Linking with Planned Missing Data:

Concurrent Calibration with Multiple Imputation

By

Min Sung Kim

University of Kansas

Submitted to the graduate degree program in Educational Psychology and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

_____

Chairperson Dr. William Skorupski

_____

Dr. Neal Kingston

_____

Dr. Bruce Frey

_____

Dr. Vicki Peyton

_____

Dr. Wei Wu

Date Defended: December 4th 2015

The Dissertation Committee for Min Sung Kim

certifies that this is the approved version of the following dissertation:

Linking with Planned Missing Data:

Concurrent Calibration with Multiple Imputation

_____

Chairperson Dr. William Skorupski

Date approved: December 4th 2015

Abstract

The purpose of this paper is to introduce a new Item Response Theory (IRT) concurrent calibration method using multiple imputation and investigate its effectiveness by comparing with other equating methods. The 3-parameter logistic (3PL) model is chosen due to its reliable performance and the 2-parameter logistic (2PL) model is also applied to compare the performance. Six equating methods were compared in simulated data studies under a common-item nonequivalent group design, and ability parameters were randomly drawn from various distributions with different combinations of mean and variance. Additionally, the effect of two anchor test lengths on parameter estimation was compared for all conditions. The main focus was on comparing concurrent calibration methods of marginal maximum likelihood estimation (MMLE) and multiple imputation (MI). For real data, PISA 2000 reading score was applied to several equating methods. Likewise the previous literatures, MI showed better or similar mean squared error (MSE) than MMLE. In addition, the usefulness of Mean Imputed Score, a byproduct of MI was proposed and compared with Observed Score Equating (OSE) and True Score Equating (TSE) results.

Key words: concurrent calibration, multiple imputation, linking, equating, IRT

Dedication

God the Father, God the Son, and God the Holy Spirit

my parents

my wife

my daughter

Acknowledgements

There are many people who have given me much more than I can ever possibly repay. I would like to express my deep gratitude to my adviser and dissertation chair, Dr. William Skorupski. My first meeting with him, which turned out to be one of the most important conversations of my life, inspired me to pursue advanced studies in psychometrics. Without his rich guidance, tremendous support, insightful comments, and great patience, this dissertation would not have been possible and my doctoral studies would never have been completed. I could not have asked for a better adviser.

Since my first step at KU, Dr. Neal Kingston has become not only a good mentor but also a great industry advisor. I am grateful to his for many things he has done for me. I will miss conversations I had with his about educational measurement.

My special thanks go to Dr. Bruce Frey who, for the last five years, has become my mentor. I would like to thank him for his strong support, both academically and emotionally, during my time at KU.

I thank Dr. Vicki Peyton for her helpful comments and great support to my dissertation. I am privileged and proud to have her in my dissertation committee. My sincere gratitude goes to Dr. Wei Wu, for her great support and joyful lecture about missing data analysis.

With all my heart, I would like to thank Dr. Marianne Perie, Director of the Center for Educational Testing and Evaluation for everything she has done for me. It was Dr. Scott Bishop who built a bridge connecting me to my success today. I could not find any better editor than my friend, Joe and could not find any word to express my thanks to him. Finally, I thank many other professors, graduate students, and friends who have enriched my life at KU.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter One - Introduction

An important area of emphasis in psychometric research over the past several decades has been the issue of equating and linking. Questions about ways of equating test forms have served, either directly or indirectly, as the predominant focus of accurate and appropriate parameter estimation. These kinds of questions have increased comparisons of methods, and efforts to introduce new methods have led to the development of equating methods designed to maximize accuracy and appropriateness.

The primary purpose of large-scale standardized testing is measuring or evaluating examinees' abilities and/or skills as fairly, equitably, and objectively as possible. To maintain the security of examinations, testing programs make every effort to ensure no one is advantaged by pre-knowledge of test questions. Therefore, most large-scale standardized testing programs develop and use multiple forms or versions of the same test.

1.1. Statement of Problem

Needs of multiple forms and equating

Multiple forms help to decrease and prevent the occurrence of unfairness, inequity, and cheating behavior. Plenty of administrative settings have been developed to obtain test security, and a variety of statistical methods have been adapted to test scores from multiple forms. Without such methods, students taking the easier test form have an advantage and achieve a higher score than an equally able student who takes the more difficult form.

Multiple test forms usually need to satisfy the specification of parallel tests, which demands equal psychometric characteristics of the test across test forms. In practice, making multiple test forms completely equivalent in reliability and difficulty is impossible. Under multiple form testing without test difficulty adjustment across test forms, examinees may have an

advantage because they take an easier or a more favorable form of a test than others. In other words, one examinee may take a form of a test with more difficult items than those taken by another examinee. Unless the fairness and equity of the test are ensured, test scores from multiple test forms cannot be utilized interchangeably.

To offset the possibility that examinees gain benefits due to different test characteristics across test forms, equating procedures are used to place scores on a common scale, both in classical test theory (CTT) and item response theory (IRT). The random group design with common items is a design mainly employed by large-scale standardized test programs because of its administrative flexibility. The two tests are linked by anchor test items, so it is important to ensure the characteristics of these items are representative of the total test. Based on the strong statistical assumptions about group separation and form differences, the selection of equating methods needs to be done carefully to find an appropriate one. Many statistical methods have been proposed to obtain fairness and objectiveness of test scores across multiple forms and, as a result, many equating and linking methods have been introduced, studied, and evaluated in a variety of conditions.

Equating in IRT frameworks

In classical test theory, examinees' skills or abilities are typically represented as number-correct scores; in IRT, however, characteristics of test takers, test items, and test sets are parameterized. As a means of measuring or evaluating a group of examinees' skills, IRT models are popular due to their usefulness in a variety of applications, like test score equating and linking, testing for differential item function (DIF), item banking, test developing, and adaptive testing and its diverse applications.

To use the IRT equating as CTT equating, a three-step process is required (Cook & Eignor, 1991): 1) selecting the design, 2) scaling, and 3) equating. After IRT model selection, large-scale test response data need to be collected under a specific plan, which is called a test design. Once a test is administered, IRT parameters are estimated by the preset rules. Each test form provides its own IRT parameters, and the multiple sets of parameters are scaled for interchangeable usage across forms. Scaled IRT parameters from multiple forms make it possible to provide number-correct scores on a unified scale. The two major methods for the last step of IRT equating are observed score equating (OSE) and true score equating (TSE).

In previous research, many efforts have been made to seek accurate and robust equating and scaling methods to enable consistent parameter estimation across test forms. Some IRT equating designs also rely on common items across forms, usually under a Non-Equivalent group with Anchor Test (NEAT) design, which is also used with CTT equating. When population distributions of the groups taking the two tests are different, estimated IRT parameters are not the same across test forms due to different IRT parameter scales. Anchor items shared by the two tests provide statistical information for IRT parameter equating formulas. Generally, the parameters of target form(s) are transformed to the scale of the base form. Linear transformation methods (mean/mean (M/M), mean/sigma (M/S)), total characteristic curve methods (Haebara, Stocking–Lord) and concurrent calibration methods have been compared under the NEAT design (Kolen & Brennan, 2014). Many researchers have concluded that the concurrent calibration method provided better accuracy of parameter recovery in the simulation studies with lower MSE or RMSE values than other equating methods (Kim & Hanson, 2002; Hanson & Beguin, 2002; Kim & Cohen, 2007; Ogasawara, 2001).

Estimation of concurrent calibration

The previous IRT concurrent estimation studies were done under the two major maximum likelihood estimation methods: Joint Maximum Likelihood Estimation (JMLE) and Marginal Maximum Likelihood Estimation (MMLE), by software packages BILOG-MG (Mislevy & Bock, 1990) and MULTILOG (Thissen, 1991). Both of these packages rely on the Expectation-Maximization (EM) algorithm (Kim & Cohen, 1998; Hanson & Beguin, 2002). These estimation methods try to find the best parameter values that fit the estimation model. Concurrent calibration of two test forms essentially has two missing data sets. In the base form, the unique items of target form are considered as missing and vice versa. In other words, concurrent calibration is conducted with missing data.

The key perspective of this study is to view concurrent calibration as a missing data analysis. Although the IRT concurrent calibration method provided better accuracy in parameter estimation in previous studies (Kim & Hanson, 2002; Hanson & Beguin, 2002; Kim & Cohen, 2007; Ogasawara, 2001), there have been no studies to my knowledge in estimation from the viewpoint of missing data.

For concurrent calibration, the NEAT design with two forms consists of three subtest sets: 1) unique test items on the base form, 2) unique test items on the target form, and 3) the common items on both base and target forms. In terms of missing data, each missing part across test forms is missing in the combined test form.

Besides ML, many missing data analysis techniques have been developed and applied to educational testing, not only for research purposes but also for practical reasons. Multiple imputation (MI) is one of the most popular estimation methods due to its distinct feature of

providing imputed data that are statistically plausible. However, MI has not been studied for the purposes of IRT concurrent calibration.

Recently, MI studies have focused on missing data analysis methods. Not many MI data augmentation have been applied to IRT, but it is done by a process similar to Bayesian estimation. Bayesian estimation involves drawing samples from a posterior distribution, based on prior distributions and a specified model. This can be accomplished using Markov Chain Monte Carlo (MCMC) methods. Similarly, in MI the imputed data are drawn from a posterior distribution through a specific model. In this study, IRT is the specific model.

Evaluation criteria

By the definition of equating, adequate evaluation criteria need to be chosen to assess the accuracy of equating results, especially in method comparison studies. Simulated data can provide the true equating relationship, and results can be evaluated by the difference between true and equated values. In the last step of equating, the number-correct score can be evaluated by the difference between true and equated number correct score. So far, the score difference between the equated score from one form and the score from another form has been utilized in OSE and TSE. This common evaluation criterion is categorized as "equating a test to itself" according to Harris & Crouse (1993). For the last step in number correct score equating, the two most common methods for equating number correct scores, OSE and TSE, were chosen. In addition, an alternative evaluation criteria, true target form equivalent number correct score (TTS), was introduced by the author. TTS was compared to the results of OSE and TSE and it provides a mean imputed score, which comes from the new estimation method in this study.

Terminology

While studying test score equating, several terms are widely used. The use of terms "linking" and "equating" need to be considered with care, with the choice depending on how comparable test scores are intended to be. Linking is less strict than equating, and is used for comparing two or more test scores under similar test constructs. Equating is used when the intent is to use scores from different forms interchangeably. On the other hand, "scaling" is the term for IRT parameter estimation. In concurrent IRT estimation, there is only one form, so the term "calibration" is commonly used instead of equating because equating is accomplished as part of the calibration.

1.2. Purpose

Concurrent calibration has been studied using ML estimation, but not using Multiple Imputation (MI). In this study, concurrent calibration with MI was conducted and its performance was compared with concurrent calibration and other linking methods in terms of estimation accuracy. The common evaluation criteria of IRT simulation studies, mean squared error (MSE) and bias values, are compared across equating methods. As the last step of the equating procedure, results of observed score equating and true score equating are performed and compared with other methods. A new score, true observed score from simulation, was introduced, and its performance was also compared across equating methods.

It is important to examine the effectiveness of new equating methods in various test conditions, including examinee distributions, the number of common items, and current linking methods. To determine the relationship between equating methods and recovery accuracy of true parameters, the true number-correct score from the simulation was evaluated in this study. The number-correct score using observed score equating (OSE) and true score equating (TSE)

methods are compared across each IRT linking method. MI can provide observed scores multiple times, so the mean of each imputed score (MIS) can be utilized as an alternative score of equated number-correct score with OSE and TSE.

1.3. Data and Variables

First of all, two kinds of data are utilized in this study, one is simulation study and the other one is PISA 2000 reading scores. To provide more generalizable results, 12 different conditions of simulation data sets were generated. Also, real test data of two forms similar to the simulation conditions were also taken from PISA 2000 reading scores of the U.S. students. The dependent variables are mean squared error and bias values. The independent variables include the individual item responses, equating method, total scores, the true ability distributions, and the number of common items and unique items across forms.

1.4. Research questions

1. With simulated data, will multiple imputation estimation produce more accurate linking results, in terms of minimizing mean squared error, than concurrent calibration using ML with the EM algorithm?
2. Will the patterns of bias of IRT person parameters (θ) be consistent both with previous studies in simulated data and PISA 2000 data and across different numbers of anchor items and population distributions?
3. Will the Mean Imputed Score (MIS) be comparable with IRT OSE and TSE results in simulated data and PISA 2000 data?
4. When the true target form equivalent score is applied in simulation study, will MIS have be comparable to OSE and TSE?

1.5. Hypotheses

Planned missing data (Graham, Taylor, Olchowski & Cumsille, 2006) designs enable researchers to study missing completely at random (MCAR) and missing at random (MAR) mechanisms in test settings and provide valuable information with increased power and accuracy of statistical inferences. These missing mechanisms are described in more detail in the literature review section later. MI allows understanding of IRT linking and equating, and illustrates the important role that these investigations can play in test development and scoring. Missing data analysis techniques have been one of many popular topics in traditional psychological measurement, with applications in psychometric, statistical, and mathematical inference. Although using MI is not fully developed in the educational measurement field, MI has the potential for various applications in the future.

1.6. Research expectations

Most psychological research on missing data analysis concludes that results of both ML and MI demonstrate good estimation in the presence of MCAR and MAR mechanisms in regression models, but ML demonstrated slightly better accuracy in previous studies (Enders, 2010). However, MI provides better estimation than ML with EM algorithm in many conditions of IRT (Finch, 2008). Also, MI estimates the expected observed scores of counter forms using mean scores across the imputed data sets. Previous concurrent calibration could only provide one set of parameter values from two forms. So, the general equated scores of OSE and TSE could not be obtained with previous concurrent calibration methods. However, MI provides imputed data sets, and the mean score of counter forms can be utilized as an equated score. The mean imputed score is compared to OSE and TSE. So far, the score difference between equivalent scores from the target and base forms has been focused on TSE and OSE results. In this

simulation study, the true observed scores of missing items in the base form, which are the unique items of target form, are imputed and the sum score of anchor items and imputed scores re compared to OSE and TSE scores of each transformation method

1.7. Summary

This research addresses the problem of current methods of concurrent calibration. Although ML estimation is mainly used for missing data, MI is also an effective data augmentation approach for missing data in the IRT framework. This simulation study provides evidence about how accurately IRT parameters are recovered using MI and how it compares with ML. One of the advantages of MI is that it provides imputed response data. The imputed data make it possible to compare imputed number-correct scores from the base form including the target form portion with true observed scores from the simulation. As a result, MI provides mean imputed scores, which are comparable with scores from observed score equating and true score equating methods.

Chapter Two - Literature Review

"Linking" is a general term used to describe the comparing of test scores. However, "equating" should be used carefully to describe test reporting where rigorous testing conditions are satisfied in order to use test scores from two or more forms interchangeably. Scaling, transformation, and calibration are the middle steps when IRT parameters are estimated and unified across test forms. This chapter discusses the literature pertaining to each of the IRT equating methods in this study. Angoff (1984) illustrated four requirements to be considered before equating. When equating, two forms of test need to: (1) measure the same abilities or skills; (2) have a common score scaling procedure that is independent from data sets; (3) have interchangeable equated scores across test forms; and (4) be equated with a symmetric transformation (i.e., regardless of which form is chosen as the base form).

2.1. IRT Equating

Scale transformation in the context of equating is a mathematical procedure to transform IRT parameters of a test onto the scale of parameters of another test under the same test construct. The "same specifications" property is an essential property of equating. The forms to be equated must test the same content and should be built to the same statistical specifications. The test information is represented as parameters in IRT framework. In terms of IRT parameter linear transformations, moment methods (Mean/Sigma and Mean/Mean), characteristic curve methods (Haebara and Stocking–Lord), and concurrent calibration are all commonly used in test development.

2.2. Equating Properties

A number of equating properties have been proposed in the previous literature (Angoff 1984; Holland and Dorans 2006; Lord 1980; Kolen & Brennan 2014). Three equating properties

were suggested by Lord (1980), and five properties of equating were summarized by Kolen and Brennan (2014). The first property is the symmetry property proposed by Lord (1980). This property describes that the inverse transformation of equating from target form score to base form score should equate from the target form to the base form. For example, if Base Score = A* Target Score + B is the equation for equating any score on the target form to a score on the base form scale, then Target Score = (Base Score – B)/A should be the function for equating any score on the base form to the scale of the target form.

The second property is the same specifications property (Kolen & Brennan, 2014). The equated tests should be developed under the same content and statistical specifications. To have an interchangeable test scores this property needs to be obtained.

The third property is the equity property (Lord, 1980). This property implies that the distribution of the equated parameters of target form is identical to that of the base form parameters. Because Lord's strong and strict equity property concept is unlikely to be met in practice, a less strict first-order equity property was introduced by Morris (1982). Strong and weak equating was summarized by Morris (1982): "Two test are strongly equated if every individual in the test population has the same probability distribution for this score on both tests. Two tests are weakly equated if each individual in the test population has the same expected score on both tests" (p.171). According to the first-order equity property, examinees with a given true score have the same mean converted score on base form as they have on target form. Linear methods have been developed that are consistent with this equity property.

The fourth property is the observed score equating property. The characteristics of score distributions are set equal for a specified population of examinees (Angoff, 1984). This is the main idea for number-correct score equating. Equipercentile equating methods ensure that scores

from different forms have the same cumulative distribution of scores. This is also applied to IRT observed score equating procedures.

The final property is the group invariance property. To meet this requirement, the equating relationship between two forms of an assessment should be the same for all groups of examinees (Kolen & Brennan, 2014).

2.3. Designs of IRT Equating

To collect data for IRT equating, the allocation of test forms needs to be predetermined with an appropriate process. Examinee scores on the two tests are typically collected according to one of the three major types of equating designs (von Davier et al., 2004; Kolen & Brennan, 2014): random group design (single group design and with counter balancing), common-item nonequivalent group design, and concurrent calibration (and with fixed parameters).

Most large scale tests are administrated using a random group design: examinees are randomly assigned to forms. There are three kinds of random group designs. The first is the Single Group (SG) design, where a random sample from a common population of examinees completes both base and target forms. The second is an Equivalent Groups (EG) design, where examinees from the same population are randomly assigned to take either a base form or a target form. The third is the non-equivalent group with anchor test (NEAT) design, where random samples (that are assumed to be) from two different examinee populations complete either the base or target form. A SG design is said to be counterbalanced (a CB design) when one examinee subgroup completes the base form first and the remaining examinees complete the target form first. NG designs, and some EG designs, make use of an anchor test consisting of items appearing on both base and target forms. The anchor test is said to be internal when the anchor items contribute to the parameter estimation and number-correct score in base and target forms,

and is said to be external when they do not contribute. In general, NEAT is a commonly used design in many high-stakes testing programs.

2.4. Anchor Test

A test with more items is usually more reliable than a shorter test according to the Spearman–Brown prophecy formula. Therefore, longer anchor tests tend to provide more reliable measures with better accuracy. Having too few common items leads to poor estimation, while a larger number of common items produces less random equating error (Petersen et al., 1983; Wingersky et al., 1987). The relationship between anchor-test length and equating error has been studied by Klein and Kolen (1985), Kaskowitz and De Ayala (2001), and Raju et al. (1983). They found that longer anchor tests provide better accuracy of equating among dissimilar groups. Klein and Kolen (1985) found that when two groups' ability distributions were similar, the number of common items did not have an effect on estimation. However, when the difference between group abilities is large, the quality of estimation is reduced. In terms of anchor-test length, the typical recommendation for the total number of common items is more than 20 items or 20% of the total number of items in a test (Angoff, 1984; Kolen & Brennan, 2014). Raju, Edwards, and Obsberg (1983) and Lord (1980) suggested that five or six good items would provide satisfactory results in IRT concurrent calibration. Wright (1977) recommended 10 to 20 common items as sufficient for most equating situations and suggested that if the items are reliable enough, 10 common items may be enough. McKinley and Reckase (1981) concluded that a five-item anchor might be adequate, but a 15-item anchor was suggested. However, Kaskowitz & De Ayala (2001) found that linking was more accurate when there was less error in the linking item parameter estimates, and Hills et al. (1988) concluded that five randomly chosen

anchor items were insufficient to perform a satisfactory estimation and that ten common items were needed to have satisfactory results when an IRT method was adopted.

2.5. IRT Equating Process

IRT equating encompasses three processes: selection of an equating design, scaling estimated IRT parameters, and equating the number-correct scores. An IRT model is supposed to be selected for IRT parameter estimation because the estimation results can be different by the selection of estimation methods. Due to the conceptual complexity of IRT, the metric of IRT scale is not easy to understand for most users of large scale test. Number-correct scores are the popular metric for educational score users, so IRT equating can be conducted to place the IRT parameter estimates onto a number-correct scale.

2.6. IRT Model selection

In previous performance comparison studies of IRT equating, 2PL (Kim & Cohen, 1998) and 3PL (Hanson & Beguin, 2002; Kim & Cohen, 2007; Kim, 2006; Lee & Ban, 2010) IRT models have been widely utilized to the parameter estimation.

The 2PL model and the 3PL model were proposed by Birnbaum (1968). In the 3PL model (Birnbaum, 1968), the probability (p_{ij}) of person *i* answering item *j* correctly is defined by the ability of the person (θ_i), the discrimination of the item (a_j), the difficulty of the item (b_j), and the "pseudo-guessing" parameter of the item (c_j) as below and the guessing parameter is omitted in 2PL model:

$$p_{ij} = p_{ij}(\theta_i, a_j, b_j, c_j) = c_j + \frac{1-c_j}{1+\exp(-a_j(\theta_i - b_j))} \quad \text{Eq.2.6.1}$$

As Holland (1990) illustrated, the 2PL IRT model is enough to describe the response pattern of examinees when the 3PL model doesn't reveal great model fit improvement over 2PL

IRT model. Also, in Haberman's (2006) study, the 2PL model described real test data as well as the 3PL model. Some researchers found that model fit statistics of the 3PL model are worse than those of the 2PL model. According to De Ayala (2009):

> If a comparison of the 2-PL model with the 3-PL model is not significant, then the additional estimation of the pseudo-guessing parameters (i.e., the increased model complexity) is not necessary to improve model-data fit over and above that obtained with the 2-PL model. (p. 140)

Due to the popularity of the 2PL and 3PL models, both are introduced in data generation and estimation. In simulation part of this study, data were drawn from the 2PL model for estimations of the 2PL model, and from the 3PL model for estimations of the 3PL model.

2.7. IRT parameter transformation methods

Equating is accomplished through scale transformations or through the calibration process. In this paper, several IRT transformation linking methods are compared using the 2PL IRT model as the scaling model. In the NEAT design, the two tests are considered two separate test forms in equating, where IRT parameters of the base form (J) are fixed in the base form and parameters of the target form (I) are equated to the base form. The common items provide the constants A and B in the equating formulas, and each equating method calculates these constants in different ways.

$$\theta_{Ji} = A\theta_{Ii} + B \qquad \text{Eq. 2.7.1}$$

$$a_{Jj} = \frac{a_{Ij}}{A} \qquad \text{Eq. 2.7.2}$$

$$b_{Jj} = Ab_{Ij} + B \qquad \text{Eq. 2.7.3}$$

$$c_{Jj} = c_{Ij} \qquad \text{Eq. 2.7.4}$$

Moment Methods: Mean/ Sigma (M/S) and Mean/Mean (M/M) Methods

The A and B constants can be calculated by the Mean/Sigma (Eq.2.7.5; Marco, 1977) or Mean/Mean (Eq.2.7.6; Loyd & Hoover, 1980) transformation methods. The parameters are used separately to estimate the equating coefficients.

$$\text{A}= \frac{\sigma(b_J)}{\sigma(b_I)}, \qquad \text{B}= \mu(b_J) - \text{A}\mu(b_I) \qquad \text{Eq. 2.7.5}$$

$$\text{A}= \frac{\mu(a_I)}{\mu(a_J)}, \qquad \text{B}= \mu(b_J) - \text{A}\mu(b_I) \qquad \text{Eq. 2.7.6}$$

The M/S method relies on IRT *b*-parameters only, but the M/M method also reflects the characteristics of the IRT *a*-parameters.

Test Characteristic Curve (TCC) Method

A test characteristic curve (TCC) method (Stocking & Lord, 1983) is used for parameter linking along with separate calibration. The TCC method is a simultaneous estimation procedure that takes into account the test information provided.

Two major characteristic curve transformation methods were introduced, by Haebara (1980) and Stocking and Lord (1983). Haebara suggests finding the two constants A and B that minimize the summed square values of the differences between item characteristic curves of two tests over items. In contrast to Haebara approach, the squared differences between sums over item are cumulated by examinees in the Stocking and Lord (S–L) function. Of these two methods, only the S–L method is applied to this study.

$$\text{SLdiff}(\theta_i) = \left[\sum_j p_{ij}(\theta_i, a_{Jj}, b_{Jj}, c_{Jj}) - \sum_j p_{ij}\left(\theta_{Ji}, \frac{a_{Ij}}{A}, Ab_{Ij} + B, c_{Ij}\right)\right]^2 \qquad \text{Eq. 2.7.7}$$

Through the estimation process, the A and B constants that minimize the function below is obtained.

$$\text{SLcrit} = \sum_i SL\ diff(\theta i) \qquad \text{Eq. 2.7.8}$$

This is a computationally intensive, iterative approach, made possible through software such as STUIRT (Kim & Kolen, 2004), the 'plink' package in R (Weeks, 2011), and jMetrik (Meyer, 2013).

