone MC format and there is a growing interest of applying the testlet format in performance assessments. Moreover, with the increasingly wide application of computer adaptive tests (CAT), using testlets instead of individual items in the adaptation process presents several benefits. Wainer et al. (2007) argued that it is advantageous to bundle items into testlets in CAT to allow the test structure to more closely match the construct. Hendrickson

(2007) also recommended using testlets as adaptation points within multistage adaptive testing since item-level adaptation can cause context effects, unbalanced content and test security and item exposure problems. Therefore we expect to see more applications of the testlet format in performance assessments and CAT.

Another trend in educational measurement is the increasing demand for growth measures that can be used to evaluate students' achievement. The No Child Left Behind Act states that "high-quality academic assessments" should be "aligned with challenging state academic standards so that students, teachers, parents, and administrators can measure progress against common expectations for student academic achievement" and that student cohorts are expected to show "adequate yearly progress"(Congress, 2001). Items and test scores on different test forms within and across grade levels often need to be on the same scale so that horizontal equating and vertical scaling are possible. As a result, scale linking and test equating have become a routine procedure with many testing programs.

TRT models have been gaining popularity within the academic community because they can account for LID that is often observed in testlet items while retaining the usual item and person parameters of the unidimensional IRT models. However, there have been very few studies on scale linking for testlet models. This can affect the broader application of TRT models in testing programs. The study extended Haebara's item characteristic curve linking method to the testlet model and demonstrated to be effective through the simulated data and the real data studies. Moreover, the algorithm of the proposed method can be implemented using the popular SAS statistical package. This allows the method to be readily applied by testing programs. The effectiveness and efficiency of the proposed testlet model scale linking method would promote the application of the TRT models for testlet-based tests.

## Appendix A  Part of the SAS NLP procedure to compute the testlet model scale linking parameter estimates

```
proc nlp data=par_est vardef=n covariance=h pcov phes;
profile a b / alpha=0.05;
min diff;
parms a=1, b=0;
bounds a>-1000;
diff=(
(
 (old_col3+(1-old_col3)/(1+exp(-old_col1*(old_col4-old_col2-
old_col5))))*old_col25
+(old_col3+(1-old_col3)/(1+exp(-old_col1*(old_col4-old_col2-
old_col6))))*old_col26
+……
+(old_col3+(1-old_col3)/(1+exp(-old_col1*(old_col4-old_col2-
old_col24))))*old_col44
)-
(
 (new_col3+(1-new_col3)/(1+exp(-(new_col1/A)*(old_col4-(A*new_col2+B)-
A*new_col5))))*new_col25
+(new_col3+(1-new_col3)/(1+exp(-(new_col1/A)*(old_col4-(A*new_col2+B)-
A*new_col6))))*new_col26
+……
+(new_col3+(1-new_col3)/(1+exp(-(new_col1/A)*(old_col4-(A*new_col2+B)-
A*new_col24))))*new_col44
)
)**2
;
ods output  ParameterEstimates(match_all=datasetnames)=testlet_coef;
run;
```

# Appendix B  Scale Linking Parameter Estimates

## *Scale linking parameter estimates using the three procedures (Var(testlet)=0)*

| | | 3-PL | | | GRM | | | Testlet | |
|---|---|---|---|---|---|---|---|---|---|
| | | $\hat{A}$ | $\hat{B}$ | | $\hat{A}$ | $\hat{B}$ | | $\hat{A}$ | $\hat{B}$ |
| Sample1 | | 1.456 | 0.442 | | 1.378 | 0.430 | | 1.410 | 0.450 |
| Sample2 | | 1.451 | 0.500 | | 1.369 | 0.525 | | 1.373 | 0.532 |
| Sample3 | | 1.418 | 0.554 | | 1.419 | 0.583 | | 1.387 | 0.570 |
| Sample4 | | 1.606 | 0.549 | | 1.591 | 0.504 | | 1.586 | 0.527 |
| Sample5 | | 1.457 | 0.406 | | 1.341 | 0.374 | | 1.435 | 0.413 |
| Sample6 | | 1.421 | 0.520 | | 1.422 | 0.492 | | 1.403 | 0.528 |
| Sample7 | | 1.440 | 0.539 | | 1.514 | 0.596 | | 1.439 | 0.558 |
| Sample8 | | 1.442 | 0.562 | | 1.473 | 0.562 | | 1.491 | 0.564 |
| Sample9 | | 1.444 | 0.434 | | 1.638 | 0.104 | | 1.502 | 0.460 |
| Sample10 | | 1.419 | 0.528 | | 1.465 | 0.499 | | 1.398 | 0.484 |
| Sample11 | | 1.350 | 0.294 | | 1.311 | 0.328 | | 1.406 | 0.351 |
| Sample12 | | 1.203 | 0.296 | | 1.179 | 0.281 | | 1.196 | 0.322 |
| Sample13 | | 1.421 | 0.444 | | 1.345 | 0.452 | | 1.445 | 0.482 |
| Sample14 | | 1.418 | 0.502 | | 1.367 | 0.493 | | 1.404 | 0.508 |
| Sample15 | | 1.291 | 0.563 | | 1.379 | 0.589 | | 1.307 | 0.579 |
| Sample16 | | 1.577 | 0.612 | | 1.455 | 0.583 | | 1.496 | 0.589 |
| Sample17 | | 1.368 | 0.536 | | 1.292 | 0.548 | | 1.338 | 0.567 |
| Sample18 | | 1.465 | 0.632 | | 1.406 | 0.659 | | 1.438 | 0.698 |
| Sample19 | | 1.369 | 0.598 | | 1.342 | 0.689 | | 1.341 | 0.656 |
| Sample20 | | 1.439 | 0.318 | | 1.308 | 0.252 | | 1.431 | 0.348 |
| Sample21 | | 1.283 | 0.469 | | 1.306 | 0.441 | | 1.242 | 0.457 |
| Sample22 | | 1.561 | 0.418 | | 1.550 | 0.422 | | 1.566 | 0.434 |
| Sample23 | | 1.525 | 0.506 | | 1.443 | 0.451 | | 1.459 | 0.477 |
| Sample24 | | 1.524 | 0.587 | | 1.512 | 0.563 | | 1.512 | 0.586 |
| Sample25 | | 1.605 | 0.488 | | 1.463 | 0.496 | | 1.532 | 0.489 |
| Sample26 | | 1.443 | 0.494 | | 1.410 | 0.479 | | 1.406 | 0.501 |
| Sample27 | | 1.488 | 0.398 | | 1.424 | 0.416 | | 1.436 | 0.417 |
| Sample28 | | 1.418 | 0.435 | | 1.446 | 0.478 | | 1.467 | 0.483 |
| Sample29 | | 1.309 | 0.372 | | 1.272 | 0.383 | | 1.321 | 0.412 |
| Sample30 | | 1.435 | 0.465 | | 1.415 | 0.549 | | 1.457 | 0.543 |
| Sample31 | | 1.493 | 0.470 | | 1.391 | 0.423 | | 1.416 | 0.475 |
| Sample32 | | 1.482 | 0.503 | | 1.374 | 0.527 | | 1.472 | 0.531 |
| Sample33 | | 1.464 | 0.497 | | 1.401 | 0.516 | | 1.444 | 0.528 |
| Sample34 | | 1.407 | 0.488 | | 1.384 | 0.497 | | 1.425 | 0.541 |
| Sample35 | | 1.285 | 0.526 | | 1.411 | 0.572 | | 1.328 | 0.567 |
| Sample36 | | 1.447 | 0.403 | | 1.465 | 0.375 | | 1.488 | 0.428 |
| Sample37 | | 1.263 | 0.444 | | 1.295 | 0.462 | | 1.254 | 0.466 |
| Sample38 | | 1.360 | 0.560 | | 1.312 | 0.525 | | 1.367 | 0.582 |
| Sample39 | | 1.416 | 0.569 | | 1.407 | 0.611 | | 1.375 | 0.608 |
| Sample40 | | 1.357 | 0.354 | | 1.305 | 0.344 | | 1.340 | 0.348 |
| Sample41 | | 1.376 | 0.375 | | 1.418 | 0.375 | | 1.413 | 0.366 |
| Sample42 | | 1.441 | 0.523 | | 1.293 | 0.478 | | 1.384 | 0.506 |
| Sample43 | | 1.424 | 0.514 | | 1.481 | 0.509 | | 1.350 | 0.522 |
| Sample44 | | 1.456 | 0.577 | | 1.409 | 0.547 | | 1.452 | 0.606 |
| Sample45 | | 1.434 | 0.691 | | 1.383 | 0.634 | | 1.402 | 0.673 |
| Sample46 | | 1.493 | 0.504 | | 1.546 | 0.539 | | 1.582 | 0.565 |
| Sample47 | | 1.475 | 0.468 | | 1.468 | 0.412 | | 1.530 | 0.456 |
| Sample48 | | 1.520 | 0.512 | | 1.477 | 0.609 | | 1.533 | 0.621 |
| Sample49 | | 1.487 | 0.427 | | 1.454 | 0.423 | | 1.434 | 0.403 |
| Sample50 | | 1.481 | 0.618 | | 1.378 | 0.613 | | 1.537 | 0.680 |
| | | | | | | | | | |
| Mean | | 1.432 | 0.490 | | 1.406 | 0.484 | | 1.423 | 0.509 |
| SD | | 0.083 | 0.086 | | 0.086 | 0.110 | | 0.084 | 0.089 |

### *Scale linking parameter estimates using the three procedures (Var(testlet)=1)*

| | | 3-PL | | | GRM | | | Testlet | |
|---|---|---|---|---|---|---|---|---|---|
| | | $\hat{A}$ | $\hat{B}$ | | $\hat{A}$ | $\hat{B}$ | | $\hat{A}$ | $\hat{B}$ |
| Sample1 | | 1.353 | 0.500 | | 1.394 | 0.542 | | 1.354 | 0.521 |
| Sample2 | | 1.516 | 0.549 | | 1.394 | 0.495 | | 1.471 | 0.514 |
| Sample3 | | 1.319 | 0.291 | | 1.318 | 0.383 | | 1.313 | 0.370 |
| Sample4 | | 1.511 | 0.542 | | 1.567 | 0.585 | | 1.472 | 0.618 |
| Sample5 | | 1.327 | 0.394 | | 1.300 | 0.423 | | 1.381 | 0.445 |
| Sample6 | | 1.365 | 0.535 | | 1.531 | 0.577 | | 1.438 | 0.559 |
| Sample7 | | 1.385 | 0.427 | | 1.384 | 0.464 | | 1.361 | 0.461 |
| Sample8 | | 1.307 | 0.386 | | 1.334 | 0.410 | | 1.376 | 0.440 |
| Sample9 | | 1.398 | 0.445 | | 1.416 | 0.487 | | 1.443 | 0.483 |
| Sample10 | | 1.575 | 0.416 | | 1.545 | 0.446 | | 1.581 | 0.489 |
| Sample11 | | 1.389 | 0.429 | | 1.398 | 0.465 | | 1.377 | 0.472 |
| Sample12 | | 1.272 | 0.500 | | 1.328 | 0.566 | | 1.358 | 0.608 |
| Sample13 | | 1.349 | 0.497 | | 1.376 | 0.579 | | 1.321 | 0.553 |
| Sample14 | | 1.554 | 0.445 | | 1.564 | 0.513 | | 1.577 | 0.502 |
| Sample15 | | 1.520 | 0.499 | | 1.492 | 0.628 | | 1.544 | 0.587 |
| Sample16 | | 1.393 | 0.490 | | 1.524 | 0.624 | | 1.403 | 0.594 |
| Sample17 | | 1.519 | 0.525 | | 1.488 | 0.483 | | 1.479 | 0.516 |
| Sample18 | | 1.418 | 0.568 | | 1.496 | 0.543 | | 1.446 | 0.600 |
| Sample19 | | 1.361 | 0.509 | | 1.439 | 0.514 | | 1.408 | 0.552 |
| Sample20 | | 1.489 | 0.473 | | 1.490 | 0.521 | | 1.495 | 0.518 |
| Sample21 | | 1.473 | 0.381 | | 1.457 | 0.407 | | 1.506 | 0.433 |
| Sample22 | | 1.308 | 0.538 | | 1.387 | 0.557 | | 1.323 | 0.545 |
| Sample23 | | 1.471 | 0.475 | | 1.576 | 0.579 | | 1.477 | 0.529 |
| Sample24 | | 1.400 | 0.492 | | 1.431 | 0.538 | | 1.443 | 0.505 |
| Sample25 | | 1.536 | 0.511 | | 1.522 | 0.549 | | 1.423 | 0.489 |
| Sample26 | | 1.501 | 0.464 | | 1.483 | 0.524 | | 1.508 | 0.527 |
| Sample27 | | 1.300 | 0.481 | | 1.397 | 0.497 | | 1.340 | 0.528 |
| Sample28 | | 1.473 | 0.488 | | 1.585 | 0.576 | | 1.482 | 0.541 |
| Sample29 | | 1.334 | 0.457 | | 1.361 | 0.539 | | 1.362 | 0.505 |
| Sample30 | | 1.324 | 0.403 | | 1.449 | 0.532 | | 1.356 | 0.475 |
| Sample31 | | 1.426 | 0.495 | | 1.418 | 0.536 | | 1.441 | 0.540 |
| Sample32 | | 1.250 | 0.611 | | 1.307 | 0.667 | | 1.307 | 0.680 |
| Sample33 | | 1.382 | 0.472 | | 1.495 | 0.569 | | 1.424 | 0.566 |
| Sample34 | | 1.352 | 0.481 | | 1.343 | 0.489 | | 1.337 | 0.501 |
| Sample35 | | 1.427 | 0.538 | | 1.398 | 0.541 | | 1.408 | 0.557 |
| Sample36 | | 1.365 | 0.370 | | 1.286 | 0.361 | | 1.280 | 0.369 |
| Sample37 | | 1.407 | 0.435 | | 1.507 | 0.442 | | 1.483 | 0.461 |
| Sample38 | | 1.538 | 0.489 | | 1.535 | 0.512 | | 1.523 | 0.501 |
| Sample39 | | 1.743 | 0.632 | | 1.729 | 0.702 | | 1.728 | 0.648 |
| Sample40 | | 1.391 | 0.592 | | 1.469 | 0.645 | | 1.388 | 0.584 |
| Sample41 | | 1.409 | 0.518 | | 1.501 | 0.587 | | 1.438 | 0.567 |
| Sample42 | | 1.343 | 0.466 | | 1.425 | 0.576 | | 1.479 | 0.568 |
| Sample43 | | 1.430 | 0.563 | | 1.404 | 0.545 | | 1.428 | 0.589 |
| Sample44 | | 1.393 | 0.357 | | 1.477 | 0.484 | | 1.474 | 0.447 |
| Sample45 | | 1.371 | 0.546 | | 1.460 | 0.597 | | 1.427 | 0.607 |
| Sample46 | | 1.289 | 0.320 | | 1.326 | 0.362 | | 1.340 | 0.357 |
| Sample47 | | 1.554 | 0.627 | | 1.484 | 0.663 | | 1.606 | 0.707 |
| Sample48 | | 1.281 | 0.452 | | 1.392 | 0.500 | | 1.361 | 0.523 |
| Sample49 | | 1.574 | 0.468 | | 1.524 | 0.495 | | 1.503 | 0.479 |
| Sample50 | | 1.250 | 0.517 | | 1.223 | 0.527 | | 1.269 | 0.552 |
| | | | | | | | | | |
| Mean | | 1.412 | 0.481 | | 1.442 | 0.527 | | 1.429 | 0.526 |
| SD | | 0.101 | 0.072 | | 0.094 | 0.076 | | 0.090 | 0.072 |

### *Scale linking parameter estimates using the three procedures (Var(testlet)=2)*

| | | 3-PL | | | GRM | | | Testlet | |
|---|---|---|---|---|---|---|---|---|---|
| | | $\hat{A}$ | $\hat{B}$ | | $\hat{A}$ | $\hat{B}$ | | $\hat{A}$ | $\hat{B}$ |
| Sample1 | | 1.380 | 0.416 | | 1.398 | 0.441 | | 1.433 | 0.481 |
| Sample2 | | 1.379 | 0.417 | | 1.415 | 0.430 | | 1.501 | 0.470 |
| Sample3 | | 1.379 | 0.419 | | 1.332 | 0.471 | | 1.304 | 0.481 |
| Sample4 | | 1.368 | 0.370 | | 1.409 | 0.413 | | 1.416 | 0.453 |
| Sample5 | | 1.190 | 0.492 | | 1.253 | 0.599 | | 1.283 | 0.577 |
| Sample6 | | 1.291 | 0.440 | | 1.348 | 0.512 | | 1.255 | 0.484 |
| Sample7 | | 1.411 | 0.429 | | 1.446 | 0.528 | | 1.424 | 0.489 |
| Sample8 | | 1.271 | 0.227 | | 1.358 | 0.324 | | 1.363 | 0.285 |
| Sample9 | | 1.563 | 0.434 | | 1.804 | 0.552 | | 1.734 | 0.512 |
| Sample10 | | 1.474 | 0.424 | | 1.538 | 0.528 | | 1.513 | 0.502 |
| Sample11 | | 1.153 | 0.378 | | 1.205 | 0.437 | | 1.191 | 0.454 |
| Sample12 | | 1.289 | 0.380 | | 1.394 | 0.455 | | 1.311 | 0.427 |
| Sample13 | | 1.537 | 0.584 | | 1.493 | 0.633 | | 1.536 | 0.650 |
| Sample14 | | 1.203 | 0.373 | | 1.309 | 0.449 | | 1.249 | 0.432 |
| Sample15 | | 1.583 | 0.482 | | 1.710 | 0.587 | | 1.664 | 0.585 |
| Sample16 | | 1.290 | 0.449 | | 1.305 | 0.498 | | 1.278 | 0.502 |
| Sample17 | | 1.393 | 0.407 | | 1.461 | 0.454 | | 1.417 | 0.451 |
| Sample18 | | 1.356 | 0.451 | | 1.415 | 0.457 | | 1.481 | 0.525 |
| Sample19 | | 1.452 | 0.390 | | 1.437 | 0.477 | | 1.404 | 0.450 |
| Sample20 | | 1.467 | 0.334 | | 1.611 | 0.367 | | 1.742 | 0.416 |
| Sample21 | | 1.309 | 0.472 | | 1.335 | 0.506 | | 1.272 | 0.462 |
| Sample22 | | 1.390 | 0.359 | | 1.491 | 0.392 | | 1.473 | 0.373 |
| Sample23 | | 1.228 | 0.362 | | 1.340 | 0.397 | | 1.393 | 0.435 |
| Sample24 | | 1.373 | 0.392 | | 1.440 | 0.502 | | 1.352 | 0.462 |
| Sample25 | | 1.379 | 0.432 | | 1.510 | 0.513 | | 1.549 | 0.494 |
| Sample26 | | 1.523 | 0.567 | | 1.649 | 0.626 | | 1.563 | 0.631 |
| Sample27 | | 1.429 | 0.585 | | 1.693 | 0.724 | | 1.730 | 0.734 |
| Sample28 | | 1.194 | 0.479 | | 1.345 | 0.619 | | 1.291 | 0.605 |
| Sample29 | | 1.587 | 0.502 | | 1.640 | 0.467 | | 1.698 | 0.547 |
| Sample30 | | 1.536 | 0.601 | | 1.603 | 0.649 | | 1.562 | 0.632 |
| Sample31 | | 1.386 | 0.333 | | 1.564 | 0.431 | | 1.485 | 0.394 |
| Sample32 | | 1.437 | 0.462 | | 1.539 | 0.537 | | 1.472 | 0.549 |
| Sample33 | | 1.301 | 0.427 | | 1.339 | 0.518 | | 1.367 | 0.533 |
| Sample34 | | 1.308 | 0.384 | | 1.393 | 0.459 | | 1.366 | 0.453 |
| Sample35 | | 1.274 | 0.481 | | 1.348 | 0.505 | | 1.328 | 0.521 |
| Sample36 | | 1.186 | 0.441 | | 1.289 | 0.476 | | 1.227 | 0.503 |
| Sample37 | | 1.392 | 0.348 | | 1.508 | 0.441 | | 1.406 | 0.408 |
| Sample38 | | 1.469 | 0.549 | | 1.520 | 0.597 | | 1.411 | 0.551 |
| Sample39 | | 1.283 | 0.370 | | 1.411 | 0.447 | | 1.381 | 0.444 |
| Sample40 | | 1.304 | 0.450 | | 1.421 | 0.489 | | 1.392 | 0.506 |
| Sample41 | | 1.319 | 0.473 | | 1.504 | 0.670 | | 1.388 | 0.612 |
| Sample42 | | 1.114 | 0.432 | | 1.243 | 0.507 | | 1.211 | 0.497 |
| Sample43 | | 1.425 | 0.545 | | 1.419 | 0.617 | | 1.451 | 0.632 |
| Sample44 | | 1.488 | 0.492 | | 1.481 | 0.610 | | 1.490 | 0.563 |
| Sample45 | | 1.183 | 0.348 | | 1.338 | 0.458 | | 1.317 | 0.443 |
| Sample46 | | 1.383 | 0.425 | | 1.463 | 0.473 | | 1.439 | 0.477 |
| Sample47 | | 1.297 | 0.352 | | 1.445 | 0.393 | | 1.442 | 0.401 |
| Sample48 | | 1.525 | 0.558 | | 1.583 | 0.637 | | 1.500 | 0.608 |
| Sample49 | | 1.669 | 0.499 | | 1.610 | 0.523 | | 1.583 | 0.496 |
| Sample50 | | 1.390 | 0.359 | | 1.606 | 0.399 | | 1.488 | 0.376 |
| | | | | | | | | | |
| Mean | | 1.370 | 0.436 | | 1.454 | 0.504 | | 1.431 | 0.499 |
| SD | | 0.124 | 0.076 | | 0.130 | 0.086 | | 0.136 | 0.084 |

# Appendix C Evaluation criteria for θ Parameter Estimates

## *Correlations of the θ estimates with the true θ values for each sample (base form)*

| | Condition 1: Var(testlet)=0 | | | Condition 2: Var(testlet)=1 | | | Condition 3: Var(testlet)=2 | | |
|---|---|---|---|---|---|---|---|---|---|
| | $r(\hat{\theta}_{3PL}, \theta)$ | $r(\hat{\theta}_{GRM}, \theta)$ | $r(\hat{\theta}_{Testlet}, \theta)$ | $r(\hat{\theta}_{3PL}, \theta)$ | $r(\hat{\theta}_{GRM}, \theta)$ | $r(\hat{\theta}_{Testlet}, \theta)$ | $r(\hat{\theta}_{3PL}, \theta)$ | $r(\hat{\theta}_{GRM}, \theta)$ | $r(\hat{\theta}_{Testlet}, \theta)$ |
| Sample1 | 0.820 | 0.808 | 0.819 | 0.775 | 0.764 | 0.773 | 0.726 | 0.728 | 0.731 |
| Sample2 | 0.837 | 0.821 | 0.837 | 0.774 | 0.753 | 0.776 | 0.738 | 0.737 | 0.742 |
| Sample3 | 0.851 | 0.839 | 0.851 | 0.769 | 0.760 | 0.770 | 0.752 | 0.755 | 0.759 |
| Sample4 | 0.846 | 0.833 | 0.844 | 0.784 | 0.774 | 0.788 | 0.747 | 0.741 | 0.759 |
| Sample5 | 0.803 | 0.776 | 0.801 | 0.775 | 0.763 | 0.775 | 0.727 | 0.725 | 0.725 |
| Sample6 | 0.849 | 0.834 | 0.849 | 0.766 | 0.756 | 0.764 | 0.695 | 0.694 | 0.699 |
| Sample7 | 0.821 | 0.803 | 0.820 | 0.796 | 0.780 | 0.793 | 0.741 | 0.738 | 0.749 |
| Sample8 | 0.827 | 0.809 | 0.827 | 0.792 | 0.785 | 0.791 | 0.696 | 0.689 | 0.699 |
| Sample9 | 0.840 | 0.833 | 0.839 | 0.767 | 0.754 | 0.766 | 0.679 | 0.674 | 0.687 |
| Sample10 | 0.811 | 0.804 | 0.811 | 0.770 | 0.759 | 0.774 | 0.720 | 0.718 | 0.728 |
| Sample11 | 0.867 | 0.837 | 0.867 | 0.771 | 0.762 | 0.769 | 0.737 | 0.730 | 0.732 |
| Sample12 | 0.865 | 0.841 | 0.865 | 0.786 | 0.769 | 0.790 | 0.698 | 0.699 | 0.699 |
| Sample13 | 0.854 | 0.838 | 0.855 | 0.798 | 0.785 | 0.798 | 0.713 | 0.693 | 0.702 |
| Sample14 | 0.848 | 0.837 | 0.848 | 0.755 | 0.752 | 0.754 | 0.740 | 0.745 | 0.751 |
| Sample15 | 0.844 | 0.839 | 0.842 | 0.786 | 0.772 | 0.785 | 0.729 | 0.724 | 0.730 |
| Sample16 | 0.864 | 0.849 | 0.864 | 0.754 | 0.745 | 0.755 | 0.753 | 0.739 | 0.762 |
| Sample17 | 0.841 | 0.828 | 0.841 | 0.766 | 0.756 | 0.770 | 0.701 | 0.694 | 0.699 |
| Sample18 | 0.817 | 0.791 | 0.819 | 0.787 | 0.769 | 0.785 | 0.734 | 0.732 | 0.739 |
| Sample19 | 0.816 | 0.805 | 0.815 | 0.783 | 0.775 | 0.783 | 0.732 | 0.729 | 0.742 |
| Sample20 | 0.815 | 0.785 | 0.814 | 0.773 | 0.763 | 0.772 | 0.721 | 0.710 | 0.724 |
| Sample21 | 0.871 | 0.854 | 0.871 | 0.775 | 0.759 | 0.774 | 0.718 | 0.709 | 0.722 |
| Sample22 | 0.812 | 0.805 | 0.812 | 0.749 | 0.747 | 0.750 | 0.732 | 0.729 | 0.729 |
| Sample23 | 0.839 | 0.821 | 0.840 | 0.769 | 0.749 | 0.766 | 0.742 | 0.731 | 0.743 |
| Sample24 | 0.826 | 0.816 | 0.826 | 0.783 | 0.789 | 0.789 | 0.726 | 0.722 | 0.735 |
| Sample25 | 0.833 | 0.821 | 0.833 | 0.766 | 0.759 | 0.765 | 0.698 | 0.693 | 0.698 |
| Sample26 | 0.846 | 0.836 | 0.846 | 0.786 | 0.773 | 0.791 | 0.707 | 0.700 | 0.712 |
| Sample27 | 0.832 | 0.810 | 0.831 | 0.787 | 0.778 | 0.788 | 0.712 | 0.708 | 0.718 |
| Sample28 | 0.853 | 0.841 | 0.852 | 0.785 | 0.773 | 0.788 | 0.709 | 0.703 | 0.715 |
| Sample29 | 0.828 | 0.813 | 0.829 | 0.780 | 0.771 | 0.781 | 0.716 | 0.711 | 0.716 |
| Sample30 | 0.849 | 0.838 | 0.849 | 0.762 | 0.751 | 0.764 | 0.708 | 0.700 | 0.714 |
| Sample31 | 0.842 | 0.824 | 0.843 | 0.779 | 0.767 | 0.778 | 0.729 | 0.710 | 0.728 |
| Sample32 | 0.850 | 0.826 | 0.850 | 0.772 | 0.764 | 0.771 | 0.714 | 0.721 | 0.726 |
| Sample33 | 0.825 | 0.809 | 0.824 | 0.759 | 0.752 | 0.765 | 0.724 | 0.708 | 0.717 |
| Sample34 | 0.870 | 0.855 | 0.869 | 0.790 | 0.783 | 0.793 | 0.711 | 0.702 | 0.708 |
| Sample35 | 0.818 | 0.804 | 0.819 | 0.768 | 0.761 | 0.764 | 0.709 | 0.689 | 0.707 |
| Sample36 | 0.837 | 0.818 | 0.837 | 0.814 | 0.801 | 0.812 | 0.735 | 0.729 | 0.744 |
| Sample37 | 0.827 | 0.801 | 0.826 | 0.760 | 0.757 | 0.762 | 0.724 | 0.716 | 0.729 |
| Sample38 | 0.849 | 0.830 | 0.849 | 0.776 | 0.766 | 0.776 | 0.716 | 0.706 | 0.720 |
| Sample39 | 0.830 | 0.818 | 0.830 | 0.774 | 0.761 | 0.773 | 0.715 | 0.706 | 0.715 |
| Sample40 | 0.825 | 0.811 | 0.824 | 0.760 | 0.752 | 0.766 | 0.715 | 0.713 | 0.720 |
| Sample41 | 0.839 | 0.829 | 0.838 | 0.764 | 0.758 | 0.769 | 0.680 | 0.674 | 0.687 |
| Sample42 | 0.828 | 0.812 | 0.827 | 0.762 | 0.752 | 0.761 | 0.746 | 0.744 | 0.749 |
| Sample43 | 0.805 | 0.796 | 0.805 | 0.773 | 0.763 | 0.778 | 0.754 | 0.747 | 0.755 |
| Sample44 | 0.859 | 0.851 | 0.858 | 0.768 | 0.740 | 0.769 | 0.720 | 0.713 | 0.719 |
| Sample45 | 0.825 | 0.805 | 0.825 | 0.755 | 0.752 | 0.755 | 0.708 | 0.703 | 0.709 |
| Sample46 | 0.837 | 0.812 | 0.838 | 0.778 | 0.763 | 0.777 | 0.712 | 0.714 | 0.719 |
| Sample47 | 0.839 | 0.827 | 0.839 | 0.778 | 0.761 | 0.776 | 0.721 | 0.724 | 0.730 |
| Sample48 | 0.833 | 0.819 | 0.833 | 0.739 | 0.729 | 0.738 | 0.739 | 0.735 | 0.746 |
| Sample49 | 0.839 | 0.822 | 0.838 | 0.773 | 0.756 | 0.775 | 0.744 | 0.735 | 0.743 |
| Sample50 | 0.811 | 0.796 | 0.811 | 0.779 | 0.773 | 0.782 | 0.711 | 0.701 | 0.713 |
| | | | | | | | | | |
| Mean | 0.836 | 0.821 | 0.836 | 0.774 | 0.763 | 0.775 | 0.722 | 0.716 | 0.725 |
| SD | 0.017 | 0.018 | 0.017 | 0.013 | 0.013 | 0.013 | 0.018 | 0.019 | 0.019 |