Concurrent Calibration

In contrast to the parameter transformation methods of Separate Calibration (SC), Concurrent Calibration (CC) estimates IRT parameters on a single calibration run as if the separate forms were a single test. In the CC procedure, the items that are not taken by one group of subjects (*i.e.*, the non-anchor items) are considered "not reached" or missing data for the other group, and the item parameters for all items on the two test forms are simultaneously estimated within a single test form. After combining response data from the two forms, missing responses are observed among forms except on the common items. The statistics of parameters can be estimated by techniques of missing data handling with negligible bias. ML with EM algorithm is the popular way to estimate the parameter statistics in IRT software like BILOG-MG and MULTILOG. In this method, Expectation and Maximization (EM) steps are iteratively computed until the difference of mean structure and variance/covariance structure satisfies the convergence threshold.

In usual 2PL IRT concurrent calibration, item (*a* and *b)* and person (θ) parameters of the base and target groups are simultaneously estimated and the mean of ability is fixed to zero and standard deviation is set to one. The ability parameters of the group taking the target form are

estimated in relation to those of the base group. Therefore, person parameters of both groups are placed on a common scale, and two separate estimation processes are not necessary.

2.8. Comparisons of scale transformation methods

Generally, TCC methods such as Haebara and S–L produce more accurate and stable results than moment methods such as M/S and M/M (Baker & Al-Karni, 1991; Hanson & Beguin, 2002; Kim & Cohen, 1992; Lee & Ban, 2010; Ogasawara, 2001). Baker and Al-Karni (1991) compared the S–L method and the M/M method in a simulation study and found that the S–L method recovered the true parameters with less error than the M/M method. M/M methods were better than M/S in Ogasawara's (2000) study. Through simulated and real data, Ogasawara concluded that the M/M method produced smaller standard errors of linking coefficients and, therefore, was superior to the M/S method.

Concurrent calibration and the TCC method were compared by many simulation studies. Petersen et al. (1983) concluded that concurrent calibration performed somewhat better than the S-L method. However, Kim and Cohen (1998) drew a somewhat different conclusion, as they found concurrent calibration less accurate because they didn't let groups have different latent mean and variance from others. On the contrary, Hanson and Beguin (2002) found that concurrent calibration is better and produces less error than separate estimation. They pointed out one limitation of the Kim and Cohen (1998) study is that different IRT software programs were used for separate and concurrent estimation, so the differences between separate calibration and concurrent calibration were confounded with the differences between computer programs. In the Hanson and Beguin study, the concurrent calibration procedures produced more accurate results than the characteristic curve methods, and the two moment methods did not perform as well as the characteristic methods or concurrent calibration. However, when response data did not fit the

model well, the S–L method exhibited better results than concurrent calibration (Beguin et al., 2000; Beguin & Hanson, 2001).

A common finding from these studies is that concurrent calibration produces better estimation with less error in simulation studies where response patterns are similar across groups and follow the IRT model. However, when response patterns have poor model fit to the IRT model and person parameter distributions are different across groups, the S–L method provides better results than concurrent calibration.

2.9. IRT ML estimation methods

The two prime IRT estimation methods are Bayesian estimation and ML. Bayesian estimation is mainly performed after ML estimation (Keller, 2000). Two common methods of ML estimation are JMLE and MMLE, which are distinguished by whether or not parameters are simultaneously estimated. Because of issues with convergence and the choice of starting values, JMLE is mainly used only with the Rasch, or 1PL IRT, model. MMLE is a more popular method with other IRT models, and Bayesian estimation adjustment provides better estimation based on prior distribution information. As De Mars (2002) pointed out in a vertical scaling simulation study, if the ability distribution is different across groups (as in vertical scaling), groups need to be taken into account, which is not required for JMLE. Bias increased when grouping was ignored in MMLE, but decreased to a similar level to that of JMLE when group was modeled (De Mars, 2002). In the 2PL IRT model study of Drasgow (1989), MMLE demonstrated better estimation, with smaller mean distance between true and estimated IRT parameters and smaller standard errors than JMLE.

2.10. Concurrent calibration estimation methods

The most popular concurrent calibration estimation method is ML estimation with EM algorithm, which has been applied to concurrent calibration methods in BILOG-MG (Mislevy & Bock, 2000), MULTILOG (Thissen, 1991) and Parscale (Muraki & Bock, 1996).

Miyazaki, Hoshino, Mayekawa and Shigemasu (2008) compared concurrent calibration of ML estimation with EM algorithm and Monte Carlo Expectation Maximization (MCEM) algorithm. MCEM is accomplished through Markov Chain Monte Carlo (MCMC) routines such as the Gibbs and Metropolis-Hastings (MH) samplers. The Akaike information criterion (AIC) provides a relationship to the optimized log-likelihood and model complexity as:

$$AIC = -2*\log L + 2*n \qquad \text{Eq. 2.10.1}$$

n= number of parameters

Like AIC, the Bayesian information criterion (BIC) uses the log-likelihood function value and penalizes more complex models. The penalty of BIC is a function of the sample size, and so is typically more severe than that of AIC. The formula for BIC is:

$$BIC = -2*\log L + n*\log(N) \qquad \text{Eq. 2.10.2}$$

N= number of observations

In simulation study of Miyazaki et al. (2008), concurrent calibration with EM algorithm and MCMC sampling were compared to see the effect of a missing at random (MAR) mechanism. In the MAR condition, MH displayed smaller AIC and BIC values than EM.

Concurrent and separate estimation procedures have been compared, but their accuracy of estimation in simulation studies illustrated disagreement due to different settings of simulations.

Hanson and Béguin (2002) used two different computer programs, BILOG-MG and MULTILOG, for CC and SC. However, they indicated the misuse of BILOG for concurrent calibration in the research of Kim and Cohen (1998): "BILOG cannot estimate the correctly specified model in which separate latent variable distributions are assumed for the groups of examinees taking the two forms" (p.4). To estimate parameters using concurrent calibration with multiple groups correctly, the latent variable mean and variance of the base group need to be fixed to 0 and 1, respectively. Also, the target group's mean and variance need to be freely estimated between different groups.

2.11. Missing Mechanism in IRT

Generally, MCAR, MAR, and MNAR are three broadly used categories of missing data mechanisms. When data are MCAR, missingness is not related to specific information, so the missingness is not predicted by the information of variables. A multivariate t-test approach to detect MCAR was proposed by Little (1988). It compares mean differences across all complete variables simultaneously and concludes whether or not the missingness is MCAR, but it cannot determine which variable predicts the missingness (if missing data are not MCAR). An MCAR mechanism also can be detected by path analysis. The significance of chi-square difference between the full model (including paths between predicting variables and missingness) and the restricted model (not including paths between predicting variables and missingness) is tested in the analysis. With an MAR mechanism, the variables that predict missingness are unrelated to the missing values after controlling for the relation between missingness and measured variables. However, with an MNAR mechanism, the predicting variables are still related to the missing values even after controlling for the relation between missingness and measured variables. In a simulation study of the IRT multiple choice model, (Wolkowitz & Skorupski, 2013), missing

mechanisms and the mean bias of item parameters were examined. In contrast to MNAR, MAR and MCAR provided relatively better results, with smaller biases in item difficulties and item-total correlations.In a single estimation, Concurrent Calibration (CC) estimates IRT parameters. Two (or more) forms of a test are regarded as one form with data missing completely at random.

The length of the anchor test determines the proportion of missing data in the concurrent calibration. According to the relationship between missing data and group membership, the missing mechanism is defined. In missing mechanisms of Missing Completely At Random (MCAR) and Missing At Random (MAR), the subject allocation to test forms is considered a random assignment.

2.12. Multiple Imputation in IRT

In the field of educational measurement, if a response is missing it is usually considered a wrong answer. Even though response data with missing data can cause biased estimates of the ability parameters, they could not have been adjusted or imputed, especially in high-stakes tests. If not reached items are imputed, to get a higher score, students may only try items which are sure to answer correctly. De Ayala, Plake, and Impara (2001) found worse estimation when omits were treated as incorrect. For better estimation, not reached items need to be handled with care from the missing data perspective.

ML with EM algorithm is one of the widely used missing data handling techniques. Another popular missing data handling technique of estimating parameters is MI (Schafer & Graham, 2002), where missing data are randomly drawn based on a model using Markov Chain Monte Carlo methods. This technique can also be applied to a concurrent calibration. Even though ML is known to provide more accurate parameter estimation than MI in many social science studies (Graham, 2009; Enders, 2010; Baraldi & Enders, 2010), MI has not yet been employed with IRT

concurrent calibration. Therefore, Concurrent Calibration with MI (CCMI) is an uncovered area of research, and a comparison study with popular methods (MM, MS, S-L, and CC with ML) has never been conducted.

MI and Full Information Maximum Likelihood (FIML) are commonly used methods to deal with the missing data for estimating the mean and variance of the virtually complete data. The efficiency and accuracy of each method have been compared by many researchers. In several social science studies, FIML provided slightly better results than MI (Acock, 2005; Enders, 2012; Yuan et al., 2012) Several imputation methods for the IRT model were applied to missing data under MAR and MNAR mechanisms with IRT estimation by Finch (2008). In the MAR mechanism condition, MI provided the most accurate estimates of item discrimination and item difficulty, and standard error and bias of item parameters estimated using MI were slightly smaller than those estimated using the EM method in all conditions. According to Finch, "in all cases, MI had the closest percent correct to that for the complete data set while EM was the furthest away from the complete data case (and always higher)" (p. 241).

2.13. Planned missing data design

A planned missing data design (Graham et al., 2006) is utilized when the missing mechanism is MCAR due to its efficiency. Concurrent calibration equating under the NEAT design has exactly the same design as planned missing. This relationship was briefly mentioned by Sinharay & Holland (2010): "The Non-Equivalent group with Anchor Test (NEAT) design involves missing data that are missing by design" (p. 309). From the perspective of missing data handling, concurrent IRT calibration has commonly used ML with EM algorithm (MULTILOG (Thissen, 1991), Parscale (Muraki & Bock, 1996), and BILOG-MG (Mislevy & Bock, 2000). In

contrast, MI generates data sets with varying estimates of missing values in an imputation step and posterior step. The process is based on Bayesian estimation.

In a planned missing design for IRT (Figure 1), which is equivalent to the IRT concurrent calibration, test items are allocated to two unique test forms (UB and UT) and a set of common items in two forms (CB and CT). Unique test items and a set of common items constitute a test form. Two different test forms are considered a planned missing data design for each group. Each counter unique form is considered as "missing" in the other form.

Form	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Base	0	0	0	0	0	0	0	0	1	1	-	-	-	-	-
Base	0	0	0	0	0	0	0	0	1	1	-	-	-	-	-
Base	1	0	0	0	1	1	1	1	0	1	-	-	-	-	-
Base	1	0	1	0	1	1	1	1	1	1	-	-	-	-	-
Base	0	0	0	0	1	0	1	1	0	1	-	-	-	-	-
Base	1	0	1	1	1	1	1	1	1	1	-	-	-	-	-
Base	0	0	0	0	1	0	0	1	1	1	-	-	-	-	-
Base	0	0	0	0	1	0	0	0	1	1	-	-	-	-	-
Target	-	-	-	-	-	0	0	1	0	1	0	1	0	1	1
Target	-	-	-	-	-	1	1	0	0	1	0	1	0	1	1
Target	-	-	-	-	-	1	0	0	0	1	0	1	0	0	1
Target	-	-	-	-	-	1	0	1	0	1	0	1	1	1	1
Target	-	-	-	-	-	0	1	0	1	1	1	1	1	0	1
Target	-	-	-	-	-	0	1	1	1	1	1	1	1	1	1
Target	-	-	-	-	-	0	1	1	1	1	0	1	0	0	1
Target	-	-	-	-	-	0	0	1	0	1	0	1	0	1	1

UB CB CT UT

Figure 1 Planned missing design for IRT equating

UB= Unique items of Base form, UT= Unique items of Target form,

CB= Common items of Base form, CT= Common items of Target form

2.14. Details of Multiple Imputation

Many missing data theorists have argued that MI and ML are equivalent in theory but not in practice. The comparison studies have been conducted in Ordinary Least-Squares Regression models, but not many in IRT models (Ender, 2010). For concurrent calibration, ML estimates can often be calculated directly from the incomplete data by specialized numerical methods, such as the EM algorithm. In some cases, MI provides better estimates than ML with EM algorithm. The EM algorithm is better at recovering the mean and variance–covariance matrix of observed scores, but MI provides better logistic parameter estimation.

Harwell, Stone, Hsu and Kirisci (1996) stated the usefulness of Monte Carlo (MC) methods in the IRT framework, and MC was better than Bayesian analysis and applied parametric modeling. For the purpose of MI, Schafer (1999) has adapted and implemented Markov Chain Monte Carlo (MCMC) methods. MI creates the probability distribution of each student's possible responses. For each student, MI draws plausible values from the probability distribution. The MCMC method of data augmentation is used to obtain this posterior probability distribution from which the imputed values for the missing observations are drawn. MI accounts for the inherent uncertainty in sampling from a population by introducing randomness to the imputations and creating *m* imputed data sets, each of which is then fitted to a 2PL IRT model. MI incorporates information from other variables into the imputation process in order to provide more accurate values. Parameter estimates are made using the Bayesian posterior distribution based upon the likelihood function of the proposed model (in this study, the 2PL IRT model), the observed response data, and the estimated IRT parameters.

Plausible values are imputed to the missing data until an adequate number of data sets are generated in the imputation stage, then parameter estimates from each data set are aggregated for

the final estimation. Generally, P(θ) represents the 2PL IRT model, and x_{ij} is the response, 1 or 0. The probability of all item data is written as

$$P = P(x_{ij} = 1 | \theta_i, a_j, b_j) \quad \text{Eq.2.13.2}$$

P is the probability of person *i* correctly answering item *j*, at the given ability parameter. If the person parameter and items parameters are given, missing item responses are considered Bernoulli trials under the IRT local independence assumption.

$$x_{ij} \sim Bernoulli(P) \quad \text{Eq.2.13.3}$$

If the probability of answering an item correctly is represented as

$$P = P(\theta)^{x_{ij}} (1 - P(\theta))^{1-x_{ij}}, \quad \text{Eq.2.13.4}$$

then the likelihood of all item data is represented as its product terms

$$p(X | \theta) = \prod_{i=1}^{N} \prod_{j=1}^{n} P(\theta)^{x_{ij}} (1 - P(\theta))^{1-x_{ij}} \quad \text{Eq.2.13.5}$$

In the imputation stage, under the given parameters θ_i, a_j, b_j, missing responses are randomly drawn from a Bernoulli distribution.

The theta parameter can be expressed as the equation

$$\overline{\theta_\iota} = \frac{1}{m} \sum_{t=1}^{m} \hat{\theta}_{it} \quad \text{Eq. 2.13.6}$$

The imputed responses are input, and $\hat{\theta}_{it}$ is estimated in IRT models at each time *t*. In the pooling stage, the MI point estimate is the arithmetic average of the *m* parameter estimates,

which are directly estimated in ML. $\hat{\theta}_{it}$ is the estimated person(i) parameter in t^{th} imputation, and m is the number of imputed datasets.

According to Rubin's (1996) procedure, error variance is estimated by a combination of the between-imputation and within-imputation variances (arithmetic average of squared S.E.), and its square root value becomes the MI standard error of the estimates.

The within-imputation variance is the arithmetic average of squared standard errors. $s(\theta)_t^2$ is the sampling variance from imputation t, and s_m^2 is the estimated value of the sampling error that would result had the data been complete.

$$s(\theta)_m^2 = \frac{1}{m}\sum_{t=1}^{m} s(\theta)_t^2 \qquad \text{Eq.2.13.7}$$

V_B is the estimate of the additional sampling error due to the missing data, and the sample variance formula is used to estimate between-imputation variance as below. The m parameter estimates serve as data points here.

$$V(\theta)_B = \frac{1}{m-1}\sum_{t=1}^{m}\left(\hat{\theta_t} - \bar{\theta}_m\right)^2 \qquad \text{Eq.2.13.8}$$

The total sampling variance (V_T) is a combination of the between-imputation (s_m^2) and within-imputation variances (V_B) as below:

$$V(\theta)_T = s(\theta)_m^2 + (1+\frac{1}{m})V(\theta)_B \qquad \text{Eq. 2.13.9}$$

The MI standard error is the square root of V_T,

$$\text{S.E. } (\theta) = \sqrt{V(\theta)_T} \qquad \text{Eq. 2.13.10}$$

The same procedure is applied to parameters *a* and *b* (Enders, 2010). Compared with other estimation methods, MI should be rich enough to preserve the associations or relationships in IRT models.

2.15. Evaluation of Equating Results

Of the two basic equating criteria (Harris & Crouse, 1993), Angoff's observed score definition, known as the equipercentile definition is more broadly applied in real equating than the equating criteria of Lord that uses the true score definition. In general simulation study generating true data, Lord's equity definition that test X and test Y are to be deemed equitable are commonly adequate (Harris & Crouse, 1993). The purpose of MC simulation study is to provide evidence of consistent results from a random sampling procedure. To evaluate the efficiency of equating, the true parameters are commonly compared with the equated parameters of transformation methods. The difference is represented by many statistics: MSE, root MSE (RMSE), root mean squared distance (RMSD), signed bias, and unsigned (absolute) bias. Even though these are slightly different from one another, the magnitude of MSE and the pattern of bias across number-correct scores point out distinct features of each linking method. MSE is a representative value of equating efficiency, but bias is not a good measure for efficiency when it is summed because the differences from true parameters may offset across examinees. However, previous linking method comparison studies denote consistent patterns of bias across number-correct scores.

2.16. Benefits of concurrent calibration by MI

MI consists of a three-phase analysis: imputation, analysis, and pooling. The imputation process is repeated several times and several datasets are generated. The inference of IRT parameter estimates comes after assembling the analysis results from multiple data sets. MI has

not been applied to CC in spite of its benefits, which include providing plausible data to replace missing responses and making item statistics more robust. From the imputed data set, the mean of imputed score can be obtained and this score is comparable to the true score equating and observed score equating score. This new proposed equated score, the procedure of mean imputed score (MIS) is described in Chapter 3, Section 9.

Chapter Three - Methods

To see how accurately the linking methods provide the information for equating, simulated data were used. Simulation studies provide known true parameters and known true number-correct scores. Therefore, simulated data are used in this study because the true equating relationship is known. Also, Monte Carlo (MC) studies are the simulation technique most frequently used in IRT studies to demonstrate whether linking methods can be applied validly to realistic datasets. The term "Monte Carlo" represents stochastic simulation which generates random numbers (Naylor, 1968). To be more specific, it is a statistical and experimental sampling technique using a computer program for studies with an assumed model. MC techniques are popularly used for many other IRT conditions for a variety of purposes, including evaluating estimation procedures or parameter recovery, evaluating statistical properties, and comparing methodologies (Harwell et al., 1996). Those studies require drawing random samples using an underlying model under the assumption that the true values are given, which is hard to know in real practice.

3.1. Usage of simulated data

In equating studies, simulated/generated data are frequently used. The question is how accurately the true equating relationship (definitive criterion) was recovered across equating methods. For instance, both real and simulated data have been utilized to compare IRT equating methods (Baker & Al-Karni, 1991; Stocking, Eignor & Cook, 1988; Way & Tang, 1991; Kim & Cohen, 1998, 2002; Ogasawara, 2000, 2001; Hanson & Beguin, 2002). However, simulation is not a flawless research method. Real data to which the study results is applied may not closely resemble simulated data from an assumed IRT model. Harris and Crouse (1993) mentioned that

when the real data are close to simulated data, the results were similar to the findings of simulation studies.

3.2. Data simulation conditions

Simulated item response data from various population distributions were utilized to compare the accuracy of estimated parameters across equating methods. Twelve different simulation studies were estimated and equated to see how the results varied across conditions. The number of unique items on each form was 40, and two different anchor test lengths were considered. Two different ability parameter means (0 and 0.5) were used for the target group prior distribution and fully crossed with three ability parameter SDs (0.5, 1.0, and 1.5) for various target prior distributions. The base group ability was set to standard normal in all conditions. The six different true person ability distributions were crossed with two anchor-test lengths (10 or 20). Therefore, a total of 12 different conditions were tested in the simulation study. These conditions are considered as missing at random (MAR) of missing mechanism because 12 conditions assume that theta distribution is related to form assignment. This missing mechanism assumption is aligned to the simulation condition and assumption of Miyazaki et.al(2009).

The number of unique items of each form was fixed to 40. To generate true θ parameters and true number-correct scores, subject ability parameters were drawn from each condition, and 90 or 100 item responses were simulated depending on anchor length (10 or 20) with the counter form's responses omitted. In concurrent calibration, these omitted responses are treated as missing data.

In each of the 12 conditions, 100 simulated data sets were generated, and IRT parameters of each missing data set were estimated. The R package 'mirt' (Chalmers, 2012) was utilized for

IRT parameter estimation and generating imputed data sets for the 2PL & 3PL IRT models. According to Graham et al. (2007), when the fraction of missing data is high, a sufficient number of imputed data sets are required to obtain better estimation. The fraction of missing data is 44.4% for a test with 10 anchor items and 40% for a test with 20 anchor items. 100 imputed data sets were required for the estimation of one simulated data set. For 100 simulated data sets per one condition, 10,000 imputed data sets were generated. In total, 120,000 imputed data sets were generated across all 12 conditions.

Table 1 Conditions of simulated person parameters

Groups	Number of Subjects	(θ) distribution(s)
Base Group (θ1)	500	N(0, 1)
Target Group (θ2)	500	N(0, 1), N(0, 0.5), N(0, 1.5), N(0.5, 0.5), N(0.5, 1), N(0.5, 1.1)

Table 2 Conditions of simulated item parameters

Item parameters	Number of Items	Item parameter distribution(s)
a (2PL& 3PL)	100 (90)	LogN(0.5, 0.25)
b (2PL& 3PL)	100 (90)	N(0, 1)
c (3PL)	100 (90)	beta(5, 17)

Table 3 Descriptive statistics of 12 simulated exams

Study	$\mu(\theta_B)$	$\mu(\theta_T)$	$\sigma^2(\theta_B)$	$\sigma^2(\theta_T)$	Anchor items	Total items
EIS	0	0	1	1	10	50
EIL	0	0	1	1	20	60
ENS	0	0	1	0.5	10	50
ENL	0	0	1	0.5	20	60
EWS	0	0	1	1.5	10	50
EWL	0	0	1	1.5	20	60
DIS	0	0.5	1	1	10	50
DIL	0	0.5	1	1	20	60
DNS	0	0.5	1	0.5	10	50
DNL	0	0.5	1	0.5	20	60
DWS	0	0.5	1	1.5	10	50
DWL	0	0.5	1	1.5	20	60

*B = base form; T = target form;

E = equal mean (μ_B=0, μ_T=0); D = different mean (μ_B=0, μ_T=0.5);

W = wider variance (σ^2_B=1.0, σ^2_T=1.5), I = identical variance (σ^2_B=1.0, σ^2_T=1.0); N = narrower variance (σ^2_B=1.0, σ^2_T= 0.5);

S = short anchor length (10), L = long anchor length (20)

3.3 Using PISA 2000 Reading data

The Program for International Student Assessment (PISA) is an international testing program to evaluate education systems worldwide by testing the skills and knowledge of

students in the key subjects of reading, mathematics, and science. The PISA test is not directly related to the school curriculum but designed to assess to what extent students can apply their knowledge to real-life situations. Since 2000, fifteen-year-old students from randomly selected schools worldwide take PISA every three years. For this study, PISA data from the U.S. in 2000 were analyzed. For comparable selection to the NEAT design, Reading tests of Book ID 1 and 5 were chosen, which consist of 63 dichotomous multiple choice items. Book ID 1 was chosen as the Base form, with 24 unique items and 16 common items. Book ID 5 was chosen as the target form, and it has 23 unique items and 16 common items. A total of 434 students took the base form and 416 took the target form. Some items were defined as unreliable items and deleted so the total number of items does not match exactly across forms.

3.4. IRT equating procedures

M/S Equating constants for the IRT parameters are expressed by the equations below.

$$A = \frac{\sigma(b_{Target})}{\sigma(b_{Base})}, \quad \text{Eq.3.5.1}$$

$$= \frac{\mu(b_{Base})}{\mu(b_{Target})}, \quad \text{Eq.3.5.2}$$

$$= \frac{\sigma(\theta_{Target})}{\sigma(\theta_{Base})} \quad \text{Eq.3.5.3}$$

$$B = \mu(b_{Target}) - A\mu(b_{Base}), \text{ and} \quad \text{Eq.3.5.4}$$

$$= \mu(\theta_{Target}) - A\mu(\theta_{Base}) \quad \text{Eq.3.5.5}$$

The R package ‘plink’ (Weeks, 2010) provided the linking constants A and B for the M/S, M/M, and S-L methods. For concurrent calibration, the mean and variance of the latent trait were freely estimated by groups. The latent mean and variance of the base group were constrained to 0 and 1.0 by default in multiple group estimation.

3.5. Concurrent Calibration in 'mirt'

The different groups share only the IRT common item parameters, not the mean or SD of ability parameters. Therefore, latent means and variances are freely estimated, but the item parameters are shared across groups. Therefore, the slopes and intercepts are fixed over groups. In the 'mirt' package (Chalmers, 2012), the "invariance" option allows this. This is the difference between concurrent and separate calibration. In separate calibration, item parameters are separately estimated.

3.6. Multiple Imputation (MI) in 'mirt'

Missing response data were imputed by the 'mirt' package, and error of the estimated person ability parameters (Θ) was compared across all equating methods (M/S, M/M, S-L, CC and CCMI) by MSE and bias. IRT parameters were generated from the distributions shown in Table 1 and Table 2. In the IRT parameter estimation, the mean of latent variable of target group was freely estimated while it is fixed as '0' in the base group by 'multipleGroup' function of 'mirt' package.

3.7. IRT True Score Equating (TSE)

Commonly, a scale linking method is applied to report test results on a metric other than the θ-scale. After scaling item parameters of two or more forms, IRT true score equating (TSE; Kolen, 1981) can be performed to provide the number-correct score on the target form corresponding to each number-correct score on the base form, $B(\theta_i)$. In the TSE, the person parameters are assumed to be equivalent across forms.

$$\text{Score\_B}(\theta_i) = \sum_{j=1}^{nB} p_{ij}(\theta_i, a_j, b_j, c_j) \qquad \text{Eq. 3.6.1}$$

$$\text{Score\_T}(\theta_i) = \sum_{j=1}^{nT} p_{ij}(\theta_i, a_j, b_j, c_j) \qquad \text{Eq. 3.6.2}$$

The total number of items, *n*, is not required to be equal; however, in this study, the total number of items is the same across forms. The process of true score equating is as below:

1. Choose a possible number-correct score, **Score_B**, from the Base form.
2. Define the corresponding **Score_B** person parameter value, θ_i, from the test characteristic curve (TCC) of the base form.
3. Find the number-correct score, equivalent **Score_T**, that corresponds with the same person parameter, θ_i, from the TCC of the target form.

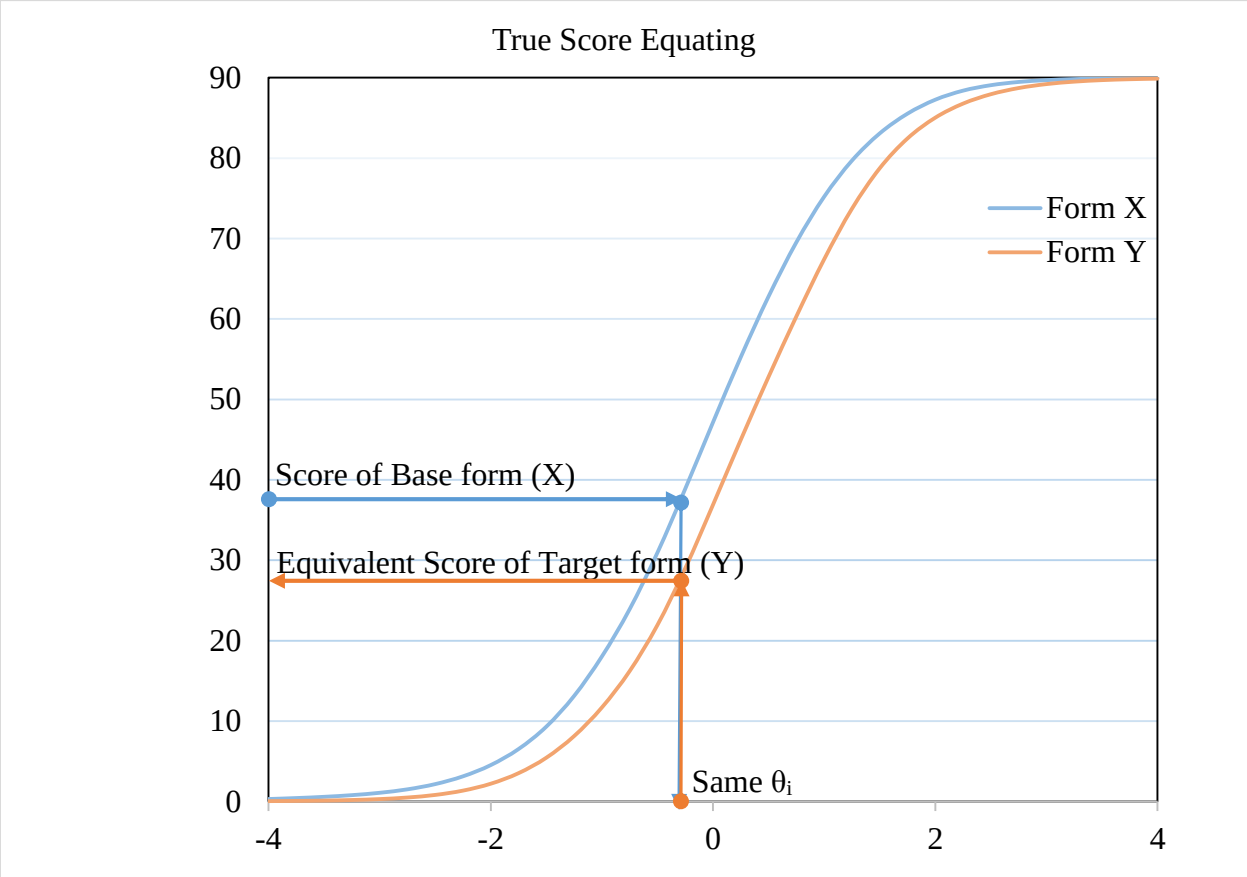

Figure 2 True score equating procedure in TCC plot

3.8. IRT Observed Score Equating (OSE)

Another IRT score equating method is Observed Score Equating (OSE). After scaling the IRT parameters, the IRT model provides estimated distributions of observed number-correct scores by each form. The observed score distribution of the target form is equated to the distribution of the base form by the conventional equipercentile method is represented as

$$f(x) = \int_\theta f(x|\theta)\varphi(\theta)d\theta \qquad \text{Eq.3.7.1}$$

and discrete ability distribution $g(x)$ is represented as

$$g(x) = \sum_i g(x|\theta i)\varphi(\theta i), \qquad \text{Eq.3.7.2}$$

where φ(θ) is the distribution of θ.

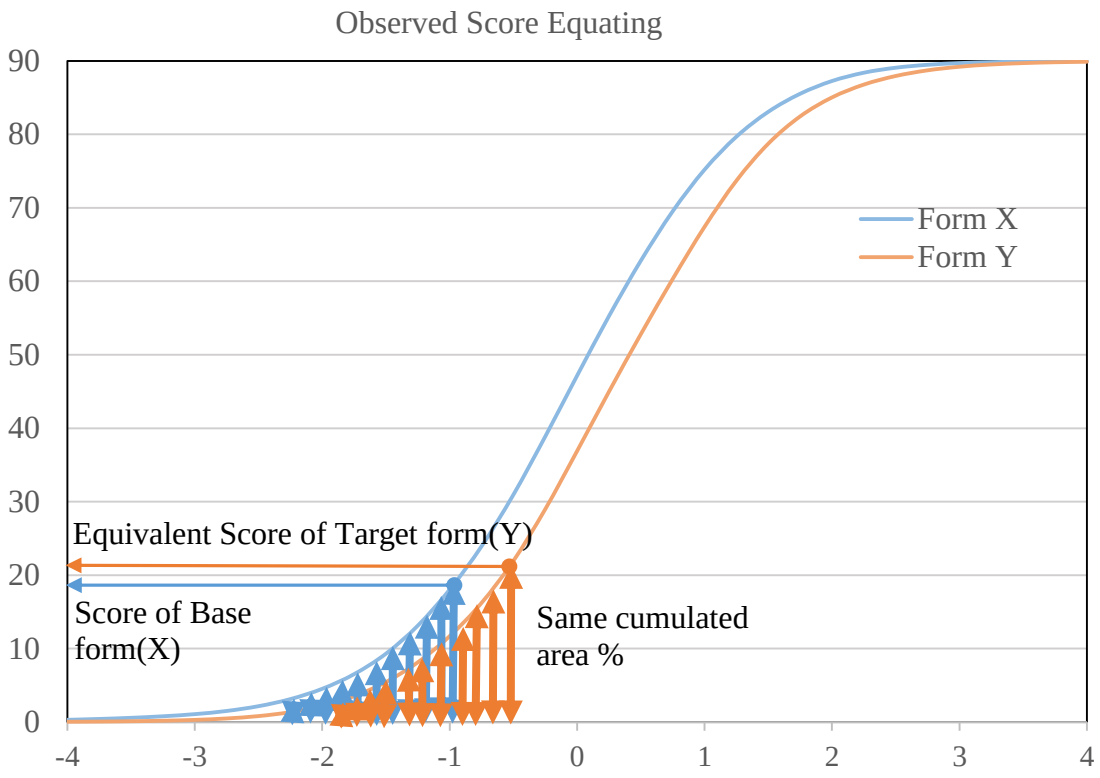

Figure 3 Observed score equating procedure in TCC plot

The observed score distributions of each form are represented as cumulative ability distributions to obtain a percentile ranking. In this research, the previous discrete ability distribution is utilized for synthetic weighting. Observed score equating relies on equipercentile equating, which is sensitive to the observed score distributions of each form. The IRT weighing procedure provides an appropriate quadrature point weight over equipercentile for calculation of a discrete ability distribution in many software programs. To obtain a synthetic population by forms and examinee groups, a synthetic weighting procedure of conventional equipercentile equating is required and provides a score table with equivalent number-correct scores across forms.

3.9. Procedures for assessing criteria for person parameters

Skaggs (1990) used an example from Livingston, Dorans, and Wright (1990) to point out the importance of selecting evaluation criteria for equating:

> These two studies do not contradict each other; they simply used different methods of evaluating the results. That the selection of a criterion can have such an impact on interpreting equating results has devastating implications for the vast majority of equating studies. (p.108)

For the evaluation of IRT parameter equating, Angoff's summary indices (Harris & Crouse, 1993) are the commonly applied equating criteria. MSE and bias of the ability parameters appear repeatedly in the equating literature. In this study, the equating methods are evaluated by these indices:

$$\text{MSE} = \frac{\sum_r \sum_j (\theta_j^{true} - \theta_{jr}^{obs})^2}{N * R} \quad \text{and} \quad \text{Eq.3.8.1}$$

$$\text{Bias} = \frac{\Sigma_r \Sigma_j (\theta_j^{true} - \theta_{jr}^{obs})}{N * R} \quad \text{Eq.3.8.2}$$

where N is the number of subjects in each condition and R is the number of replications.

3.10. Applying Lord's generated data criterion to OSE and TSE

After the IRT parameter scaling by equating methods, true score equating and observed score equating were also performed. Even though the direct comparison is impossible, previous research of OSE and TSE focused on the score difference between the equivalent target form score on the base form and original base form score. This difference guides how similar the score distributions are between forms. However, using simulated data provides an absolute criterion that explains the true equating relationship between different test forms. This absolute criterion needs to be applied to evaluate the efficiency of the new statistic from CCMI. As with evaluation of IRT parameter recovery, the counter equivalent score of the target form on the base form is compared with the simulated true target form score. Instead of the score difference between the base form and equated target form, the true observed score (which is drawn by MC simulation) is compared with the equated score. This difference represents how accurately each equating method recovers the portion of the base form score in the target group.

nU is the total number of unique items on the base and target forms. In this study, the two forms have the same number of unique items. *nC* is the total number of common items on the base and target forms.

$$\text{UB}_\text{i} = \sum_{j=1}^{nU} UB_{ij} \quad \text{Eq. 3.9.1}$$

$$\text{UT}_\text{i} = \sum_{j=1}^{nU} UT_{ij} \quad \text{Eq. 3.9.2}$$

$$\mathrm{CB_i} = \sum_{j=1}^{nC} CB_{ij} \qquad \text{Eq. 3.9.3}$$

$$\mathrm{CT_i} = \sum_{j=1}^{nC} CT_{ij} \qquad \text{Eq. 3.9.4}$$

The sum of UB_i and CB_i is the base form score BS_i of examinee i. The sum of UT_i and CT_i is the target form score TS_i.

$$\mathrm{BS_i} = \mathrm{UB_i} + \mathrm{CB_i} \qquad \text{Eq. 3.9.5}$$

$$\mathrm{TS_i} = \mathrm{UT_i} + \mathrm{CT_i} \qquad \text{Eq. 3.9.6}$$

Form	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Base	0	0	0	0	0	0	0	0	1	1	imp	imp	imp	imp	imp
Base	0	0	0	0	0	0	0	0	1	1	imp	imp	imp	imp	imp
Base	1	0	0	0	1	1	1	1	0	1	imp	imp	imp	imp	imp
Base	1	0	1	0	1	1	1	1	1	1	imp	imp	imp	imp	imp
Base	0	0	0	0	1	0	1	1	0	1	imp	imp	imp	imp	imp
Base	1	0	1	1	1	1	1	1	1	1	imp	imp	imp	imp	imp
Base	0	0	0	0	1	0	0	1	1	1	imp	imp	imp	imp	imp
Base	0	0	0	0	1	0	0	0	1	1	imp	imp	imp	imp	imp
Target	-	-	-	-	-	0	0	1	0	1	0	1	0	1	1
Target	-	-	-	-	-	1	1	0	0	1	0	1	0	1	1
Target	-	-	-	-	-	1	0	0	0	1	0	1	0	0	1
Target	-	-	-	-	-	1	0	1	0	1	0	1	1	1	1
Target	-	-	-	-	-	0	1	0	1	1	1	1	1	0	1
Target	-	-	-	-	-	0	1	1	1	1	1	1	1	1	1
Target	-	-	-	-	-	0	1	1	1	1	0	1	0	0	1
Target	-	-	-	-	-	0	0	1	0	1	0	1	0	1	1

UB CB IB CT UT

Figure 4 Unique items, common items and imputed items in the base form and target form

3.11. Introduction of Mean Imputed Score (MIS)

Through MI, the counterpart of each form is imputed *P* times, where *P* is the number of imputed data sets. Item responses from the base form portion are imputed for each base form examinee *i* whose score is represented as *IBi*.

$$\mathrm{IB_i} = \sum_{j=1}^{nU} IB_{ij} \qquad \text{Eq. 3.9.7}$$

Imputed equivalent test scores of the base group score on target form portion at the p^{th} imputation can be represented as

$$\mathrm{IEB_{ip}} = \mathrm{CB_i} + \mathrm{IB_{ip}} \qquad \text{Eq. 3.9.8}$$

Given θ, when this imputation is repeated *P* times, the mean imputed base form score (MIS) is represented as below.

$$\mathrm{MIS\_i} = \frac{\sum_{p=1}^{P} IEB_{ip}}{P} \qquad \text{Eq. 3.9.9}$$

The base form score (BS) and MIS can be matched to provide a score comparison or conversion table, which can be considered equivalent target form scores for base form scores. This score can be compared to TSE and OSE results.

Before creating the missing data, the original data set of each replication is saved, and the observed score IT_i can be regarded as the true observed score. Each examinee's original responses on the base form in the target form portion, which are deleted for the planned missing design, is OBT_{ir} (area equivalent to 'IB' in Figure 4). At each replication *r*, the true observed score on the base form in the target form portion can be represented as the sum of common item

score (CB_i) and original base form score of target form portion (OBT_i). When this score is averaged around base form score, then base form score *b* has an equivalent score of target form, EB_b.

$$EB_b = \frac{\sum (CBi + OBTi)_{B=b}}{\sum_{i=1}^{nB} i_{B=b}} \qquad \text{Eq. 3.9.10}$$

The true relationship between base form score and target form score can be obtained by the table B vs. EB. MIBS is comparable to the true values EB and OSE. TSE score is compared in terms of recovery accuracy (MSE and bias). Target equivalent minus base form score across raw scores on the base form are also compared to see how consistent the results are with previous studies of OSE and TSE.

Chapter Four - Results

This chapter presents the results obtained from this study. For completeness and convenience, the results are presented in both numerical and graphical forms. The chapter begins with a brief review of the research purpose and questions, followed by a review of the evaluation criteria used to evaluate equating results. A general framework for presenting the results is described next. After that, the results are presented. The chapter concludes with a brief summary of the main findings.

4.1. Review of research purpose and questions

This study was designed to compare the performance of a newly introduced equating method with commonly used equating methods in the NEAT design using both simulated and real data. The real data were from two forms of U.S. PISA 2000 reading tests. The methods under investigation are two moment methods (M/S, M/M), a TCC method (S/L) in the IRT true score equating (TSE), and the IRT observed score equating (OSE). In particular, the study aimed to address several research questions on how those methods perform across all study conditions and how differences between tests being equated and differences between groups taking the test forms affect the equating results. After the IRT parameter transformation, IRT observed score equating and IRT true score equating were performed.

4.2. Review of evaluation indices

The most advantageous aspect of simulation study is that it provides true values of ability as well as true observed scores. True ability has been popularly introduced in IRT framework so far but the true observed score has not been utilized much. In real data equating of TSE and OSE, results provide score conversion tables across multiple forms. However, simulated data can provide true scores, which can be used to evaluate the accuracy of OSE and TSE. Therefore, in

the simulation study, those results were compared across transformation methods. In particular, mean imputed scores from concurrent calibration with multiple imputation were examined to see whether or not they provide an accurate conversion table over concurrent calibration with EM.

Especially for the IRT TSE and OSE, the difference between the equated number-correct score of target form usually decimal points, while the number-correct score of base form is an integer. The patterns of this difference across number-correct scores don't explain the relationship between true score and equating results. The difference explains how the scoring patterns of the two forms are similar, but not the true relationship. We are interested in how accurately the equating works comparing the true observed number-correct score. The ability distributions of two groups were designed to vary in both mean and variance, so the patterns of score difference are also expected to be dissimilar.

Based on those reasons, two evaluation criteria were developed and used in this study. First, the relationship between true number correct scores of target form and base form were examined. From the simulation, the data were created and intentionally erased to generate planned missing data design. The true score difference through IRT OSE and TSE result were obtained. After this step, the differences between equated scores of target form and base form were obtained from IRT OSE and TSE results. These values were then compared to the true relationship between forms and the value from the first step.

4.3. General framework for presenting the results

Simulation results are presented graphically, and the patterns of MSE and BIAS are mainly graphically presented. The numbers are summarized in **Error! Reference source not found. ~Error! Reference source not found.** in Appendix A, separately. The horizontal axis

represents ability and number-correct score, and the vertical axis represents MSE and BIAS value by transformation methods. Loess lines are also added to see the difference across transformation methods. A major focus of this study is on the performance comparison of two estimation methods (EM, MI) of concurrent calibration, which are compared in a separate bar chart.

4.4. Results of 2PL model

In the planned missing data design (Graham et al. 2006), the missing mechanism is regarded as Missing Completely At Random (MCAR) and its ML estimation is comparable to the estimation of MI. CCMI provides MSE and BIAS values as low as CC because the NEAT condition satisfies this. From Study 1 to 6, the mean of target group ability distribution is set to 0. The standard deviation and length of anchor test were different from each study. The descriptive statistics of each simulation study are summarized in Table 3. To evaluate θ estimation, MSE values were compared among distinct conditions.

4.4.1. Comparison of two anchor lengths (2PL). The pattern of MSE values for all conditions of study was generally similar: better θ estimation was observed with 20 anchor items than 10. Average MSE of anchor length 20, was 15% (6.72% to 22.36%) smaller than average MSE of 10 anchor items. Figure 5 and Figure 6 illustrates this pattern clearly. The x-axis represents MSE. The improvement was large when the standard deviation of the target population was the same as the base group.

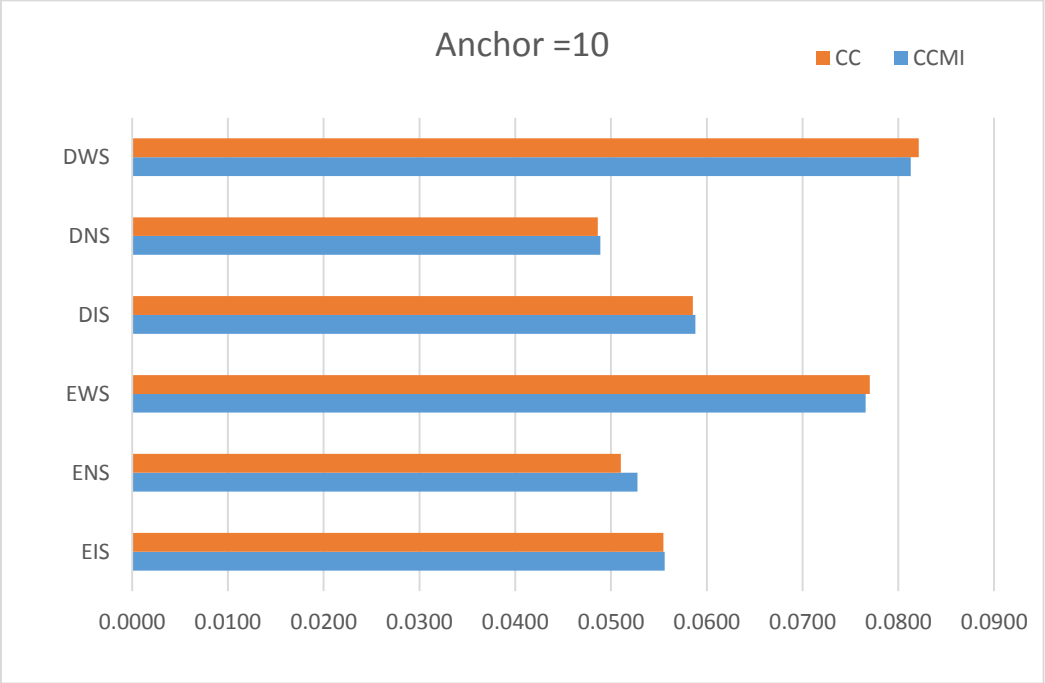

Figure 5 EM & MI comparison in equal anchor length (2PL, anchor =10)

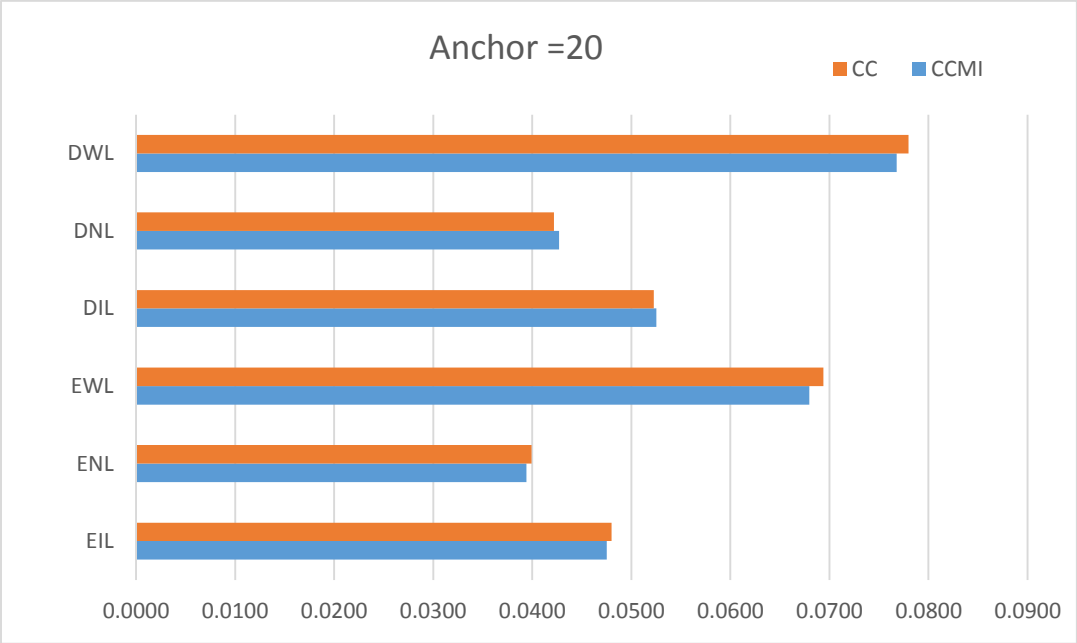

Figure 6 EM & MI comparison in anchor equal length (2PL, anchor =20)

4.4.2. Comparison of three different standard deviations of ability distribution

(2PL) Within the same anchor length and the same mean ability distribution, θ estimation was best in terms of MSE when standard deviation of target ability distribution is 1, followed by SD of 0.5 and 1.5. The EIS, EIL, DIS, and DIL conditions had target distributions with SD 1, and CCMI, CC, and S-L methods displayed moderately smaller average MSE values (0.0536) than other methods. The ENS, ENL, DNS and DNL conditions had 0.5 SD of target person parameter distribution. Among these studies, CC displayed the smallest average value of MSE (0.0454). When the target person parameter prior had wider distribution than the base group, S-L provided the best recovery, with the smallest average MSE (0.0748). When the variance of target group θ was narrower, in Figure 8, CC revealed better performance, with smaller MSE than CCMI. However, CCMI provided better accuracy in θ parameter recovery when the variance of target group θ was wider than base group, as Figure 9 indicates.