### *Correlations of the θ estimates with the true θ values for each sample (new form)*

| | Condition 1: Var(testlet)=0 | | | Condition 2: Var(testlet)=1 | | | Condition 3: Var(testlet)=2 | | |
|---|---|---|---|---|---|---|---|---|---|
| | r($\hat{\theta}_{3PL}$, $\theta$) | r($\hat{\theta}_{GRM}$, $\theta$) | r($\hat{\theta}_{Testlet}$, $\theta$) | r($\hat{\theta}_{3PL}$, $\theta$) | r($\hat{\theta}_{GRM}$, $\theta$) | r($\hat{\theta}_{Testlet}$ $\theta$) | r($\hat{\theta}_{3PL}$, $\theta$) | r($\hat{\theta}_{GRM}$, $\theta$) | r($\hat{\theta}_{Testlet}$, $\theta$) |
| Sample1 | 0.907 | 0.899 | 0.906 | 0.875 | 0.871 | 0.878 | 0.841 | 0.837 | 0.843 |
| Sample2 | 0.910 | 0.904 | 0.911 | 0.880 | 0.869 | 0.882 | 0.849 | 0.844 | 0.851 |
| Sample3 | 0.911 | 0.901 | 0.910 | 0.869 | 0.862 | 0.869 | 0.818 | 0.813 | 0.821 |
| Sample4 | 0.910 | 0.902 | 0.910 | 0.869 | 0.860 | 0.869 | 0.847 | 0.836 | 0.852 |
| Sample5 | 0.887 | 0.872 | 0.886 | 0.865 | 0.856 | 0.864 | 0.829 | 0.824 | 0.833 |
| Sample6 | 0.899 | 0.887 | 0.898 | 0.874 | 0.860 | 0.874 | 0.834 | 0.834 | 0.839 |
| Sample7 | 0.911 | 0.901 | 0.912 | 0.886 | 0.879 | 0.886 | 0.854 | 0.848 | 0.855 |
| Sample8 | 0.902 | 0.890 | 0.902 | 0.868 | 0.861 | 0.869 | 0.838 | 0.838 | 0.843 |
| Sample9 | 0.908 | 0.899 | 0.908 | 0.876 | 0.867 | 0.875 | 0.826 | 0.818 | 0.828 |
| Sample10 | 0.894 | 0.887 | 0.894 | 0.873 | 0.869 | 0.874 | 0.851 | 0.845 | 0.853 |
| Sample11 | 0.891 | 0.885 | 0.890 | 0.873 | 0.863 | 0.874 | 0.826 | 0.818 | 0.828 |
| Sample12 | 0.895 | 0.883 | 0.895 | 0.859 | 0.845 | 0.859 | 0.820 | 0.816 | 0.818 |
| Sample13 | 0.909 | 0.899 | 0.909 | 0.868 | 0.861 | 0.869 | 0.825 | 0.816 | 0.826 |
| Sample14 | 0.908 | 0.897 | 0.908 | 0.882 | 0.877 | 0.883 | 0.832 | 0.830 | 0.832 |
| Sample15 | 0.888 | 0.877 | 0.888 | 0.870 | 0.863 | 0.871 | 0.841 | 0.832 | 0.838 |
| Sample16 | 0.909 | 0.900 | 0.909 | 0.877 | 0.870 | 0.879 | 0.834 | 0.831 | 0.835 |
| Sample17 | 0.906 | 0.891 | 0.906 | 0.881 | 0.874 | 0.883 | 0.843 | 0.840 | 0.843 |
| Sample18 | 0.902 | 0.889 | 0.902 | 0.866 | 0.855 | 0.866 | 0.847 | 0.842 | 0.851 |
| Sample19 | 0.895 | 0.885 | 0.895 | 0.860 | 0.854 | 0.860 | 0.850 | 0.834 | 0.849 |
| Sample20 | 0.900 | 0.885 | 0.900 | 0.877 | 0.869 | 0.876 | 0.847 | 0.839 | 0.847 |
| Sample21 | 0.903 | 0.887 | 0.903 | 0.865 | 0.856 | 0.866 | 0.854 | 0.842 | 0.853 |
| Sample22 | 0.898 | 0.893 | 0.897 | 0.875 | 0.870 | 0.877 | 0.842 | 0.840 | 0.843 |
| Sample23 | 0.908 | 0.898 | 0.908 | 0.859 | 0.849 | 0.858 | 0.840 | 0.835 | 0.841 |
| Sample24 | 0.904 | 0.897 | 0.904 | 0.871 | 0.860 | 0.871 | 0.838 | 0.830 | 0.842 |
| Sample25 | 0.916 | 0.901 | 0.916 | 0.864 | 0.851 | 0.864 | 0.847 | 0.838 | 0.850 |
| Sample26 | 0.902 | 0.894 | 0.901 | 0.852 | 0.846 | 0.853 | 0.841 | 0.834 | 0.842 |
| Sample27 | 0.915 | 0.904 | 0.915 | 0.862 | 0.859 | 0.863 | 0.846 | 0.841 | 0.848 |
| Sample28 | 0.907 | 0.895 | 0.906 | 0.876 | 0.868 | 0.875 | 0.841 | 0.836 | 0.844 |
| Sample29 | 0.905 | 0.895 | 0.906 | 0.859 | 0.852 | 0.860 | 0.851 | 0.843 | 0.851 |
| Sample30 | 0.907 | 0.899 | 0.907 | 0.877 | 0.866 | 0.877 | 0.847 | 0.846 | 0.848 |
| Sample31 | 0.909 | 0.899 | 0.909 | 0.875 | 0.869 | 0.875 | 0.845 | 0.841 | 0.849 |
| Sample32 | 0.916 | 0.902 | 0.916 | 0.854 | 0.853 | 0.855 | 0.846 | 0.842 | 0.845 |
| Sample33 | 0.905 | 0.894 | 0.905 | 0.875 | 0.860 | 0.876 | 0.818 | 0.812 | 0.822 |
| Sample34 | 0.902 | 0.891 | 0.902 | 0.861 | 0.851 | 0.861 | 0.830 | 0.824 | 0.833 |
| Sample35 | 0.891 | 0.876 | 0.891 | 0.861 | 0.852 | 0.862 | 0.818 | 0.808 | 0.819 |
| Sample36 | 0.914 | 0.902 | 0.914 | 0.863 | 0.858 | 0.862 | 0.814 | 0.807 | 0.813 |
| Sample37 | 0.907 | 0.896 | 0.907 | 0.866 | 0.862 | 0.868 | 0.824 | 0.814 | 0.827 |
| Sample38 | 0.902 | 0.892 | 0.902 | 0.880 | 0.875 | 0.880 | 0.828 | 0.819 | 0.828 |
| Sample39 | 0.915 | 0.903 | 0.915 | 0.877 | 0.872 | 0.877 | 0.830 | 0.821 | 0.833 |
| Sample40 | 0.895 | 0.890 | 0.894 | 0.869 | 0.860 | 0.869 | 0.833 | 0.826 | 0.834 |
| Sample41 | 0.915 | 0.910 | 0.916 | 0.849 | 0.847 | 0.854 | 0.820 | 0.813 | 0.825 |
| Sample42 | 0.891 | 0.882 | 0.891 | 0.877 | 0.874 | 0.878 | 0.841 | 0.837 | 0.847 |
| Sample43 | 0.888 | 0.880 | 0.888 | 0.866 | 0.855 | 0.865 | 0.856 | 0.849 | 0.856 |
| Sample44 | 0.912 | 0.897 | 0.911 | 0.873 | 0.869 | 0.873 | 0.831 | 0.821 | 0.832 |
| Sample45 | 0.901 | 0.889 | 0.901 | 0.867 | 0.860 | 0.866 | 0.822 | 0.827 | 0.827 |
| Sample46 | 0.909 | 0.897 | 0.909 | 0.867 | 0.861 | 0.868 | 0.833 | 0.828 | 0.835 |
| Sample47 | 0.906 | 0.894 | 0.905 | 0.874 | 0.867 | 0.874 | 0.849 | 0.842 | 0.851 |
| Sample48 | 0.908 | 0.900 | 0.908 | 0.850 | 0.840 | 0.852 | 0.844 | 0.843 | 0.847 |
| Sample49 | 0.911 | 0.897 | 0.910 | 0.881 | 0.875 | 0.882 | 0.826 | 0.817 | 0.827 |
| Sample50 | 0.892 | 0.882 | 0.892 | 0.873 | 0.863 | 0.872 | 0.826 | 0.820 | 0.830 |
| | | | | | | | | | |
| Mean | 0.904 | 0.893 | 0.904 | 0.869 | 0.862 | 0.870 | 0.837 | 0.831 | 0.838 |
| SD | 0.008 | 0.008 | 0.008 | 0.009 | 0.009 | 0.009 | 0.011 | 0.012 | 0.011 |

***MAD of θ estimates for each sample (base form)***

| | Condition 1: Var(testlet)=0 | | | Condition 2: Var(testlet)=1 | | | Condition 3: Var(testlet)=2 | | |
|---|---|---|---|---|---|---|---|---|---|
| | 3-PL | GRM | Testlet | 3-PL | GRM | Testlet | 3-PL | GRM | Testlet |
| Sample1 | 0.149 | 0.286 | 0.129 | 0.184 | 0.249 | 0.245 | 1.161 | 1.497 | 1.500 |
| Sample2 | 0.158 | 0.391 | 0.191 | 0.831 | 0.866 | 0.758 | 0.057 | 0.269 | 0.313 |
| Sample3 | 0.258 | 0.268 | 0.271 | 0.342 | 0.576 | 0.484 | 0.561 | 0.674 | 0.568 |
| Sample4 | 0.665 | 0.859 | 0.700 | 0.248 | 0.364 | 0.315 | 0.372 | 0.264 | 0.394 |
| Sample5 | 0.006 | 0.104 | 0.003 | 0.915 | 0.988 | 0.892 | 1.246 | 1.382 | 1.230 |
| Sample6 | 0.689 | 0.655 | 0.708 | 0.026 | 0.022 | 0.069 | 0.202 | 0.111 | 0.160 |
| Sample7 | 0.402 | 0.368 | 0.435 | 1.371 | 1.347 | 1.401 | 0.926 | 0.741 | 0.781 |
| Sample8 | 1.170 | 0.849 | 1.117 | 0.720 | 0.958 | 0.737 | 0.360 | 0.313 | 0.247 |
| Sample9 | 0.099 | 0.134 | 0.131 | 0.462 | 0.420 | 0.406 | 0.546 | 0.527 | 0.544 |
| Sample10 | 0.951 | 0.867 | 0.951 | 0.955 | 0.706 | 0.893 | 1.124 | 1.082 | 1.048 |
| Sample11 | 0.311 | 0.442 | 0.295 | 0.888 | 0.823 | 0.869 | 0.162 | 0.019 | 0.151 |
| Sample12 | 0.011 | 0.135 | 0.018 | 0.587 | 0.666 | 0.560 | 0.747 | 0.624 | 0.629 |
| Sample13 | 1.149 | 1.052 | 1.143 | 0.282 | 0.158 | 0.368 | 0.412 | 0.483 | 0.645 |
| Sample14 | 0.107 | 0.063 | 0.116 | 1.089 | 1.176 | 1.093 | 0.716 | 0.817 | 0.892 |
| Sample15 | 0.347 | 0.143 | 0.358 | 0.918 | 0.930 | 0.956 | 0.876 | 0.568 | 0.567 |
| Sample16 | 0.840 | 0.680 | 0.793 | 0.661 | 0.289 | 0.463 | 0.153 | 0.120 | 0.082 |
| Sample17 | 0.230 | 0.120 | 0.240 | 0.063 | 0.115 | 0.190 | 0.909 | 0.851 | 0.735 |
| Sample18 | 0.115 | 0.035 | 0.145 | 0.595 | 0.539 | 0.469 | 1.378 | 1.059 | 1.183 |
| Sample19 | 0.044 | 0.036 | 0.050 | 1.159 | 0.898 | 1.196 | 1.423 | 1.370 | 1.269 |
| Sample20 | 0.671 | 0.379 | 0.687 | 0.235 | 0.215 | 0.077 | 0.848 | 1.062 | 1.086 |
| Sample21 | 0.037 | 0.210 | 0.041 | 0.872 | 1.069 | 0.986 | 0.535 | 0.524 | 0.539 |
| Sample22 | 0.129 | 0.094 | 0.149 | 0.428 | 0.147 | 0.352 | 0.868 | 0.782 | 0.853 |
| Sample23 | 0.104 | 0.069 | 0.141 | 0.767 | 0.659 | 0.754 | 0.847 | 1.150 | 0.945 |
| Sample24 | 0.293 | 0.410 | 0.258 | 0.009 | 0.068 | 0.041 | 0.428 | 0.138 | 0.250 |
| Sample25 | 0.485 | 0.267 | 0.480 | 0.068 | 0.396 | 0.097 | 0.190 | 0.217 | 0.340 |
| Sample26 | 0.252 | 0.439 | 0.244 | 0.260 | 0.324 | 0.202 | 0.903 | 0.980 | 0.986 |
| Sample27 | 1.264 | 0.705 | 1.274 | 1.002 | 1.058 | 0.748 | 0.063 | 0.064 | 0.086 |
| Sample28 | 0.609 | 0.598 | 0.602 | 0.008 | 0.171 | 0.095 | 0.409 | 0.164 | 0.257 |
| Sample29 | 0.324 | 0.389 | 0.337 | 0.145 | 0.405 | 0.148 | 0.033 | 0.036 | 0.086 |
| Sample30 | 0.192 | 0.215 | 0.142 | 0.077 | 0.189 | 0.117 | 0.381 | 0.339 | 0.286 |
| Sample31 | 0.759 | 0.876 | 0.746 | 0.903 | 0.801 | 0.711 | 1.282 | 1.142 | 1.210 |
| Sample32 | 0.153 | 0.021 | 0.173 | 0.234 | 0.418 | 0.445 | 1.073 | 0.875 | 0.927 |
| Sample33 | 0.369 | 0.391 | 0.371 | 0.148 | 0.408 | 0.160 | 0.260 | 0.243 | 0.319 |
| Sample34 | 0.049 | 0.129 | 0.051 | 0.617 | 0.474 | 0.575 | 0.019 | 0.163 | 0.092 |
| Sample35 | 0.755 | 0.848 | 0.725 | 0.733 | 0.716 | 0.730 | 0.156 | 0.170 | 0.156 |
| Sample36 | 0.263 | 0.322 | 0.282 | 0.313 | 0.282 | 0.412 | 0.530 | 0.132 | 0.293 |
| Sample37 | 0.775 | 0.680 | 0.745 | 1.439 | 1.468 | 1.537 | 0.970 | 0.806 | 1.072 |
| Sample38 | 0.108 | 0.264 | 0.071 | 0.938 | 0.987 | 1.085 | 0.196 | 0.100 | 0.245 |
| Sample39 | 0.308 | 0.481 | 0.316 | 0.680 | 0.858 | 0.712 | 0.067 | 0.316 | 0.210 |
| Sample40 | 0.062 | 0.131 | 0.040 | 0.773 | 0.570 | 0.766 | 0.084 | 0.151 | 0.084 |
| Sample41 | 0.389 | 0.294 | 0.398 | 0.313 | 0.350 | 0.343 | 1.244 | 0.948 | 0.910 |
| Sample42 | 0.019 | 0.114 | 0.096 | 0.425 | 0.053 | 0.380 | 0.161 | 0.258 | 0.220 |
| Sample43 | 0.465 | 0.278 | 0.489 | 0.378 | 0.293 | 0.202 | 0.254 | 0.290 | 0.428 |
| Sample44 | 0.565 | 0.794 | 0.646 | 0.069 | 0.066 | 0.033 | 0.232 | 0.203 | 0.411 |
| Sample45 | 0.484 | 0.394 | 0.475 | 1.013 | 0.924 | 1.167 | 0.586 | 0.656 | 0.355 |
| Sample46 | 0.726 | 1.025 | 0.740 | 0.305 | 0.286 | 0.315 | 0.239 | 0.294 | 0.354 |
| Sample47 | 0.762 | 1.000 | 0.742 | 0.084 | 0.234 | 0.056 | 0.567 | 0.317 | 0.396 |
| Sample48 | 0.741 | 0.920 | 0.741 | 0.485 | 0.155 | 0.289 | 1.286 | 1.237 | 1.084 |
| Sample49 | 0.716 | 0.985 | 0.749 | 0.653 | 0.722 | 0.657 | 0.856 | 0.685 | 0.719 |
| Sample50 | 0.526 | 0.433 | 0.562 | 0.965 | 1.047 | 1.122 | 0.791 | 0.979 | 0.782 |
| | | | | | | | | | |
| Mean | 0.421 | 0.433 | 0.426 | 0.553 | 0.558 | 0.554 | 0.594 | 0.564 | 0.578 |
| SD | 0.331 | 0.313 | 0.326 | 0.381 | 0.372 | 0.388 | 0.418 | 0.416 | 0.388 |

### *MAD of rescaled θ estimates for each sample (new form)*

| | Condition 1: Var(testlet)=0 | | | Condition 2: Var(testlet)=1 | | | Condition 3: Var(testlet)=2 | | |
|---|---|---|---|---|---|---|---|---|---|
| | 3-PL | GRM | Testlet | 3-PL | GRM | Testlet | 3-PL | GRM | Testlet |
| Sample1 | 0.499 | 0.523 | 0.500 | 0.580 | 0.587 | 0.575 | 0.621 | 0.630 | 0.619 |
| Sample2 | 0.465 | 0.478 | 0.463 | 0.563 | 0.601 | 0.560 | 0.639 | 0.655 | 0.636 |
| Sample3 | 0.484 | 0.512 | 0.484 | 0.584 | 0.593 | 0.578 | 0.683 | 0.689 | 0.675 |
| Sample4 | 0.535 | 0.552 | 0.532 | 0.582 | 0.607 | 0.588 | 0.664 | 0.686 | 0.645 |
| Sample5 | 0.525 | 0.556 | 0.528 | 0.597 | 0.618 | 0.585 | 0.669 | 0.679 | 0.661 |
| Sample6 | 0.504 | 0.530 | 0.506 | 0.589 | 0.621 | 0.587 | 0.678 | 0.677 | 0.679 |
| Sample7 | 0.497 | 0.532 | 0.497 | 0.568 | 0.591 | 0.569 | 0.658 | 0.657 | 0.650 |
| Sample8 | 0.511 | 0.539 | 0.514 | 0.579 | 0.594 | 0.574 | 0.707 | 0.686 | 0.682 |
| Sample9 | 0.495 | 0.515 | 0.495 | 0.596 | 0.619 | 0.591 | 0.693 | 0.739 | 0.706 |
| Sample10 | 0.537 | 0.556 | 0.539 | 0.602 | 0.608 | 0.597 | 0.642 | 0.645 | 0.635 |
| Sample11 | 0.551 | 0.568 | 0.538 | 0.592 | 0.617 | 0.592 | 0.694 | 0.707 | 0.693 |
| Sample12 | 0.570 | 0.599 | 0.566 | 0.587 | 0.611 | 0.586 | 0.677 | 0.672 | 0.676 |
| Sample13 | 0.486 | 0.519 | 0.483 | 0.606 | 0.617 | 0.609 | 0.670 | 0.684 | 0.671 |
| Sample14 | 0.477 | 0.507 | 0.476 | 0.575 | 0.587 | 0.569 | 0.632 | 0.629 | 0.630 |
| Sample15 | 0.554 | 0.572 | 0.552 | 0.609 | 0.626 | 0.611 | 0.676 | 0.705 | 0.688 |
| Sample16 | 0.490 | 0.509 | 0.486 | 0.576 | 0.593 | 0.574 | 0.650 | 0.658 | 0.654 |
| Sample17 | 0.489 | 0.527 | 0.488 | 0.565 | 0.583 | 0.561 | 0.646 | 0.653 | 0.647 |
| Sample18 | 0.522 | 0.559 | 0.539 | 0.600 | 0.614 | 0.604 | 0.620 | 0.628 | 0.611 |
| Sample19 | 0.508 | 0.543 | 0.513 | 0.621 | 0.630 | 0.620 | 0.650 | 0.669 | 0.646 |
| Sample20 | 0.514 | 0.569 | 0.509 | 0.562 | 0.575 | 0.562 | 0.637 | 0.652 | 0.667 |
| Sample21 | 0.503 | 0.539 | 0.508 | 0.586 | 0.599 | 0.582 | 0.654 | 0.681 | 0.670 |
| Sample22 | 0.540 | 0.548 | 0.544 | 0.586 | 0.594 | 0.588 | 0.653 | 0.653 | 0.650 |
| Sample23 | 0.492 | 0.514 | 0.492 | 0.600 | 0.627 | 0.603 | 0.651 | 0.657 | 0.636 |
| Sample24 | 0.512 | 0.528 | 0.514 | 0.574 | 0.598 | 0.577 | 0.637 | 0.650 | 0.631 |
| Sample25 | 0.490 | 0.526 | 0.487 | 0.605 | 0.622 | 0.603 | 0.617 | 0.645 | 0.621 |
| Sample26 | 0.505 | 0.532 | 0.507 | 0.617 | 0.625 | 0.615 | 0.654 | 0.680 | 0.655 |
| Sample27 | 0.474 | 0.504 | 0.472 | 0.605 | 0.607 | 0.602 | 0.650 | 0.692 | 0.701 |
| Sample28 | 0.479 | 0.511 | 0.482 | 0.588 | 0.606 | 0.589 | 0.639 | 0.647 | 0.637 |
| Sample29 | 0.538 | 0.567 | 0.526 | 0.597 | 0.606 | 0.591 | 0.669 | 0.685 | 0.667 |
| Sample30 | 0.508 | 0.523 | 0.505 | 0.590 | 0.597 | 0.584 | 0.659 | 0.662 | 0.655 |
| Sample31 | 0.482 | 0.499 | 0.479 | 0.577 | 0.589 | 0.577 | 0.675 | 0.674 | 0.657 |
| Sample32 | 0.455 | 0.489 | 0.460 | 0.608 | 0.612 | 0.611 | 0.638 | 0.644 | 0.635 |
| Sample33 | 0.507 | 0.536 | 0.509 | 0.595 | 0.631 | 0.594 | 0.672 | 0.692 | 0.668 |
| Sample34 | 0.495 | 0.524 | 0.499 | 0.594 | 0.613 | 0.596 | 0.678 | 0.680 | 0.666 |
| Sample35 | 0.541 | 0.569 | 0.538 | 0.615 | 0.635 | 0.612 | 0.677 | 0.693 | 0.678 |
| Sample36 | 0.487 | 0.519 | 0.487 | 0.619 | 0.638 | 0.625 | 0.634 | 0.644 | 0.639 |
| Sample37 | 0.507 | 0.527 | 0.503 | 0.578 | 0.586 | 0.576 | 0.684 | 0.694 | 0.680 |
| Sample38 | 0.522 | 0.550 | 0.523 | 0.581 | 0.590 | 0.574 | 0.656 | 0.673 | 0.653 |
| Sample39 | 0.480 | 0.518 | 0.485 | 0.620 | 0.623 | 0.614 | 0.674 | 0.677 | 0.659 |
| Sample40 | 0.511 | 0.524 | 0.513 | 0.573 | 0.598 | 0.572 | 0.649 | 0.663 | 0.645 |
| Sample41 | 0.477 | 0.494 | 0.474 | 0.612 | 0.616 | 0.605 | 0.683 | 0.691 | 0.669 |
| Sample42 | 0.504 | 0.524 | 0.503 | 0.580 | 0.581 | 0.575 | 0.672 | 0.661 | 0.651 |
| Sample43 | 0.523 | 0.543 | 0.524 | 0.590 | 0.609 | 0.591 | 0.629 | 0.648 | 0.633 |
| Sample44 | 0.487 | 0.531 | 0.486 | 0.588 | 0.589 | 0.576 | 0.696 | 0.709 | 0.692 |
| Sample45 | 0.511 | 0.537 | 0.509 | 0.583 | 0.600 | 0.586 | 0.692 | 0.668 | 0.668 |
| Sample46 | 0.499 | 0.536 | 0.507 | 0.629 | 0.633 | 0.615 | 0.677 | 0.685 | 0.669 |
| Sample47 | 0.510 | 0.540 | 0.512 | 0.603 | 0.620 | 0.612 | 0.637 | 0.641 | 0.624 |
| Sample48 | 0.523 | 0.542 | 0.523 | 0.629 | 0.638 | 0.616 | 0.634 | 0.639 | 0.625 |
| Sample49 | 0.520 | 0.557 | 0.524 | 0.598 | 0.606 | 0.596 | 0.693 | 0.684 | 0.662 |
| Sample50 | 0.519 | 0.548 | 0.532 | 0.606 | 0.637 | 0.607 | 0.678 | 0.697 | 0.669 |
| | | | | | | | | | |
| Mean | 0.506 | 0.533 | 0.507 | 0.593 | 0.608 | 0.591 | 0.660 | 0.670 | 0.657 |
| SD | 0.024 | 0.024 | 0.024 | 0.017 | 0.017 | 0.017 | 0.023 | 0.024 | 0.022 |

### *RMSD of θ estimates for each sample (base form)*

| | **Condition 1: Var(testlet)=0** | | | **Condition 2: Var(testlet)=1** | | | **Condition 3: Var(testlet)=2** | | |
|---|---|---|---|---|---|---|---|---|---|
| | 3-PL | GRM | Testlet | 3-PL | GRM | Testlet | 3-PL | GRM | Testlet |
| Sample1 | 0.559 | 0.576 | 0.559 | 0.627 | 0.638 | 0.627 | 0.724 | 0.720 | 0.717 |
| Sample2 | 0.527 | 0.549 | 0.526 | 0.639 | 0.662 | 0.633 | 0.713 | 0.713 | 0.706 |
| Sample3 | 0.545 | 0.562 | 0.544 | 0.637 | 0.644 | 0.632 | 0.687 | 0.684 | 0.680 |
| Sample4 | 0.514 | 0.532 | 0.515 | 0.611 | 0.619 | 0.602 | 0.699 | 0.706 | 0.686 |
| Sample5 | 0.591 | 0.626 | 0.594 | 0.657 | 0.672 | 0.657 | 0.720 | 0.719 | 0.719 |
| Sample6 | 0.521 | 0.545 | 0.522 | 0.626 | 0.633 | 0.625 | 0.721 | 0.715 | 0.711 |
| Sample7 | 0.569 | 0.594 | 0.570 | 0.620 | 0.638 | 0.620 | 0.687 | 0.683 | 0.672 |
| Sample8 | 0.559 | 0.584 | 0.559 | 0.619 | 0.627 | 0.619 | 0.719 | 0.717 | 0.707 |
| Sample9 | 0.541 | 0.551 | 0.541 | 0.642 | 0.652 | 0.638 | 0.731 | 0.724 | 0.712 |
| Sample10 | 0.584 | 0.593 | 0.584 | 0.607 | 0.614 | 0.597 | 0.700 | 0.695 | 0.684 |
| Sample11 | 0.496 | 0.545 | 0.497 | 0.647 | 0.657 | 0.648 | 0.698 | 0.699 | 0.698 |
| Sample12 | 0.539 | 0.579 | 0.540 | 0.619 | 0.637 | 0.612 | 0.723 | 0.714 | 0.714 |
| Sample13 | 0.549 | 0.576 | 0.549 | 0.625 | 0.642 | 0.625 | 0.694 | 0.705 | 0.696 |
| Sample14 | 0.556 | 0.573 | 0.557 | 0.656 | 0.656 | 0.654 | 0.668 | 0.652 | 0.646 |
| Sample15 | 0.538 | 0.546 | 0.541 | 0.619 | 0.635 | 0.619 | 0.691 | 0.687 | 0.680 |
| Sample16 | 0.501 | 0.525 | 0.502 | 0.664 | 0.672 | 0.660 | 0.674 | 0.685 | 0.659 |
| Sample17 | 0.541 | 0.559 | 0.540 | 0.652 | 0.663 | 0.646 | 0.722 | 0.721 | 0.716 |
| Sample18 | 0.592 | 0.628 | 0.589 | 0.618 | 0.639 | 0.619 | 0.658 | 0.649 | 0.642 |
| Sample19 | 0.579 | 0.594 | 0.580 | 0.632 | 0.641 | 0.631 | 0.687 | 0.683 | 0.669 |
| Sample20 | 0.579 | 0.619 | 0.580 | 0.636 | 0.645 | 0.635 | 0.691 | 0.694 | 0.680 |
| Sample21 | 0.495 | 0.524 | 0.495 | 0.640 | 0.658 | 0.641 | 0.697 | 0.699 | 0.685 |
| Sample22 | 0.575 | 0.585 | 0.574 | 0.663 | 0.662 | 0.657 | 0.705 | 0.703 | 0.702 |
| Sample23 | 0.532 | 0.558 | 0.530 | 0.636 | 0.657 | 0.638 | 0.678 | 0.683 | 0.671 |
| Sample24 | 0.557 | 0.571 | 0.557 | 0.607 | 0.596 | 0.596 | 0.705 | 0.703 | 0.689 |
| Sample25 | 0.538 | 0.555 | 0.538 | 0.646 | 0.651 | 0.644 | 0.709 | 0.707 | 0.703 |
| Sample26 | 0.526 | 0.541 | 0.526 | 0.619 | 0.632 | 0.609 | 0.709 | 0.709 | 0.697 |
| Sample27 | 0.532 | 0.561 | 0.532 | 0.639 | 0.651 | 0.638 | 0.694 | 0.686 | 0.676 |
| Sample28 | 0.512 | 0.531 | 0.514 | 0.640 | 0.655 | 0.636 | 0.716 | 0.711 | 0.699 |
| Sample29 | 0.572 | 0.595 | 0.572 | 0.656 | 0.669 | 0.656 | 0.687 | 0.684 | 0.678 |
| Sample30 | 0.522 | 0.540 | 0.523 | 0.651 | 0.663 | 0.647 | 0.715 | 0.715 | 0.701 |
| Sample31 | 0.531 | 0.558 | 0.529 | 0.644 | 0.659 | 0.645 | 0.692 | 0.709 | 0.690 |
| Sample32 | 0.525 | 0.563 | 0.525 | 0.623 | 0.627 | 0.618 | 0.700 | 0.682 | 0.677 |
| Sample33 | 0.556 | 0.577 | 0.556 | 0.647 | 0.653 | 0.638 | 0.695 | 0.704 | 0.695 |
| Sample34 | 0.511 | 0.537 | 0.513 | 0.607 | 0.614 | 0.601 | 0.704 | 0.702 | 0.697 |
| Sample35 | 0.577 | 0.596 | 0.575 | 0.653 | 0.661 | 0.657 | 0.704 | 0.715 | 0.697 |
| Sample36 | 0.558 | 0.587 | 0.558 | 0.601 | 0.620 | 0.605 | 0.678 | 0.678 | 0.662 |
| Sample37 | 0.579 | 0.616 | 0.581 | 0.649 | 0.649 | 0.643 | 0.690 | 0.694 | 0.681 |
| Sample38 | 0.529 | 0.557 | 0.529 | 0.631 | 0.640 | 0.628 | 0.717 | 0.721 | 0.707 |
| Sample39 | 0.560 | 0.577 | 0.560 | 0.634 | 0.648 | 0.633 | 0.698 | 0.697 | 0.688 |
| Sample40 | 0.570 | 0.590 | 0.571 | 0.638 | 0.642 | 0.626 | 0.711 | 0.705 | 0.697 |
| Sample41 | 0.555 | 0.569 | 0.556 | 0.661 | 0.668 | 0.657 | 0.727 | 0.723 | 0.711 |
| Sample42 | 0.553 | 0.577 | 0.554 | 0.659 | 0.667 | 0.656 | 0.688 | 0.689 | 0.683 |
| Sample43 | 0.594 | 0.606 | 0.595 | 0.631 | 0.640 | 0.622 | 0.672 | 0.675 | 0.666 |
| Sample44 | 0.509 | 0.522 | 0.510 | 0.635 | 0.665 | 0.632 | 0.685 | 0.683 | 0.676 |
| Sample45 | 0.594 | 0.622 | 0.595 | 0.640 | 0.638 | 0.635 | 0.716 | 0.716 | 0.710 |
| Sample46 | 0.550 | 0.587 | 0.549 | 0.666 | 0.685 | 0.668 | 0.694 | 0.681 | 0.676 |
| Sample47 | 0.529 | 0.546 | 0.528 | 0.626 | 0.643 | 0.626 | 0.668 | 0.651 | 0.646 |
| Sample48 | 0.565 | 0.586 | 0.567 | 0.677 | 0.687 | 0.676 | 0.671 | 0.669 | 0.657 |
| Sample49 | 0.557 | 0.583 | 0.558 | 0.642 | 0.662 | 0.639 | 0.684 | 0.691 | 0.682 |
| Sample50 | 0.573 | 0.593 | 0.572 | 0.635 | 0.641 | 0.630 | 0.701 | 0.705 | 0.694 |
| | | | | | | | | | |
| Mean | 0.548 | 0.571 | 0.549 | 0.637 | 0.648 | 0.634 | 0.698 | 0.697 | 0.688 |
| SD | 0.027 | 0.027 | 0.027 | 0.017 | 0.018 | 0.018 | 0.017 | 0.019 | 0.019 |

### *RMSD of rescaled θ estimates for each sample (new form)*

| | Condition 1: Var(testlet)=0 | | | Condition 2: Var(testlet)=1 | | | Condition 3: Var(testlet)=2 | | |
|---|---|---|---|---|---|---|---|---|---|
| | 3-PL | GRM | Testlet | 3-PL | GRM | Testlet | 3-PL | GRM | Testlet |
| Sample1 | 0.629 | 0.661 | 0.633 | 0.727 | 0.739 | 0.721 | 0.798 | 0.806 | 0.793 |
| Sample2 | 0.602 | 0.624 | 0.603 | 0.728 | 0.769 | 0.723 | 0.807 | 0.821 | 0.797 |
| Sample3 | 0.622 | 0.654 | 0.623 | 0.751 | 0.766 | 0.744 | 0.854 | 0.865 | 0.851 |
| Sample4 | 0.682 | 0.700 | 0.678 | 0.731 | 0.760 | 0.739 | 0.820 | 0.846 | 0.804 |
| Sample5 | 0.670 | 0.716 | 0.675 | 0.760 | 0.784 | 0.749 | 0.850 | 0.861 | 0.835 |
| Sample6 | 0.642 | 0.677 | 0.644 | 0.742 | 0.772 | 0.736 | 0.854 | 0.857 | 0.860 |
| Sample7 | 0.634 | 0.673 | 0.634 | 0.720 | 0.743 | 0.723 | 0.813 | 0.821 | 0.806 |
| Sample8 | 0.650 | 0.684 | 0.651 | 0.739 | 0.756 | 0.726 | 0.889 | 0.870 | 0.861 |
| Sample9 | 0.624 | 0.659 | 0.627 | 0.760 | 0.783 | 0.755 | 0.863 | 0.915 | 0.875 |
| Sample10 | 0.688 | 0.707 | 0.692 | 0.755 | 0.760 | 0.750 | 0.817 | 0.824 | 0.805 |
| Sample11 | 0.704 | 0.721 | 0.686 | 0.752 | 0.782 | 0.752 | 0.869 | 0.889 | 0.871 |
| Sample12 | 0.729 | 0.768 | 0.724 | 0.748 | 0.780 | 0.749 | 0.856 | 0.856 | 0.858 |
| Sample13 | 0.619 | 0.659 | 0.616 | 0.766 | 0.780 | 0.766 | 0.836 | 0.846 | 0.833 |
| Sample14 | 0.612 | 0.647 | 0.611 | 0.716 | 0.726 | 0.710 | 0.790 | 0.787 | 0.788 |
| Sample15 | 0.710 | 0.728 | 0.708 | 0.764 | 0.790 | 0.770 | 0.843 | 0.879 | 0.860 |
| Sample16 | 0.627 | 0.647 | 0.620 | 0.721 | 0.741 | 0.715 | 0.825 | 0.840 | 0.834 |
| Sample17 | 0.624 | 0.675 | 0.625 | 0.727 | 0.749 | 0.722 | 0.822 | 0.829 | 0.823 |
| Sample18 | 0.666 | 0.714 | 0.685 | 0.767 | 0.792 | 0.771 | 0.791 | 0.804 | 0.777 |
| Sample19 | 0.655 | 0.693 | 0.659 | 0.787 | 0.799 | 0.786 | 0.809 | 0.836 | 0.808 |
| Sample20 | 0.655 | 0.730 | 0.651 | 0.717 | 0.733 | 0.717 | 0.790 | 0.815 | 0.824 |
| Sample21 | 0.645 | 0.687 | 0.652 | 0.737 | 0.757 | 0.734 | 0.819 | 0.853 | 0.841 |
| Sample22 | 0.678 | 0.694 | 0.680 | 0.726 | 0.737 | 0.726 | 0.829 | 0.831 | 0.827 |
| Sample23 | 0.629 | 0.661 | 0.629 | 0.762 | 0.796 | 0.764 | 0.826 | 0.829 | 0.805 |
| Sample24 | 0.643 | 0.661 | 0.644 | 0.725 | 0.758 | 0.729 | 0.808 | 0.825 | 0.804 |
| Sample25 | 0.624 | 0.673 | 0.621 | 0.756 | 0.783 | 0.753 | 0.780 | 0.804 | 0.779 |
| Sample26 | 0.634 | 0.660 | 0.637 | 0.790 | 0.799 | 0.785 | 0.822 | 0.846 | 0.822 |
| Sample27 | 0.597 | 0.630 | 0.596 | 0.764 | 0.765 | 0.760 | 0.820 | 0.867 | 0.870 |
| Sample28 | 0.612 | 0.647 | 0.615 | 0.737 | 0.761 | 0.739 | 0.806 | 0.815 | 0.800 |
| Sample29 | 0.706 | 0.743 | 0.690 | 0.767 | 0.783 | 0.761 | 0.843 | 0.865 | 0.845 |
| Sample30 | 0.645 | 0.666 | 0.639 | 0.746 | 0.759 | 0.737 | 0.832 | 0.837 | 0.831 |
| Sample31 | 0.613 | 0.640 | 0.611 | 0.738 | 0.759 | 0.740 | 0.854 | 0.852 | 0.837 |
| Sample32 | 0.581 | 0.628 | 0.586 | 0.771 | 0.778 | 0.777 | 0.799 | 0.808 | 0.799 |
| Sample33 | 0.645 | 0.683 | 0.648 | 0.753 | 0.794 | 0.752 | 0.849 | 0.866 | 0.841 |
| Sample34 | 0.640 | 0.675 | 0.643 | 0.746 | 0.774 | 0.749 | 0.854 | 0.861 | 0.842 |
| Sample35 | 0.693 | 0.721 | 0.688 | 0.777 | 0.802 | 0.775 | 0.851 | 0.872 | 0.849 |
| Sample36 | 0.615 | 0.656 | 0.613 | 0.784 | 0.818 | 0.804 | 0.818 | 0.828 | 0.821 |
| Sample37 | 0.662 | 0.681 | 0.659 | 0.725 | 0.737 | 0.721 | 0.855 | 0.871 | 0.846 |
| Sample38 | 0.656 | 0.690 | 0.656 | 0.721 | 0.733 | 0.717 | 0.829 | 0.853 | 0.827 |
| Sample39 | 0.610 | 0.652 | 0.616 | 0.783 | 0.796 | 0.776 | 0.851 | 0.860 | 0.835 |
| Sample40 | 0.642 | 0.663 | 0.645 | 0.728 | 0.756 | 0.729 | 0.824 | 0.837 | 0.820 |
| Sample41 | 0.607 | 0.622 | 0.601 | 0.780 | 0.787 | 0.768 | 0.857 | 0.867 | 0.840 |
| Sample42 | 0.651 | 0.683 | 0.651 | 0.734 | 0.736 | 0.722 | 0.839 | 0.829 | 0.813 |
| Sample43 | 0.666 | 0.692 | 0.669 | 0.734 | 0.763 | 0.737 | 0.790 | 0.817 | 0.801 |
| Sample44 | 0.637 | 0.692 | 0.636 | 0.748 | 0.746 | 0.733 | 0.858 | 0.881 | 0.853 |
| Sample45 | 0.655 | 0.685 | 0.653 | 0.744 | 0.761 | 0.747 | 0.874 | 0.845 | 0.845 |
| Sample46 | 0.632 | 0.675 | 0.644 | 0.784 | 0.793 | 0.771 | 0.843 | 0.853 | 0.836 |
| Sample47 | 0.640 | 0.673 | 0.643 | 0.762 | 0.788 | 0.774 | 0.798 | 0.802 | 0.781 |
| Sample48 | 0.657 | 0.687 | 0.655 | 0.786 | 0.799 | 0.772 | 0.791 | 0.800 | 0.782 |
| Sample49 | 0.656 | 0.702 | 0.661 | 0.755 | 0.776 | 0.754 | 0.867 | 0.856 | 0.831 |
| Sample50 | 0.672 | 0.704 | 0.691 | 0.766 | 0.807 | 0.769 | 0.855 | 0.870 | 0.844 |
| | | | | | | | | | |
| Mean | 0.646 | 0.679 | 0.646 | 0.750 | 0.770 | 0.748 | 0.831 | 0.843 | 0.827 |
| SD | 0.031 | 0.031 | 0.030 | 0.021 | 0.023 | 0.022 | 0.026 | 0.027 | 0.026 |

## *Mean TIF of θ estimates for each sample (base form)*

| | Condition 1: Var(testlet)=0 | | | Condition 2: Var(testlet)=1 | | | Condition 3: Var(testlet)=2 | | |
|---|---|---|---|---|---|---|---|---|---|
| | 3-PL | GRM | Testlet | 3-PL | GRM | Testlet | 3-PL | GRM | Testlet |
| Sample1 | 3.319 | 2.985 | 3.016 | 3.256 | 2.395 | 2.408 | 3.395 | 2.102 | 2.173 |
| Sample2 | 3.638 | 3.201 | 3.271 | 3.681 | 2.651 | 2.762 | 3.440 | 2.227 | 2.181 |
| Sample3 | 3.653 | 3.391 | 3.256 | 3.447 | 2.479 | 2.564 | 3.297 | 2.166 | 2.185 |
| Sample4 | 3.748 | 3.434 | 3.306 | 3.883 | 2.780 | 2.898 | 3.426 | 2.218 | 2.241 |
| Sample5 | 2.797 | 2.551 | 2.532 | 3.538 | 2.600 | 2.636 | 3.307 | 2.191 | 2.215 |
| Sample6 | 3.628 | 3.293 | 3.253 | 3.168 | 2.221 | 2.412 | 2.845 | 1.963 | 2.119 |
| Sample7 | 3.309 | 2.990 | 2.985 | 3.916 | 2.867 | 2.844 | 3.634 | 2.149 | 2.232 |
| Sample8 | 3.293 | 2.966 | 2.932 | 3.430 | 2.496 | 2.499 | 3.038 | 2.132 | 2.189 |
| Sample9 | 3.431 | 3.151 | 3.058 | 3.399 | 2.395 | 2.389 | 2.968 | 1.866 | 1.921 |
| Sample10 | 3.219 | 2.901 | 2.888 | 3.196 | 2.296 | 2.424 | 3.228 | 2.147 | 2.141 |
| Sample11 | 3.931 | 3.457 | 3.505 | 3.429 | 2.441 | 2.545 | 3.756 | 2.448 | 2.458 |
| Sample12 | 4.237 | 3.551 | 3.754 | 3.376 | 2.458 | 2.445 | 3.103 | 2.082 | 2.135 |
| Sample13 | 3.985 | 3.684 | 3.514 | 3.720 | 2.795 | 2.781 | 2.984 | 1.882 | 2.001 |
| Sample14 | 3.627 | 3.342 | 3.217 | 3.155 | 2.302 | 2.367 | 3.627 | 2.284 | 2.302 |
| Sample15 | 3.706 | 3.504 | 3.273 | 3.365 | 2.510 | 2.543 | 3.715 | 2.327 | 2.328 |
| Sample16 | 3.992 | 3.435 | 3.537 | 3.192 | 2.299 | 2.419 | 3.641 | 2.172 | 2.267 |
| Sample17 | 3.650 | 3.387 | 3.218 | 3.176 | 2.418 | 2.464 | 2.982 | 1.972 | 1.988 |
| Sample18 | 3.391 | 3.000 | 2.998 | 3.256 | 2.267 | 2.491 | 3.230 | 2.072 | 2.050 |
| Sample19 | 3.334 | 3.005 | 2.979 | 3.331 | 2.445 | 2.552 | 3.336 | 2.128 | 2.139 |
| Sample20 | 3.153 | 2.768 | 2.877 | 3.071 | 2.252 | 2.349 | 3.139 | 1.947 | 2.038 |
| Sample21 | 4.238 | 3.771 | 3.759 | 3.151 | 2.287 | 2.392 | 3.168 | 2.088 | 2.240 |
| Sample22 | 3.104 | 2.810 | 2.739 | 3.139 | 2.387 | 2.392 | 3.553 | 2.309 | 2.280 |
| Sample23 | 3.777 | 3.279 | 3.374 | 2.995 | 2.170 | 2.396 | 3.405 | 2.215 | 2.244 |
| Sample24 | 3.433 | 3.103 | 3.043 | 3.440 | 2.503 | 2.538 | 3.403 | 2.160 | 2.211 |
| Sample25 | 3.385 | 3.105 | 3.021 | 3.205 | 2.352 | 2.491 | 2.771 | 1.850 | 1.967 |
| Sample26 | 3.746 | 3.279 | 3.355 | 3.680 | 2.548 | 2.606 | 2.951 | 1.896 | 1.976 |
| Sample27 | 3.512 | 2.967 | 3.164 | 3.523 | 2.478 | 2.577 | 3.325 | 2.024 | 2.046 |
| Sample28 | 3.724 | 3.322 | 3.270 | 3.492 | 2.529 | 2.666 | 3.504 | 2.133 | 2.220 |
| Sample29 | 3.371 | 3.168 | 3.012 | 3.297 | 2.404 | 2.450 | 3.014 | 1.944 | 2.064 |
| Sample30 | 3.696 | 3.432 | 3.310 | 3.087 | 2.387 | 2.382 | 3.221 | 2.033 | 2.151 |
| Sample31 | 3.810 | 3.548 | 3.332 | 3.445 | 2.597 | 2.685 | 3.104 | 2.017 | 2.090 |
| Sample32 | 3.890 | 3.463 | 3.425 | 3.643 | 2.659 | 2.631 | 3.360 | 2.155 | 2.153 |
| Sample33 | 3.615 | 3.204 | 3.233 | 3.167 | 2.418 | 2.459 | 3.351 | 2.207 | 2.272 |
| Sample34 | 4.297 | 3.794 | 3.759 | 3.524 | 2.533 | 2.697 | 3.317 | 2.164 | 2.221 |
| Sample35 | 3.321 | 3.071 | 2.936 | 3.100 | 2.333 | 2.439 | 2.994 | 1.935 | 2.049 |
| Sample36 | 3.780 | 3.361 | 3.354 | 3.760 | 2.664 | 2.695 | 3.237 | 2.078 | 2.210 |
| Sample37 | 3.400 | 2.973 | 3.014 | 3.355 | 2.505 | 2.521 | 3.047 | 2.070 | 2.200 |
| Sample38 | 3.564 | 3.176 | 3.172 | 3.405 | 2.400 | 2.496 | 3.257 | 2.003 | 2.173 |
| Sample39 | 3.760 | 3.391 | 3.280 | 3.135 | 2.340 | 2.479 | 3.178 | 2.028 | 2.060 |
| Sample40 | 3.411 | 3.179 | 3.040 | 3.265 | 2.369 | 2.479 | 3.551 | 2.278 | 2.231 |
| Sample41 | 3.614 | 3.280 | 3.153 | 2.989 | 2.261 | 2.324 | 2.814 | 1.822 | 1.930 |
| Sample42 | 3.382 | 3.028 | 3.082 | 3.476 | 2.576 | 2.653 | 3.353 | 2.167 | 2.281 |
| Sample43 | 3.015 | 2.760 | 2.725 | 3.441 | 2.558 | 2.661 | 3.583 | 2.306 | 2.299 |
| Sample44 | 4.036 | 3.424 | 3.566 | 3.077 | 2.196 | 2.369 | 3.143 | 2.066 | 2.149 |
| Sample45 | 3.061 | 2.668 | 2.762 | 3.273 | 2.442 | 2.470 | 2.891 | 1.927 | 2.039 |
| Sample46 | 3.730 | 3.219 | 3.310 | 3.621 | 2.634 | 2.754 | 3.250 | 2.119 | 2.186 |
| Sample47 | 3.578 | 3.340 | 3.218 | 3.335 | 2.403 | 2.507 | 3.230 | 1.988 | 2.062 |
| Sample48 | 3.179 | 3.010 | 2.862 | 2.919 | 2.200 | 2.365 | 3.308 | 2.144 | 2.165 |
| Sample49 | 3.661 | 3.167 | 3.277 | 3.310 | 2.398 | 2.497 | 3.407 | 2.130 | 2.112 |
| Sample50 | 3.080 | 2.780 | 2.769 | 3.484 | 2.586 | 2.604 | 2.876 | 1.925 | 2.093 |
| | | | | | | | | | |
| Mean | 3.564 | 3.201 | 3.174 | 3.354 | 2.450 | 2.529 | 3.253 | 2.093 | 2.154 |
| SD | 0.326 | 0.277 | 0.272 | 0.230 | 0.158 | 0.137 | 0.245 | 0.137 | 0.111 |