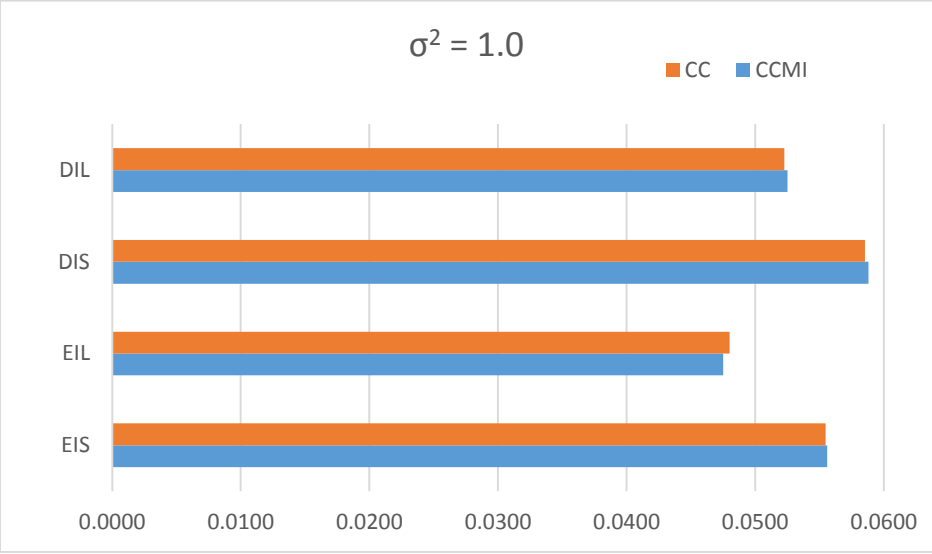

Figure 7 EM & MI comparison in equal target group theta σ (2PL, σ of target θ = 1.0)

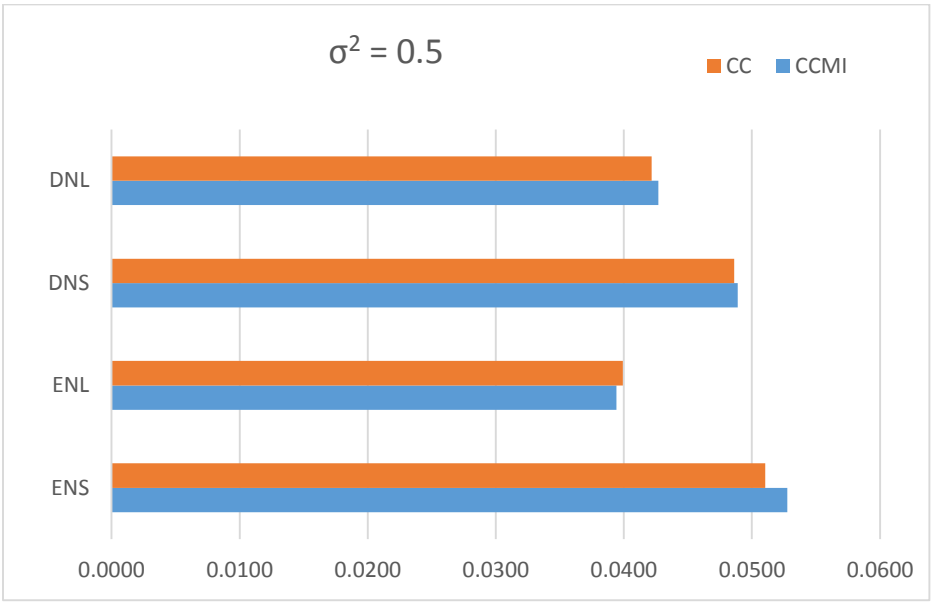

Figure 8 EM & MI comparison in equal target group theta σ (2PL, σ of target θ = 0.5)

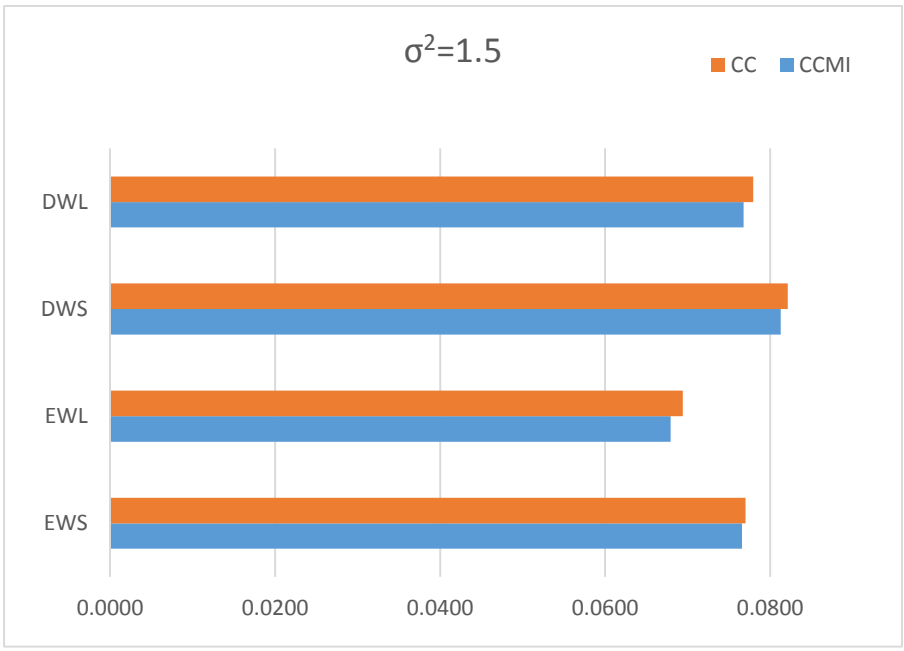

Figure 9 EM & MI comparison in equal target group theta σ (2PL, σ of target θ = 1.5)

4.4.3. Comparison of two different means of ability distribution (2PL) In conditions 1 through 6, the means of theta distributions were all 0. Among these six simulations, CCMI revealed the best estimation in cases of EIL, ENL and EWS, while CC was superior in cases of EIS, ENS. S-L obtained the smallest MSE in EWL. The average MSE of CCMI (0.0566) was the smallest value among the five methods when the mean prior distribution of the target population was 0. When the mean of the prior target person parameter distribution was higher than that of the base population by 0.5, S-L obtained the smallest average MSE value (0.0595).

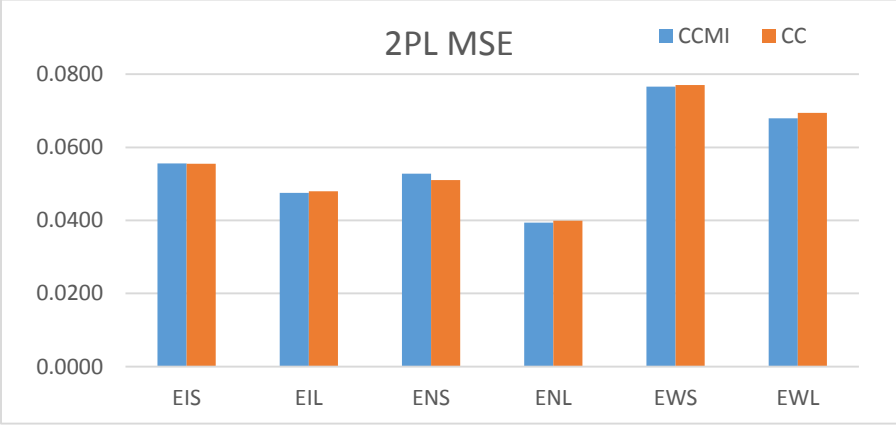

Figure 10 EM & MI comparison in equal target group mean theta (2PL, μ of target θ =0.0)

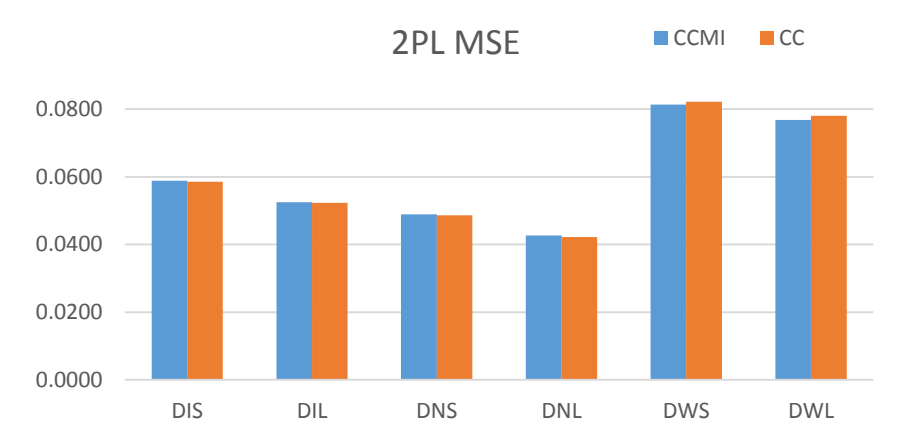

Figure 11 EM & MI comparison in equal target group mean theta (2PL, μ of target θ =0.5)

4.4.4. Comparison of five linking methods (2PL) In five simulation conditions (EIS, ENS, DIS, DNS, and DNL), CC displayed the smallest MSE result. S-L exhibited best accuracy in parameter recovery in four conditions (EWL, DIL, DWS and DWL). In the EIL, ENL and EWS conditions, CCMI showed the best ability parameter estimation. The moment methods (M/M, M/S) didn't provide any better estimation than other methods in any of the 12 conditions.

4.5. Results of 3PL model

The descriptive statistics of each simulation study are summarized in Table 3.

4.5.1. Comparison of two anchor lengths (3PL) Tests with a 20-item anchor had smaller MSE than the 10-item anchor by an average of 35.2% (11.9% to 70.3%) with the 3PL model. Compared with the average value of 2PL model of 15% improvement, 3PL illustrated a better improvement by increasing the number of anchor items. The improvement was large when the standard deviation of target population was narrower than base group. MSE was improved by 72.1% from DNS to DNL. From ENS to ENL, MSE decreases by 73.3%. This large drop can be clearly observed Figure 12and Figure 13. Results from these comparisons suggest that, in general, when the number of common item is small, concurrent calibration runs may be preferable to CCMI.

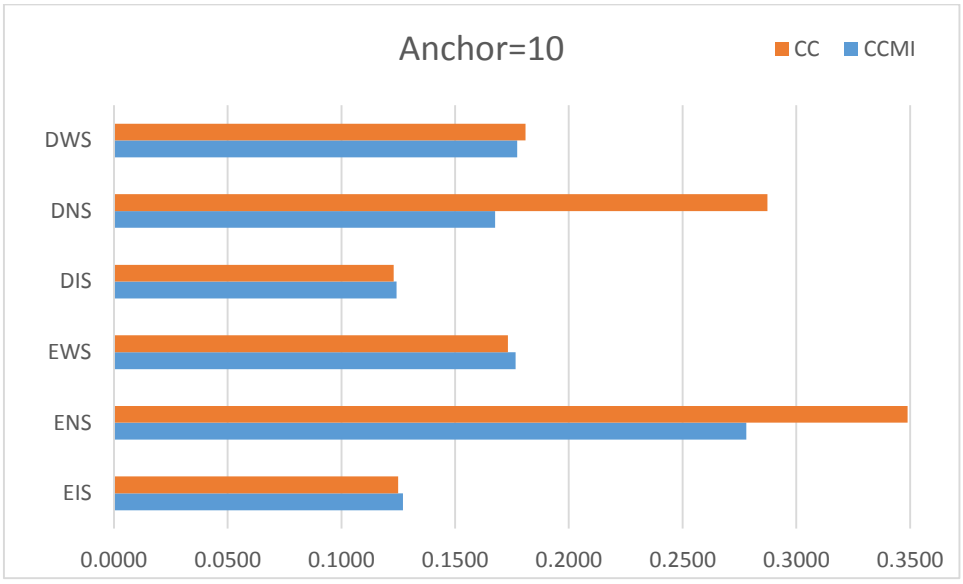

Figure 12 EM & MI comparison in equal anchor length (3PL, anchor =10)

A noteworthy observation was that when the number of common items was large, both estimation methods appear to function similarly. This unique feature of 3PL model was not observed in 2PL model results. The performance of the two estimation methods was remarkably similar in the 2PL model short anchor test conditions.

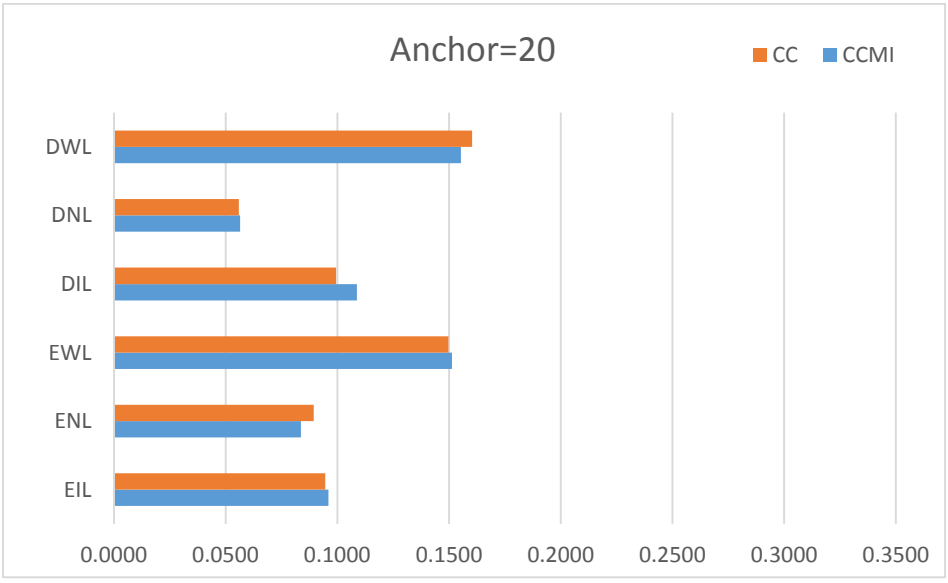

Figure 13 EM & MI comparison in equal anchor length (3PL, anchor =20)

4.5.2. Comparison of three different standard deviations of ability distribution (3PL) Within the same SD of θ between base forma and target form, the MSE values were similar when variance of target θ was 1.0 and 1.5. This can be observed in Figure 14. However, when the target θ distribution was narrower, anchor length played an important role of reducing the MSE (Figure 15). CCMI was also more accurate in this condition. When the θ variance of the target group was wider than that of the base group, the larger anchor length provided better parameter estimation, with smaller MSE. Overall pattern of MSE values in Figure 16 were accordant with major comparisons: anchor length, and theta mean of target group.

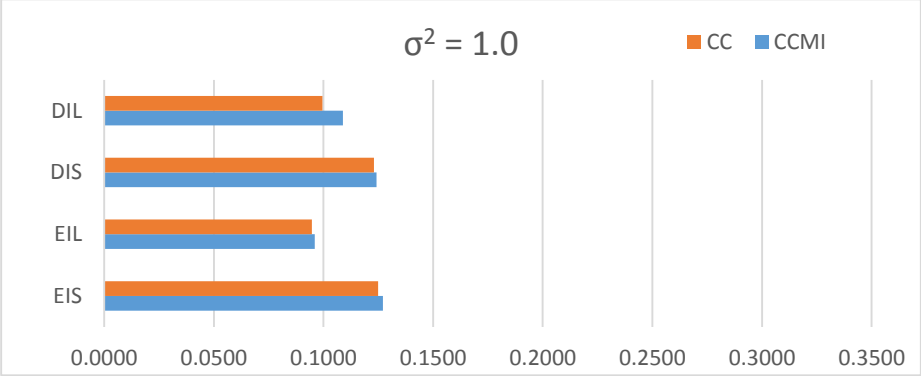

Figure 14 EM & MI comparison in equal target group theta σ (3PL, σ of target θ = 1.0)

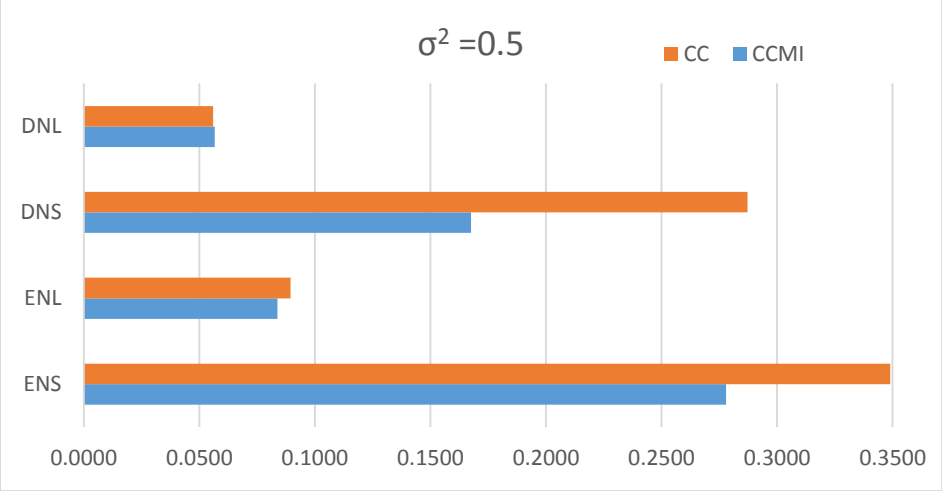

Figure 15 EM & MI comparison in equal target group theta σ (3PL, σ of target θ =0.5)

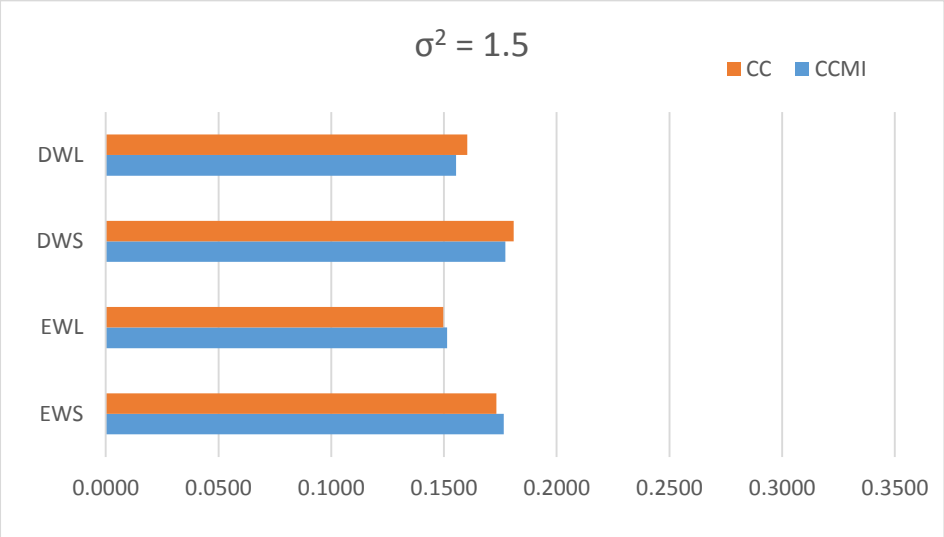

Figure 16 EM & MI comparison in equal target group theta σ (3PL, σ of target θ = 1.5)

4.5.3. Comparison of two different means of target group ability distribution (3PL)

When the means of θ distributions were the same across base and target groups, an inferior MSE for CC equating was observed in the ENS condition. This huge difference was caused by some extreme IRT parameter estimation in a few iterations. In spite of the prior distribution setting, this MSE inflation of CC happened when the θ variance of the target group was narrow and the anchor length was short. Aside from those conditions, the patterns and amounts of MSE were very similar across different theta mean distributions.

Another noteworthy result was that when the variance of target group ability was narrower than the base group and the anchor length was short, both CC and CCMI provided somewhat large MSE values, and CCMI was preferred to CC. This unique feature was not observed in 2PL model results. The performances of the two estimation methods were remarkably similar to each other in the 2PL model short anchor test condition.

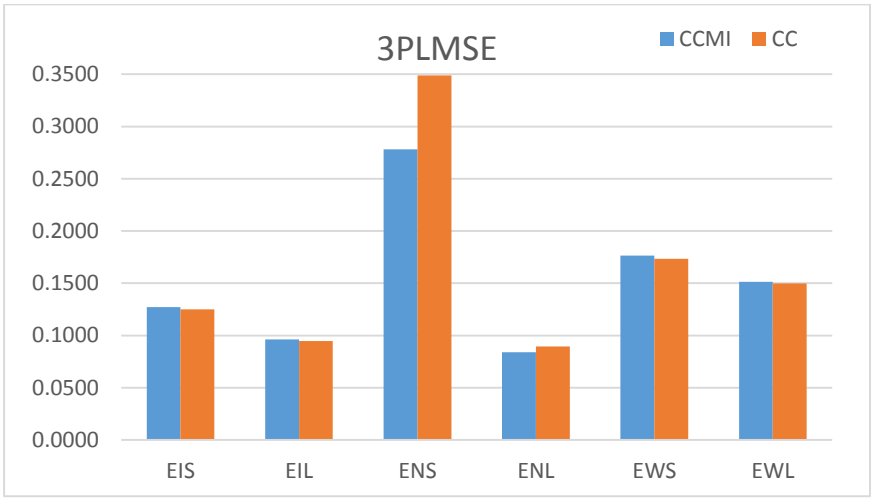

Figure 17 EM & MI comparison in equal target group mean theta (3PL, μ of target θ =0.0)

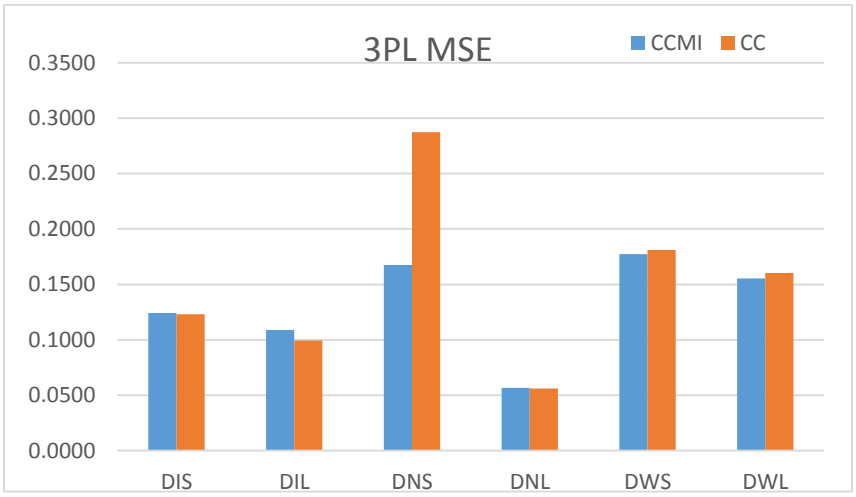

Figure 18 EM & MI comparison in equal target group mean theta (3PL, μ of target θ =0.5)

4.6 Comparison 2PL and 3PL

The result of the 3PL model is not different from the 2PL model. One distinct feature was that the stability of item parameters was little better in 2PL model due to guessing parameter estimation. Even though the prior distribution was applied, the estimated guessing parameters were little bit bigger than true *c* parameter. Table 4 displays the comparison of CCMI and CC. In 2PL model simulation study, CCMI performed better in terms of MSE in 6 conditions. With the

3PL model, CCMI performed better in 5 of the 12 conditions. In the other 7 conditions, the performance patterns were consistent across IRT models.

Table 4 Lower MSE between CC and CCMI in 2PL and 3PL IRT models by cases

	lower MSE (2PL)		lower MSE (3PL)		
case	CCMI	CC	CCMI	CC	consistent?
EIS		v		v	yes
EIL	v			v	no
ENS		v	v		no
ENL	v		v		yes
EWS	v			v	no
EWL	v			v	no
DIS		v		v	yes
DIL		v		v	yes
DNS		v	v		no
DNL		v		v	yes
DWS	v		v		yes
DWL	v		v		yes
count	6	6	5	7	7

4.7. General description of bias in 12 conditions (2PL and 3PL)

In Appendix B, patterns of MSE and bias across score are graphically presented. The 2PL model is summarized in Figure 39 through Figure 62, and the 3PL model is summarized in Figure 63 trough Figure 86. In 2PL, the mean of equated θ bias was large for very low and very high raw scores through all linking methods. The signs of bias are mostly positive for low scores, and negative for high scores. Those patterns were consistent with previous simulation studies

(Jurich et al., 2012; Huggins 2013). When the values are summed, the degree of estimation accuracy offsets each other because of the signed value.

4.8. OSE and TSE results of EIS & EIL (3PL only)

Two conditions were selected to see the result of OSE and TSE to see those patterns by linking methods. Each linking method demonstrated different patterns of relationship between base form score and equated equivalent form score in the EIS and EIL conditions. M/M and M/S illustrated similarity in patterns. Also, CC and CCMI expressed also pattern similarity, and the S–L pattern was closer to the patterns of CC and CCMI.

4.8.1 OSE result of EIS. When it comes to EIS, the large gaps in low score range were observed in S-L OSE and TSE. CC and CCMI illustrated very similar patterns across all score ranges in Figure 19. By and large, M/S and M/M expressed a large gap in the middle of score range.