## *Mean TIF of θ estimates for each sample (new form)*

| | Condition 1: Var(testlet)=0 | | | Condition 2: Var(testlet)=1 | | | Condition 3: Var(testlet)=2 | | |
|---|---|---|---|---|---|---|---|---|---|
| | 3-PL | GRM | Testlet | 3-PL | GRM | Testlet | 3-PL | GRM | Testlet |
| Sample1 | 5.733 | 5.243 | 4.929 | 5.559 | 4.148 | 3.938 | 4.991 | 3.328 | 3.252 |
| Sample2 | 6.225 | 5.848 | 5.309 | 5.781 | 4.051 | 4.058 | 5.557 | 3.614 | 3.413 |
| Sample3 | 6.339 | 5.867 | 5.313 | 5.644 | 4.223 | 3.985 | 4.335 | 2.994 | 3.015 |
| Sample4 | 6.219 | 5.945 | 5.271 | 5.647 | 4.054 | 3.982 | 4.822 | 3.190 | 3.196 |
| Sample5 | 5.360 | 5.043 | 4.600 | 5.179 | 3.809 | 3.658 | 5.038 | 3.365 | 3.291 |
| Sample6 | 5.720 | 5.084 | 4.912 | 5.279 | 4.179 | 3.988 | 4.736 | 3.149 | 3.177 |
| Sample7 | 5.764 | 5.263 | 4.911 | 5.796 | 4.274 | 4.027 | 5.554 | 3.504 | 3.400 |
| Sample8 | 5.912 | 5.195 | 5.041 | 5.186 | 3.848 | 3.646 | 4.987 | 3.235 | 3.196 |
| Sample9 | 6.337 | 4.170 | 5.358 | 5.558 | 4.103 | 3.981 | 4.607 | 3.168 | 3.226 |
| Sample10 | 5.167 | 4.822 | 4.439 | 5.525 | 4.195 | 4.024 | 5.268 | 3.450 | 3.316 |
| Sample11 | 5.123 | 4.453 | 4.410 | 5.319 | 3.901 | 3.855 | 4.418 | 2.875 | 2.874 |
| Sample12 | 5.151 | 4.635 | 4.446 | 4.653 | 3.613 | 3.521 | 4.690 | 3.129 | 3.050 |
| Sample13 | 6.289 | 5.411 | 5.356 | 5.353 | 3.881 | 3.784 | 4.666 | 3.215 | 3.236 |
| Sample14 | 5.932 | 5.528 | 5.055 | 5.036 | 3.793 | 3.732 | 5.017 | 3.243 | 3.083 |
| Sample15 | 5.025 | 4.756 | 4.289 | 6.060 | 4.535 | 4.301 | 5.125 | 3.476 | 3.321 |
| Sample16 | 5.992 | 5.575 | 5.061 | 5.684 | 4.296 | 4.033 | 4.937 | 3.147 | 3.065 |
| Sample17 | 5.929 | 5.396 | 5.031 | 5.709 | 4.277 | 4.100 | 4.690 | 3.182 | 3.148 |
| Sample18 | 5.693 | 4.963 | 4.868 | 5.352 | 3.996 | 3.844 | 5.280 | 3.540 | 3.332 |
| Sample19 | 5.365 | 4.864 | 4.572 | 5.094 | 4.008 | 3.861 | 4.758 | 3.167 | 3.122 |
| Sample20 | 5.294 | 4.769 | 4.593 | 5.524 | 4.171 | 4.123 | 5.258 | 3.556 | 3.460 |
| Sample21 | 5.694 | 5.381 | 4.872 | 4.849 | 3.912 | 3.812 | 5.000 | 3.267 | 3.248 |
| Sample22 | 5.432 | 5.114 | 4.672 | 5.509 | 4.031 | 3.926 | 5.456 | 3.572 | 3.470 |
| Sample23 | 6.752 | 6.010 | 5.675 | 4.650 | 3.678 | 3.590 | 5.204 | 3.466 | 3.327 |
| Sample24 | 6.320 | 5.970 | 5.326 | 5.577 | 4.036 | 4.106 | 4.901 | 3.223 | 3.135 |
| Sample25 | 7.117 | 5.798 | 5.821 | 5.047 | 3.931 | 3.872 | 5.202 | 3.299 | 3.217 |
| Sample26 | 5.507 | 5.140 | 4.689 | 5.155 | 3.902 | 3.941 | 5.125 | 3.432 | 3.336 |
| Sample27 | 5.900 | 5.551 | 4.990 | 5.033 | 3.754 | 3.654 | 5.583 | 3.584 | 3.522 |
| Sample28 | 5.902 | 5.298 | 4.993 | 5.306 | 4.167 | 4.008 | 5.352 | 3.216 | 3.158 |
| Sample29 | 5.955 | 5.394 | 5.060 | 4.974 | 3.640 | 3.787 | 5.075 | 3.445 | 3.331 |
| Sample30 | 6.101 | 5.627 | 5.152 | 5.565 | 4.148 | 4.115 | 5.039 | 3.432 | 3.403 |
| Sample31 | 5.816 | 5.241 | 4.929 | 5.250 | 4.010 | 4.009 | 4.951 | 3.388 | 3.191 |
| Sample32 | 6.346 | 5.704 | 5.374 | 4.485 | 3.607 | 3.497 | 5.232 | 3.310 | 3.173 |
| Sample33 | 6.155 | 5.547 | 5.256 | 5.534 | 4.116 | 4.086 | 4.720 | 3.073 | 3.189 |
| Sample34 | 5.879 | 5.500 | 5.020 | 4.828 | 3.567 | 3.544 | 5.056 | 3.421 | 3.290 |
| Sample35 | 5.024 | 4.640 | 4.314 | 5.335 | 4.079 | 4.054 | 4.475 | 2.929 | 2.977 |
| Sample36 | 6.791 | 6.088 | 5.750 | 4.864 | 3.655 | 3.534 | 4.371 | 2.899 | 2.886 |
| Sample37 | 5.348 | 5.050 | 4.628 | 5.340 | 4.072 | 3.943 | 4.530 | 3.076 | 2.948 |
| Sample38 | 5.749 | 5.106 | 4.883 | 5.645 | 4.348 | 4.247 | 4.428 | 2.967 | 2.996 |
| Sample39 | 5.948 | 5.414 | 5.073 | 5.461 | 4.204 | 4.183 | 4.830 | 3.215 | 3.206 |
| Sample40 | 5.192 | 4.785 | 4.446 | 5.285 | 4.068 | 3.901 | 4.694 | 3.038 | 2.971 |
| Sample41 | 6.699 | 6.111 | 5.629 | 5.120 | 3.916 | 3.819 | 4.623 | 3.118 | 3.052 |
| Sample42 | 5.219 | 4.756 | 4.475 | 5.817 | 4.424 | 4.374 | 5.220 | 3.372 | 3.277 |
| Sample43 | 5.387 | 5.006 | 4.619 | 4.934 | 3.632 | 3.672 | 5.323 | 3.490 | 3.346 |
| Sample44 | 6.587 | 5.680 | 5.536 | 4.962 | 3.865 | 3.785 | 4.768 | 3.245 | 3.244 |
| Sample45 | 5.703 | 4.994 | 4.909 | 5.323 | 4.182 | 4.019 | 4.999 | 3.231 | 3.130 |
| Sample46 | 6.076 | 5.518 | 5.116 | 5.149 | 3.966 | 3.852 | 5.022 | 3.373 | 3.368 |
| Sample47 | 5.644 | 5.240 | 4.868 | 4.881 | 3.707 | 3.741 | 5.319 | 3.363 | 3.325 |
| Sample48 | 5.987 | 5.322 | 5.044 | 4.788 | 3.766 | 3.813 | 4.774 | 3.206 | 3.110 |
| Sample49 | 6.242 | 5.544 | 5.283 | 5.933 | 4.453 | 4.411 | 4.527 | 3.115 | 3.037 |
| Sample50 | 5.443 | 5.131 | 4.648 | 5.385 | 4.036 | 4.015 | 4.819 | 3.159 | 3.136 |
| | | | | | | | | | |
| Mean | 5.850 | 5.290 | 4.976 | 5.298 | 4.005 | 3.915 | 4.947 | 3.269 | 3.203 |
| SD | 0.494 | 0.439 | 0.380 | 0.355 | 0.236 | 0.214 | 0.324 | 0.187 | 0.153 |

## Appendix D Evaluation Criteria for Item Parameter Estimates

### *Correlations of the 3-PL and testlet model estimated item parameters with the true parameter values (base form, Var(testlet)=0)*

| | 3-PL | | | Testlet | | |
|---|---|---|---|---|---|---|
| | $r(\hat{a}_{3PL}, a)$ | $r|(\hat{b}_{3PL}, b)$ | $r(\hat{c}_{3PL}, c)$ | $r(\hat{a}_{testlet}, a)$ | $r(\hat{b}_{testlet}, b)$ | $r(\hat{c}_{testlet}, c)$ |
| Sample1 | 0.868 | 0.928 | 0.176 | 0.858 | 0.941 | 0.184 |
| Sample2 | 0.865 | 0.954 | 0.402 | 0.891 | 0.963 | 0.413 |
| Sample3 | 0.902 | 0.959 | 0.121 | 0.917 | 0.968 | 0.138 |
| Sample4 | 0.855 | 0.930 | 0.189 | 0.887 | 0.940 | 0.166 |
| Sample5 | 0.837 | 0.934 | -0.052 | 0.848 | 0.931 | -0.069 |
| Sample6 | 0.899 | 0.928 | 0.221 | 0.911 | 0.935 | 0.204 |
| Sample7 | 0.872 | 0.905 | 0.114 | 0.883 | 0.917 | 0.091 |
| Sample8 | 0.886 | 0.961 | 0.173 | 0.907 | 0.970 | 0.271 |
| Sample9 | 0.678 | 0.920 | 0.347 | 0.686 | 0.928 | 0.362 |
| Sample10 | 0.843 | 0.947 | 0.453 | 0.804 | 0.948 | 0.483 |
| Sample11 | 0.948 | 0.926 | 0.294 | 0.943 | 0.932 | 0.223 |
| Sample12 | 0.918 | 0.960 | 0.588 | 0.907 | 0.967 | 0.621 |
| Sample13 | 0.874 | 0.954 | 0.149 | 0.889 | 0.956 | 0.135 |
| Sample14 | 0.806 | 0.969 | 0.226 | 0.827 | 0.974 | 0.233 |
| Sample15 | 0.866 | 0.966 | 0.568 | 0.875 | 0.963 | 0.529 |
| Sample16 | 0.906 | 0.979 | 0.455 | 0.899 | 0.982 | 0.435 |
| Sample17 | 0.833 | 0.938 | 0.411 | 0.871 | 0.937 | 0.421 |
| Sample18 | 0.832 | 0.976 | 0.430 | 0.848 | 0.976 | 0.344 |
| Sample19 | 0.876 | 0.975 | 0.525 | 0.898 | 0.973 | 0.496 |
| Sample20 | 0.824 | 0.969 | -0.063 | 0.856 | 0.971 | -0.058 |
| Sample21 | 0.910 | 0.945 | 0.250 | 0.905 | 0.958 | 0.296 |
| Sample22 | 0.620 | 0.881 | 0.212 | 0.697 | 0.882 | 0.212 |
| Sample23 | 0.858 | 0.960 | 0.282 | 0.866 | 0.961 | 0.280 |
| Sample24 | 0.882 | 0.972 | 0.278 | 0.860 | 0.977 | 0.270 |
| Sample25 | 0.873 | 0.969 | 0.296 | 0.892 | 0.973 | 0.353 |
| Sample26 | 0.811 | 0.928 | 0.261 | 0.793 | 0.933 | 0.237 |
| Sample27 | 0.873 | 0.965 | 0.425 | 0.890 | 0.964 | 0.435 |
| Sample28 | 0.812 | 0.959 | 0.186 | 0.815 | 0.956 | 0.165 |
| Sample29 | 0.762 | 0.977 | 0.408 | 0.764 | 0.979 | 0.425 |
| Sample30 | 0.827 | 0.960 | 0.492 | 0.830 | 0.968 | 0.579 |
| Sample31 | 0.642 | 0.961 | 0.340 | 0.710 | 0.968 | 0.376 |
| Sample32 | 0.916 | 0.959 | 0.268 | 0.915 | 0.969 | 0.253 |
| Sample33 | 0.841 | 0.949 | 0.175 | 0.875 | 0.955 | 0.130 |
| Sample34 | 0.849 | 0.966 | 0.358 | 0.843 | 0.969 | 0.348 |
| Sample35 | 0.787 | 0.927 | 0.162 | 0.802 | 0.942 | 0.126 |
| Sample36 | 0.862 | 0.972 | 0.538 | 0.864 | 0.978 | 0.535 |
| Sample37 | 0.878 | 0.938 | 0.298 | 0.873 | 0.930 | 0.271 |
| Sample38 | 0.896 | 0.927 | 0.014 | 0.916 | 0.944 | 0.017 |
| Sample39 | 0.851 | 0.941 | 0.369 | 0.856 | 0.953 | 0.337 |
| Sample40 | 0.778 | 0.953 | 0.438 | 0.800 | 0.955 | 0.385 |
| Sample41 | 0.918 | 0.946 | 0.368 | 0.922 | 0.959 | 0.399 |
| Sample42 | 0.852 | 0.916 | 0.104 | 0.875 | 0.926 | 0.060 |
| Sample43 | 0.489 | 0.950 | 0.419 | 0.553 | 0.954 | 0.391 |
| Sample44 | 0.941 | 0.968 | 0.532 | 0.942 | 0.969 | 0.498 |
| Sample45 | 0.932 | 0.960 | 0.373 | 0.901 | 0.966 | 0.366 |
| Sample46 | 0.836 | 0.956 | 0.505 | 0.850 | 0.962 | 0.526 |
| Sample47 | 0.880 | 0.967 | 0.139 | 0.864 | 0.969 | 0.158 |
| Sample48 | 0.743 | 0.941 | -0.053 | 0.777 | 0.943 | -0.097 |
| Sample49 | 0.903 | 0.969 | 0.550 | 0.888 | 0.973 | 0.564 |
| Sample50 | 0.852 | 0.949 | 0.449 | 0.849 | 0.961 | 0.473 |
| | | | | | | |
| Mean | 0.841 | 0.950 | 0.303 | 0.852 | 0.955 | 0.300 |
| SD | 0.085 | 0.021 | 0.168 | 0.072 | 0.020 | 0.175 |

***Correlations of the 3-PL and testlet model estimated item parameters with the true parameter values (base form, Var(testlet)=1)***

| | 3-PL | | | Testlet | | |
|---|---|---|---|---|---|---|
| | $r(\hat{a}_{3PL}, a)$ | $rl(\hat{b}_{3PL}, b)$ | $r(\hat{c}_{3PL}, c)$ | $r(\hat{a}_{testlet}, a)$ | $r(\hat{b}_{testlet}, b)$ | $r(\hat{c}_{testlet}, c)$ |
| Sample1 | 0.834 | 0.973 | 0.542 | 0.861 | 0.967 | 0.479 |
| Sample2 | 0.809 | 0.958 | 0.367 | 0.900 | 0.965 | 0.257 |
| Sample3 | 0.829 | 0.955 | 0.284 | 0.886 | 0.970 | 0.423 |
| Sample4 | 0.872 | 0.924 | -0.015 | 0.927 | 0.950 | 0.173 |
| Sample5 | 0.708 | 0.975 | 0.075 | 0.834 | 0.977 | 0.238 |
| Sample6 | 0.825 | 0.908 | 0.261 | 0.882 | 0.921 | 0.134 |
| Sample7 | 0.801 | 0.927 | 0.235 | 0.862 | 0.933 | 0.215 |
| Sample8 | 0.717 | 0.957 | 0.232 | 0.742 | 0.953 | 0.186 |
| Sample9 | 0.827 | 0.948 | 0.217 | 0.848 | 0.953 | 0.323 |
| Sample10 | 0.724 | 0.958 | 0.224 | 0.744 | 0.967 | 0.234 |
| Sample11 | 0.860 | 0.961 | 0.460 | 0.901 | 0.956 | 0.429 |
| Sample12 | 0.666 | 0.948 | 0.477 | 0.749 | 0.960 | 0.382 |
| Sample13 | 0.738 | 0.972 | 0.270 | 0.755 | 0.977 | 0.285 |
| Sample14 | 0.755 | 0.927 | 0.050 | 0.807 | 0.948 | 0.258 |
| Sample15 | 0.812 | 0.966 | 0.426 | 0.844 | 0.963 | 0.337 |
| Sample16 | 0.816 | 0.958 | 0.453 | 0.874 | 0.957 | 0.442 |
| Sample17 | 0.580 | 0.937 | 0.096 | 0.828 | 0.953 | 0.102 |
| Sample18 | 0.898 | 0.975 | 0.285 | 0.918 | 0.979 | 0.337 |
| Sample19 | 0.780 | 0.934 | 0.349 | 0.834 | 0.948 | 0.392 |
| Sample20 | 0.637 | 0.967 | 0.055 | 0.623 | 0.966 | 0.053 |
| Sample21 | 0.885 | 0.929 | 0.382 | 0.936 | 0.941 | 0.456 |
| Sample22 | 0.792 | 0.942 | 0.468 | 0.749 | 0.939 | 0.280 |
| Sample23 | 0.797 | 0.921 | -0.284 | 0.858 | 0.939 | -0.009 |
| Sample24 | 0.761 | 0.936 | 0.261 | 0.826 | 0.944 | 0.245 |
| Sample25 | 0.700 | 0.977 | 0.535 | 0.712 | 0.973 | 0.508 |
| Sample26 | 0.877 | 0.975 | 0.412 | 0.879 | 0.982 | 0.549 |
| Sample27 | 0.688 | 0.956 | 0.024 | 0.856 | 0.966 | 0.057 |
| Sample28 | 0.715 | 0.962 | 0.283 | 0.729 | 0.971 | 0.178 |
| Sample29 | 0.755 | 0.960 | 0.341 | 0.837 | 0.971 | 0.534 |
| Sample30 | 0.735 | 0.937 | 0.304 | 0.802 | 0.945 | 0.262 |
| Sample31 | 0.799 | 0.969 | 0.516 | 0.820 | 0.961 | 0.490 |
| Sample32 | 0.687 | 0.969 | 0.226 | 0.755 | 0.975 | 0.284 |
| Sample33 | 0.777 | 0.919 | 0.529 | 0.840 | 0.940 | 0.477 |
| Sample34 | 0.714 | 0.957 | -0.002 | 0.792 | 0.952 | -0.041 |
| Sample35 | 0.747 | 0.903 | 0.283 | 0.769 | 0.903 | 0.230 |
| Sample36 | 0.831 | 0.945 | -0.023 | 0.893 | 0.963 | 0.289 |
| Sample37 | 0.841 | 0.951 | 0.627 | 0.837 | 0.950 | 0.640 |
| Sample38 | 0.732 | 0.968 | 0.335 | 0.901 | 0.973 | 0.583 |
| Sample39 | 0.453 | 0.935 | -0.090 | 0.665 | 0.949 | -0.048 |
| Sample40 | 0.818 | 0.918 | 0.267 | 0.883 | 0.931 | 0.271 |
| Sample41 | 0.704 | 0.960 | 0.290 | 0.778 | 0.964 | 0.244 |
| Sample42 | 0.800 | 0.922 | 0.083 | 0.835 | 0.929 | 0.157 |
| Sample43 | 0.674 | 0.972 | 0.501 | 0.852 | 0.976 | 0.634 |
| Sample44 | 0.840 | 0.936 | 0.469 | 0.810 | 0.940 | 0.439 |
| Sample45 | 0.825 | 0.951 | 0.411 | 0.822 | 0.955 | 0.417 |
| Sample46 | 0.753 | 0.962 | 0.160 | 0.821 | 0.966 | 0.054 |
| Sample47 | 0.875 | 0.967 | 0.472 | 0.917 | 0.962 | 0.477 |
| Sample48 | 0.673 | 0.926 | 0.005 | 0.777 | 0.927 | 0.086 |
| Sample49 | 0.848 | 0.932 | 0.495 | 0.850 | 0.949 | 0.558 |
| Sample50 | 0.693 | 0.970 | 0.251 | 0.776 | 0.975 | 0.377 |
| | | | | | | |
| Mean | 0.766 | 0.949 | 0.278 | 0.824 | 0.955 | 0.307 |
| SD | 0.084 | 0.020 | 0.195 | 0.067 | 0.017 | 0.176 |

### *Correlations of the 3-PL and testlet model estimated item parameters with the true parameter values (base form, Var(testlet)=2)*

| | 3-PL | | | Testlet | | |
|---|---|---|---|---|---|---|
| | $r(\hat{a}_{3PL}, a)$ | $r_l(\hat{b}_{3PL}, b)$ | $r(\hat{c}_{3PL}, c)$ | $r(\hat{a}_{testlet}, a)$ | $r(\hat{b}_{testlet}, b)$ | $r(\hat{c}_{testlet}, c)$ |
| Sample1 | 0.812 | 0.961 | 0.573 | 0.895 | 0.963 | 0.539 |
| Sample2 | 0.686 | 0.956 | 0.463 | 0.717 | 0.973 | 0.299 |
| Sample3 | 0.509 | 0.958 | 0.140 | 0.625 | 0.972 | 0.304 |
| Sample4 | 0.480 | 0.928 | 0.415 | 0.809 | 0.940 | 0.449 |
| Sample5 | 0.658 | 0.963 | 0.234 | 0.755 | 0.968 | 0.285 |
| Sample6 | 0.635 | 0.911 | 0.265 | 0.730 | 0.909 | 0.137 |
| Sample7 | 0.819 | 0.953 | 0.458 | 0.886 | 0.962 | 0.394 |
| Sample8 | 0.742 | 0.959 | 0.156 | 0.655 | 0.964 | 0.105 |
| Sample9 | 0.697 | 0.972 | 0.391 | 0.897 | 0.974 | 0.300 |
| Sample10 | 0.717 | 0.972 | 0.716 | 0.796 | 0.976 | 0.614 |
| Sample11 | 0.633 | 0.963 | 0.324 | 0.683 | 0.972 | 0.433 |
| Sample12 | 0.798 | 0.936 | 0.140 | 0.735 | 0.937 | 0.164 |
| Sample13 | 0.803 | 0.946 | 0.258 | 0.858 | 0.920 | 0.115 |
| Sample14 | 0.660 | 0.928 | 0.325 | 0.807 | 0.942 | 0.420 |
| Sample15 | 0.838 | 0.943 | 0.430 | 0.834 | 0.953 | 0.361 |
| Sample16 | 0.666 | 0.945 | 0.410 | 0.845 | 0.957 | 0.405 |
| Sample17 | 0.218 | 0.934 | 0.482 | 0.426 | 0.951 | 0.363 |
| Sample18 | 0.752 | 0.911 | 0.330 | 0.899 | 0.921 | 0.285 |
| Sample19 | 0.636 | 0.945 | 0.154 | 0.811 | 0.942 | 0.351 |
| Sample20 | 0.348 | 0.928 | 0.349 | 0.714 | 0.962 | 0.511 |
| Sample21 | 0.814 | 0.971 | 0.371 | 0.840 | 0.971 | 0.462 |
| Sample22 | 0.819 | 0.935 | 0.265 | 0.810 | 0.937 | 0.418 |
| Sample23 | 0.726 | 0.971 | 0.269 | 0.825 | 0.982 | 0.376 |
| Sample24 | 0.850 | 0.897 | 0.233 | 0.943 | 0.918 | 0.295 |
| Sample25 | 0.489 | 0.965 | 0.285 | 0.597 | 0.966 | 0.370 |
| Sample26 | 0.835 | 0.966 | 0.410 | 0.802 | 0.964 | 0.170 |
| Sample27 | 0.533 | 0.922 | 0.206 | 0.699 | 0.946 | 0.510 |
| Sample28 | 0.719 | 0.934 | 0.490 | 0.822 | 0.933 | 0.350 |
| Sample29 | 0.684 | 0.950 | 0.090 | 0.716 | 0.958 | 0.376 |
| Sample30 | 0.787 | 0.894 | 0.195 | 0.876 | 0.891 | 0.223 |
| Sample31 | 0.794 | 0.952 | 0.466 | 0.820 | 0.962 | 0.557 |
| Sample32 | 0.678 | 0.923 | 0.062 | 0.885 | 0.918 | 0.075 |
| Sample33 | 0.781 | 0.954 | 0.361 | 0.888 | 0.968 | 0.401 |
| Sample34 | 0.828 | 0.910 | 0.090 | 0.872 | 0.937 | 0.254 |
| Sample35 | 0.726 | 0.967 | 0.284 | 0.857 | 0.967 | 0.065 |
| Sample36 | 0.688 | 0.948 | 0.313 | 0.786 | 0.955 | 0.261 |
| Sample37 | 0.679 | 0.915 | 0.131 | 0.813 | 0.927 | 0.037 |
| Sample38 | 0.743 | 0.963 | 0.532 | 0.886 | 0.951 | 0.545 |
| Sample39 | 0.775 | 0.967 | 0.464 | 0.693 | 0.973 | 0.336 |
| Sample40 | 0.790 | 0.931 | 0.057 | 0.756 | 0.939 | 0.297 |
| Sample41 | 0.797 | 0.920 | 0.140 | 0.788 | 0.934 | 0.112 |
| Sample42 | 0.629 | 0.951 | 0.164 | 0.710 | 0.956 | 0.451 |
| Sample43 | 0.329 | 0.959 | 0.449 | 0.826 | 0.959 | 0.674 |
| Sample44 | 0.640 | 0.950 | 0.548 | 0.737 | 0.941 | 0.485 |
| Sample45 | 0.624 | 0.955 | 0.330 | 0.710 | 0.952 | 0.365 |
| Sample46 | 0.673 | 0.904 | 0.184 | 0.879 | 0.936 | 0.308 |
| Sample47 | 0.734 | 0.957 | 0.269 | 0.842 | 0.974 | 0.607 |
| Sample48 | 0.846 | 0.912 | 0.313 | 0.881 | 0.922 | 0.297 |
| Sample49 | 0.733 | 0.939 | 0.427 | 0.858 | 0.962 | 0.600 |
| Sample50 | 0.777 | 0.907 | 0.306 | 0.864 | 0.921 | 0.334 |
| | | | | | | |
| Mean | 0.693 | 0.942 | 0.314 | 0.793 | 0.949 | 0.349 |
| SD | 0.137 | 0.022 | 0.147 | 0.095 | 0.021 | 0.153 |

***Correlations of the 3-PL and testlet model estimated item parameters with the true parameter values (new Form, var(testlet)=0)***

| | **3-PL** | | | **Testlet** | | |
|---|---|---|---|---|---|---|
| | $r(\hat{a}_{3PL}, a)$ | $r(\hat{b}_{3PL}, b)$ | $r(\hat{c}_{3PL}, c)$ | $r(\hat{a}_{testlet}, a)$ | $r(\hat{b}_{testlet}, b)$ | $r(\hat{c}_{testlet}, c)$ |
| Sample1 | 0.932 | 0.975 | 0.225 | 0.930 | 0.980 | 0.237 |
| Sample2 | 0.898 | 0.939 | 0.491 | 0.901 | 0.946 | 0.463 |
| Sample3 | 0.925 | 0.965 | 0.250 | 0.917 | 0.971 | 0.236 |
| Sample4 | 0.856 | 0.972 | 0.613 | 0.856 | 0.970 | 0.581 |
| Sample5 | 0.901 | 0.951 | 0.306 | 0.902 | 0.963 | 0.306 |
| Sample6 | 0.920 | 0.953 | 0.352 | 0.918 | 0.951 | 0.317 |
| Sample7 | 0.943 | 0.941 | 0.397 | 0.948 | 0.954 | 0.453 |
| Sample8 | 0.906 | 0.962 | 0.096 | 0.921 | 0.967 | 0.193 |
| Sample9 | 0.900 | 0.957 | 0.243 | 0.908 | 0.963 | 0.245 |
| Sample10 | 0.763 | 0.921 | 0.391 | 0.740 | 0.937 | 0.430 |
| Sample11 | 0.917 | 0.944 | 0.269 | 0.911 | 0.947 | 0.372 |
| Sample12 | 0.910 | 0.962 | 0.580 | 0.913 | 0.970 | 0.547 |
| Sample13 | 0.956 | 0.931 | -0.013 | 0.953 | 0.945 | -0.067 |
| Sample14 | 0.925 | 0.965 | 0.225 | 0.921 | 0.972 | 0.272 |
| Sample15 | 0.761 | 0.969 | 0.309 | 0.754 | 0.969 | 0.293 |
| Sample16 | 0.944 | 0.967 | 0.090 | 0.948 | 0.972 | 0.148 |
| Sample17 | 0.906 | 0.965 | 0.439 | 0.913 | 0.970 | 0.431 |
| Sample18 | 0.961 | 0.924 | 0.575 | 0.959 | 0.929 | 0.554 |
| Sample19 | 0.885 | 0.940 | 0.174 | 0.896 | 0.946 | 0.253 |
| Sample20 | 0.922 | 0.938 | 0.343 | 0.926 | 0.945 | 0.365 |
| Sample21 | 0.946 | 0.954 | 0.292 | 0.948 | 0.962 | 0.311 |
| Sample22 | 0.749 | 0.946 | 0.357 | 0.759 | 0.952 | 0.368 |
| Sample23 | 0.874 | 0.955 | 0.260 | 0.875 | 0.959 | 0.177 |
| Sample24 | 0.893 | 0.960 | 0.350 | 0.894 | 0.965 | 0.305 |
| Sample25 | 0.952 | 0.954 | 0.358 | 0.933 | 0.959 | 0.426 |
| Sample26 | 0.922 | 0.955 | 0.090 | 0.925 | 0.960 | 0.118 |
| Sample27 | 0.939 | 0.918 | 0.287 | 0.952 | 0.940 | 0.328 |
| Sample28 | 0.929 | 0.954 | 0.426 | 0.929 | 0.963 | 0.495 |
| Sample29 | 0.902 | 0.966 | 0.499 | 0.910 | 0.971 | 0.497 |
| Sample30 | 0.926 | 0.947 | 0.367 | 0.922 | 0.953 | 0.346 |
| Sample31 | 0.903 | 0.968 | 0.265 | 0.902 | 0.969 | 0.297 |
| Sample32 | 0.941 | 0.967 | 0.259 | 0.943 | 0.969 | 0.298 |
| Sample33 | 0.953 | 0.913 | 0.500 | 0.951 | 0.926 | 0.521 |
| Sample34 | 0.905 | 0.956 | 0.269 | 0.914 | 0.962 | 0.303 |
| Sample35 | 0.801 | 0.966 | 0.270 | 0.805 | 0.968 | 0.186 |
| Sample36 | 0.951 | 0.976 | 0.578 | 0.945 | 0.967 | 0.538 |
| Sample37 | 0.916 | 0.907 | 0.283 | 0.929 | 0.927 | 0.354 |
| Sample38 | 0.918 | 0.920 | 0.266 | 0.917 | 0.921 | 0.258 |
| Sample39 | 0.932 | 0.939 | 0.356 | 0.929 | 0.956 | 0.351 |
| Sample40 | 0.928 | 0.975 | 0.552 | 0.923 | 0.974 | 0.512 |
| Sample41 | 0.869 | 0.981 | 0.463 | 0.875 | 0.984 | 0.464 |
| Sample42 | 0.943 | 0.915 | 0.150 | 0.938 | 0.924 | 0.158 |
| Sample43 | 0.959 | 0.953 | 0.168 | 0.963 | 0.957 | 0.219 |
| Sample44 | 0.956 | 0.955 | 0.272 | 0.963 | 0.956 | 0.275 |
| Sample45 | 0.957 | 0.965 | 0.294 | 0.958 | 0.967 | 0.344 |
| Sample46 | 0.919 | 0.971 | 0.474 | 0.903 | 0.974 | 0.400 |
| Sample47 | 0.871 | 0.965 | 0.018 | 0.864 | 0.967 | 0.041 |
| Sample48 | 0.924 | 0.973 | 0.362 | 0.919 | 0.978 | 0.335 |
| Sample49 | 0.899 | 0.959 | 0.522 | 0.898 | 0.962 | 0.555 |
| Sample50 | 0.870 | 0.947 | 0.178 | 0.870 | 0.958 | 0.113 |
| | | | | | | |
| Mean | 0.908 | 0.953 | 0.323 | 0.908 | 0.958 | 0.330 |
| SD | 0.049 | 0.018 | 0.148 | 0.050 | 0.015 | 0.142 |

***Correlations of the 3-PL and testlet model estimated item parameters with the true parameter values (new form, Var(testlet)=1)***

| | **3-PL** | | | **Testlet** | | |
|---|---|---|---|---|---|---|
| | $r(\hat{a}_{3PL}, a)$ | $r(\hat{b}_{3PL}, b)$ | $r(\hat{c}_{3PL}, c)$ | $r(\hat{a}_{testlet}, a)$ | $r(\hat{b}_{testlet}, b)$ | $r(\hat{c}_{testlet}, c)$ |
| Sample1 | 0.905 | 0.969 | 0.442 | 0.912 | 0.963 | 0.486 |
| Sample2 | 0.892 | 0.950 | 0.098 | 0.920 | 0.960 | 0.150 |
| Sample3 | 0.898 | 0.942 | 0.404 | 0.918 | 0.961 | 0.465 |
| Sample4 | 0.896 | 0.940 | 0.418 | 0.913 | 0.947 | 0.509 |
| Sample5 | 0.839 | 0.968 | 0.515 | 0.846 | 0.965 | 0.473 |
| Sample6 | 0.862 | 0.972 | 0.378 | 0.871 | 0.967 | 0.256 |
| Sample7 | 0.930 | 0.918 | 0.292 | 0.926 | 0.940 | 0.349 |
| Sample8 | 0.886 | 0.958 | 0.391 | 0.891 | 0.967 | 0.525 |
| Sample9 | 0.831 | 0.937 | 0.432 | 0.878 | 0.940 | 0.466 |
| Sample10 | 0.940 | 0.945 | 0.434 | 0.942 | 0.945 | 0.385 |
| Sample11 | 0.905 | 0.886 | 0.185 | 0.930 | 0.887 | 0.125 |
| Sample12 | 0.870 | 0.964 | 0.439 | 0.900 | 0.965 | 0.406 |
| Sample13 | 0.877 | 0.946 | 0.184 | 0.899 | 0.963 | 0.321 |
| Sample14 | 0.883 | 0.948 | 0.164 | 0.907 | 0.957 | 0.241 |
| Sample15 | 0.894 | 0.964 | 0.254 | 0.896 | 0.972 | 0.261 |
| Sample16 | 0.862 | 0.955 | 0.337 | 0.911 | 0.974 | 0.454 |
| Sample17 | 0.908 | 0.951 | 0.404 | 0.904 | 0.960 | 0.365 |
| Sample18 | 0.808 | 0.980 | 0.296 | 0.860 | 0.986 | 0.465 |
| Sample19 | 0.885 | 0.963 | 0.263 | 0.909 | 0.975 | 0.348 |
| Sample20 | 0.884 | 0.934 | 0.230 | 0.870 | 0.943 | 0.373 |
| Sample21 | 0.742 | 0.939 | 0.367 | 0.794 | 0.952 | 0.504 |
| Sample22 | 0.873 | 0.984 | 0.553 | 0.904 | 0.985 | 0.478 |
| Sample23 | 0.781 | 0.942 | 0.024 | 0.795 | 0.945 | 0.042 |
| Sample24 | 0.950 | 0.939 | 0.121 | 0.961 | 0.951 | 0.106 |
| Sample25 | 0.854 | 0.955 | 0.153 | 0.799 | 0.960 | 0.219 |
| Sample26 | 0.657 | 0.921 | 0.575 | 0.780 | 0.939 | 0.587 |
| Sample27 | 0.843 | 0.938 | 0.484 | 0.911 | 0.954 | 0.475 |
| Sample28 | 0.854 | 0.979 | 0.639 | 0.868 | 0.981 | 0.566 |
| Sample29 | 0.931 | 0.939 | 0.010 | 0.939 | 0.945 | 0.260 |
| Sample30 | 0.897 | 0.917 | 0.511 | 0.925 | 0.932 | 0.542 |
| Sample31 | 0.860 | 0.961 | 0.415 | 0.878 | 0.966 | 0.313 |
| Sample32 | 0.831 | 0.966 | 0.416 | 0.865 | 0.978 | 0.549 |
| Sample33 | 0.896 | 0.951 | 0.513 | 0.910 | 0.958 | 0.460 |
| Sample34 | 0.876 | 0.934 | -0.016 | 0.899 | 0.944 | -0.024 |
| Sample35 | 0.855 | 0.985 | 0.625 | 0.900 | 0.983 | 0.594 |
| Sample36 | 0.673 | 0.952 | 0.045 | 0.677 | 0.959 | 0.079 |
| Sample37 | 0.801 | 0.977 | 0.318 | 0.856 | 0.979 | 0.357 |
| Sample38 | 0.868 | 0.972 | 0.578 | 0.854 | 0.973 | 0.523 |
| Sample39 | 0.876 | 0.962 | 0.474 | 0.916 | 0.970 | 0.506 |
| Sample40 | 0.898 | 0.936 | 0.389 | 0.927 | 0.944 | 0.385 |
| Sample41 | 0.835 | 0.951 | 0.276 | 0.832 | 0.957 | 0.223 |
| Sample42 | 0.802 | 0.951 | 0.504 | 0.887 | 0.956 | 0.529 |
| Sample43 | 0.885 | 0.950 | 0.315 | 0.929 | 0.958 | 0.403 |
| Sample44 | 0.843 | 0.946 | 0.419 | 0.879 | 0.963 | 0.456 |
| Sample45 | 0.843 | 0.953 | 0.455 | 0.844 | 0.960 | 0.385 |
| Sample46 | 0.778 | 0.954 | 0.071 | 0.884 | 0.980 | 0.226 |
| Sample47 | 0.895 | 0.943 | 0.318 | 0.885 | 0.947 | 0.262 |
| Sample48 | 0.890 | 0.942 | 0.384 | 0.899 | 0.947 | 0.442 |
| Sample49 | 0.835 | 0.971 | 0.192 | 0.845 | 0.974 | 0.185 |
| Sample50 | 0.850 | 0.966 | 0.166 | 0.865 | 0.978 | 0.309 |
| | | | | | | |
| Mean | 0.859 | 0.951 | 0.337 | 0.882 | 0.959 | 0.367 |
| SD | 0.058 | 0.019 | 0.168 | 0.049 | 0.017 | 0.152 |

### *Correlations of the 3-PL and testlet model estimated item parameters with the true parameter values (new form, Var(testlet)=2)*

| | **3-PL** | | | **Testlet** | | |
|---|---|---|---|---|---|---|
| | $r(\hat{a}_{3PL}, a)$ | $r_l(\hat{b}_{3PL}, b)$ | $r(\hat{c}_{3PL}, c)$ | $r(\hat{a}_{testlet}, a)$ | $r(\hat{b}_{testlet}, b)$ | $r(\hat{c}_{testlet}, c)$ |
| Sample1 | 0.879 | 0.942 | 0.329 | 0.858 | 0.960 | 0.455 |
| Sample2 | 0.740 | 0.947 | 0.415 | 0.805 | 0.953 | 0.384 |
| Sample3 | 0.821 | 0.958 | -0.045 | 0.830 | 0.951 | -0.167 |
| Sample4 | 0.875 | 0.945 | 0.309 | 0.930 | 0.947 | 0.398 |
| Sample5 | 0.769 | 0.979 | 0.447 | 0.883 | 0.985 | 0.493 |
| Sample6 | 0.859 | 0.926 | 0.295 | 0.900 | 0.940 | 0.295 |
| Sample7 | 0.788 | 0.969 | 0.261 | 0.894 | 0.968 | 0.489 |
| Sample8 | 0.883 | 0.920 | 0.268 | 0.913 | 0.935 | 0.362 |
| Sample9 | 0.909 | 0.923 | 0.404 | 0.942 | 0.916 | 0.505 |
| Sample10 | 0.790 | 0.962 | 0.508 | 0.751 | 0.962 | 0.635 |
| Sample11 | 0.958 | 0.898 | 0.111 | 0.954 | 0.912 | 0.208 |
| Sample12 | 0.799 | 0.951 | 0.467 | 0.712 | 0.952 | 0.372 |
| Sample13 | 0.861 | 0.898 | 0.233 | 0.912 | 0.913 | 0.264 |
| Sample14 | 0.864 | 0.917 | 0.462 | 0.915 | 0.933 | 0.475 |
| Sample15 | 0.849 | 0.927 | 0.535 | 0.910 | 0.936 | 0.671 |
| Sample16 | 0.660 | 0.914 | 0.411 | 0.767 | 0.933 | 0.423 |
| Sample17 | 0.519 | 0.929 | 0.075 | 0.558 | 0.944 | 0.064 |
| Sample18 | 0.758 | 0.933 | 0.314 | 0.925 | 0.920 | 0.305 |
| Sample19 | 0.753 | 0.949 | 0.357 | 0.913 | 0.965 | 0.491 |
| Sample20 | 0.860 | 0.914 | 0.161 | 0.801 | 0.913 | 0.140 |
| Sample21 | 0.813 | 0.946 | 0.418 | 0.871 | 0.962 | 0.551 |
| Sample22 | 0.797 | 0.920 | 0.217 | 0.831 | 0.926 | 0.132 |
| Sample23 | 0.714 | 0.974 | 0.672 | 0.888 | 0.987 | 0.618 |
| Sample24 | 0.841 | 0.922 | 0.141 | 0.937 | 0.934 | 0.198 |
| Sample25 | 0.746 | 0.955 | 0.607 | 0.882 | 0.964 | 0.725 |
| Sample26 | 0.741 | 0.948 | 0.295 | 0.870 | 0.948 | 0.381 |
| Sample27 | 0.746 | 0.972 | 0.594 | 0.798 | 0.974 | 0.585 |
| Sample28 | 0.859 | 0.947 | 0.282 | 0.850 | 0.953 | 0.402 |
| Sample29 | 0.909 | 0.925 | 0.159 | 0.917 | 0.932 | 0.338 |
| Sample30 | 0.892 | 0.976 | 0.555 | 0.909 | 0.983 | 0.696 |
| Sample31 | 0.780 | 0.975 | 0.518 | 0.853 | 0.982 | 0.557 |
| Sample32 | 0.838 | 0.914 | 0.157 | 0.922 | 0.928 | 0.242 |
| Sample33 | 0.754 | 0.976 | 0.667 | 0.825 | 0.985 | 0.762 |
| Sample34 | 0.873 | 0.968 | 0.607 | 0.914 | 0.957 | 0.473 |
| Sample35 | 0.830 | 0.953 | 0.452 | 0.902 | 0.958 | 0.455 |
| Sample36 | 0.896 | 0.954 | 0.322 | 0.872 | 0.952 | 0.427 |
| Sample37 | 0.798 | 0.949 | 0.165 | 0.842 | 0.969 | 0.369 |
| Sample38 | 0.773 | 0.937 | 0.250 | 0.781 | 0.956 | 0.333 |
| Sample39 | 0.836 | 0.954 | 0.421 | 0.897 | 0.953 | 0.418 |
| Sample40 | 0.806 | 0.972 | 0.340 | 0.834 | 0.975 | 0.451 |
| Sample41 | 0.886 | 0.954 | 0.435 | 0.889 | 0.960 | 0.554 |
| Sample42 | 0.749 | 0.951 | 0.521 | 0.899 | 0.960 | 0.612 |
| Sample43 | 0.831 | 0.927 | 0.371 | 0.844 | 0.935 | 0.593 |
| Sample44 | 0.868 | 0.943 | 0.548 | 0.950 | 0.957 | 0.521 |
| Sample45 | 0.786 | 0.905 | -0.059 | 0.645 | 0.895 | -0.053 |
| Sample46 | 0.799 | 0.933 | 0.181 | 0.832 | 0.939 | 0.280 |
| Sample47 | 0.920 | 0.953 | 0.488 | 0.897 | 0.963 | 0.579 |
| Sample48 | 0.846 | 0.950 | 0.312 | 0.886 | 0.969 | 0.462 |
| Sample49 | 0.709 | 0.907 | 0.506 | 0.780 | 0.923 | 0.466 |
| Sample50 | 0.759 | 0.774 | -0.004 | 0.907 | 0.880 | 0.136 |
| | | | | | | |
| Mean | 0.812 | 0.939 | 0.349 | 0.860 | 0.948 | 0.410 |
| SD | 0.076 | 0.032 | 0.180 | 0.077 | 0.024 | 0.192 |

***MAD of the item parameter estimates (base form, Var(testlet)=0)***

| | 3-PL | | | Testlet | | |
|---|---|---|---|---|---|---|
| | $\hat{a}$ | $\hat{b}$ | $\hat{c}$ | $\hat{a}$ | $\hat{b}$ | $\hat{c}$ |
| Sample1 | 0.147 | 0.345 | 0.075 | 0.139 | 0.288 | 0.059 |
| Sample2 | 0.145 | 0.353 | 0.059 | 0.132 | 0.327 | 0.049 |
| Sample3 | 0.153 | 0.261 | 0.064 | 0.137 | 0.217 | 0.051 |
| Sample4 | 0.116 | 0.280 | 0.056 | 0.101 | 0.261 | 0.050 |
| Sample5 | 0.112 | 0.317 | 0.059 | 0.107 | 0.330 | 0.050 |
| Sample6 | 0.125 | 0.361 | 0.070 | 0.115 | 0.305 | 0.051 |
| Sample7 | 0.161 | 0.403 | 0.075 | 0.149 | 0.341 | 0.060 |
| Sample8 | 0.142 | 0.316 | 0.070 | 0.133 | 0.246 | 0.053 |
| Sample9 | 0.162 | 0.314 | 0.068 | 0.152 | 0.289 | 0.054 |
| Sample10 | 0.142 | 0.345 | 0.068 | 0.146 | 0.295 | 0.052 |
| Sample11 | 0.107 | 0.293 | 0.070 | 0.100 | 0.243 | 0.055 |
| Sample12 | 0.203 | 0.284 | 0.062 | 0.198 | 0.234 | 0.047 |
| Sample13 | 0.159 | 0.245 | 0.051 | 0.139 | 0.239 | 0.045 |
| Sample14 | 0.136 | 0.271 | 0.063 | 0.120 | 0.201 | 0.053 |
| Sample15 | 0.133 | 0.267 | 0.058 | 0.121 | 0.237 | 0.045 |
| Sample16 | 0.134 | 0.225 | 0.062 | 0.132 | 0.192 | 0.049 |
| Sample17 | 0.133 | 0.222 | 0.047 | 0.116 | 0.219 | 0.037 |
| Sample18 | 0.194 | 0.286 | 0.065 | 0.184 | 0.254 | 0.053 |
| Sample19 | 0.144 | 0.309 | 0.062 | 0.126 | 0.281 | 0.047 |
| Sample20 | 0.131 | 0.265 | 0.056 | 0.127 | 0.279 | 0.049 |
| Sample21 | 0.161 | 0.259 | 0.060 | 0.159 | 0.236 | 0.046 |
| Sample22 | 0.154 | 0.313 | 0.070 | 0.136 | 0.288 | 0.056 |
| Sample23 | 0.132 | 0.258 | 0.058 | 0.135 | 0.254 | 0.048 |
| Sample24 | 0.121 | 0.285 | 0.067 | 0.116 | 0.241 | 0.053 |
| Sample25 | 0.120 | 0.215 | 0.053 | 0.106 | 0.186 | 0.037 |
| Sample26 | 0.170 | 0.284 | 0.063 | 0.164 | 0.247 | 0.046 |
| Sample27 | 0.133 | 0.255 | 0.061 | 0.127 | 0.229 | 0.049 |
| Sample28 | 0.151 | 0.273 | 0.068 | 0.147 | 0.251 | 0.055 |
| Sample29 | 0.154 | 0.205 | 0.050 | 0.141 | 0.179 | 0.038 |
| Sample30 | 0.128 | 0.250 | 0.056 | 0.125 | 0.222 | 0.041 |
| Sample31 | 0.209 | 0.328 | 0.077 | 0.181 | 0.299 | 0.059 |
| Sample32 | 0.150 | 0.326 | 0.069 | 0.146 | 0.262 | 0.053 |
| Sample33 | 0.163 | 0.306 | 0.066 | 0.147 | 0.274 | 0.054 |
| Sample34 | 0.190 | 0.232 | 0.052 | 0.188 | 0.224 | 0.045 |
| Sample35 | 0.140 | 0.339 | 0.057 | 0.127 | 0.310 | 0.047 |
| Sample36 | 0.150 | 0.242 | 0.053 | 0.142 | 0.193 | 0.041 |
| Sample37 | 0.156 | 0.366 | 0.078 | 0.155 | 0.338 | 0.067 |
| Sample38 | 0.141 | 0.309 | 0.075 | 0.117 | 0.241 | 0.054 |
| Sample39 | 0.155 | 0.258 | 0.046 | 0.152 | 0.236 | 0.040 |
| Sample40 | 0.152 | 0.286 | 0.063 | 0.133 | 0.230 | 0.048 |
| Sample41 | 0.129 | 0.277 | 0.066 | 0.110 | 0.227 | 0.046 |
| Sample42 | 0.130 | 0.433 | 0.085 | 0.120 | 0.368 | 0.070 |
| Sample43 | 0.158 | 0.356 | 0.071 | 0.137 | 0.316 | 0.054 |
| Sample44 | 0.141 | 0.242 | 0.054 | 0.136 | 0.187 | 0.043 |
| Sample45 | 0.124 | 0.372 | 0.062 | 0.119 | 0.321 | 0.045 |
| Sample46 | 0.177 | 0.297 | 0.068 | 0.176 | 0.239 | 0.052 |
| Sample47 | 0.105 | 0.267 | 0.064 | 0.109 | 0.211 | 0.048 |
| Sample48 | 0.148 | 0.340 | 0.059 | 0.139 | 0.313 | 0.050 |
| Sample49 | 0.151 | 0.305 | 0.063 | 0.139 | 0.258 | 0.048 |
| Sample50 | 0.117 | 0.312 | 0.053 | 0.115 | 0.266 | 0.038 |
| | | | | | | |
| Mean | 0.146 | 0.295 | 0.063 | 0.136 | 0.258 | 0.050 |
| SD | 0.023 | 0.049 | 0.008 | 0.022 | 0.045 | 0.007 |