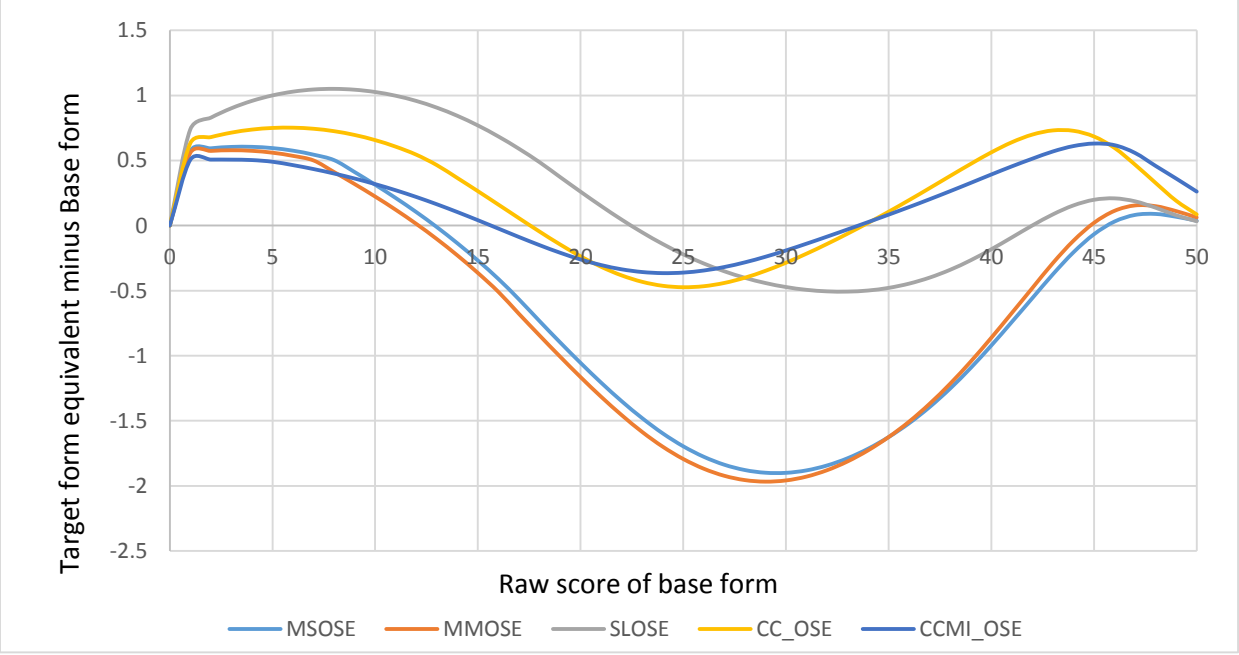

Figure 19 Estimated relationships for EIS 3PL observed score equating

4.8.2. TSE result of EIS. Overall, EIL displayed a larger gap than EIS. This seems to be caused by the larger number of test items in Figure 20 Estimated relationships for EIS 3PL true score equating. Similarly to EIS, M/S and M/M denoted a large gap in the middle of score range. S-L, CC, and CCMI had very similar patterns across scores. Around the score of 10, there was a peak in all equating methods, which seems to be caused by rare data under total score of 10 in 3PL model simulation data. When *c* parameters are summed in 3PL TCC, its simulation generates response data to have lowest score of the sum of *c* in the low theta range. In the 3PL model, very low true scores are not available because when θ come near to -∞, p(θ) approaches c. When 50 items have *c* with mean 0.2, 10 is the lower asymptote value in TCC. Below 10, a linear relationship was generated by TSE without data.

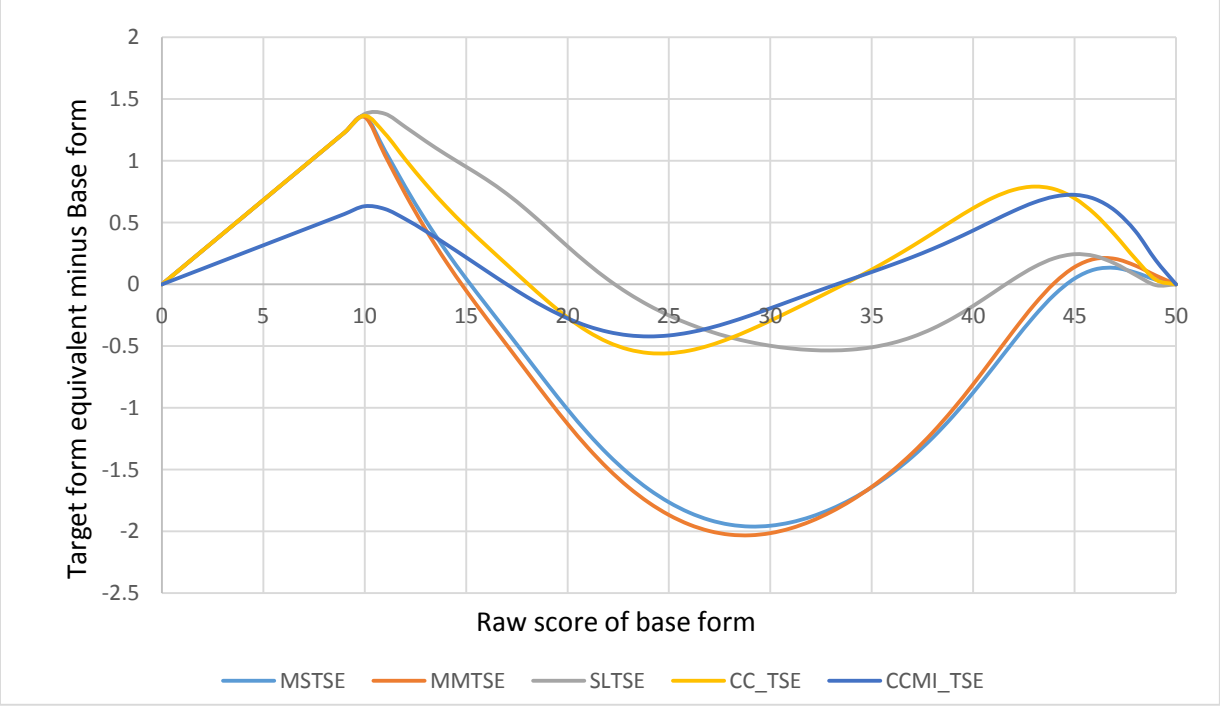

Figure 20 Estimated relationships for EIS 3PL true score equating

4.8.3. OSE result of EIL. In Figure 21, M/S and M/M showed a large gap in the middle of score range, and those methods underestimated the equated score in OSE and TSE. The similarity of CC and CCMI increased more than EIS, and the pattern of S-L was similar in both EIS and EIL

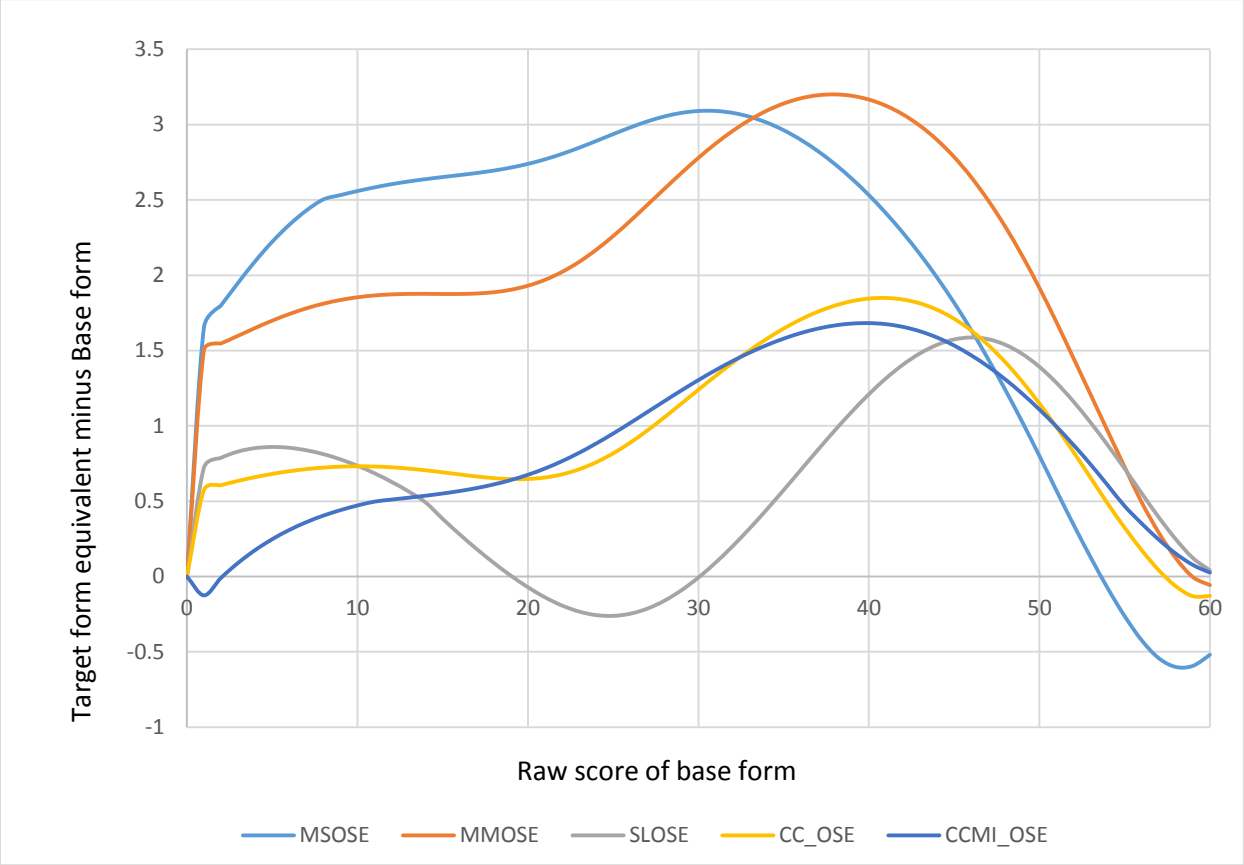

Figure 21 Estimated relationships for EIL 3PL observed score equating

4.8.4. TSE result of EIL. Figure 22 presents similar patterns of TSE over OSE. Due to the longer test length in EIL over EIS, the score difference also increased. This does not mean that the accuracy of equating was reduced in longer anchor test. Those graphs represent the score relationship between forms after the equating procedure.Similar to EIS, there was a peak around the score of 12 for all linking methods, which also may be happened due to a lack of total scores below 12 with the 3PL model. When c parameters of 60 items have mean 0.2, 12 is the lower asymptote value of the TCC. The peak at score 12 comes from this lower bound in simulated data.

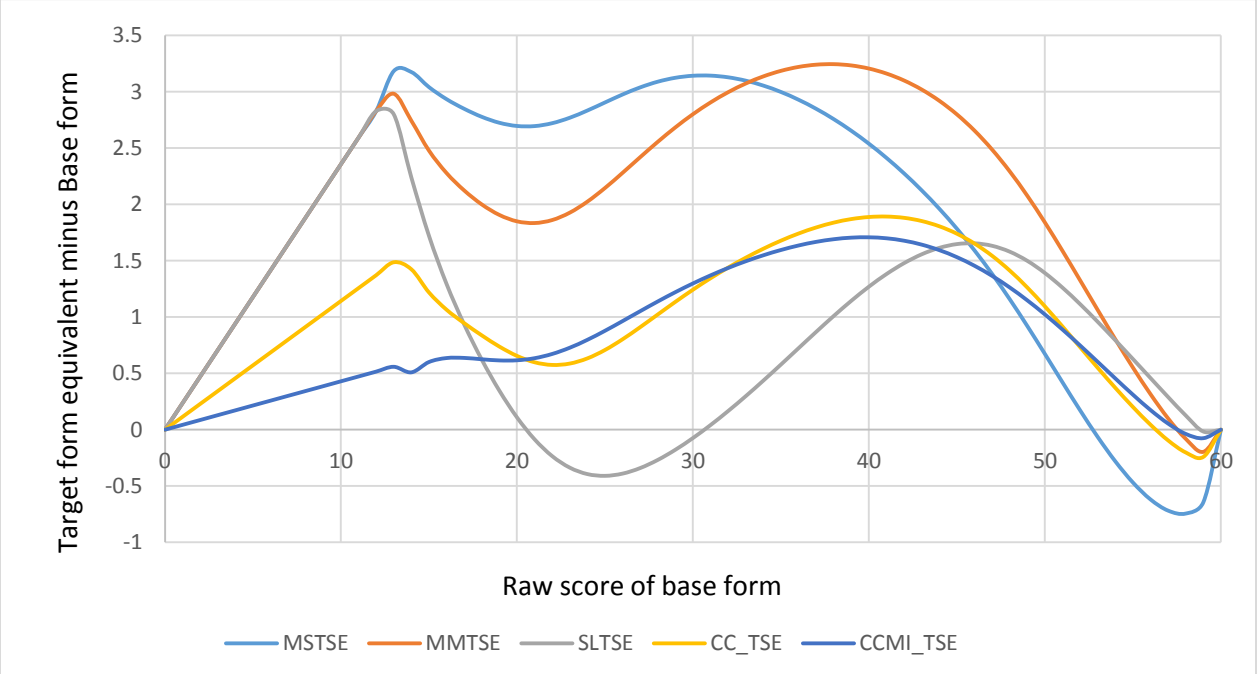

Figure 22 Estimated relationships for EIL 3PL true score equating

4.9. Is MIS noise or a genuine relationship between base and target form scores?

To see the usefulness of MIS, this value was applied to the OSE and TSE for the EIS and EIL conditions. In Figure 23 through Figure 26, in the low score range, MIS value was very far

from other equated scores, and its value fluctuated across the base score range. It looked like a useless byproduct of statistics from multiple imputation estimation.

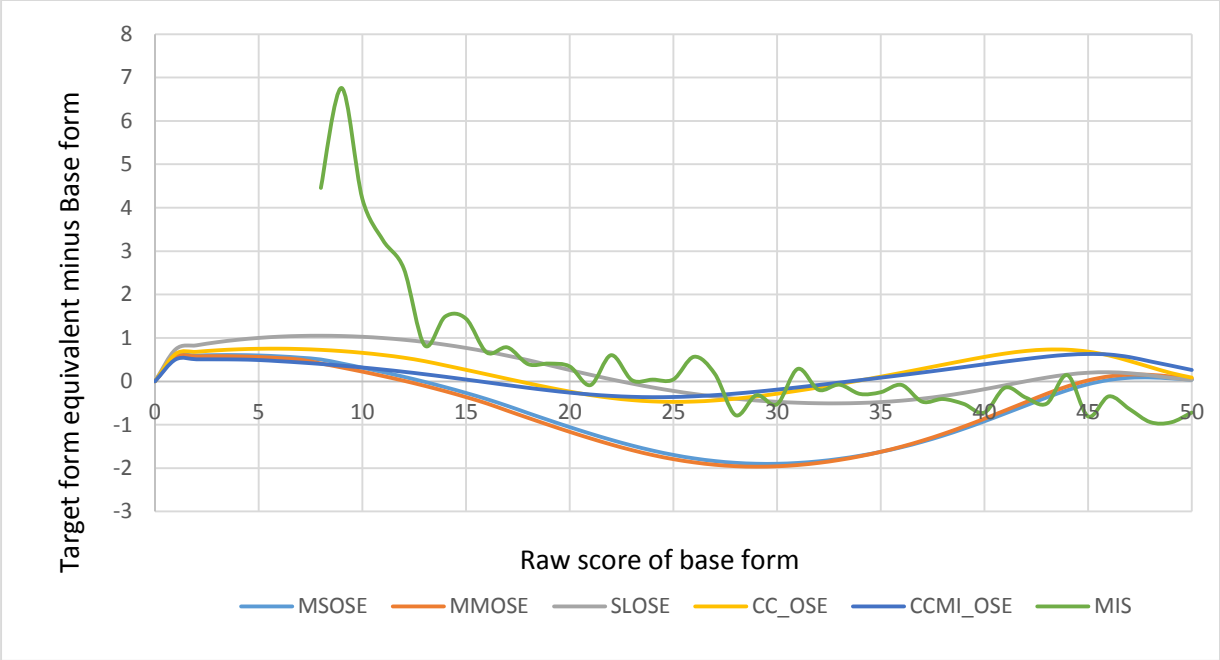

Figure 23 OSE result of EIS including MIS

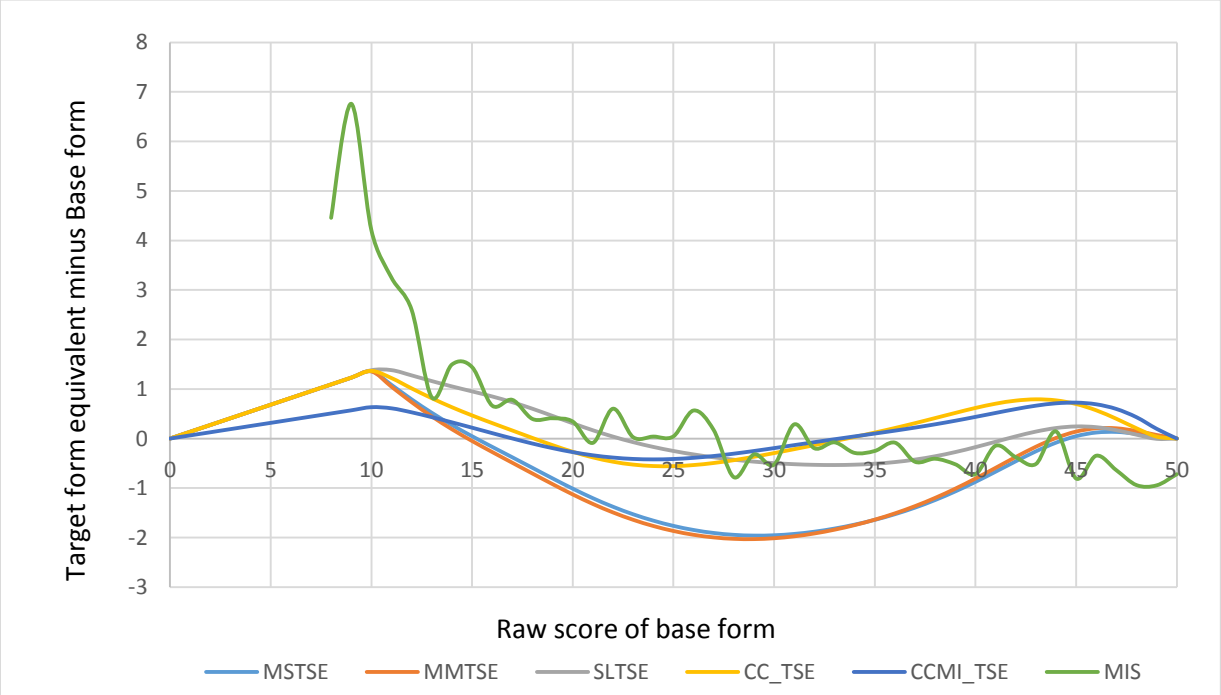

Figure 24 TSE result of EIS including MIS

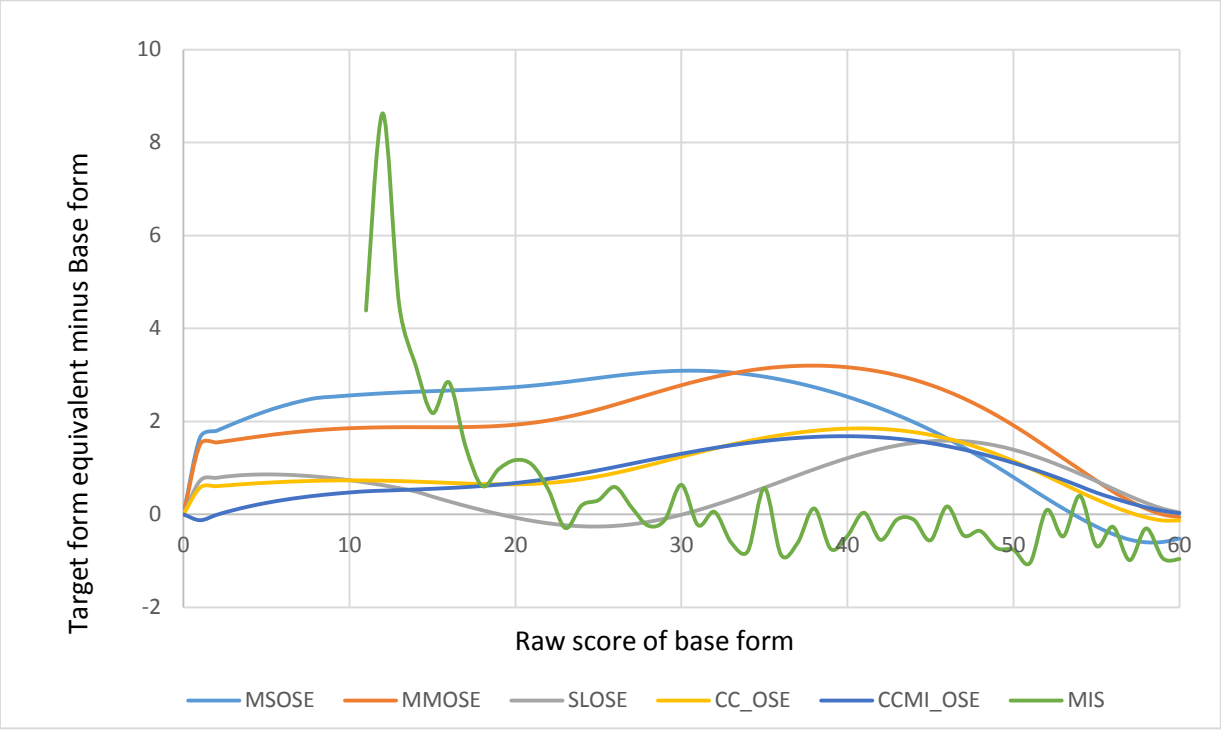

Figure 25 OSE result of EIL including MIS

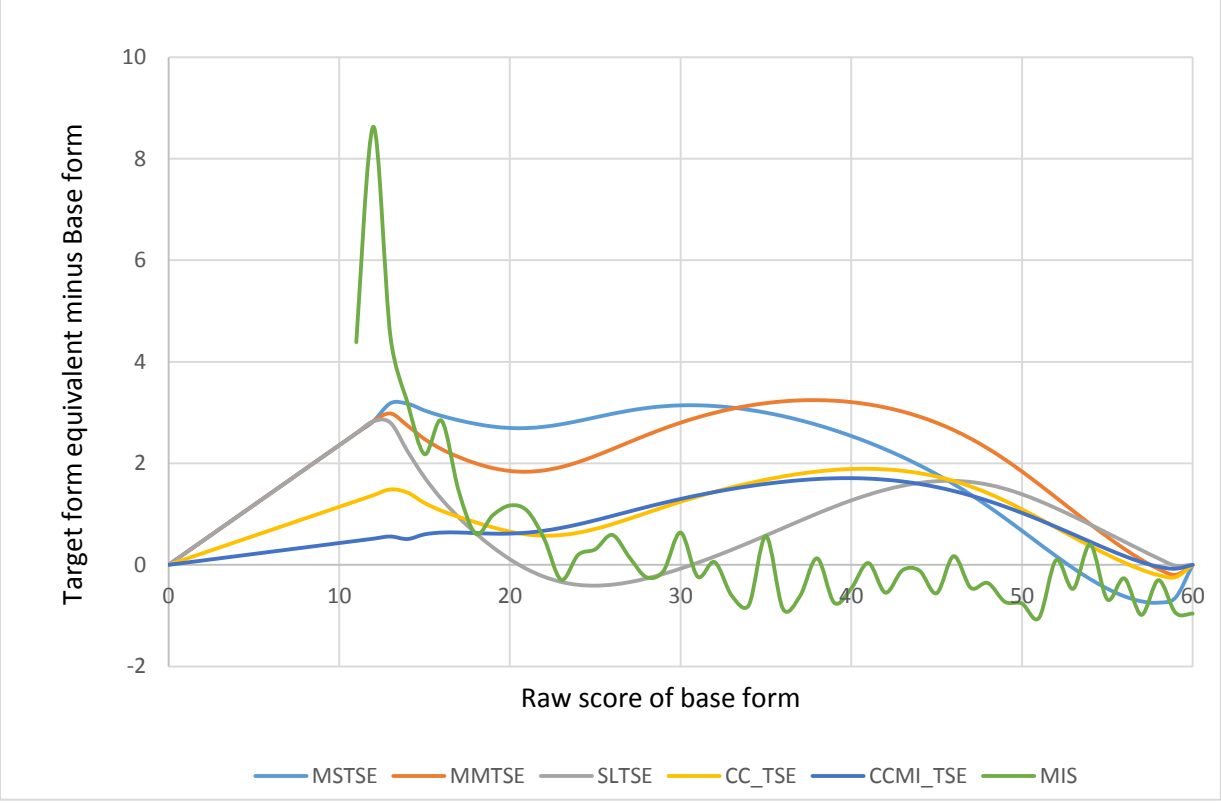

Figure 26 TSE result of EIL including MIS

Equivalent target form scores of base form examinees are missing in the NEAT design, but these missing scores can be derived in simulation study, as average value of replication. This value represents the relationship between base form score and hypothetical complete data (HCD) of target form score as a number-correct score. Figure 27 illustrates the relationship between target form scores of HCD and base form scores in EIL. At the low score range, a non-linear pattern was observed in both cases (EIS and EIL).

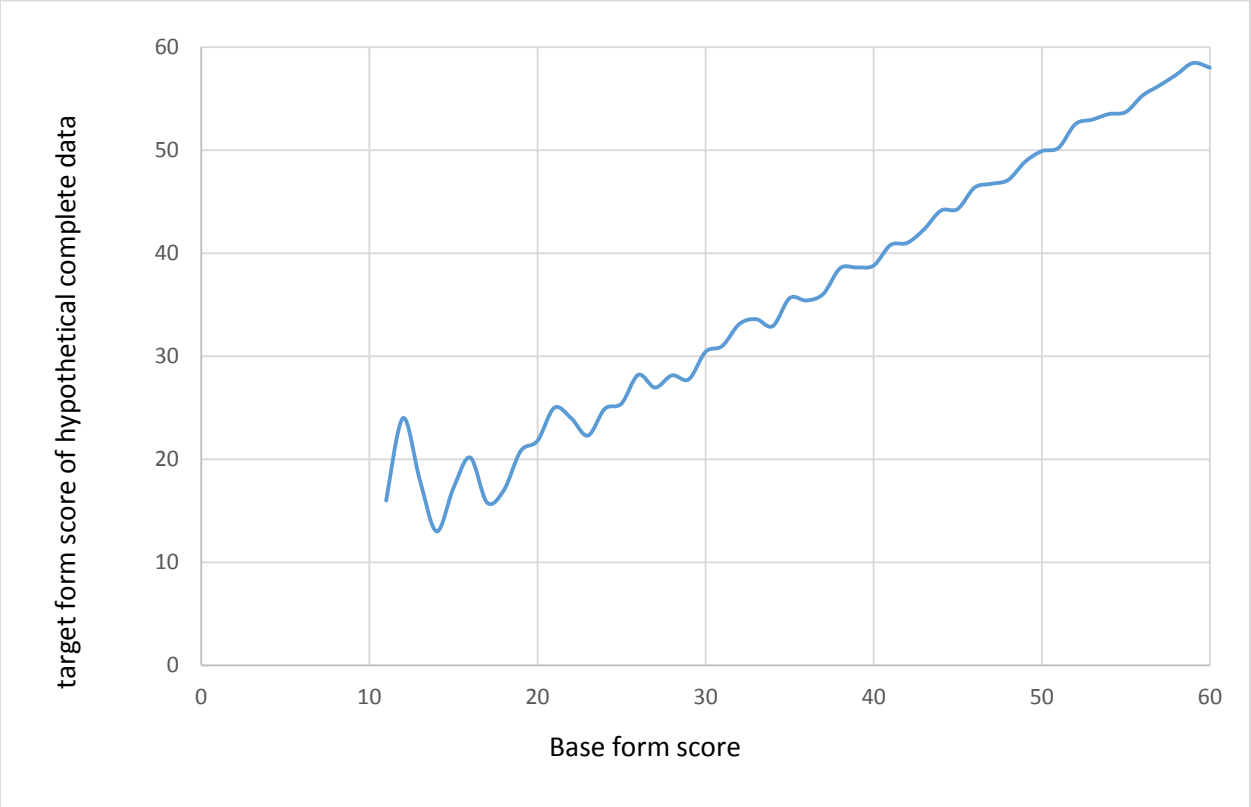

Figure 27 Base form score vs. HCD target form score (EIL, 3PL)

As the simulation data provide the equivalent target form score, MI is able to provide equivalent target form scores with the mean value of imputed scores. When seeing the relationship between MIS score and base form score, a similar pattern of the Figure 27 Base

form score vs. HCD target form score (EIL, 3PL) is observed in the Figure 28 MIS vs. base form score (EIL,3PL) around the low score range.

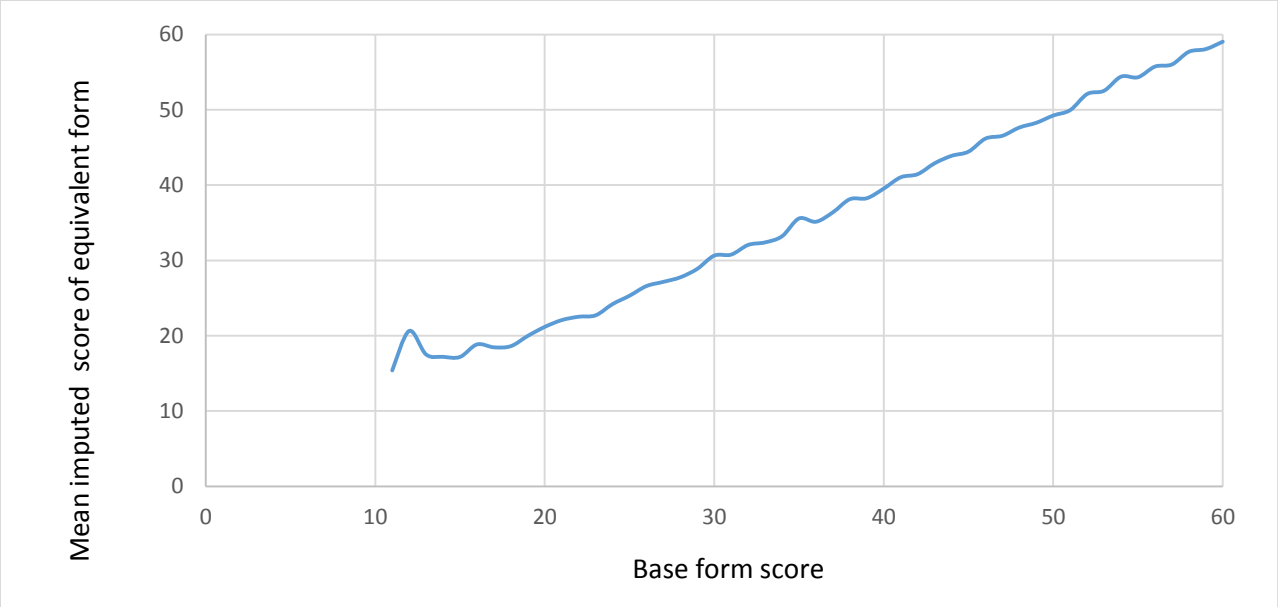

Figure 28 MIS vs. base form score (EIL,3PL)

When the MIS is applied to the previous evaluation criterion (the relationship between equated target form score (OSE or TSE) and base form score), in Figure 29 the sampling error of MIS can be regarded as useless results. However, if we change the evaluation criterion to HCD target form score minus base form score, the similarity of patterns of MIS can be observed in the Figure 30 and it describes MIS values are more closely to the HCD target form score. When the graph is closer to zero, it represents the more accurate estimation to the HCD target form score. MIS is displaying the superior estimation features over TSE of other methods, which is more stabilized around 0 over the whole score range in Figure 30.

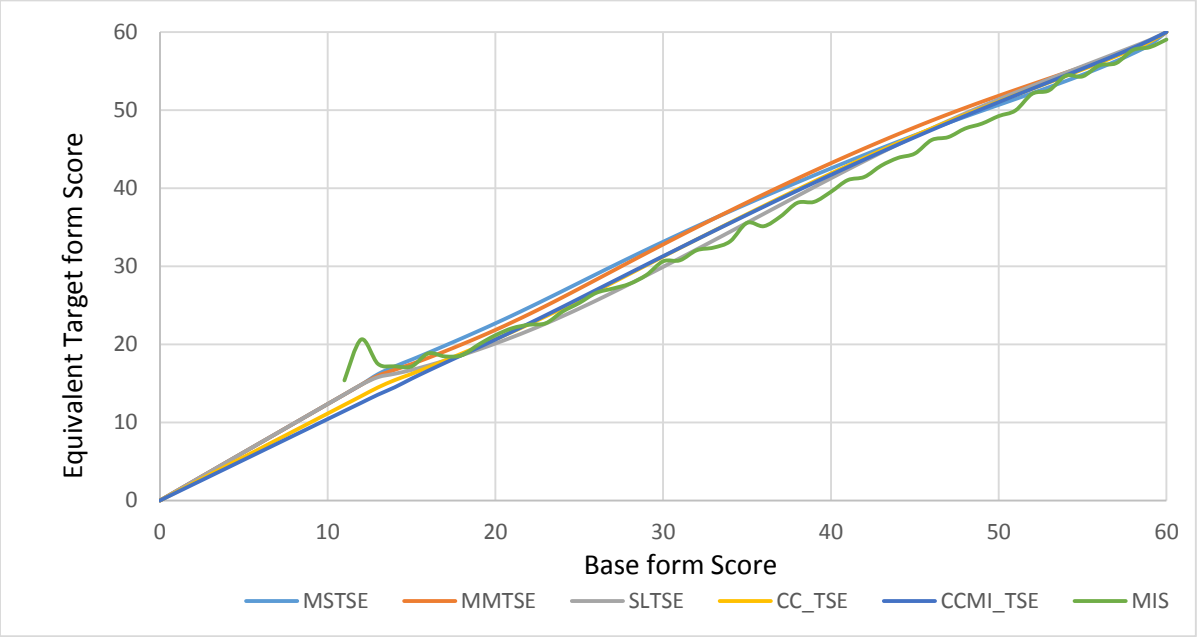

Figure 29 TSE Results base form and HCD target form (EIL,3PL)

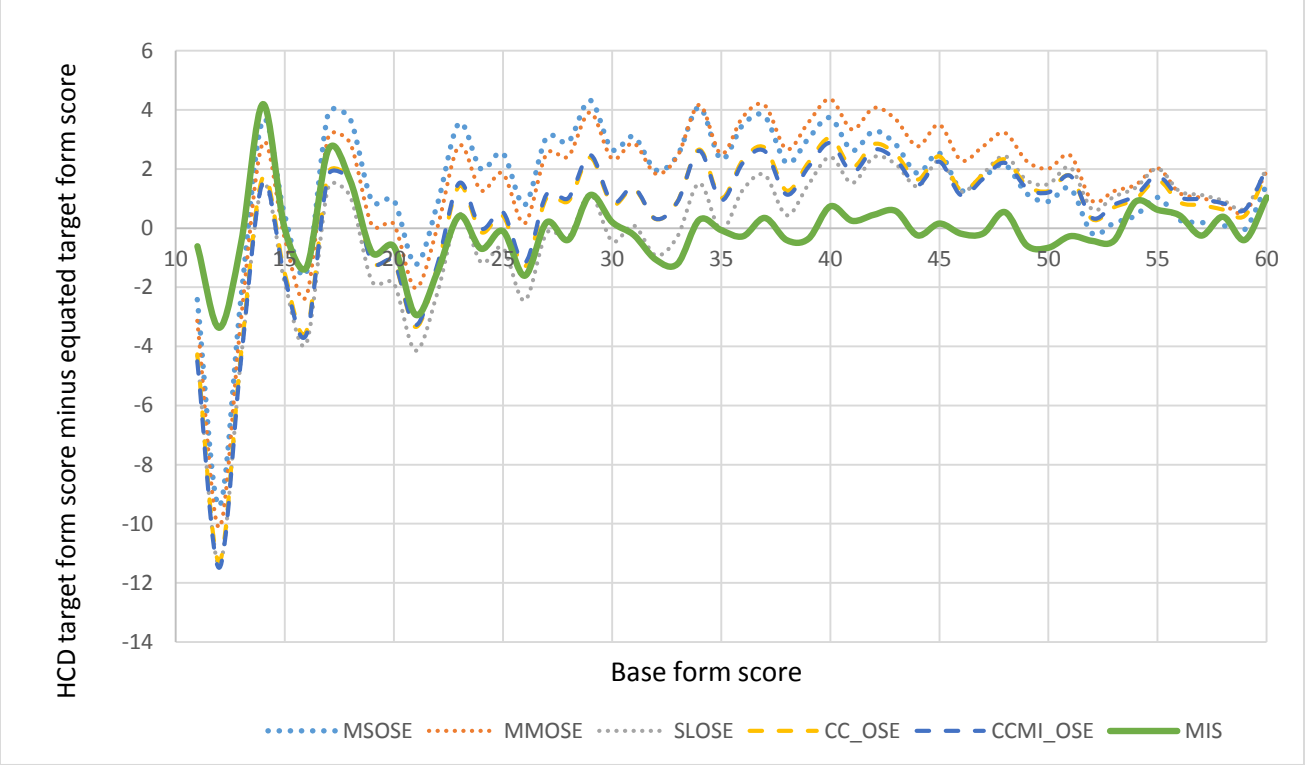

Figure 30 Evaluating number correct scores by equating methods in a HCD criteria (EIL,3PL)

4.10. PISA linking results with OSE and TSE (mainly) including MIS

Two forms of PISA 2000 reading score were linked by the methods above. A total of 850 students took the test, which including 16 anchor items, 63 items were asked across forms. Form A was defined as the base form and Form B was assigned as the target form. Due to elimination of some items at each form, the number of items does not match. However, similar response patterns and descriptive statistics were observed in the frequency chart Figure 31and Table 5

Table 5 PISA 2000 basic descriptive statistics of test score and subjects by test form

Form	Subjects	Items	Mean	SD	Min	Max
A (base)	416	40	24.29	8.31	4	40
B (target)	434	39	23.88	8.37	2	38

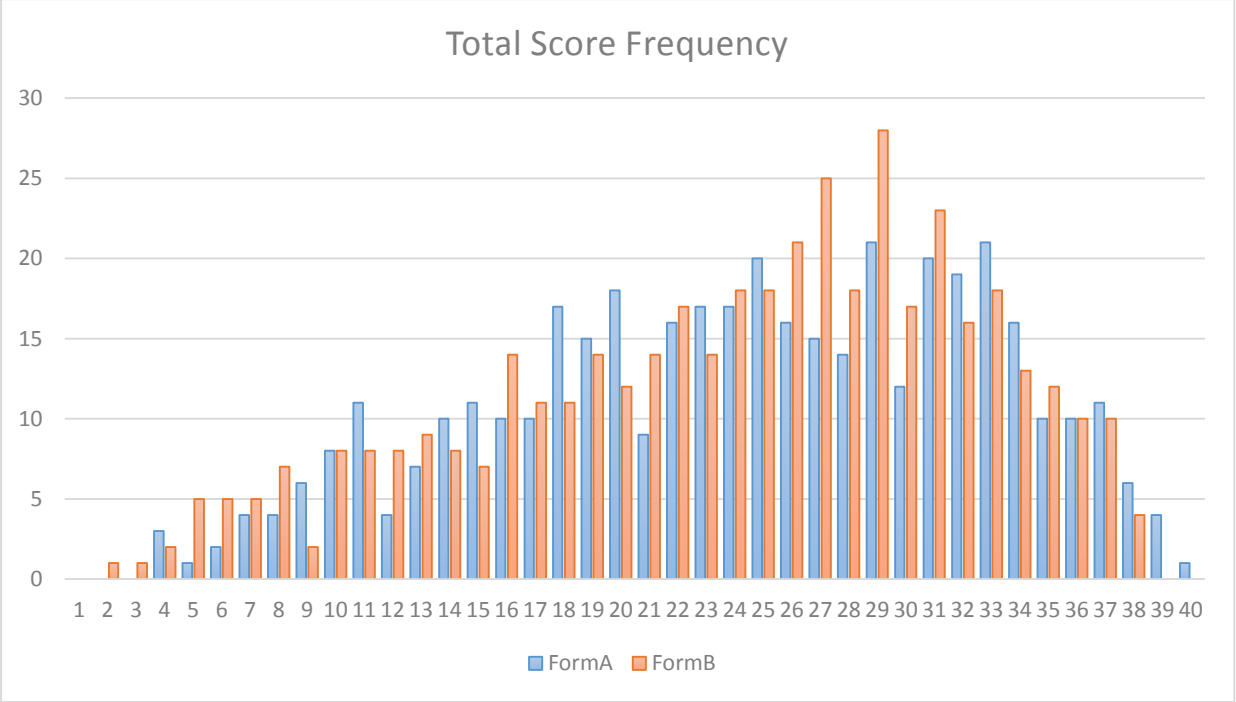

Figure 31 Histogram of PISA 2000 reading (U.S.) test scores by test form (1 and 9)

The results of CCMI and CC were similar in OSE and TSE performance. Base form and equivalent target form are supposed to have a one point difference at the end of score range due to their design. This is clearly represented in the TSE graph.

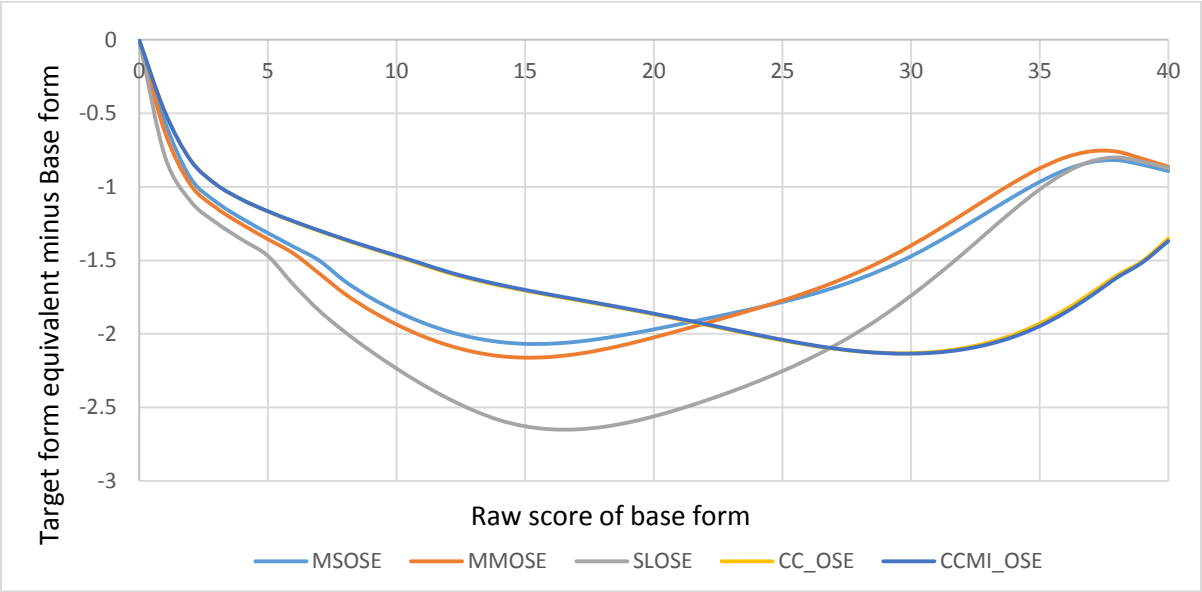

Figure 32 Estimated relationships for PISA 2000 3PL observed score equating

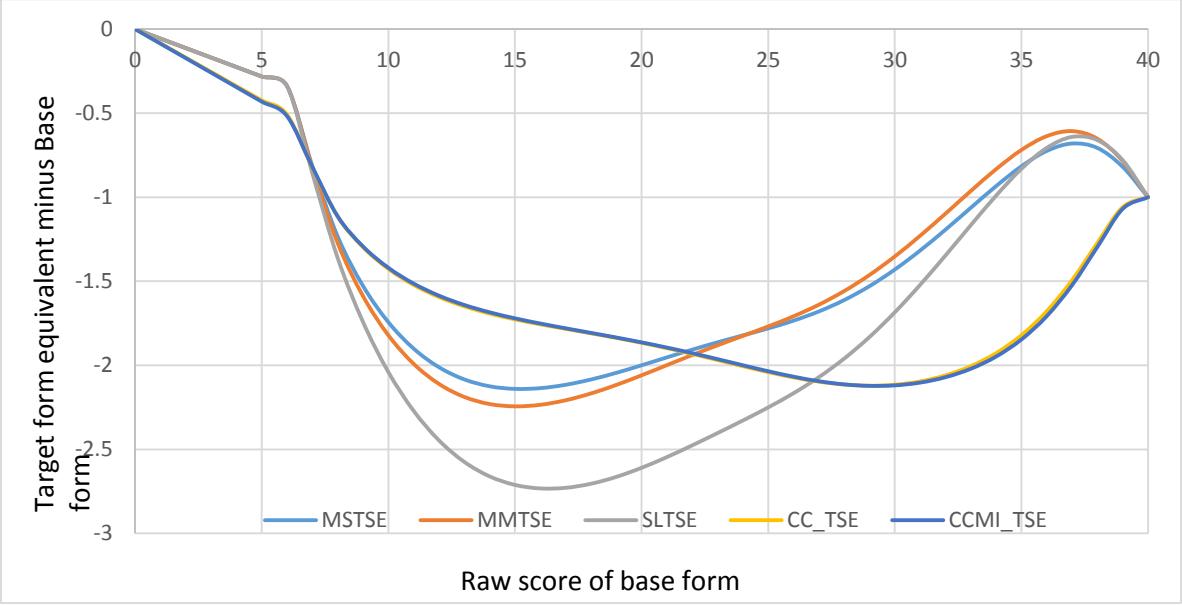

Figure 33 Estimated relationships for PISA 2000 3PL true score equating

The non-monotonically increasing pattern in the low score range was also observed in the relationship between target form Mean Imputed Score and base form. This pattern was observed in all 12 simulation conditions.

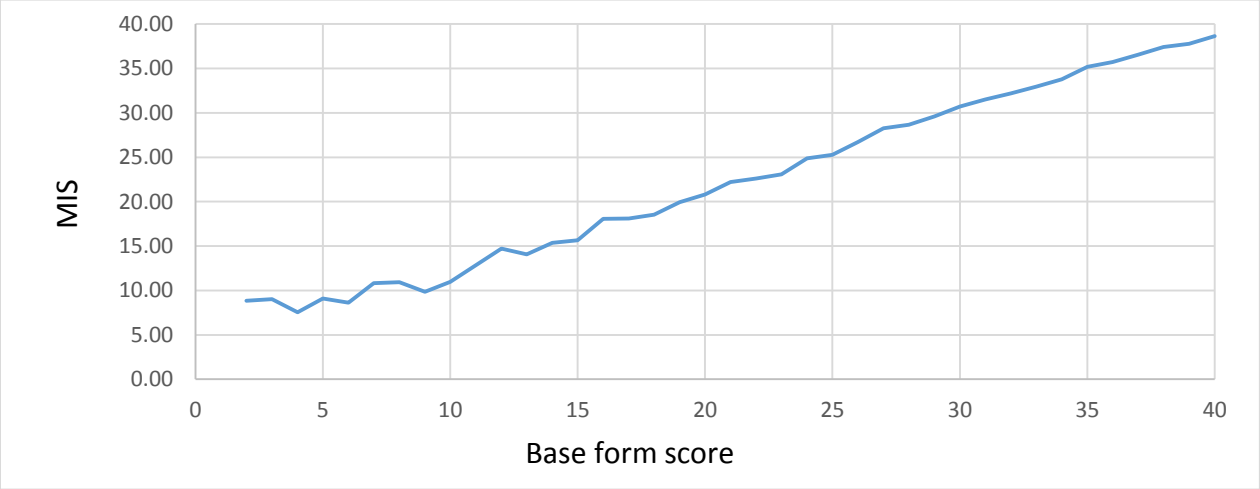

Figure 34 MIS vs. base form score (PISA 2000)

MIS values were also added to OSE and TSE results to compare with other equated scores in Figure 35 and Figure 36.

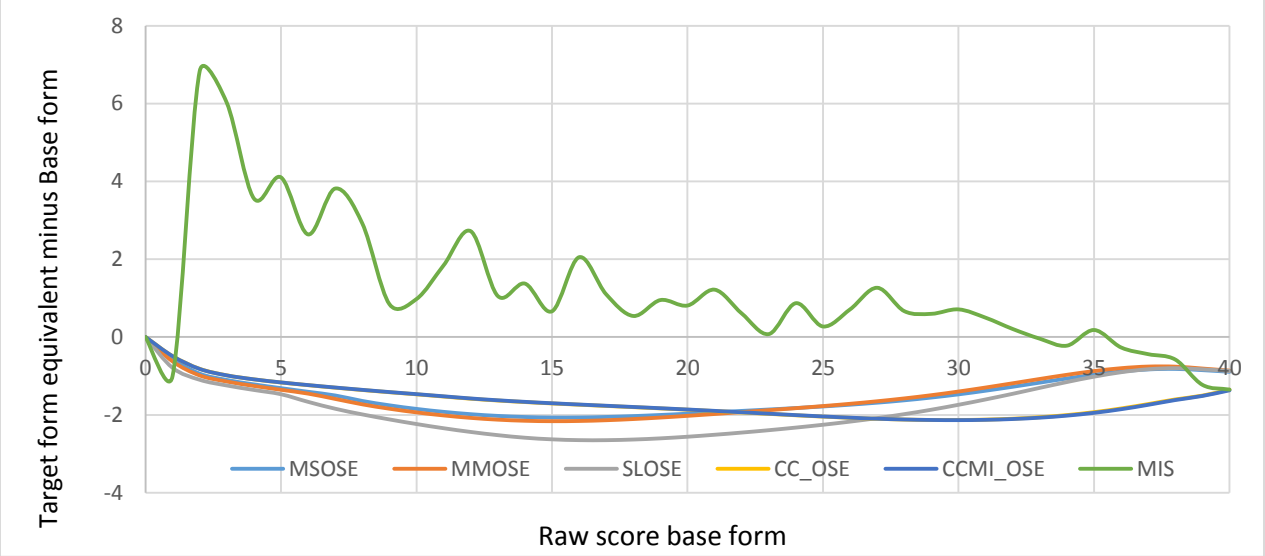

Figure 35 Estimated relationships of 3PL observed score equating including MIS

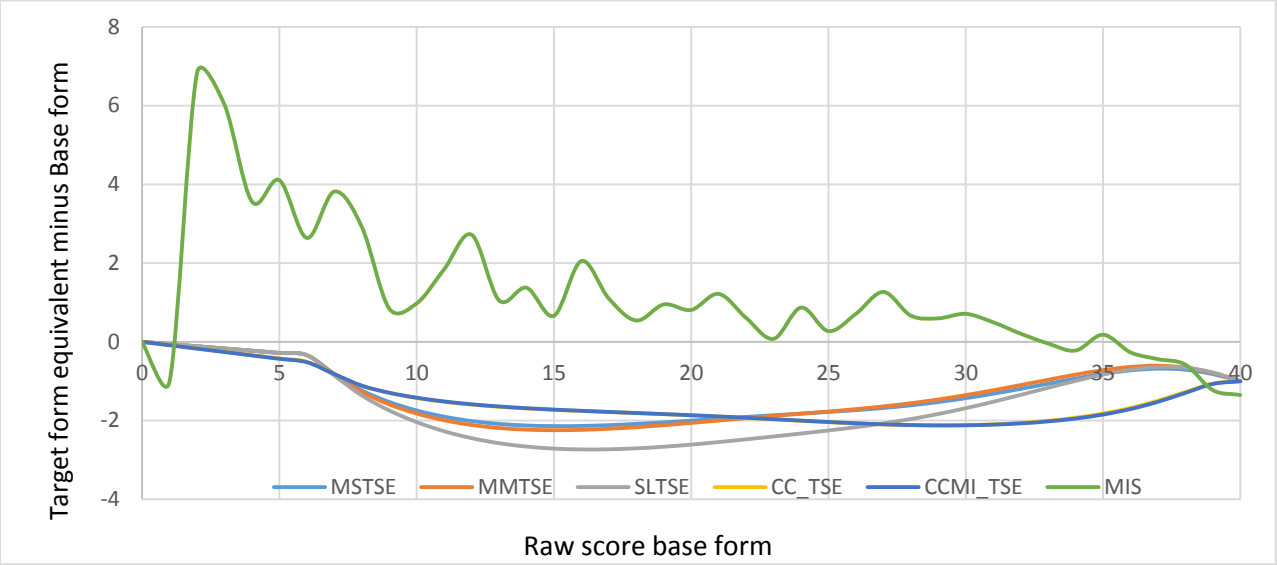

Figure 36 Estimated relationships of 3PL true score equating including MIS

Similar to the simulation results, the lower bound of real scores also had a linear relationship in the low score range of OSE and TSE in Figure 35 and Figure 36. The MIS value has a peak in the low score range, and this is comparable in Figure 35and Figure 36. However, in the real test score, HCD cannot be applied to show the hypothetical score relationship between base and target score.