***MAD of the item parameter estimates (base form, Var(testlet)=1)***

| | **3-PL** | | | **Testlet** | | |
|---|---|---|---|---|---|---|
| | $\hat{a}$ | $\hat{b}$ | $\hat{c}$ | $\hat{a}$ | $\hat{b}$ | $\hat{c}$ |
| Sample1 | 0.118 | 0.284 | 0.060 | 0.109 | 0.225 | 0.046 |
| Sample2 | 0.200 | 0.261 | 0.054 | 0.154 | 0.216 | 0.038 |
| Sample3 | 0.126 | 0.220 | 0.063 | 0.113 | 0.177 | 0.048 |
| Sample4 | 0.113 | 0.272 | 0.056 | 0.102 | 0.204 | 0.040 |
| Sample5 | 0.145 | 0.253 | 0.061 | 0.162 | 0.216 | 0.049 |
| Sample6 | 0.131 | 0.433 | 0.079 | 0.135 | 0.406 | 0.061 |
| Sample7 | 0.194 | 0.356 | 0.074 | 0.168 | 0.305 | 0.055 |
| Sample8 | 0.130 | 0.299 | 0.067 | 0.135 | 0.257 | 0.053 |
| Sample9 | 0.151 | 0.286 | 0.067 | 0.145 | 0.270 | 0.053 |
| Sample10 | 0.158 | 0.263 | 0.058 | 0.152 | 0.247 | 0.046 |
| Sample11 | 0.101 | 0.276 | 0.050 | 0.097 | 0.258 | 0.044 |
| Sample12 | 0.182 | 0.273 | 0.063 | 0.173 | 0.257 | 0.054 |
| Sample13 | 0.185 | 0.226 | 0.053 | 0.194 | 0.180 | 0.046 |
| Sample14 | 0.126 | 0.262 | 0.057 | 0.132 | 0.230 | 0.049 |
| Sample15 | 0.148 | 0.387 | 0.073 | 0.159 | 0.323 | 0.054 |
| Sample16 | 0.144 | 0.317 | 0.069 | 0.163 | 0.259 | 0.051 |
| Sample17 | 0.170 | 0.361 | 0.073 | 0.137 | 0.272 | 0.051 |
| Sample18 | 0.144 | 0.263 | 0.062 | 0.159 | 0.241 | 0.048 |
| Sample19 | 0.138 | 0.372 | 0.070 | 0.152 | 0.291 | 0.052 |
| Sample20 | 0.135 | 0.250 | 0.058 | 0.151 | 0.250 | 0.050 |
| Sample21 | 0.117 | 0.313 | 0.062 | 0.120 | 0.276 | 0.048 |
| Sample22 | 0.136 | 0.376 | 0.061 | 0.144 | 0.338 | 0.040 |
| Sample23 | 0.148 | 0.436 | 0.078 | 0.149 | 0.376 | 0.065 |
| Sample24 | 0.127 | 0.307 | 0.064 | 0.120 | 0.235 | 0.048 |
| Sample25 | 0.157 | 0.247 | 0.058 | 0.177 | 0.251 | 0.047 |
| Sample26 | 0.123 | 0.222 | 0.055 | 0.147 | 0.172 | 0.042 |
| Sample27 | 0.205 | 0.307 | 0.068 | 0.199 | 0.223 | 0.050 |
| Sample28 | 0.200 | 0.319 | 0.073 | 0.225 | 0.263 | 0.059 |
| Sample29 | 0.171 | 0.236 | 0.061 | 0.164 | 0.196 | 0.038 |
| Sample30 | 0.143 | 0.232 | 0.050 | 0.146 | 0.206 | 0.042 |
| Sample31 | 0.141 | 0.289 | 0.062 | 0.152 | 0.253 | 0.047 |
| Sample32 | 0.138 | 0.354 | 0.060 | 0.142 | 0.266 | 0.044 |
| Sample33 | 0.160 | 0.253 | 0.054 | 0.175 | 0.236 | 0.046 |
| Sample34 | 0.140 | 0.353 | 0.081 | 0.149 | 0.316 | 0.067 |
| Sample35 | 0.146 | 0.358 | 0.061 | 0.167 | 0.354 | 0.056 |
| Sample36 | 0.149 | 0.275 | 0.066 | 0.150 | 0.198 | 0.041 |
| Sample37 | 0.091 | 0.239 | 0.047 | 0.101 | 0.216 | 0.036 |
| Sample38 | 0.170 | 0.241 | 0.066 | 0.129 | 0.225 | 0.048 |
| Sample39 | 0.208 | 0.374 | 0.072 | 0.179 | 0.318 | 0.057 |
| Sample40 | 0.142 | 0.281 | 0.061 | 0.131 | 0.239 | 0.044 |
| Sample41 | 0.137 | 0.328 | 0.067 | 0.124 | 0.307 | 0.055 |
| Sample42 | 0.143 | 0.280 | 0.054 | 0.145 | 0.232 | 0.041 |
| Sample43 | 0.170 | 0.235 | 0.052 | 0.132 | 0.199 | 0.034 |
| Sample44 | 0.129 | 0.297 | 0.068 | 0.150 | 0.286 | 0.052 |
| Sample45 | 0.138 | 0.325 | 0.065 | 0.142 | 0.308 | 0.052 |
| Sample46 | 0.193 | 0.299 | 0.061 | 0.181 | 0.244 | 0.049 |
| Sample47 | 0.121 | 0.293 | 0.059 | 0.121 | 0.241 | 0.047 |
| Sample48 | 0.151 | 0.287 | 0.059 | 0.134 | 0.274 | 0.049 |
| Sample49 | 0.117 | 0.265 | 0.052 | 0.134 | 0.243 | 0.041 |
| Sample50 | 0.168 | 0.263 | 0.062 | 0.168 | 0.204 | 0.044 |
| | | | | | | |
| Mean | 0.148 | 0.296 | 0.063 | 0.148 | 0.256 | 0.048 |
| SD | 0.027 | 0.053 | 0.008 | 0.025 | 0.050 | 0.007 |

### *MAD of the item parameter estimates (base form, Var(testlet)=2)*

| | 3-PL | | | Testlet | | |
|---|---|---|---|---|---|---|
| | $\hat{a}$ | $\hat{b}$ | $\hat{c}$ | $\hat{a}$ | $\hat{b}$ | $\hat{c}$ |
| Sample1 | 0.135 | 0.310 | 0.051 | 0.176 | 0.243 | 0.035 |
| Sample2 | 0.138 | 0.254 | 0.067 | 0.173 | 0.229 | 0.054 |
| Sample3 | 0.176 | 0.295 | 0.055 | 0.196 | 0.209 | 0.040 |
| Sample4 | 0.225 | 0.327 | 0.055 | 0.183 | 0.266 | 0.042 |
| Sample5 | 0.185 | 0.412 | 0.077 | 0.173 | 0.336 | 0.057 |
| Sample6 | 0.158 | 0.397 | 0.056 | 0.180 | 0.374 | 0.052 |
| Sample7 | 0.134 | 0.293 | 0.063 | 0.151 | 0.239 | 0.045 |
| Sample8 | 0.125 | 0.269 | 0.057 | 0.170 | 0.252 | 0.052 |
| Sample9 | 0.134 | 0.316 | 0.074 | 0.138 | 0.314 | 0.058 |
| Sample10 | 0.134 | 0.244 | 0.050 | 0.127 | 0.178 | 0.033 |
| Sample11 | 0.161 | 0.302 | 0.051 | 0.188 | 0.208 | 0.041 |
| Sample12 | 0.105 | 0.283 | 0.062 | 0.146 | 0.251 | 0.055 |
| Sample13 | 0.121 | 0.308 | 0.057 | 0.122 | 0.328 | 0.043 |
| Sample14 | 0.164 | 0.295 | 0.063 | 0.199 | 0.231 | 0.047 |
| Sample15 | 0.158 | 0.314 | 0.070 | 0.178 | 0.226 | 0.049 |
| Sample16 | 0.220 | 0.285 | 0.058 | 0.204 | 0.214 | 0.043 |
| Sample17 | 0.155 | 0.447 | 0.073 | 0.189 | 0.380 | 0.060 |
| Sample18 | 0.141 | 0.367 | 0.067 | 0.129 | 0.314 | 0.050 |
| Sample19 | 0.240 | 0.277 | 0.053 | 0.175 | 0.240 | 0.038 |
| Sample20 | 0.230 | 0.352 | 0.079 | 0.182 | 0.255 | 0.054 |
| Sample21 | 0.146 | 0.335 | 0.068 | 0.199 | 0.290 | 0.052 |
| Sample22 | 0.121 | 0.319 | 0.065 | 0.176 | 0.275 | 0.048 |
| Sample23 | 0.154 | 0.307 | 0.058 | 0.140 | 0.185 | 0.042 |
| Sample24 | 0.135 | 0.310 | 0.060 | 0.125 | 0.267 | 0.047 |
| Sample25 | 0.176 | 0.454 | 0.080 | 0.201 | 0.379 | 0.060 |
| Sample26 | 0.119 | 0.390 | 0.077 | 0.142 | 0.303 | 0.059 |
| Sample27 | 0.254 | 0.338 | 0.072 | 0.211 | 0.295 | 0.045 |
| Sample28 | 0.143 | 0.362 | 0.070 | 0.145 | 0.288 | 0.055 |
| Sample29 | 0.149 | 0.323 | 0.072 | 0.162 | 0.261 | 0.048 |
| Sample30 | 0.142 | 0.336 | 0.056 | 0.150 | 0.299 | 0.046 |
| Sample31 | 0.122 | 0.287 | 0.053 | 0.185 | 0.218 | 0.040 |
| Sample32 | 0.192 | 0.416 | 0.080 | 0.128 | 0.315 | 0.055 |
| Sample33 | 0.143 | 0.398 | 0.074 | 0.134 | 0.279 | 0.047 |
| Sample34 | 0.159 | 0.351 | 0.080 | 0.163 | 0.264 | 0.057 |
| Sample35 | 0.163 | 0.316 | 0.072 | 0.143 | 0.270 | 0.057 |
| Sample36 | 0.148 | 0.359 | 0.073 | 0.184 | 0.285 | 0.055 |
| Sample37 | 0.150 | 0.306 | 0.064 | 0.134 | 0.271 | 0.046 |
| Sample38 | 0.191 | 0.314 | 0.045 | 0.152 | 0.289 | 0.035 |
| Sample39 | 0.112 | 0.314 | 0.061 | 0.155 | 0.244 | 0.054 |
| Sample40 | 0.150 | 0.359 | 0.079 | 0.168 | 0.269 | 0.059 |
| Sample41 | 0.128 | 0.341 | 0.070 | 0.151 | 0.310 | 0.060 |
| Sample42 | 0.202 | 0.252 | 0.067 | 0.226 | 0.233 | 0.049 |
| Sample43 | 0.227 | 0.307 | 0.065 | 0.176 | 0.271 | 0.040 |
| Sample44 | 0.164 | 0.309 | 0.065 | 0.166 | 0.281 | 0.047 |
| Sample45 | 0.128 | 0.279 | 0.061 | 0.142 | 0.309 | 0.049 |
| Sample46 | 0.163 | 0.411 | 0.071 | 0.134 | 0.344 | 0.054 |
| Sample47 | 0.139 | 0.286 | 0.077 | 0.153 | 0.232 | 0.057 |
| Sample48 | 0.123 | 0.384 | 0.067 | 0.123 | 0.342 | 0.046 |
| Sample49 | 0.186 | 0.257 | 0.070 | 0.181 | 0.230 | 0.056 |
| Sample50 | 0.140 | 0.322 | 0.074 | 0.135 | 0.288 | 0.056 |
| | | | | | | |
| Mean | 0.158 | 0.328 | 0.066 | 0.163 | 0.273 | 0.049 |
| SD | 0.035 | 0.050 | 0.009 | 0.026 | 0.047 | 0.007 |

***MAD of the rescaled item parameter estimates (new form, Var(testlet)=0)***

| | **3-PL** | | | **Testlet** | | |
|---|---|---|---|---|---|---|
| | $\hat{a}$ | $\hat{b}$ | $\hat{c}$ | $\hat{a}$ | $\hat{b}$ | $\hat{c}$ |
| Sample1 | 0.101 | 0.251 | 0.057 | 0.108 | 0.207 | 0.046 |
| Sample2 | 0.104 | 0.270 | 0.053 | 0.106 | 0.252 | 0.044 |
| Sample3 | 0.123 | 0.301 | 0.057 | 0.114 | 0.231 | 0.046 |
| Sample4 | 0.093 | 0.199 | 0.038 | 0.094 | 0.198 | 0.035 |
| Sample5 | 0.101 | 0.221 | 0.055 | 0.103 | 0.205 | 0.046 |
| Sample6 | 0.109 | 0.255 | 0.055 | 0.104 | 0.256 | 0.043 |
| Sample7 | 0.118 | 0.359 | 0.050 | 0.092 | 0.316 | 0.037 |
| Sample8 | 0.128 | 0.271 | 0.053 | 0.105 | 0.248 | 0.041 |
| Sample9 | 0.134 | 0.214 | 0.053 | 0.104 | 0.227 | 0.044 |
| Sample10 | 0.135 | 0.259 | 0.050 | 0.134 | 0.226 | 0.041 |
| Sample11 | 0.144 | 0.268 | 0.076 | 0.114 | 0.242 | 0.055 |
| Sample12 | 0.209 | 0.292 | 0.054 | 0.178 | 0.250 | 0.044 |
| Sample13 | 0.125 | 0.265 | 0.056 | 0.111 | 0.218 | 0.046 |
| Sample14 | 0.081 | 0.223 | 0.056 | 0.068 | 0.192 | 0.051 |
| Sample15 | 0.198 | 0.341 | 0.059 | 0.161 | 0.288 | 0.047 |
| Sample16 | 0.083 | 0.221 | 0.053 | 0.090 | 0.225 | 0.046 |
| Sample17 | 0.119 | 0.232 | 0.038 | 0.105 | 0.212 | 0.035 |
| Sample18 | 0.084 | 0.327 | 0.047 | 0.084 | 0.284 | 0.039 |
| Sample19 | 0.159 | 0.376 | 0.074 | 0.146 | 0.358 | 0.058 |
| Sample20 | 0.116 | 0.326 | 0.057 | 0.111 | 0.313 | 0.049 |
| Sample21 | 0.116 | 0.234 | 0.051 | 0.118 | 0.191 | 0.039 |
| Sample22 | 0.095 | 0.274 | 0.053 | 0.091 | 0.244 | 0.041 |
| Sample23 | 0.133 | 0.201 | 0.033 | 0.137 | 0.205 | 0.036 |
| Sample24 | 0.111 | 0.268 | 0.056 | 0.110 | 0.278 | 0.050 |
| Sample25 | 0.096 | 0.224 | 0.043 | 0.103 | 0.203 | 0.036 |
| Sample26 | 0.080 | 0.251 | 0.055 | 0.082 | 0.210 | 0.044 |
| Sample27 | 0.074 | 0.215 | 0.050 | 0.066 | 0.185 | 0.036 |
| Sample28 | 0.113 | 0.260 | 0.045 | 0.091 | 0.231 | 0.032 |
| Sample29 | 0.157 | 0.224 | 0.040 | 0.122 | 0.195 | 0.031 |
| Sample30 | 0.103 | 0.211 | 0.046 | 0.081 | 0.217 | 0.037 |
| Sample31 | 0.089 | 0.270 | 0.053 | 0.091 | 0.225 | 0.046 |
| Sample32 | 0.108 | 0.294 | 0.053 | 0.101 | 0.268 | 0.041 |
| Sample33 | 0.091 | 0.265 | 0.043 | 0.088 | 0.229 | 0.037 |
| Sample34 | 0.134 | 0.257 | 0.053 | 0.105 | 0.225 | 0.041 |
| Sample35 | 0.172 | 0.293 | 0.054 | 0.131 | 0.239 | 0.048 |
| Sample36 | 0.105 | 0.172 | 0.041 | 0.085 | 0.201 | 0.041 |
| Sample37 | 0.175 | 0.309 | 0.062 | 0.143 | 0.258 | 0.050 |
| Sample38 | 0.125 | 0.301 | 0.055 | 0.100 | 0.259 | 0.045 |
| Sample39 | 0.101 | 0.253 | 0.052 | 0.101 | 0.197 | 0.039 |
| Sample40 | 0.102 | 0.288 | 0.050 | 0.098 | 0.237 | 0.043 |
| Sample41 | 0.168 | 0.215 | 0.057 | 0.129 | 0.165 | 0.045 |
| Sample42 | 0.074 | 0.299 | 0.057 | 0.076 | 0.257 | 0.048 |
| Sample43 | 0.076 | 0.256 | 0.051 | 0.093 | 0.233 | 0.047 |
| Sample44 | 0.115 | 0.229 | 0.049 | 0.102 | 0.194 | 0.043 |
| Sample45 | 0.107 | 0.314 | 0.052 | 0.109 | 0.234 | 0.042 |
| Sample46 | 0.103 | 0.261 | 0.044 | 0.089 | 0.288 | 0.037 |
| Sample47 | 0.101 | 0.255 | 0.047 | 0.112 | 0.223 | 0.042 |
| Sample48 | 0.111 | 0.251 | 0.053 | 0.088 | 0.238 | 0.040 |
| Sample49 | 0.111 | 0.242 | 0.044 | 0.124 | 0.219 | 0.034 |
| Sample50 | 0.099 | 0.339 | 0.052 | 0.086 | 0.277 | 0.040 |
| | | | | | | |
| Mean | 0.116 | 0.264 | 0.052 | 0.106 | 0.236 | 0.042 |
| SD | 0.030 | 0.043 | 0.008 | 0.022 | 0.037 | 0.006 |

***MAD of the rescaled item parameter estimates (new form, Var(testlet)=1)***

| | 3-PL | | | Testlet | | |
|---|---|---|---|---|---|---|
| | $\hat{a}$ | $\hat{b}$ | $\hat{c}$ | $\hat{a}$ | $\hat{b}$ | $\hat{c}$ |
| Sample1 | 0.091 | 0.259 | 0.055 | 0.117 | 0.221 | 0.042 |
| Sample2 | 0.114 | 0.284 | 0.052 | 0.112 | 0.251 | 0.039 |
| Sample3 | 0.110 | 0.238 | 0.059 | 0.118 | 0.185 | 0.047 |
| Sample4 | 0.114 | 0.326 | 0.057 | 0.119 | 0.265 | 0.036 |
| Sample5 | 0.117 | 0.220 | 0.046 | 0.124 | 0.232 | 0.037 |
| Sample6 | 0.129 | 0.274 | 0.056 | 0.111 | 0.228 | 0.045 |
| Sample7 | 0.096 | 0.291 | 0.053 | 0.108 | 0.240 | 0.041 |
| Sample8 | 0.124 | 0.249 | 0.048 | 0.122 | 0.217 | 0.040 |
| Sample9 | 0.115 | 0.273 | 0.056 | 0.110 | 0.254 | 0.043 |
| Sample10 | 0.069 | 0.255 | 0.050 | 0.073 | 0.229 | 0.044 |
| Sample11 | 0.109 | 0.315 | 0.063 | 0.120 | 0.263 | 0.039 |
| Sample12 | 0.105 | 0.283 | 0.052 | 0.087 | 0.288 | 0.041 |
| Sample13 | 0.124 | 0.267 | 0.050 | 0.157 | 0.211 | 0.039 |
| Sample14 | 0.091 | 0.301 | 0.054 | 0.092 | 0.268 | 0.042 |
| Sample15 | 0.104 | 0.325 | 0.062 | 0.122 | 0.266 | 0.052 |
| Sample16 | 0.107 | 0.319 | 0.064 | 0.108 | 0.243 | 0.041 |
| Sample17 | 0.122 | 0.272 | 0.044 | 0.095 | 0.245 | 0.037 |
| Sample18 | 0.152 | 0.227 | 0.037 | 0.138 | 0.208 | 0.028 |
| Sample19 | 0.109 | 0.356 | 0.052 | 0.098 | 0.249 | 0.039 |
| Sample20 | 0.085 | 0.288 | 0.054 | 0.120 | 0.252 | 0.043 |
| Sample21 | 0.111 | 0.292 | 0.058 | 0.109 | 0.270 | 0.044 |
| Sample22 | 0.143 | 0.287 | 0.033 | 0.118 | 0.208 | 0.031 |
| Sample23 | 0.097 | 0.381 | 0.060 | 0.105 | 0.349 | 0.048 |
| Sample24 | 0.101 | 0.346 | 0.057 | 0.083 | 0.267 | 0.045 |
| Sample25 | 0.119 | 0.259 | 0.048 | 0.124 | 0.224 | 0.039 |
| Sample26 | 0.114 | 0.351 | 0.062 | 0.110 | 0.291 | 0.044 |
| Sample27 | 0.115 | 0.318 | 0.053 | 0.093 | 0.243 | 0.041 |
| Sample28 | 0.108 | 0.226 | 0.038 | 0.112 | 0.195 | 0.033 |
| Sample29 | 0.096 | 0.320 | 0.056 | 0.133 | 0.263 | 0.038 |
| Sample30 | 0.136 | 0.275 | 0.058 | 0.125 | 0.246 | 0.046 |
| Sample31 | 0.128 | 0.291 | 0.052 | 0.120 | 0.226 | 0.042 |
| Sample32 | 0.155 | 0.391 | 0.058 | 0.118 | 0.292 | 0.035 |
| Sample33 | 0.129 | 0.286 | 0.055 | 0.119 | 0.282 | 0.045 |
| Sample34 | 0.100 | 0.391 | 0.077 | 0.106 | 0.292 | 0.057 |
| Sample35 | 0.101 | 0.171 | 0.046 | 0.096 | 0.182 | 0.040 |
| Sample36 | 0.128 | 0.237 | 0.052 | 0.168 | 0.219 | 0.046 |
| Sample37 | 0.086 | 0.181 | 0.047 | 0.085 | 0.166 | 0.043 |
| Sample38 | 0.087 | 0.223 | 0.045 | 0.085 | 0.195 | 0.043 |
| Sample39 | 0.132 | 0.405 | 0.051 | 0.103 | 0.327 | 0.036 |
| Sample40 | 0.099 | 0.268 | 0.051 | 0.083 | 0.228 | 0.040 |
| Sample41 | 0.123 | 0.299 | 0.061 | 0.122 | 0.253 | 0.047 |
| Sample42 | 0.147 | 0.271 | 0.053 | 0.104 | 0.223 | 0.036 |
| Sample43 | 0.109 | 0.301 | 0.052 | 0.080 | 0.239 | 0.038 |
| Sample44 | 0.107 | 0.227 | 0.049 | 0.086 | 0.231 | 0.039 |
| Sample45 | 0.119 | 0.391 | 0.058 | 0.129 | 0.289 | 0.040 |
| Sample46 | 0.177 | 0.274 | 0.056 | 0.138 | 0.192 | 0.042 |
| Sample47 | 0.078 | 0.328 | 0.054 | 0.082 | 0.291 | 0.040 |
| Sample48 | 0.139 | 0.316 | 0.053 | 0.112 | 0.263 | 0.044 |
| Sample49 | 0.100 | 0.212 | 0.053 | 0.109 | 0.169 | 0.045 |
| Sample50 | 0.167 | 0.293 | 0.054 | 0.183 | 0.199 | 0.039 |
| | | | | | | |
| Mean | 0.115 | 0.289 | 0.053 | 0.112 | 0.243 | 0.041 |
| SD | 0.022 | 0.053 | 0.007 | 0.022 | 0.039 | 0.005 |

### *MAD of the rescaled item parameter estimates (new form, Var(testlet)=2)*

| | **3-PL** | | | **Testlet** | | |
|---|---|---|---|---|---|---|
| | $\hat{a}$ | $\hat{b}$ | $\hat{c}$ | $\hat{a}$ | $\hat{b}$ | $\hat{c}$ |
| Sample1 | 0.088 | 0.274 | 0.054 | 0.122 | 0.203 | 0.036 |
| Sample2 | 0.176 | 0.252 | 0.054 | 0.163 | 0.260 | 0.044 |
| Sample3 | 0.102 | 0.294 | 0.055 | 0.163 | 0.279 | 0.049 |
| Sample4 | 0.105 | 0.289 | 0.054 | 0.099 | 0.228 | 0.036 |
| Sample5 | 0.187 | 0.399 | 0.063 | 0.155 | 0.237 | 0.043 |
| Sample6 | 0.135 | 0.383 | 0.049 | 0.164 | 0.319 | 0.042 |
| Sample7 | 0.135 | 0.277 | 0.069 | 0.133 | 0.215 | 0.039 |
| Sample8 | 0.131 | 0.293 | 0.059 | 0.122 | 0.292 | 0.054 |
| Sample9 | 0.086 | 0.312 | 0.054 | 0.086 | 0.294 | 0.039 |
| Sample10 | 0.108 | 0.244 | 0.048 | 0.133 | 0.243 | 0.035 |
| Sample11 | 0.124 | 0.373 | 0.059 | 0.163 | 0.283 | 0.041 |
| Sample12 | 0.115 | 0.257 | 0.051 | 0.157 | 0.230 | 0.044 |
| Sample13 | 0.101 | 0.424 | 0.052 | 0.082 | 0.351 | 0.041 |
| Sample14 | 0.156 | 0.264 | 0.055 | 0.161 | 0.230 | 0.048 |
| Sample15 | 0.105 | 0.291 | 0.049 | 0.091 | 0.259 | 0.034 |
| Sample16 | 0.150 | 0.335 | 0.051 | 0.189 | 0.269 | 0.044 |
| Sample17 | 0.141 | 0.382 | 0.060 | 0.142 | 0.308 | 0.052 |
| Sample18 | 0.164 | 0.279 | 0.050 | 0.106 | 0.260 | 0.045 |
| Sample19 | 0.149 | 0.252 | 0.053 | 0.128 | 0.186 | 0.035 |
| Sample20 | 0.135 | 0.278 | 0.056 | 0.165 | 0.248 | 0.049 |
| Sample21 | 0.117 | 0.313 | 0.049 | 0.185 | 0.224 | 0.038 |
| Sample22 | 0.127 | 0.251 | 0.057 | 0.130 | 0.217 | 0.045 |
| Sample23 | 0.166 | 0.299 | 0.050 | 0.094 | 0.159 | 0.033 |
| Sample24 | 0.136 | 0.273 | 0.060 | 0.111 | 0.257 | 0.045 |
| Sample25 | 0.126 | 0.373 | 0.068 | 0.105 | 0.278 | 0.045 |
| Sample26 | 0.138 | 0.329 | 0.052 | 0.098 | 0.282 | 0.039 |
| Sample27 | 0.151 | 0.323 | 0.054 | 0.148 | 0.296 | 0.038 |
| Sample28 | 0.129 | 0.296 | 0.056 | 0.187 | 0.259 | 0.040 |
| Sample29 | 0.097 | 0.358 | 0.067 | 0.089 | 0.304 | 0.045 |
| Sample30 | 0.096 | 0.305 | 0.045 | 0.084 | 0.202 | 0.032 |
| Sample31 | 0.126 | 0.263 | 0.049 | 0.114 | 0.182 | 0.039 |
| Sample32 | 0.111 | 0.299 | 0.062 | 0.087 | 0.221 | 0.041 |
| Sample33 | 0.157 | 0.375 | 0.076 | 0.126 | 0.238 | 0.042 |
| Sample34 | 0.126 | 0.282 | 0.054 | 0.123 | 0.238 | 0.046 |
| Sample35 | 0.164 | 0.371 | 0.061 | 0.138 | 0.267 | 0.042 |
| Sample36 | 0.125 | 0.352 | 0.062 | 0.174 | 0.303 | 0.049 |
| Sample37 | 0.101 | 0.253 | 0.051 | 0.104 | 0.231 | 0.036 |
| Sample38 | 0.150 | 0.366 | 0.045 | 0.140 | 0.271 | 0.036 |
| Sample39 | 0.129 | 0.285 | 0.047 | 0.120 | 0.263 | 0.040 |
| Sample40 | 0.118 | 0.285 | 0.060 | 0.117 | 0.221 | 0.039 |
| Sample41 | 0.130 | 0.275 | 0.059 | 0.111 | 0.256 | 0.040 |
| Sample42 | 0.226 | 0.291 | 0.055 | 0.196 | 0.215 | 0.032 |
| Sample43 | 0.117 | 0.373 | 0.060 | 0.127 | 0.324 | 0.042 |
| Sample44 | 0.104 | 0.347 | 0.065 | 0.103 | 0.256 | 0.038 |
| Sample45 | 0.158 | 0.327 | 0.070 | 0.149 | 0.329 | 0.063 |
| Sample46 | 0.115 | 0.360 | 0.062 | 0.117 | 0.284 | 0.046 |
| Sample47 | 0.134 | 0.309 | 0.060 | 0.127 | 0.251 | 0.039 |
| Sample48 | 0.128 | 0.282 | 0.049 | 0.090 | 0.198 | 0.036 |
| Sample49 | 0.160 | 0.287 | 0.047 | 0.123 | 0.268 | 0.039 |
| Sample50 | 0.137 | 0.332 | 0.063 | 0.101 | 0.280 | 0.050 |
| | | | | | | |
| Mean | 0.132 | 0.312 | 0.056 | 0.129 | 0.255 | 0.042 |
| SD | 0.027 | 0.046 | 0.007 | 0.031 | 0.040 | 0.006 |