4.11. MSE and bias of 12 conditions across number-correct score in 2PL and 3PL

In Appendix B, the patterns of two evaluation criteria are graphically presented. MSE and bias of 12 conditions (2PL) were illustrated from Figure 39 to Figure 62. Those values in 3PL model are given in Figure 63 to Figure 86. From all the 2PL graphs, the patterns were very similar across equating methods. Although the patterns within CC and CCMI tended to be very similar, the large CC MSE values were captured in Figure 68 and Figure 80. Both conditions had the two things in common: narrower variance of target group θ and short anchor length.

Chapter Five - Discussion

The concurrent calibration method demonstrated its usefulness in the NEAT design in a number of studies. Despite the popularity of the NEAT design with concurrent calibration, there have not been previous efforts to use multiple imputation techniques for equating. For this Monte Carlo study, 100 replications were performed for each condition. For one replication, 100 imputed data sets were drawn. The purpose of this study was to introduce a new linking method, concurrent calibration with multiple imputation, in planned missing data design. Compared with concurrent calibration with MMLE, MI provides imputed data sets. Generally, observed or true score equating (OSE/TSE) is performed after the IRT parameter scaling, which requires some computational procedure (Kolen & Brennan, 2014). However, MI provides multiple sets of observed scores. The mean scores of target forms can be calculated from multiple data sets and can be utilized to estimate observed scores for the target form test takers.

There are two main purposes of this study: one is to evaluate the performance of a newly introduced estimation method, multiple imputation in concurrent calibration, compared with the previous prevalent estimation method, EM. To see under which circumstances each method can exhibit better accuracy in IRT person parameter recovery, 12 conditions were tested. Theoretically it is known that the two estimation methods produce the same or similar estimation results in psychology researches. However in IRT frame work, a lower estimation bias of MI was obtained in previous results by De Ayala et al. (2001) and Finch (2008). Both EM and MI displayed similar results in the 12 conditions of 2PL. However, compared with the average value of MSE in 3PL, MI had lower bias (mean=0.1316, SD=0.0574) than EM (mean=0.1512, SD= 0.0847). The effect size, Cohen's d, was 0.27, which can be categorized as medium effect size according to Cohen's rules of thumb, (0.1 = small, 0.3 = medium, 0.5 = large). EM had slightly

larger bias than MI. Finch (2008, p. 243) pointed out that the multivariate normality assumption of the EM approach is one possible explanations, because it could clearly not apply to dichotomous data. With MI, it was possible to use the number-correct scores which work for categorical data.

According to the Table *6* Relative MSE to CCMI(2PL) the differences between EM and MI seems to be ignorable but not in the cases of ENS and DNS in 3PL. In those cases, the differences of MSE values were 26% (ENS) and 71% (DNS) and MI provided far more robust results. In two conditions of 3PL (ENS & DNS), large MSE occurred in CC when the variance of target group θ was narrower than the base group with short anchor test (Figure 68 and Figure 80). This occurred even after changing the random seed several times. The worse response simulations in two conditions may be caused by the relatively small number of subjects (500) against the number of items (50) and short anchor length (10) and seems to make model estimation harder, especially in 3PL. However, there are many possible scenarios that those test administration settings can happen. When it comes to the size of group, typically smaller subject groups are regarded as target groups. According to central limit theorem, randomly selected smaller subject group tends to have a smaller score variance and it can be applied to theta estimate from pattern scoring. Also, the length of anchor test is typically supposed to be short especially for test security reasons. When test conditions of ENS and DNS are observed in real test administration, MI needs to be implemented against EM for 3PL theta estimation.

The other purpose was to evaluate the possibility of mean imputed score utilization. Mean imputed score is a byproduct of multiple imputation estimation. There is no study on mean imputed scores yet, which can explain the hypothetical features between base form score and target form score. Number-correct score is commonly used for practical reasons, because

extreme values of θ often occur in the estimation so OSE and TSE scores. However, this equating is based on the assumption that θ and number-correct scores are monotonically increasing over the score range. However, at the extremes score, sometimes odd things can happen in TCC equating methods. Although TCC θ estimation is relatively unbiased, conditional SEMs of number-correct scores are too large in lower and upper ends of θ. However, expected *a posterior* θ estimates based on the number-correct scores (EAP; Thissen & Orlanso, 2001) are currently used to avoid CSEM amplification. TCC equating methods rely on the assumption of monotonic increasing relationship between theta and number correct score. Some features violating this assumption were observed in the hypothetical complete data sets. Hypothetical target form equivalent scores of base group were obtained in low theta range from simulation of 3PL EIS & EIL.

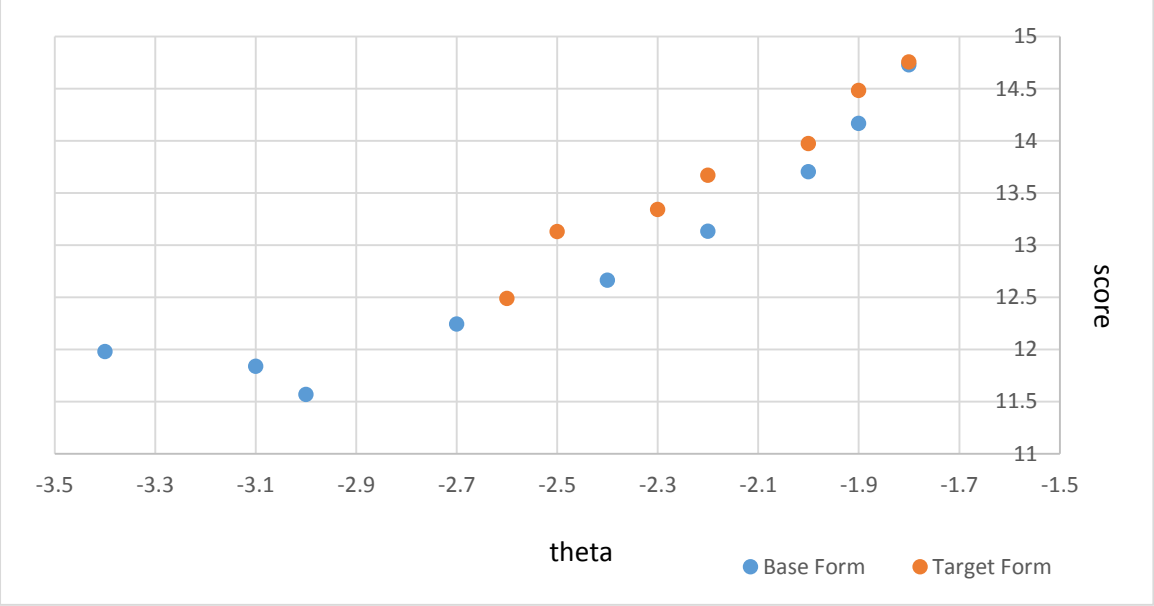

Figure 37 Non-monotonic relationship between theta and number correct score (EIS)

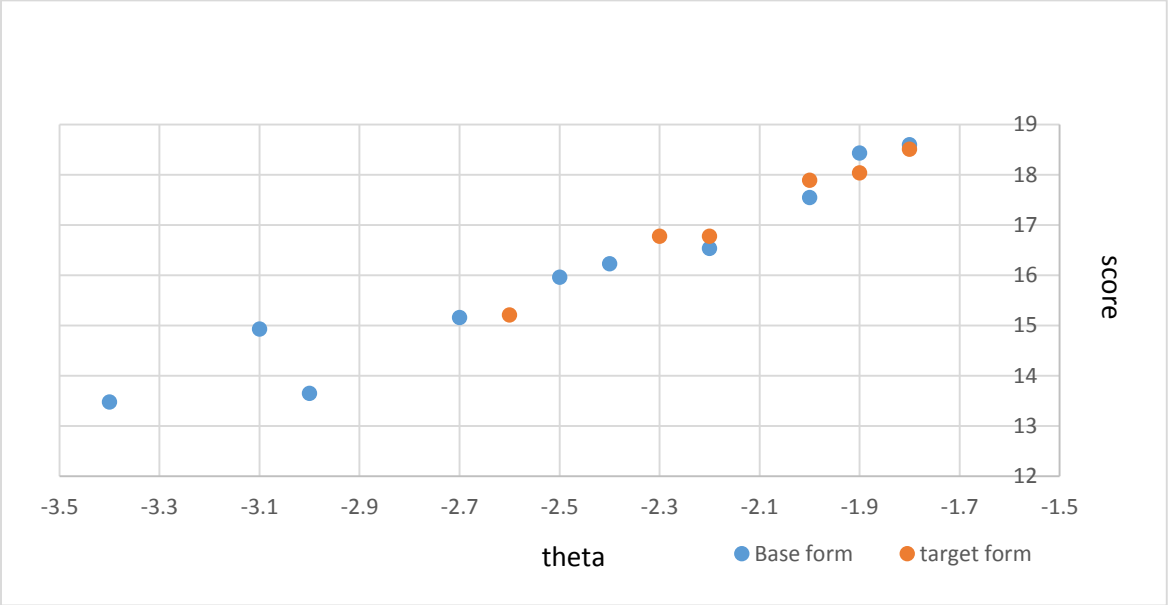

Figure 38 Non-monotonic relationship between theta and number correct score (EIL)

The non-monotonic relationships were observed in Figure 37 Non-monotonic relationship between theta and number correct score (EIS) Figure 38 Non-monotonic relationship between theta and number correct score (EIL).MI has lots of things in common with Bayesian estimation (Enders, 2010), posterior distribution is obtained by drawing samples from a prior distribution based on a specified model. Imputed scores can be drawn from the prior distribution through the converged IRT model. Within Bayes' theorem, the distribution of number-correct scores can be obtained. TCC is defined by a likelihood function of θ under a given number-correct score which is represented as p(θ|y). However, to have the distribution of number correct score p(ỹ) in the given estimated θ distribution, p(ỹ| θ) needs to be. In the Equation 5.1., the posterior distribution, p(ỹ|θ) is represented by p(θ|y), likelihood function (3PL model here), p(y), prior distribution of score and p(θ), marginal distribution in MI.

$$p(\tilde{y}|\theta) = p(\theta|y)*p(y) /p(\theta) \qquad \text{Eq. 5.1.}$$

While imputed scores are drawn multiple times from the prior distribution p(y), mean imputed scores are obtained from the posterior distribution. This is the difference between Bayesian approach and TCC equating approaches (OSE & TSE). When TCC approach is represented in a Bayesian way, estimated theta, $E(\theta_{base}|y_{base})$ is linearly transformed to target theta, and number correct score is estimated by target theta and it can be shown as $E(y_{target}|\theta_{target})$. However, the $p(y|\theta)$ does not have to be linear to the $p(\theta|y)$, according to Bayesian theorem. Therefore, monotonically increasing number correct score vs theta across forms is no longer needed under Bayesian estimation and MI.

Fortunately, the simulation study could provide hypothetical complete data set to have number-correct scores of the target form portion in the base form, which are shown as IB in Figure 4 Unique items, common items and imputed items in the base form and target form. This made it possible to compare the base form score and target form score of HCD. This is hypothetical relationship between the number-correct score of base form and target form, so the difference from this true value and the results of OSE, TSE, and mean imputed score could be possibly compared by linking methods. The common evaluation criterion for OSE and TSE is the difference between target form equivalent score and base form score of each equating methods (Kolen & Brenan, 2014, pp. 213 & 217).

Previous studies have described inaccurate equating results when two groups differ in performance (Livingston, Dorans, & Wright, 1990; Wright & Dorans, 1993). In terms of group mean difference, the accuracy of parameter recovery was not affected much using the 2PL and 3PL models. When the variance of the target group θ was wider than the base group in the 2PL model (Figure 10 and Figure 11) and 3PL model (Figure 17and Figure 18), the MSE value increased on both models. On the contrary, when the variance of the target group was narrower

than base group, MSE values were lower in 2PL and larger in 3PL. In IRT estimation of this study, the latent mean of base group was fixed at 0, and variance was also set to 1.0 and the target group's mean and variance were freely estimated. This was different from previous studies, which used population-dependent equating and violated a statistical assumption (Powers & Kolen, 2014).

5.1. Limitations and considerations for future research

In the simulation study and real data of PISA 2000, the effect of the anchor test was ignored. The anchors were pre-assigned to the NEAT design, but the qualities and statistics of the anchor were not intentionally controlled, which can play an important role in linking performance. Also, the effect of Differential Item Function (DIF) was not considered in anchor items or unique items of base form and target form. In the same vein, Raykov (2012) mentioned measurement invariance as a key factor of anchor items in NEAT design. As the difference in test performance of groups increases, the bias of equating results also increased in the study of Powers and Kolen (2014). Their results also showed that the difference in group size is also associated with equating error and methods. Huggins (2014) studied the effect of anchor item DIF on equating and found when the amount of DIF and number of DIF items increase, population invariance of equating can be compromised.

Simulated data and real data were assumed to be within a unidimensional framework, with items measuring a single latent ability. In practice, tests tend to have multidimensionality. For multidimensional equating, other equating methods need to be separately introduced. Under the multidimensionality assumption, an equating framework which takes into account the nature of multidimensionality need to be utilized.

In both the simulated and PISA 2000 data, only dichotomous data were utilized, but mixed format tests are popular in real practice. For more generalizable results, mixed format test design will also be required in future studies.

The sizes of base and target groups were well balanced in the simulation (500/500) and PISA 2000 data (416/434). The only group difference considered between base form and target form test takers was in mean and SD of ability population. Although smaller variance in the target group θ (ENS, ENL, DNS, and DNL) may be regarded as a smaller number of subject groups, various pairs of subject numbers between base and target groups need to be studied for better accuracy. The effect of group size difference on equating is also required to study in future research.

References

Acock, A.C. (2005). Working with Missing Data. *Journal of Marriage and Family, 67*, 1012-1028.

Angoff, W. H. (1984). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), Educational measurement (2nd ed., pp. 508-600). Washington, DC: American Council on Education.

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

Baraldi, A.N. & Enders, C.K. (2010). An introduction to modern missing data analysis. *Journal of School Psychology, 48,* 5-37.

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

Beguin, A. A.., & Hanson, B. A. (2001). Effect of noncompensatory multidimensionality on separate and concurrent estimation in IRT observed score equating. CitoGroep 2001-2, Arnhem, Maart: CitoGroep.

Birnbaum, A. (1968). Some latent trait models. In F.M. Lord & M.R. Novick, (Eds.), *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Chalmers, R.P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software, 48*(6), 1-29.

Cook, L. L., & Eignor, D. R. (1991). An NCME module on IRT Equating methods. *Educational Measurement: Issues and Practice, 10*(3), 191-199.

De Ayala, R. J., Plake, B., & Impara, J. C. (2001). The effect of omitted responses on the accuracy of ability estimation in item response theory. *Journal of Educational Measurement, 38,* 213-234.

De Ayala, R.J. (2009). The Theory and Practice of Item Response Theory. NY: The Guildford Press.

DeMars, C. (2002). Incomplete data and item parameter estimates under JMLE and MML. *Applied Measurement in Education, 15,* 15-31.

Drasgow, F. (1989). An evaluation of marginal maximum likelihood estimation for the two parameter model. *Applied Psychological Measurement, 13,* 77–90.

Enders, C.K. (2010). *Applied missing data analysis.* New York: Guilford Press.

Finch, H. (2008). Estimation of Item Response Theory Parameters in the Presence of Missing Data. *Journal of Educational Measurement, 45,* 225– 245.

Graham, J.W. (2009). Missing Data Analysis: Making it Work in the Real World. *Annual Review of Psychology, 60,* 549-576.

Graham, J. W., Taylor, B. J., Olchowski, A. E., & Cumsille, P. E. (2006). Planned missing data designs in psychological research. *Psychological Methods, 11*, 323−343.

Graham, J.W., Olchowski, A.E., & Gilreath, T.D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Research, 8,* 206-213.

Haberman, S. J. (2006). An elementary test of the normal 2PL model against the normal 3PL alternative (ETS Research Rep. No. RR-06-14). Princeton, NJ: ETS.

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

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), 3-24.

Harris, D.J., & Crouse, J.D. (1993). A study of criteria used in equating. *Applied Measurement in Education, 6*(3), 195-240.

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

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

Holland, P. W. (1990). The Dutch identity: A new tool for the study of item response models. *Psychometrika, 55,* 5–18.

Holland, P. W., & Dorans, N. J. (2006). *Linking and equating.* In R. L. Brennan (Ed.), Educational measurement (4th ed., pp. 187-220).Westport, CT: Praeger.

Huggins, A. C. (2014). The effect of differential item functioning in anchor items on populationinvariance of equating. *Educational and Psychological Measurement, 74*(4), 627–658.

Kaskowitz, G.S. and De Ayala, R.J. (2001). The Effect of Error in Item Parameter Estimates on the Test Response Function Method of Linking. *Applied Psychological Measurement, 25* (1), 39-52.

Keller, L.A. (2000): *Ability Estimation Procedures in Computerized Adaptive Testing* (AICPA technical report). Ewing: The American Institute of Certified Public Accountants. AICPA.

Kim, S., & Hanson, B. A. (2002). Test equating under the multiple-choice model. *Applied Psychological Measurement, 26*, 255-270.

Kim, S., & Cohen, A. S. (1992). Effects of linking methods on detection of DIF. *Applied Psychological Measurement, 29*(1), 51-56.

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

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

Kim, S., & Kolen, M. J. (2004). STUIRT: A computer program for scale transformation under unidimensional item response theory models (Version 1.0). Iowa Testing Programs, University of Iowa.

Klein, L. W., & Kolen, M. J. (1985, April). *Effect of number of common items in common-item equating with nonrandom groups.* Paper presented at the meeting of the American Educational Research Association, Chicago.

Kolen, M. J. (1981). Comparison of traditional and item response theory methods for equating tests. *Journal of Educational Measurement, 18*(1), 1-11.

Kolen, M. J., & Brennan, R. L. (2014). Test equating, scaling, and linking. Methods and practices (3rd Ed.), New York, NY. Springer.

Lee, W., & Ban, J. C. (2010). A comparison of IRT linking procedures. *Applied Measurement in Education, 23,* 23-48.

Little, R.J.A. (1988). A Test of Missing Completely at Random for Multivariate Data with Missing Values. *Journal of the American Statistical Association, 83,* 1198-1202.

Livingston, S. A., Dorans, N. J., & Wright, N. K. (1990). What combination of sampling and equating methods works best? *Applied Measurement in Education, 3*(1), 73–95.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings." *Applied Psychological Measurement, 8,* 453-461.

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems.* New Jersey: Lawrence Erlbaum.

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

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

McKinley, R. L., & Reckase, M. D. (1981*). A comparison of procedures for constructing large item pools* (Research Report 81-3). Columbia MO: University of Missouri, Department of Educational Psychology.

Meyer, J.P. (2013). jMetrik version 3 [computer software]. Retrieved from www.ItemAnalysis.com.

Mislevy, R. J., & Bock, R. D. (2000). BILOG 3.2: Item analysis and test scoring with binary logistic models [Computer program]. Mooresville, IN: Scientific Software, Inc.

Miyazaki, K., Hoshino, T., Mayekawa, S. & Shigemasu, K. (2009). A new concurrent calibration method for nonequivalent group design under nonrandom assignment. *Psychometrika, 74*, 1-19.

Morris, G. M. (1982). On the foundations of test equating. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 169–191). New York: Academic Press.

Muraki, E., & Bock, R. D. (1996*). PARSCALE: IRT based test scoring and item analysis for graded open-ended exercises and performance tasks (Version 3)* [Computer software]. Chicago: Scientific Software.

Naylor, T. H., Balintfy, J. L., Burdick, D. S., & Chu, K. (1968). *Computer simulation techniques*. New York: Wiley.

Ogasawara, H. (2000). Asymptotic standard errors of IRT equating coefficients using moments. *Economic Review* (Otaru University of Commerce), 51, 1–23

Ogasawara, H. (2001). Standard Errors of Item Response Theory Equating/Linking by Response Function Methods. Applied Psychological Measurement, 25 (1), 53 67.

Petersen, N. S., Cook, L. L., & Stocking, M. L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational Statistics, 8*, 137-156.

Powers, S. J., & Kolen, M. J. (2014). Evaluating equating accuracy and assumptions for groups that differ in performance. *Journal of Educational Measurement, 51*(1), 39-56.

Raju, N. S., Edwards, J. E., & Osberg, D. W. (1983, April). *The effect of anchor test size in vertical equating with the Rasch and three-parameter models.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Montreal.

Raykov, T. (2012). Estimation of Latent Construct Correlations in the Presence of Missing Data: A Note on a Latent Variable Modeling Approach. *British Journal of Mathematical and Statistical Psychology, 65,* 19-31.

Rubin, D.B. (1996) Multiple imputation after 18+ years (with discussion). *Journal of the American Statistical Association, 91*, 473-489.

Schafer, J. L. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research, 8,* 3–15.

Schafer, J.L. & J.W. Graham (2002). "Missing Data: Our View of the State of the Art. *Psychological Methods, 7*(2), 147-177.

Stocking, M.L., & Lord, F.M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7,* 201–210.

Stocking, M. L., Eignor, D. R., & Cook, L. L. (1988*). Factors affecting the sample invariant properties of linear and curvilinear observed- and true-score equating procedures* (ETS Research Report RR-88-41). Princeton, NJ: Educational Testing Service.

Thissen, D., Chen, W-H, & Bock, R.D. (2003). Multilog (version 7) [Computer sotware]. Lincolnwood, IL: Scientific Software International.

Thissen, D., & Orlando, M. (2001). Item response theory for items scored in two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (p. 73-140). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

von Davier, A.A., & Wilson, C. (2007). IRT true-score test equating: A guide through assumptions and applications. *Educational and Psychological Measurement, 67,* 940-957

Way, W. D., and 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.

Weeks, J.P. (2010). plink: An R Package for Linking Mixed-Format Tests Using IRT-Based Methods. *Journal of Statistical Software, 35*(12), 1-33.

Wingersky, M. S., Cook, L. L. & Eignor, D. R. (1987). *Specifying the characteristics of linking items used for item response theory item calibration.* (ETS research Report 87 – 24). Princeton, NJ: Educational Testing Service.

Wright, B. D. (1977) Solving measurement problems with the Rasch model. Journal of Educational Measurement, 14(2), 97-116.

Wright, N. K., & Dorans, N. J. (1993). *Using the selection variable for matching or equating* (Research Report 93-4). Princeton, NJ: Educational Testing Service.

Yuan, K.-H., Yang-Wallentin, F., & Bentler, P. M. (2012). ML versus MI for Missing Data with Violation of Distribution Conditions. *Sociological Methods & Research, 41*(4), 598–629.

Appendices

Appendix A.