### *RMSD of the item parameter estimates (base form, Var(testlet)=0)*

| | **3-PL** | | | **Testlet** | | |
|---|---|---|---|---|---|---|
| | $\hat{a}$ | $\hat{b}$ | $\hat{c}$ | $\hat{a}$ | $\hat{b}$ | $\hat{c}$ |
| Sample1 | 0.170 | 0.459 | 0.093 | 0.167 | 0.394 | 0.074 |
| Sample2 | 0.173 | 0.430 | 0.077 | 0.158 | 0.387 | 0.062 |
| Sample3 | 0.192 | 0.355 | 0.085 | 0.166 | 0.288 | 0.068 |
| Sample4 | 0.155 | 0.393 | 0.068 | 0.128 | 0.353 | 0.058 |
| Sample5 | 0.132 | 0.366 | 0.072 | 0.125 | 0.375 | 0.061 |
| Sample6 | 0.159 | 0.494 | 0.085 | 0.147 | 0.444 | 0.066 |
| Sample7 | 0.206 | 0.521 | 0.093 | 0.189 | 0.468 | 0.074 |
| Sample8 | 0.193 | 0.408 | 0.088 | 0.182 | 0.324 | 0.069 |
| Sample9 | 0.203 | 0.421 | 0.085 | 0.195 | 0.385 | 0.068 |
| Sample10 | 0.194 | 0.416 | 0.082 | 0.229 | 0.391 | 0.064 |
| Sample11 | 0.135 | 0.374 | 0.081 | 0.142 | 0.320 | 0.066 |
| Sample12 | 0.233 | 0.348 | 0.073 | 0.238 | 0.287 | 0.056 |
| Sample13 | 0.205 | 0.346 | 0.071 | 0.181 | 0.328 | 0.060 |
| Sample14 | 0.179 | 0.319 | 0.077 | 0.154 | 0.261 | 0.063 |
| Sample15 | 0.178 | 0.325 | 0.071 | 0.163 | 0.300 | 0.058 |
| Sample16 | 0.181 | 0.283 | 0.078 | 0.195 | 0.242 | 0.062 |
| Sample17 | 0.189 | 0.306 | 0.061 | 0.155 | 0.289 | 0.047 |
| Sample18 | 0.246 | 0.348 | 0.078 | 0.231 | 0.313 | 0.064 |
| Sample19 | 0.174 | 0.368 | 0.077 | 0.153 | 0.328 | 0.063 |
| Sample20 | 0.175 | 0.351 | 0.076 | 0.165 | 0.353 | 0.067 |
| Sample21 | 0.209 | 0.363 | 0.074 | 0.225 | 0.318 | 0.057 |
| Sample22 | 0.233 | 0.383 | 0.084 | 0.196 | 0.358 | 0.067 |
| Sample23 | 0.198 | 0.331 | 0.071 | 0.199 | 0.327 | 0.058 |
| Sample24 | 0.153 | 0.356 | 0.082 | 0.161 | 0.326 | 0.067 |
| Sample25 | 0.149 | 0.296 | 0.069 | 0.129 | 0.248 | 0.050 |
| Sample26 | 0.205 | 0.369 | 0.077 | 0.200 | 0.319 | 0.058 |
| Sample27 | 0.166 | 0.336 | 0.072 | 0.156 | 0.315 | 0.060 |
| Sample28 | 0.213 | 0.334 | 0.083 | 0.201 | 0.326 | 0.069 |
| Sample29 | 0.184 | 0.246 | 0.059 | 0.171 | 0.223 | 0.047 |
| Sample30 | 0.167 | 0.298 | 0.064 | 0.171 | 0.263 | 0.047 |
| Sample31 | 0.279 | 0.406 | 0.096 | 0.231 | 0.384 | 0.074 |
| Sample32 | 0.189 | 0.377 | 0.079 | 0.183 | 0.313 | 0.060 |
| Sample33 | 0.205 | 0.387 | 0.084 | 0.178 | 0.360 | 0.067 |
| Sample34 | 0.243 | 0.307 | 0.063 | 0.247 | 0.303 | 0.054 |
| Sample35 | 0.182 | 0.450 | 0.073 | 0.167 | 0.400 | 0.059 |
| Sample36 | 0.185 | 0.301 | 0.067 | 0.179 | 0.239 | 0.052 |
| Sample37 | 0.200 | 0.467 | 0.092 | 0.200 | 0.440 | 0.080 |
| Sample38 | 0.182 | 0.440 | 0.096 | 0.154 | 0.353 | 0.075 |
| Sample39 | 0.233 | 0.363 | 0.068 | 0.238 | 0.343 | 0.056 |
| Sample40 | 0.177 | 0.393 | 0.072 | 0.161 | 0.349 | 0.056 |
| Sample41 | 0.185 | 0.400 | 0.085 | 0.159 | 0.333 | 0.063 |
| Sample42 | 0.171 | 0.536 | 0.102 | 0.154 | 0.465 | 0.085 |
| Sample43 | 0.231 | 0.440 | 0.086 | 0.196 | 0.375 | 0.065 |
| Sample44 | 0.173 | 0.307 | 0.068 | 0.175 | 0.252 | 0.054 |
| Sample45 | 0.139 | 0.493 | 0.073 | 0.154 | 0.429 | 0.057 |
| Sample46 | 0.223 | 0.375 | 0.083 | 0.227 | 0.319 | 0.063 |
| Sample47 | 0.157 | 0.333 | 0.076 | 0.179 | 0.275 | 0.058 |
| Sample48 | 0.179 | 0.409 | 0.073 | 0.162 | 0.388 | 0.062 |
| Sample49 | 0.184 | 0.366 | 0.073 | 0.170 | 0.312 | 0.055 |
| Sample50 | 0.145 | 0.400 | 0.064 | 0.148 | 0.335 | 0.048 |
| | | | | | | |
| Mean | 0.188 | 0.378 | 0.078 | 0.179 | 0.336 | 0.062 |
| SD | 0.030 | 0.062 | 0.010 | 0.030 | 0.058 | 0.008 |

### *RMSD of the item parameter estimates (base form, Var(testlet)=1)*

| | **3-PL** | | | **Testlet** | | |
|---|---|---|---|---|---|---|
| | $\hat{a}$ | $\hat{b}$ | $\hat{c}$ | $\hat{a}$ | $\hat{b}$ | $\hat{c}$ |
| Sample1 | 0.148 | 0.330 | 0.068 | 0.137 | 0.277 | 0.055 |
| Sample2 | 0.247 | 0.351 | 0.068 | 0.200 | 0.282 | 0.051 |
| Sample3 | 0.173 | 0.313 | 0.080 | 0.153 | 0.253 | 0.060 |
| Sample4 | 0.179 | 0.365 | 0.072 | 0.148 | 0.280 | 0.051 |
| Sample5 | 0.213 | 0.307 | 0.075 | 0.215 | 0.248 | 0.058 |
| Sample6 | 0.175 | 0.552 | 0.092 | 0.175 | 0.513 | 0.074 |
| Sample7 | 0.243 | 0.486 | 0.089 | 0.218 | 0.429 | 0.069 |
| Sample8 | 0.175 | 0.375 | 0.080 | 0.174 | 0.329 | 0.066 |
| Sample9 | 0.177 | 0.387 | 0.080 | 0.178 | 0.376 | 0.065 |
| Sample10 | 0.245 | 0.346 | 0.073 | 0.231 | 0.301 | 0.058 |
| Sample11 | 0.135 | 0.374 | 0.061 | 0.126 | 0.362 | 0.055 |
| Sample12 | 0.234 | 0.340 | 0.075 | 0.234 | 0.316 | 0.064 |
| Sample13 | 0.229 | 0.300 | 0.068 | 0.258 | 0.242 | 0.055 |
| Sample14 | 0.166 | 0.391 | 0.077 | 0.161 | 0.324 | 0.062 |
| Sample15 | 0.181 | 0.446 | 0.085 | 0.210 | 0.411 | 0.068 |
| Sample16 | 0.191 | 0.389 | 0.079 | 0.209 | 0.345 | 0.063 |
| Sample17 | 0.274 | 0.482 | 0.097 | 0.193 | 0.380 | 0.068 |
| Sample18 | 0.207 | 0.323 | 0.075 | 0.182 | 0.289 | 0.057 |
| Sample19 | 0.172 | 0.454 | 0.086 | 0.198 | 0.360 | 0.063 |
| Sample20 | 0.193 | 0.335 | 0.071 | 0.206 | 0.329 | 0.063 |
| Sample21 | 0.148 | 0.366 | 0.071 | 0.142 | 0.337 | 0.059 |
| Sample22 | 0.163 | 0.465 | 0.073 | 0.193 | 0.451 | 0.053 |
| Sample23 | 0.182 | 0.556 | 0.095 | 0.181 | 0.465 | 0.076 |
| Sample24 | 0.165 | 0.372 | 0.076 | 0.158 | 0.291 | 0.058 |
| Sample25 | 0.188 | 0.293 | 0.068 | 0.231 | 0.298 | 0.056 |
| Sample26 | 0.174 | 0.280 | 0.068 | 0.209 | 0.223 | 0.051 |
| Sample27 | 0.275 | 0.382 | 0.090 | 0.278 | 0.303 | 0.063 |
| Sample28 | 0.265 | 0.411 | 0.089 | 0.301 | 0.345 | 0.071 |
| Sample29 | 0.222 | 0.313 | 0.074 | 0.213 | 0.253 | 0.049 |
| Sample30 | 0.192 | 0.343 | 0.069 | 0.201 | 0.304 | 0.057 |
| Sample31 | 0.171 | 0.344 | 0.074 | 0.192 | 0.334 | 0.058 |
| Sample32 | 0.178 | 0.416 | 0.072 | 0.183 | 0.316 | 0.052 |
| Sample33 | 0.234 | 0.330 | 0.066 | 0.226 | 0.320 | 0.057 |
| Sample34 | 0.207 | 0.409 | 0.100 | 0.212 | 0.365 | 0.081 |
| Sample35 | 0.176 | 0.451 | 0.078 | 0.192 | 0.433 | 0.068 |
| Sample36 | 0.205 | 0.358 | 0.078 | 0.201 | 0.275 | 0.051 |
| Sample37 | 0.113 | 0.316 | 0.060 | 0.127 | 0.299 | 0.049 |
| Sample38 | 0.220 | 0.320 | 0.086 | 0.157 | 0.287 | 0.062 |
| Sample39 | 0.314 | 0.478 | 0.097 | 0.248 | 0.387 | 0.070 |
| Sample40 | 0.183 | 0.404 | 0.074 | 0.158 | 0.358 | 0.057 |
| Sample41 | 0.159 | 0.427 | 0.077 | 0.163 | 0.385 | 0.064 |
| Sample42 | 0.183 | 0.377 | 0.072 | 0.182 | 0.328 | 0.057 |
| Sample43 | 0.208 | 0.301 | 0.064 | 0.166 | 0.266 | 0.045 |
| Sample44 | 0.162 | 0.395 | 0.082 | 0.205 | 0.374 | 0.067 |
| Sample45 | 0.171 | 0.456 | 0.081 | 0.192 | 0.412 | 0.063 |
| Sample46 | 0.242 | 0.366 | 0.075 | 0.248 | 0.316 | 0.062 |
| Sample47 | 0.177 | 0.332 | 0.067 | 0.166 | 0.287 | 0.052 |
| Sample48 | 0.187 | 0.374 | 0.078 | 0.168 | 0.348 | 0.061 |
| Sample49 | 0.146 | 0.340 | 0.065 | 0.176 | 0.306 | 0.052 |
| Sample50 | 0.223 | 0.345 | 0.075 | 0.232 | 0.278 | 0.055 |
| | | | | | | |
| Mean | 0.196 | 0.379 | 0.077 | 0.194 | 0.332 | 0.060 |
| SD | 0.040 | 0.064 | 0.010 | 0.037 | 0.061 | 0.008 |

***RMSD of the item parameter estimates (base form, Var(testlet)=2)***

| | **3-PL** | | | **Testlet** | | |
|---|---|---|---|---|---|---|
| | $\hat{a}$ | $\hat{b}$ | $\hat{c}$ | $\hat{a}$ | $\hat{b}$ | $\hat{c}$ |
| Sample1 | 0.167 | 0.358 | 0.060 | 0.203 | 0.304 | 0.044 |
| Sample2 | 0.177 | 0.338 | 0.079 | 0.249 | 0.275 | 0.068 |
| Sample3 | 0.232 | 0.377 | 0.076 | 0.246 | 0.273 | 0.054 |
| Sample4 | 0.285 | 0.415 | 0.072 | 0.227 | 0.358 | 0.055 |
| Sample5 | 0.239 | 0.531 | 0.092 | 0.236 | 0.406 | 0.069 |
| Sample6 | 0.201 | 0.517 | 0.075 | 0.258 | 0.506 | 0.070 |
| Sample7 | 0.168 | 0.385 | 0.073 | 0.204 | 0.309 | 0.055 |
| Sample8 | 0.165 | 0.351 | 0.075 | 0.225 | 0.325 | 0.068 |
| Sample9 | 0.192 | 0.367 | 0.084 | 0.162 | 0.368 | 0.070 |
| Sample10 | 0.177 | 0.316 | 0.060 | 0.167 | 0.221 | 0.040 |
| Sample11 | 0.194 | 0.372 | 0.066 | 0.224 | 0.262 | 0.050 |
| Sample12 | 0.136 | 0.389 | 0.082 | 0.182 | 0.333 | 0.071 |
| Sample13 | 0.141 | 0.421 | 0.068 | 0.159 | 0.488 | 0.061 |
| Sample14 | 0.221 | 0.371 | 0.083 | 0.270 | 0.290 | 0.062 |
| Sample15 | 0.194 | 0.399 | 0.087 | 0.218 | 0.320 | 0.066 |
| Sample16 | 0.294 | 0.353 | 0.071 | 0.277 | 0.297 | 0.053 |
| Sample17 | 0.223 | 0.538 | 0.082 | 0.249 | 0.470 | 0.071 |
| Sample18 | 0.191 | 0.466 | 0.083 | 0.155 | 0.395 | 0.069 |
| Sample19 | 0.315 | 0.336 | 0.065 | 0.238 | 0.351 | 0.047 |
| Sample20 | 0.369 | 0.442 | 0.100 | 0.262 | 0.321 | 0.065 |
| Sample21 | 0.172 | 0.390 | 0.079 | 0.243 | 0.363 | 0.062 |
| Sample22 | 0.150 | 0.374 | 0.081 | 0.219 | 0.328 | 0.061 |
| Sample23 | 0.190 | 0.394 | 0.073 | 0.200 | 0.249 | 0.052 |
| Sample24 | 0.192 | 0.408 | 0.081 | 0.142 | 0.355 | 0.065 |
| Sample25 | 0.220 | 0.517 | 0.093 | 0.248 | 0.461 | 0.075 |
| Sample26 | 0.149 | 0.458 | 0.090 | 0.193 | 0.367 | 0.070 |
| Sample27 | 0.388 | 0.440 | 0.100 | 0.268 | 0.377 | 0.057 |
| Sample28 | 0.213 | 0.443 | 0.080 | 0.214 | 0.407 | 0.065 |
| Sample29 | 0.218 | 0.401 | 0.085 | 0.231 | 0.325 | 0.058 |
| Sample30 | 0.181 | 0.475 | 0.074 | 0.179 | 0.464 | 0.063 |
| Sample31 | 0.156 | 0.356 | 0.067 | 0.245 | 0.315 | 0.048 |
| Sample32 | 0.243 | 0.472 | 0.095 | 0.162 | 0.401 | 0.066 |
| Sample33 | 0.176 | 0.451 | 0.084 | 0.173 | 0.345 | 0.060 |
| Sample34 | 0.209 | 0.412 | 0.093 | 0.214 | 0.337 | 0.071 |
| Sample35 | 0.233 | 0.403 | 0.085 | 0.197 | 0.365 | 0.074 |
| Sample36 | 0.197 | 0.400 | 0.087 | 0.255 | 0.348 | 0.069 |
| Sample37 | 0.183 | 0.400 | 0.076 | 0.163 | 0.356 | 0.057 |
| Sample38 | 0.272 | 0.411 | 0.058 | 0.200 | 0.381 | 0.044 |
| Sample39 | 0.140 | 0.398 | 0.076 | 0.207 | 0.304 | 0.068 |
| Sample40 | 0.186 | 0.434 | 0.096 | 0.215 | 0.371 | 0.071 |
| Sample41 | 0.156 | 0.418 | 0.083 | 0.177 | 0.386 | 0.069 |
| Sample42 | 0.267 | 0.334 | 0.083 | 0.295 | 0.303 | 0.063 |
| Sample43 | 0.304 | 0.358 | 0.079 | 0.223 | 0.343 | 0.049 |
| Sample44 | 0.215 | 0.388 | 0.078 | 0.204 | 0.366 | 0.062 |
| Sample45 | 0.175 | 0.363 | 0.072 | 0.191 | 0.380 | 0.062 |
| Sample46 | 0.219 | 0.553 | 0.086 | 0.172 | 0.441 | 0.066 |
| Sample47 | 0.185 | 0.365 | 0.090 | 0.217 | 0.285 | 0.066 |
| Sample48 | 0.150 | 0.488 | 0.085 | 0.181 | 0.444 | 0.064 |
| Sample49 | 0.241 | 0.325 | 0.082 | 0.230 | 0.292 | 0.066 |
| Sample50 | 0.173 | 0.418 | 0.087 | 0.161 | 0.372 | 0.066 |
| | | | | | | |
| Mean | 0.209 | 0.410 | 0.080 | 0.213 | 0.354 | 0.062 |
| SD | 0.056 | 0.057 | 0.010 | 0.037 | 0.062 | 0.008 |

## *RMSD of the rescaled item parameter estimates (new form, Var(testlet)=0)*

| | **3-PL** | | | **Testlet** | | |
|---|---|---|---|---|---|---|
| | $\hat{a}$ | $\hat{b}$ | $\hat{c}$ | $\hat{a}$ | $\hat{b}$ | $\hat{c}$ |
| Sample1 | 0.135 | 0.307 | 0.074 | 0.147 | 0.248 | 0.058 |
| Sample2 | 0.124 | 0.338 | 0.064 | 0.127 | 0.304 | 0.054 |
| Sample3 | 0.164 | 0.379 | 0.069 | 0.171 | 0.304 | 0.056 |
| Sample4 | 0.121 | 0.275 | 0.047 | 0.124 | 0.250 | 0.046 |
| Sample5 | 0.139 | 0.280 | 0.071 | 0.135 | 0.241 | 0.060 |
| Sample6 | 0.132 | 0.346 | 0.064 | 0.132 | 0.349 | 0.053 |
| Sample7 | 0.143 | 0.486 | 0.066 | 0.114 | 0.416 | 0.050 |
| Sample8 | 0.159 | 0.371 | 0.068 | 0.129 | 0.323 | 0.053 |
| Sample9 | 0.156 | 0.320 | 0.067 | 0.126 | 0.300 | 0.055 |
| Sample10 | 0.166 | 0.374 | 0.066 | 0.169 | 0.304 | 0.052 |
| Sample11 | 0.181 | 0.334 | 0.090 | 0.153 | 0.325 | 0.069 |
| Sample12 | 0.244 | 0.358 | 0.065 | 0.215 | 0.305 | 0.054 |
| Sample13 | 0.173 | 0.375 | 0.071 | 0.195 | 0.319 | 0.059 |
| Sample14 | 0.103 | 0.291 | 0.069 | 0.097 | 0.235 | 0.058 |
| Sample15 | 0.255 | 0.424 | 0.077 | 0.217 | 0.364 | 0.063 |
| Sample16 | 0.104 | 0.298 | 0.066 | 0.114 | 0.286 | 0.056 |
| Sample17 | 0.154 | 0.315 | 0.054 | 0.143 | 0.274 | 0.044 |
| Sample18 | 0.105 | 0.431 | 0.060 | 0.105 | 0.393 | 0.049 |
| Sample19 | 0.196 | 0.449 | 0.087 | 0.182 | 0.413 | 0.072 |
| Sample20 | 0.131 | 0.424 | 0.068 | 0.129 | 0.418 | 0.058 |
| Sample21 | 0.149 | 0.294 | 0.064 | 0.147 | 0.241 | 0.050 |
| Sample22 | 0.125 | 0.328 | 0.064 | 0.121 | 0.335 | 0.052 |
| Sample23 | 0.169 | 0.255 | 0.049 | 0.170 | 0.234 | 0.043 |
| Sample24 | 0.140 | 0.354 | 0.069 | 0.138 | 0.329 | 0.060 |
| Sample25 | 0.131 | 0.308 | 0.058 | 0.164 | 0.273 | 0.049 |
| Sample26 | 0.107 | 0.302 | 0.066 | 0.107 | 0.251 | 0.052 |
| Sample27 | 0.086 | 0.299 | 0.061 | 0.078 | 0.245 | 0.044 |
| Sample28 | 0.137 | 0.362 | 0.059 | 0.127 | 0.330 | 0.041 |
| Sample29 | 0.194 | 0.268 | 0.048 | 0.154 | 0.240 | 0.039 |
| Sample30 | 0.131 | 0.337 | 0.062 | 0.100 | 0.319 | 0.048 |
| Sample31 | 0.102 | 0.331 | 0.064 | 0.107 | 0.286 | 0.055 |
| Sample32 | 0.129 | 0.357 | 0.066 | 0.122 | 0.335 | 0.051 |
| Sample33 | 0.118 | 0.419 | 0.063 | 0.110 | 0.381 | 0.052 |
| Sample34 | 0.157 | 0.335 | 0.066 | 0.131 | 0.298 | 0.053 |
| Sample35 | 0.213 | 0.363 | 0.067 | 0.169 | 0.308 | 0.056 |
| Sample36 | 0.145 | 0.224 | 0.054 | 0.127 | 0.268 | 0.051 |
| Sample37 | 0.198 | 0.402 | 0.074 | 0.166 | 0.336 | 0.058 |
| Sample38 | 0.160 | 0.406 | 0.067 | 0.137 | 0.360 | 0.056 |
| Sample39 | 0.134 | 0.310 | 0.061 | 0.135 | 0.242 | 0.047 |
| Sample40 | 0.136 | 0.347 | 0.058 | 0.137 | 0.286 | 0.050 |
| Sample41 | 0.204 | 0.263 | 0.072 | 0.164 | 0.212 | 0.058 |
| Sample42 | 0.091 | 0.418 | 0.072 | 0.095 | 0.342 | 0.058 |
| Sample43 | 0.091 | 0.355 | 0.069 | 0.110 | 0.324 | 0.059 |
| Sample44 | 0.152 | 0.272 | 0.061 | 0.145 | 0.249 | 0.053 |
| Sample45 | 0.146 | 0.394 | 0.068 | 0.169 | 0.313 | 0.055 |
| Sample46 | 0.129 | 0.348 | 0.055 | 0.129 | 0.363 | 0.047 |
| Sample47 | 0.140 | 0.337 | 0.060 | 0.151 | 0.310 | 0.052 |
| Sample48 | 0.155 | 0.372 | 0.073 | 0.139 | 0.333 | 0.055 |
| Sample49 | 0.148 | 0.301 | 0.051 | 0.164 | 0.271 | 0.041 |
| Sample50 | 0.136 | 0.410 | 0.061 | 0.130 | 0.349 | 0.048 |
| | | | | | | |
| Mean | 0.147 | 0.345 | 0.065 | 0.139 | 0.307 | 0.053 |
| SD | 0.036 | 0.055 | 0.008 | 0.029 | 0.051 | 0.007 |

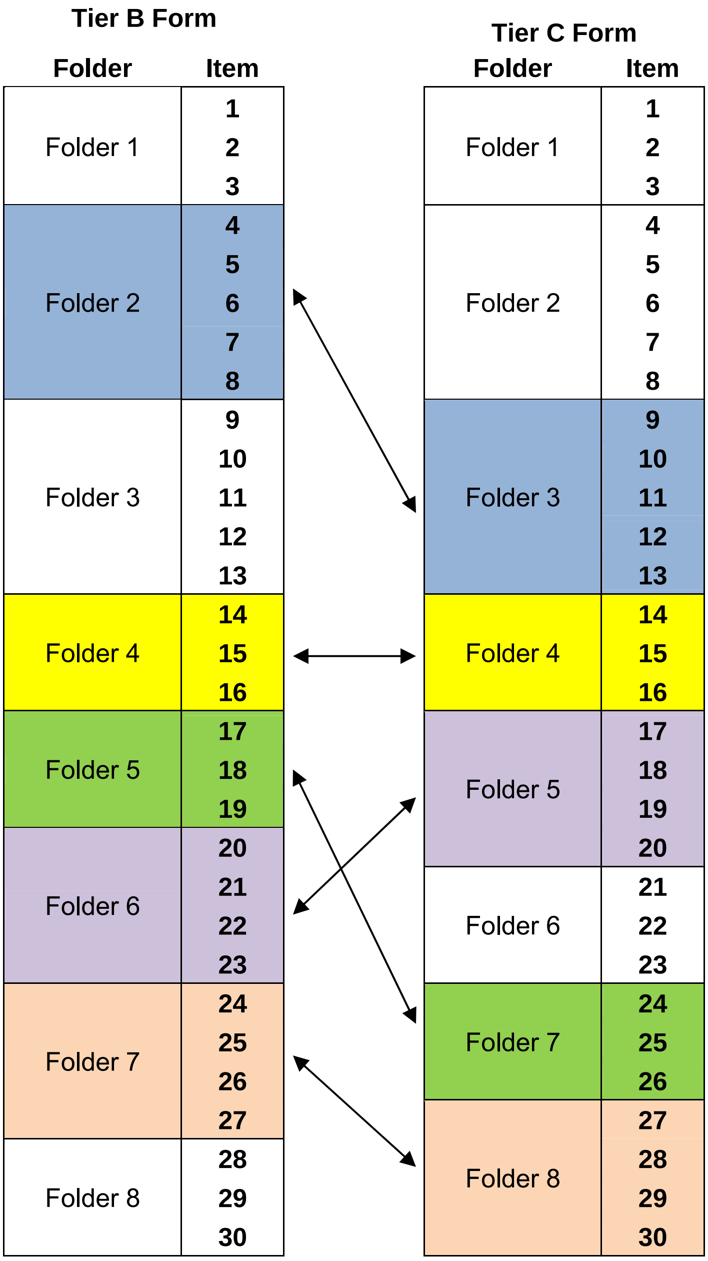
***RMSD of the rescaled item parameter estimates (new form, Var(testlet)=1)***

| | **3-PL** | | | **Testlet** | | |
|---|---|---|---|---|---|---|
| | $\hat{a}$ | $\hat{b}$ | $\hat{c}$ | $\hat{a}$ | $\hat{b}$ | $\hat{c}$ |
| Sample1 | 0.123 | 0.314 | 0.067 | 0.158 | 0.277 | 0.052 |
| Sample2 | 0.154 | 0.362 | 0.070 | 0.159 | 0.297 | 0.054 |
| Sample3 | 0.154 | 0.309 | 0.070 | 0.148 | 0.240 | 0.053 |
| Sample4 | 0.151 | 0.395 | 0.065 | 0.169 | 0.343 | 0.045 |
| Sample5 | 0.151 | 0.275 | 0.057 | 0.165 | 0.303 | 0.045 |
| Sample6 | 0.173 | 0.360 | 0.067 | 0.166 | 0.307 | 0.057 |
| Sample7 | 0.128 | 0.407 | 0.069 | 0.139 | 0.338 | 0.051 |
| Sample8 | 0.153 | 0.315 | 0.070 | 0.169 | 0.277 | 0.055 |
| Sample9 | 0.159 | 0.349 | 0.073 | 0.140 | 0.329 | 0.056 |
| Sample10 | 0.088 | 0.315 | 0.064 | 0.094 | 0.284 | 0.052 |
| Sample11 | 0.141 | 0.513 | 0.077 | 0.154 | 0.452 | 0.054 |
| Sample12 | 0.141 | 0.333 | 0.061 | 0.117 | 0.337 | 0.051 |
| Sample13 | 0.170 | 0.337 | 0.067 | 0.225 | 0.265 | 0.051 |
| Sample14 | 0.119 | 0.392 | 0.069 | 0.108 | 0.358 | 0.052 |
| Sample15 | 0.129 | 0.413 | 0.080 | 0.154 | 0.339 | 0.065 |
| Sample16 | 0.138 | 0.399 | 0.081 | 0.124 | 0.295 | 0.053 |
| Sample17 | 0.165 | 0.378 | 0.058 | 0.141 | 0.311 | 0.045 |
| Sample18 | 0.250 | 0.278 | 0.048 | 0.210 | 0.256 | 0.037 |
| Sample19 | 0.129 | 0.398 | 0.064 | 0.124 | 0.291 | 0.050 |
| Sample20 | 0.113 | 0.390 | 0.071 | 0.173 | 0.339 | 0.054 |
| Sample21 | 0.148 | 0.349 | 0.072 | 0.144 | 0.328 | 0.055 |
| Sample22 | 0.171 | 0.346 | 0.045 | 0.160 | 0.263 | 0.038 |
| Sample23 | 0.115 | 0.477 | 0.078 | 0.125 | 0.444 | 0.064 |
| Sample24 | 0.158 | 0.419 | 0.069 | 0.124 | 0.340 | 0.056 |
| Sample25 | 0.155 | 0.332 | 0.062 | 0.167 | 0.291 | 0.051 |
| Sample26 | 0.163 | 0.519 | 0.080 | 0.143 | 0.457 | 0.056 |
| Sample27 | 0.134 | 0.398 | 0.065 | 0.119 | 0.308 | 0.049 |
| Sample28 | 0.152 | 0.275 | 0.047 | 0.137 | 0.231 | 0.040 |
| Sample29 | 0.130 | 0.378 | 0.071 | 0.166 | 0.327 | 0.049 |
| Sample30 | 0.163 | 0.358 | 0.070 | 0.162 | 0.312 | 0.054 |
| Sample31 | 0.156 | 0.380 | 0.068 | 0.149 | 0.317 | 0.057 |
| Sample32 | 0.180 | 0.482 | 0.071 | 0.144 | 0.382 | 0.044 |
| Sample33 | 0.164 | 0.351 | 0.072 | 0.167 | 0.342 | 0.057 |
| Sample34 | 0.127 | 0.506 | 0.094 | 0.128 | 0.393 | 0.070 |
| Sample35 | 0.123 | 0.221 | 0.058 | 0.124 | 0.218 | 0.049 |
| Sample36 | 0.164 | 0.309 | 0.061 | 0.204 | 0.292 | 0.056 |
| Sample37 | 0.128 | 0.225 | 0.060 | 0.116 | 0.217 | 0.053 |
| Sample38 | 0.123 | 0.254 | 0.057 | 0.116 | 0.240 | 0.052 |
| Sample39 | 0.181 | 0.494 | 0.066 | 0.142 | 0.443 | 0.043 |
| Sample40 | 0.143 | 0.385 | 0.060 | 0.112 | 0.325 | 0.050 |
| Sample41 | 0.156 | 0.393 | 0.076 | 0.158 | 0.334 | 0.060 |
| Sample42 | 0.189 | 0.344 | 0.067 | 0.121 | 0.296 | 0.047 |
| Sample43 | 0.129 | 0.402 | 0.066 | 0.103 | 0.327 | 0.050 |
| Sample44 | 0.134 | 0.327 | 0.061 | 0.116 | 0.303 | 0.047 |
| Sample45 | 0.153 | 0.488 | 0.070 | 0.168 | 0.374 | 0.053 |
| Sample46 | 0.233 | 0.369 | 0.076 | 0.175 | 0.237 | 0.052 |
| Sample47 | 0.102 | 0.430 | 0.067 | 0.112 | 0.392 | 0.051 |
| Sample48 | 0.166 | 0.378 | 0.073 | 0.142 | 0.336 | 0.058 |
| Sample49 | 0.136 | 0.266 | 0.062 | 0.162 | 0.225 | 0.054 |
| Sample50 | 0.203 | 0.359 | 0.069 | 0.267 | 0.247 | 0.049 |
| | | | | | | |
| Mean | 0.150 | 0.369 | 0.067 | 0.148 | 0.316 | 0.052 |
| SD | 0.029 | 0.071 | 0.009 | 0.032 | 0.059 | 0.006 |

***RMSD of the rescaled item parameter estimates (new form, Var(testlet)=2)***

| | **3-PL** | | | **Testlet** | | |
|---|---|---|---|---|---|---|
| | $\hat{a}$ | $\hat{b}$ | $\hat{c}$ | $\hat{a}$ | $\hat{b}$ | $\hat{c}$ |
| Sample1 | 0.103 | 0.337 | 0.064 | 0.151 | 0.251 | 0.045 |
| Sample2 | 0.333 | 0.305 | 0.072 | 0.281 | 0.313 | 0.055 |
| Sample3 | 0.124 | 0.390 | 0.070 | 0.207 | 0.394 | 0.065 |
| Sample4 | 0.133 | 0.370 | 0.066 | 0.126 | 0.313 | 0.048 |
| Sample5 | 0.218 | 0.466 | 0.076 | 0.199 | 0.286 | 0.053 |
| Sample6 | 0.167 | 0.491 | 0.064 | 0.195 | 0.414 | 0.055 |
| Sample7 | 0.170 | 0.363 | 0.077 | 0.161 | 0.288 | 0.048 |
| Sample8 | 0.167 | 0.377 | 0.079 | 0.153 | 0.354 | 0.062 |
| Sample9 | 0.123 | 0.398 | 0.069 | 0.109 | 0.449 | 0.050 |
| Sample10 | 0.149 | 0.311 | 0.062 | 0.174 | 0.307 | 0.041 |
| Sample11 | 0.138 | 0.516 | 0.073 | 0.201 | 0.431 | 0.054 |
| Sample12 | 0.139 | 0.353 | 0.063 | 0.199 | 0.305 | 0.057 |
| Sample13 | 0.160 | 0.513 | 0.065 | 0.118 | 0.426 | 0.050 |
| Sample14 | 0.183 | 0.372 | 0.070 | 0.196 | 0.291 | 0.057 |
| Sample15 | 0.171 | 0.367 | 0.061 | 0.128 | 0.323 | 0.041 |
| Sample16 | 0.189 | 0.440 | 0.065 | 0.229 | 0.364 | 0.054 |
| Sample17 | 0.171 | 0.493 | 0.078 | 0.177 | 0.424 | 0.065 |
| Sample18 | 0.224 | 0.357 | 0.062 | 0.137 | 0.345 | 0.057 |
| Sample19 | 0.233 | 0.326 | 0.065 | 0.151 | 0.242 | 0.044 |
| Sample20 | 0.211 | 0.335 | 0.070 | 0.249 | 0.340 | 0.062 |
| Sample21 | 0.148 | 0.412 | 0.061 | 0.223 | 0.304 | 0.045 |
| Sample22 | 0.182 | 0.309 | 0.073 | 0.182 | 0.269 | 0.059 |
| Sample23 | 0.206 | 0.378 | 0.062 | 0.124 | 0.204 | 0.043 |
| Sample24 | 0.177 | 0.363 | 0.074 | 0.145 | 0.315 | 0.059 |
| Sample25 | 0.177 | 0.452 | 0.083 | 0.127 | 0.381 | 0.051 |
| Sample26 | 0.165 | 0.408 | 0.063 | 0.119 | 0.346 | 0.048 |
| Sample27 | 0.208 | 0.410 | 0.063 | 0.197 | 0.409 | 0.044 |
| Sample28 | 0.156 | 0.403 | 0.070 | 0.235 | 0.327 | 0.051 |
| Sample29 | 0.122 | 0.456 | 0.083 | 0.107 | 0.433 | 0.056 |
| Sample30 | 0.138 | 0.376 | 0.056 | 0.121 | 0.251 | 0.040 |
| Sample31 | 0.150 | 0.333 | 0.061 | 0.139 | 0.229 | 0.048 |
| Sample32 | 0.140 | 0.395 | 0.072 | 0.107 | 0.324 | 0.051 |
| Sample33 | 0.201 | 0.428 | 0.086 | 0.175 | 0.301 | 0.047 |
| Sample34 | 0.152 | 0.335 | 0.062 | 0.167 | 0.283 | 0.053 |
| Sample35 | 0.220 | 0.446 | 0.072 | 0.171 | 0.336 | 0.051 |
| Sample36 | 0.148 | 0.455 | 0.077 | 0.206 | 0.397 | 0.062 |
| Sample37 | 0.140 | 0.313 | 0.065 | 0.126 | 0.280 | 0.046 |
| Sample38 | 0.218 | 0.461 | 0.060 | 0.207 | 0.354 | 0.046 |
| Sample39 | 0.152 | 0.364 | 0.057 | 0.160 | 0.312 | 0.047 |
| Sample40 | 0.166 | 0.342 | 0.074 | 0.171 | 0.271 | 0.049 |
| Sample41 | 0.160 | 0.355 | 0.070 | 0.149 | 0.322 | 0.048 |
| Sample42 | 0.274 | 0.366 | 0.068 | 0.244 | 0.271 | 0.043 |
| Sample43 | 0.133 | 0.449 | 0.073 | 0.169 | 0.401 | 0.050 |
| Sample44 | 0.144 | 0.447 | 0.079 | 0.129 | 0.338 | 0.046 |
| Sample45 | 0.212 | 0.394 | 0.087 | 0.234 | 0.400 | 0.077 |
| Sample46 | 0.142 | 0.435 | 0.077 | 0.176 | 0.366 | 0.057 |
| Sample47 | 0.157 | 0.378 | 0.079 | 0.187 | 0.299 | 0.051 |
| Sample48 | 0.169 | 0.355 | 0.058 | 0.117 | 0.255 | 0.046 |
| Sample49 | 0.200 | 0.361 | 0.057 | 0.149 | 0.341 | 0.052 |
| Sample50 | 0.193 | 0.633 | 0.091 | 0.133 | 0.462 | 0.063 |
| | | | | | | |
| Mean | 0.173 | 0.398 | 0.070 | 0.169 | 0.333 | 0.052 |
| SD | 0.041 | 0.064 | 0.008 | 0.042 | 0.062 | 0.007 |

## Appendix E  The Original Item Structure of Tier B Form and Tier C Form of Grade Cluster 3-5 Reading Tests of 2004-2005 ACCESS for ELLs®



**Tier B Form**

| Folder | Item |
|---|---|
| Folder 1 | 1 |
| | 2 |
| | 3 |
| Folder 2 | 4 |
| | 5 |
| | 6 |
| | 7 |
| | 8 |
| Folder 3 | 9 |
| | 10 |
| | 11 |
| | 12 |
| | 13 |
| Folder 4 | 14 |
| | 15 |
| | 16 |
| Folder 5 | 17 |
| | 18 |
| | 19 |
| Folder 6 | 20 |
| | 21 |
| | 22 |
| | 23 |
| Folder 7 | 24 |
| | 25 |
| | 26 |
| | 27 |
| Folder 8 | 28 |
| | 29 |
| | 30 |

**Tier C Form**

| Folder | Item |
|---|---|
| Folder 1 | 1 |
| | 2 |
| | 3 |
| Folder 2 | 4 |
| | 5 |
| | 6 |
| | 7 |
| | 8 |
| Folder 3 | 9 |
| | 10 |
| | 11 |
| | 12 |
| | 13 |
| Folder 4 | 14 |
| | 15 |
| | 16 |
| Folder 5 | 17 |
| | 18 |
| | 19 |
| | 20 |
| Folder 6 | 21 |
| | 22 |
| | 23 |
| Folder 7 | 24 |
| | 25 |
| | 26 |
| Folder 8 | 27 |
| | 28 |
| | 29 |
| | 30 |

Note: The two headed arrows point to the common folders that are shared by the two test forms. For example, Folder 2 in the Tier B form is Folder 3 in the Tier C form.

## References

Ackerman, P. L. (1987). Individual differences in skill learning: an integration of psychometric and information processing perspectives. *Psychological Bulletin, 102*(1), 3-27.

Allen, S., & Sudweeks, R. R. (2001, April 10-14). *Identifying and managing local item dependence in context-dependent item sets.* Paper presented at the Paper presented at the Annual Meeting of the American Educational Research Association, Seattle, WA.

Andrich, D. (1985). A latent trait model for items with response dependencies: Implications for test construction and analysis. *Test Design: Developments in Psychology and Psychometrics*, 245-275.

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

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

Bishop, N. S., & Omar, M. H. (2002, April 2-4). *Comparing vertical scales derived from dichotomous and polytomous IRT models for a test composed of testlets.* Paper presented at the Annual Meeting of the National Council on Educational Measurement, New Orleans, LA.

Bowman, C. M., & Peng, S. S. (1972). *A preliminary investigation of recent advanced psychology tests in the GRE program-an application of a cognitive classification system*. Princeton, NJ.

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

Camilli, G., Yamamoto, K., & Wang, M. (1993). Scale shrinkage in vertical equating. *Applied Psychological Measurement, 17*(4), 379.

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.

Chib, S., & Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *American Statistician, 49*, 327-335.

Congress, U. S. (2001). No child left behind act of 2001. *Public Law*, 107-110.

Cook, K. F., Dodd, B. G., & Fitzpatrick, S. J. (1999). A comparison of three polytomous Item Response Theory models in the context of testlet scoring. *Journal of Outcome Measurement, 3*(1), 1.

Crehan, K. D. (1993, April 12-16). *A comparison of testlet reliability for polytomous scoring methods.* Paper presented at the Annual Meeting of the American Educational Research Association, Atlanta, GA.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334.

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.

Embretson, S., & Gorin, J. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement, 38*(4), 343-368.

Embretson, S., & Yang, X. (2006). Item Response Theory. In J. L. Green & P. B. Elmore (Eds.), *Handbook of Complementary Methods in Education Research*. Mahwah, N.J: Lawrence Erlbaum Associates.

Feldt, L. S. (2002). Estimating the internal consistency reliability of tests composed of testlets varying in length. *Applied Measurement in Education, 15*(1), 33.

Fennessy, L. M. (1995). *The impact of local dependencies on various irt outcomes.* University of Massachusetts at Amherst.

Frederiksen, N. (1984). The real test bias: influences of testing on teaching and learning. *The American Psychologist, 39*(3), 193-202.

Frisbie, D. A., & Druva, C. A. (1986). Estimating the reliability of multiple true-false tests. *Journal of Educational Measurement, 23*(2), 99-105.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*(4), 457.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 6*(6), 721-741.

Glas, C. A. W., Wainer, H., & Bradlow, E. T. (2000). MML and EAP estimation in testlet-based adaptive testing In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized Adaptive testing: Theory and practice* (5th ed., pp. 271-287). Boston, MA: Kluwer.

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

Haertel, E. (2004). *The behavior of linking items in test equating*: Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education & Information Studies, University of California, Los Angeles.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items*. Mahwah, N.J: Lawrence Erlbaum Assoc Inc.

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

Hanson, B., & Zeng, L. (2004). *ST: A Computer Program for IRT Scale Transformation.* Iowa Testing Program, University of Iowa.

Harwell, M.R., & Janosky, J.E. (1991). An empirical study of the effects of small datasets and varying prior variances on item parameter estimation in BILOG. *Applied Psychological Measurement, 15,* 279-291

Hendrickson, A. (2007). An NCME instructional module on multistage testing. *Educational Measurement: Issues and Practice, 26*(2), 44-52.

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

Kim, S., & Kolen, M. J. (2003). *POLYST: A Computer Program for Polytomous IRT Scale Transformation*. Iowa City, IA: Iowa Testing Programs, University of Iowa.

Kingston, N. M., & Dorans, N. J. (1982). *The feasibility of using item response theory as a psychometric model for the GRE aptitude test*: Graduate Record Examinations Board, Princeton N. J.

Kingston, N. M., & Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adaptive testing. *Applied Psychological Measurement, 8*(2), 147.

Kolen, M. J., & Brennan, R. L. (2004). *Test Equating, Scaling, and Linking: Methods and Practices*. New York, NY: Springer.

Lee, G. (1999, April 19-23). *Conditional standard errors of measurement for tests composed of testlets.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Montreal, Canada.

Lee, G., Brennan, R. L., & Frisbie, D. A. (2000). Incorporating the testlet concept in test score analyses. *Educational Measurement: Issues and Practice, 19*(4), 9.

Lee, G., Kolen, M. J., Frisbie, D. A., & Ankenmann, R. D. (1998, April 13-17). *Equating test forms composed of testlets using dichotomous and polytomous IRT models.* Paper presented at the Annual Meeting of the American Educational Research Association San Diego, CA.

Li, D., & Yin, P. (2008). *Equating with polytomous IRT models under the common item nonequivalent groups design*. Paper presented at the 2008 annual conference of NCME.

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

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

MacGreger, D., Louguit, M., Ryu, J. R., Li, D., & Kenyon, D. M. (2008). *WIDA Annual Technical Report for ACCESS for ELLs English Language Proficiency Test, Series 102, 2006-2007 Administration*. Washington D.C.: Center for Applied Linguistics.

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

Masters, G. N. (1988). Item discrimination: when more is worse. *Journal of Educational Measurement, 25*(1), 15-29.

Maydeu-Olivares, A., Drasgow, F., & Mead, A. D. (1994). Distinguishing among Paranletric Item Response Models for polychotomous ordered data. *Applied Psychological Measurement, 18*(3), 245.

Mehrens, W. A., & Lehmann, I. J. (1984). *Measurement and evaluation in education and psychology*. New York, NY: Holt, Rinehart and Winston Inc.

Millman, J., & Greene, J. (1989). The specification and development of tests of achievement and ability. *Educational measurement, 3*, 335-366.

Mislevy, R. J. (1994). Evidence and inference in educational assessment. *Psychometrika, 59*(4), 439-483.

Mislevy, R. J., & Bock, R. D. (1990). *BILOG 3: item analysis and test scoring with binary logistic models*: Mooresville, IN: Scientific Software.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (1999). *On the Roles of Task Model Variables in Assessment Design*: National Center for Research on Evaluation, Standards, and Student Testing, Center for the Study of Evaluation, Graduate School of Education & Information Studies, University of California, Los Angeles; US Dept. of Education, Office of Educational Research and Improvement, Educational Resources Information Center.

Morgenstern, C. F., & Renner, J. W. (1984). Measuring thinking with standardized science tests. *Journal of Research in Science Teaching, 21*(6), 639-648.

Muraki, E. (1992). A generalized partial credit model: application of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159.

Muraki, E., & Bock, R. D. (1997). *PARSCALE: IRT Item Analysis and Test Scoring for Rating-scale Data*. Chicago, IL: Scientific Software International.

Nickerson, R. S. (1989). New directions in educational assessment. *Educational Researcher, 18*(9), 3-7.

Patz, R. J., & Junker, B. W. (1999). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics, 24*(4), 342.

Reese, L. M. (1999). *A Classical Test Theory Perspective on LSAT Local Item Dependence*. Newtown, PA: Law School Admission Council.

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 Chicago, IL: Psychometric Society.

Sinharay, S., & Educational Testing, S. (2003). *Assessing Convergence of the Markov Chain Monte Carlo Algorithms: A Review*. Princeton, NJ: Educational Testing Service.

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

Spearman, C. (1904). " General intelligence," objectively determined and measured. *The American Journal of Psychology, 16*(2).

Spiegelhalter, D. J., Thomas, A., Best, N. G., & Lunn, D. (2003). *WinBUGS Version 1.4 User Manual*. Cambridge, UK: MRC Biostatistics Unit.

Steinberg, L., & Thissen, D. (1996). Uses of item response theory and the testlet concept in the measurement of psychopathology. *Psychological Methods, 1*(1), 81-97.

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

Tang, K. L., & Eignor, D. R. (1997). *Concurrent Calibration of Dichotomously and Polytomously Scored TOEFL Items Using IRT Models*: Educational Testing Service.

Thissen, D., Billeaud, K., McLeod, L., & Nelson, L. (1997). *A brief introduction to item response theory for items scored in more than two categories*. Paper presented at the National Assessment Governing Board Achievement Levels Workshop.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51*(4), 567-577.

Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: a use of multiple-categorical-response models. *Journal of Educational Measurement, 26*(3), 247.

van den Wollenberg, A. L. (1982). Two new test statistics for the Rasch model. *Psychometrika, 47*(2), 123-140.

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

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet Response Theory and its Applications*. New York: Cambridge University Press.

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

Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: toward a Marxist theory of test construction. *Applied Measurement in Education, 6*(2), 103-118.

Wainer, H., & Wang, X. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement, 37*(3), 203.

Wang, W. C., Cheng, Y. Y., & Wilson, M. (2005). Local item dependence for items across tests connected by common stimuli. *Educational and Psychological Measurement, 65*(1), 5.

Wang, X., Bradlow, E. T., & Wainer, H. (2002). A general Bayesian model for testlets: theory and applications. *Applied Psychological Measurement, 26*(1), 109.

Warren, G. (1979). Essay versus multiple choice tests. *Journal of Research in Science Teaching, 16*(6), 563-567.

Way, W. D., & Tang, K. L. (1991, April 4-6). *A comparison of four logistic model equating methods.* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

WIDA. (2008). *ACCESS for ELLS Listening, Reading, Writing and Speaking sample items*. Madison, WI: World-class Instructional Design and Assessment Consortium.

Xiao, B. (1999). Strategies for computerized adaptive grading testing. *Applied Psychological Measurement, 23*(2), 136.

Yen, W. M. (1980). The extent, causes and importance of context effects on item parameters for two latent trait models. *Journal of Educational Measurement, 17*(4), 297.

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*(2), 125-145.

Yen, W. M. (1985). Increasing item complexity: a possible cause of scale shrinkage for unidimensional item response theory. *Psychometrika, 50*(4), 399-410.

Yen, W. M. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement*, 299-325.

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

Zeng, L., & Kolen, M. J. (1994). *IRT scale transformations using numerical integration.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans.

Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (2001). *Effects of local item dependence on the validity of IRT item, test, and ability statistics*. Washington DC: MCAT section, Association of American Medical Colleges.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2005). *BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items*. Lincolnwood, IL: Scientific Software International, Inc.