Table 6 Relative MSE to CCMI(2PL)

case	CCMI	Relative MSE to CCMI				
		CC	S-L	M/S	M/M	μ(MSE)
EIS	0.0556	1.00	1.00	1.04	1.02	1.01
EIL	0.0475	1.01	1.01	1.03	1.02	1.01
ENS	0.0528	0.97	0.98	1.01	1.00	0.99
ENL	0.0394	1.01	1.02	1.07	1.06	1.03
EWS	0.0766	1.01	1.00	1.03	1.01	1.01
EWL	0.068	1.02	1.00	1.01	1.01	1.01
DIS	0.0588	0.99	1.00	1.02	1.02	1.01
DIL	0.0525	1.00	0.99	1.01	1.00	1.00
DNS	0.0489	0.99	1.01	1.16	1.95	1.22
DNL	0.0427	0.99	0.99	1.22	1.38	1.12
DWS	0.0813	1.01	0.99	1.01	1.00	1.00
DWL	0.0768	1.02	0.97	0.98	0.98	0.99
average	0.0602	1.00	0.99	1.05	1.17	1.04

Table 7 BIAS (2PL)

		BIAS				
case	CCMI	CC	S-L	M/S	M/M	μ(BIAS)
EIS	-0.0093	-0.0105	-0.0059	-0.007	-0.0096	-0.0085
EIL	-0.0222	-0.0311	-0.0247	-0.0251	-0.0261	-0.0258
ENS	-0.0799	-0.0735	-0.0778	-0.0773	-0.0757	-0.0768
ENL	-0.0212	-0.0311	-0.024	-0.0237	-0.0192	-0.0238
EWS	-0.0101	-0.012	-0.0071	-0.0073	-0.0086	-0.009
EWL	-0.0232	-0.0312	-0.0247	-0.0248	-0.0259	-0.026
DIS	-0.017	-0.0151	-0.0103	-0.0115	-0.0106	-0.0129
DIL	-0.0404	-0.0379	-0.0306	-0.0273	-0.031	-0.0335
DNS	-0.0178	-0.0136	-0.0067	-0.0109	-0.0181	-0.0134
DNL	-0.0437	-0.0385	-0.033	-0.046	0.0013	-0.032
DWS	-0.0237	-0.0237	-0.0078	-0.008	-0.0072	-0.0141
DWL	-0.0454	-0.0454	-0.0306	-0.0273	-0.0326	-0.0363
average	-0.0313	-0.029	-0.0198	-0.0218	-0.0164	-0.0237

Table 8 Relative MSE to CCMI(3PL)

case	CCMI	Relative MSE to CCMI				
		CC	S-L	M/S	M/M	μ(MSE)
EIS	0.1271	0.98	0.82	1.22	1.52	1.11
EIL	0.096	0.99	0.93	1.23	1.17	1.06
ENS	0.278	1.26	0.22	0.29	0.32	0.62
ENL	0.0838	1.07	0.78	0.96	4.20	1.60
EWS	0.1766	0.98	0.98	1.24	1.38	1.12
EWL	0.1514	0.99	1.00	1.20	1.08	1.06
DIS	0.1242	0.99	0.84	1.17	1.61	1.12
DIL	0.1088	0.91	0.94	1.17	1.03	1.01
DNS	0.1675	1.71	0.47	0.92	4.45	1.71
DNL	0.0566	0.99	0.99	1.21	1.64	1.17
DWS	0.1773	1.02	0.97	1.24	1.26	1.10
DWL	0.1553	1.03	1.01	1.18	1.14	1.07
average	0.1316	1.15	0.85	1.14	1.97	1.22

Table 9 BIAS (3PL)

		BIAS				
case	CCMI	CC	S-L	M/S	M/M	μ(BIAS)
EIS	-0.0712	-0.0751	-0.0216	-0.0451	-0.0766	-0.0579
EIL	-0.0743	-0.0741	-0.0168	-0.0596	-0.0553	-0.056
ENS	-0.3325	-0.3423	0.6923	0.1067	0.201	0.0651
ENL	-0.1386	-0.1503	0.0734	-0.0471	-0.0111	-0.0547
EWS	-0.0392	-0.0409	-0.042	-0.0338	-0.028	-0.0368
EWL	-0.0555	-0.0662	-0.0464	-0.056	-0.0552	-0.0559
DIS	-0.1033	-0.1006	-0.0299	-0.0938	-0.0942	-0.0844
DIL	-0.1349	-0.1132	-0.0529	-0.1446	-0.1024	-0.1096
DNS	-0.365	-0.3629	0.0613	-0.0324	0.431	-0.0536
DNL	-0.0244	-0.0167	0.0423	-0.0449	0.1093	0.0131
DWS	-0.0619	-0.076	-0.049	-0.0457	-0.0507	-0.0567
DWL	-0.066	-0.0816	-0.0693	-0.0691	-0.0727	-0.0717
average	-0.1259	-0.1252	-0.0162	-0.0717	0.0367	-0.0605

Table 10 Mean Squared Errors (2PL) by forms

case	Form	Avg. Score μ(σ)	Avg. Anchor Score μ(σ)	MSE				
				CC	CCMI	MM	MS	SL
EIS	Base	30.78 (9.15)	5.70 (1.53)	0.0952	0.0956	0.0954	0.0954	0.0954
	Target	30.49 (8.91)	5.67 (1.47)	0.1249	0.1271	0.1934	0.1548	0.1039
EIL	Base	37.14 (11.13)	12.67 (3.49)	0.0836	0.0844	0.0841	0.0841	0.0841
	Target	37.58 (10.76)	12.54 (3.24)	0.0947	0.096	0.1126	0.1181	0.0893
ENS	Base	30.49 (8.61)	5.56 (1.87)	0.1433	0.1062	0.1137	0.1137	0.1137
	Target	30.23 (5.64)	5.53 (1.13)	0.349	0.278	0.0895	0.0806	0.0617
ENL	Base	37.10 (11.13)	12.66 (3.49)	0.0834	0.0842	0.0834	0.0834	0.0834
	Target	37.74 (6.31)	12.55 (1.84)	0.0894	0.0838	0.3519	0.0804	0.0653
EWS	Base	30.78 (9.15)	5.70 (1.53)	0.0953	0.0955	0.0956	0.0956	0.0956
	Target	30.42 (11.35)	5.71 (1.98)	0.1733	0.1766	0.2435	0.2188	0.1732
EWL	Base	37.09 (11.11)	12.66 (3.49)	0.0825	0.0828	0.0834	0.0834	0.0834
	Target	37.32 (13.60)	12.49 (4.21)	0.1498	0.1514	0.1639	0.182	0.1519

Table 11 BIAS (2PL) by forms

case	Form	Avg. Score μ(σ)	Avg. Anchor Score μ(σ)	BIAS				
				CC	CCMI	MM	MS	SL
EIS	Base	30.78 (9.15)	5.70 (1.53)	-0.0246	-0.0226	-0.0226	-0.0226	-0.0226
	Target	30.49 (8.91)	5.67 (1.47)	-0.0751	-0.0712	-0.0766	-0.0451	-0.0216
EIL	Base	37.14 (11.13)	12.67 (3.49)	-0.0283	-0.0294	-0.0251	-0.0251	-0.0251
	Target	37.58 (10.76)	12.54 (3.24)	-0.0741	-0.0743	-0.0553	-0.0596	-0.0168
ENS	Base	30.49 (8.61)	5.56 (1.87)	-0.0293	-0.0237	-0.0226	-0.0226	-0.0226
	Target	30.23 (5.64)	5.53 (1.13)	-0.3423	-0.3325	0.201	0.1067	0.6923
ENL	Base	37.10 (11.13)	12.66 (3.49)	-0.0315	-0.0283	-0.0251	-0.0251	-0.0251
	Target	37.74 (6.31)	12.55 (1.84)	-0.1503	-0.1386	-0.0111	-0.0471	0.0734
EWS	Base	30.78 (9.15)	5.70 (1.53)	-0.0245	-0.0224	-0.0226	-0.0226	-0.0226
	Target	30.42 (11.35)	5.71 (1.98)	-0.0409	-0.0392	-0.028	-0.0338	-0.042
EWL	Base	37.09 (11.11)	12.66 (3.49)	-0.028	-0.0248	-0.025	-0.025	-0.025
	Target	37.32 (13.60)	12.49 (4.21)	-0.0662	-0.0555	-0.0552	-0.056	-0.0464

Table 12 Mean Squared Errors (3PL) by forms

case	Form	Avg. Score μ(σ)	Avg. Anchor Score μ(σ)	MSE				
				CC	CCMI	MM	MS	SL
DIS	Base	30.77 (9.19)	5.70 (1.52)	0.0952	0.0964	0.0957	0.0957	0.0957
	Target	34.95 (8.46)	6.40 (1.47)	0.123	0.1242	0.2004	0.1459	0.1047
DIL	Base	37.11 (11.12)	12.68 (3.47)	0.0834	0.0892	0.084	0.084	0.084
	Target	43.08 (10.16)	14.22 (3.12)	0.0995	0.1088	0.1125	0.1278	0.102
DNS	Base	30.77 (9.16)	5.70 (1.52)	0.1071	0.095	0.0955	0.0955	0.0955
	Target	35.68 (4.74)	6.42 (0.78)	0.2872	0.1675	0.7457	0.1538	0.0792
DNL	Base	37.06 (10.65)	12.63 (3.87)	0.0882	0.0892	0.0893	0.0893	0.0893
	Target	43.22 (6.34)	15.05 (2.19)	0.056	0.0566	0.0928	0.0685	0.0562
DWS	Base	30.78 (9.15)	5.70 (1.53)	0.0954	0.0957	0.0956	0.0956	0.0956
	Target	34.17 (11.00)	6.38 (1.98)	0.181	0.1773	0.2236	0.2205	0.1723
DWL	Base	37.10 (11.14)	12.66 (3.48)	0.083	0.0834	0.0837	0.0837	0.0837
	Target	41.98 (13.12)	13.94 (4.09)	0.1603	0.1553	0.1774	0.1837	0.1567

Table 13 BIAS (3PL) by forms

case	Form	Avg. Score μ(σ)	Avg. Anchor Score μ(σ)	BIAS				
				CC	CCMI	MM	MS	SL
DIS	Base	30.77 (9.19)	5.70 (1.52)	-0.0251	-0.0225	-0.0227	-0.0227	-0.0227
	Target	34.95 (8.46)	6.40 (1.47)	-0.1006	-0.1033	-0.0942	-0.0938	-0.0299
DIL	Base	37.11 (11.12)	12.68 (3.47)	-0.0281	-0.0479	-0.0251	-0.0251	-0.0251
	Target	43.08 (10.16)	14.22 (3.12)	-0.1132	-0.1349	-0.1024	-0.1446	-0.0529
DNS	Base	30.77 (9.16)	5.70 (1.52)	-0.03	-0.0298	-0.0226	-0.0226	-0.0226
	Target	35.68 (4.74)	6.42 (0.78)	-0.3629	-0.365	0.431	-0.0324	0.0613
DNL	Base	37.06 (10.65)	12.63 (3.87)	0.0356	0.0399	0.0411	0.0411	0.0411
	Target	43.22 (6.34)	15.05 (2.19)	-0.0167	-0.0244	0.1093	-0.0449	0.0423
DWS	Base	30.78 (9.15)	5.70 (1.53)	-0.026	-0.0237	-0.0227	-0.0227	-0.0227
	Target	34.17 (11.00)	6.38 (1.98)	-0.076	-0.0619	-0.0507	-0.0457	-0.049
DWL	Base	37.10 (11.14)	12.66 (3.48)	-0.0287	-0.0261	-0.0251	-0.0251	-0.0251
	Target	41.98 (13.12)	13.94 (4.09)	-0.0816	-0.066	-0.0727	-0.0691	-0.0693

Appendix B.

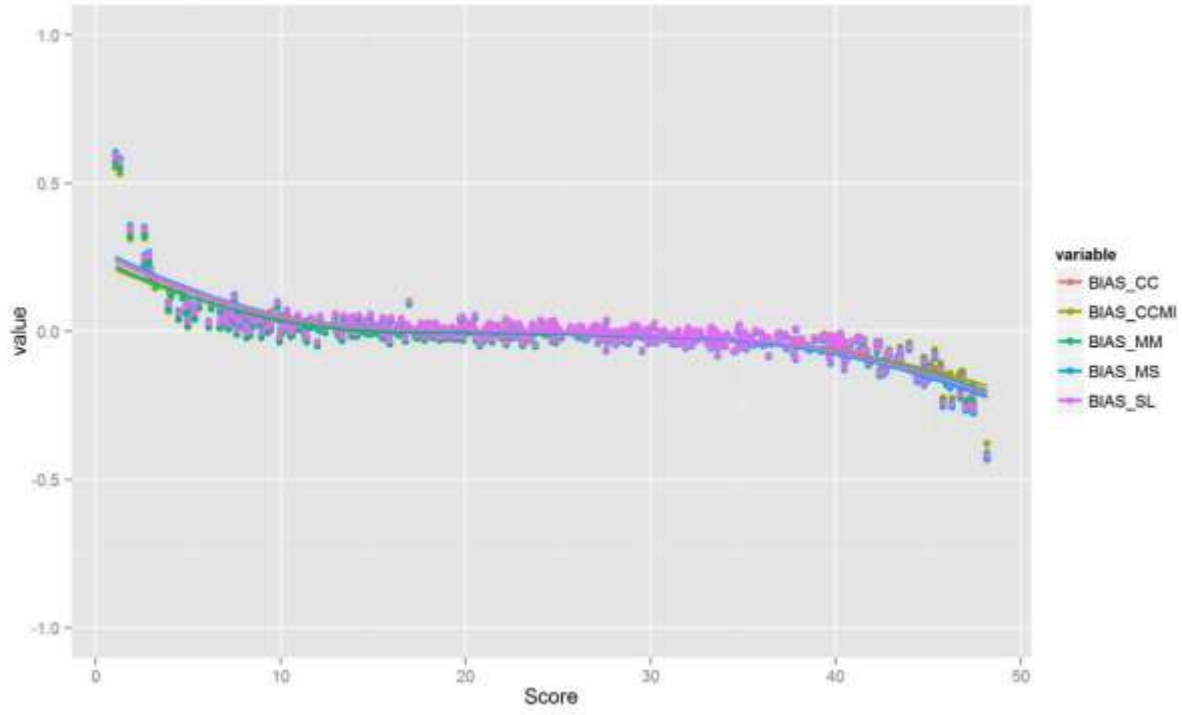

Figure 39 Bias result of condition EIS 2PL Model

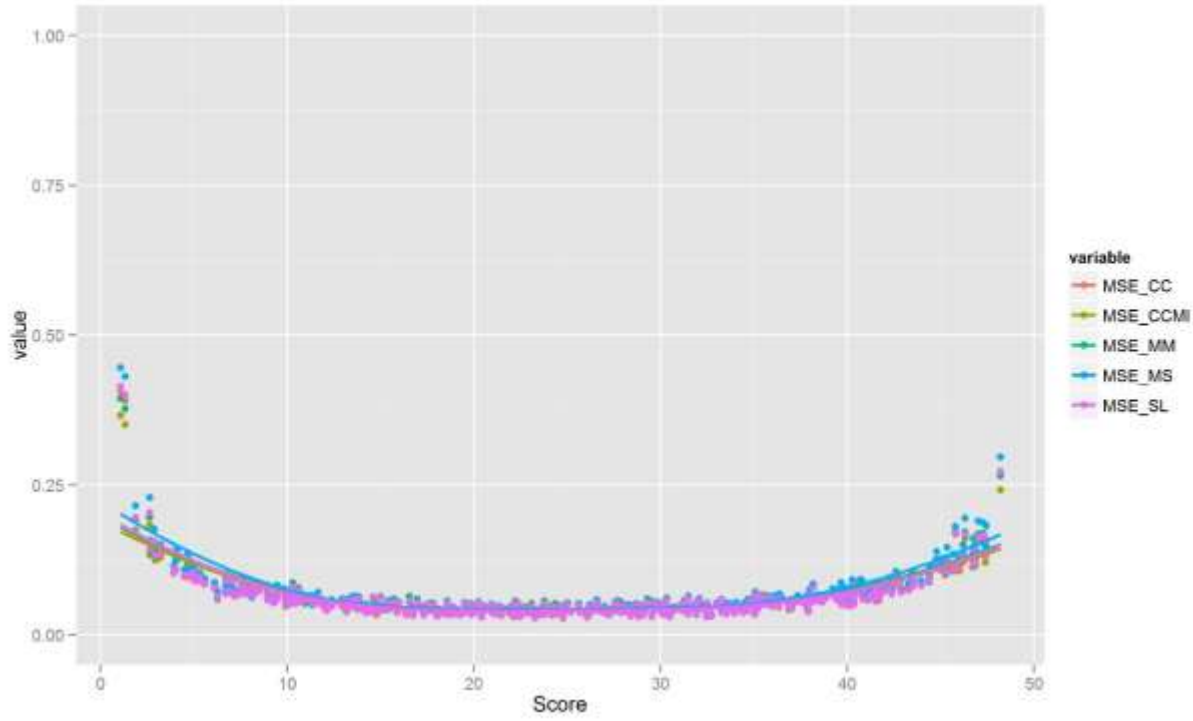

Figure 40 MSE result of condition EIS 2PL Model

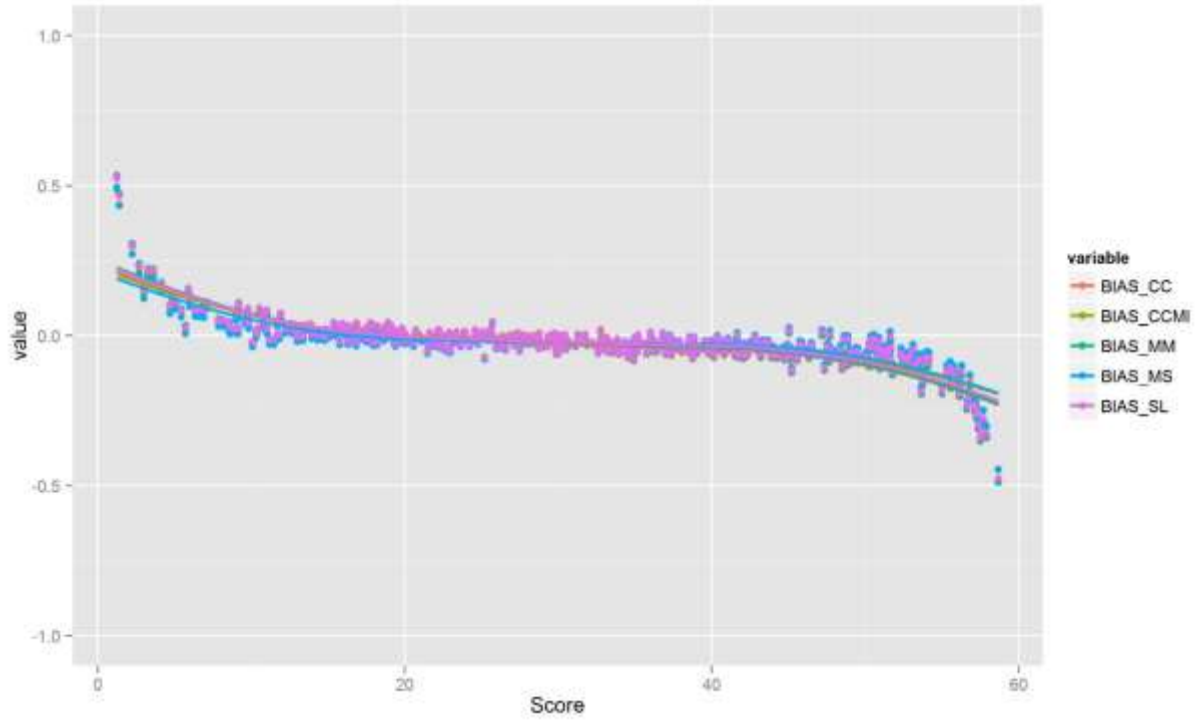

Figure 41 Bias result of condition EIL 2PL Model

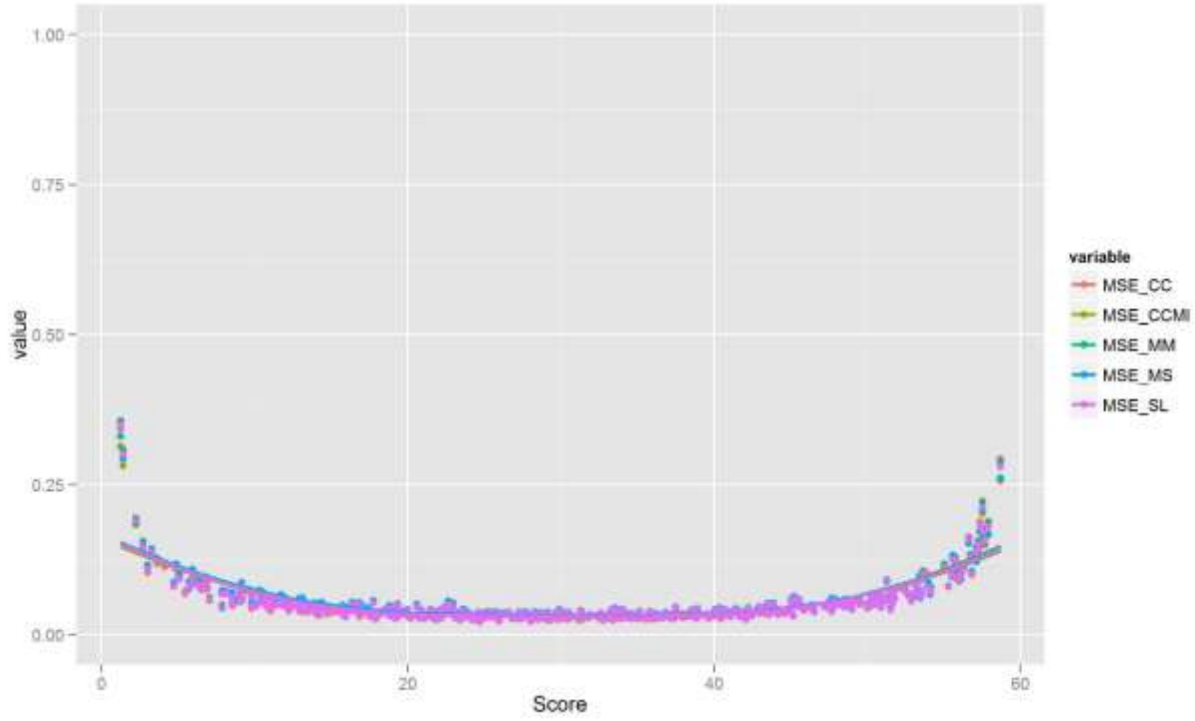

Figure 42 MSE result of condition EIL 2PL Model

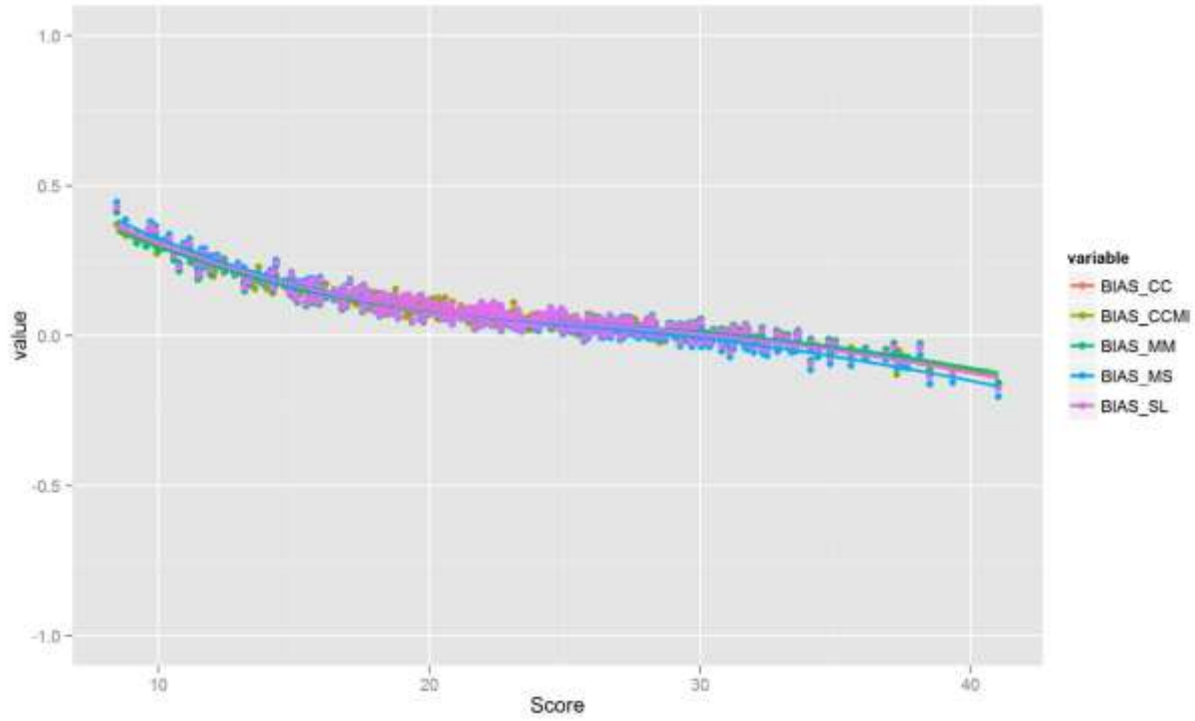

Figure 43 Bias result of condition ENS 2PL Model

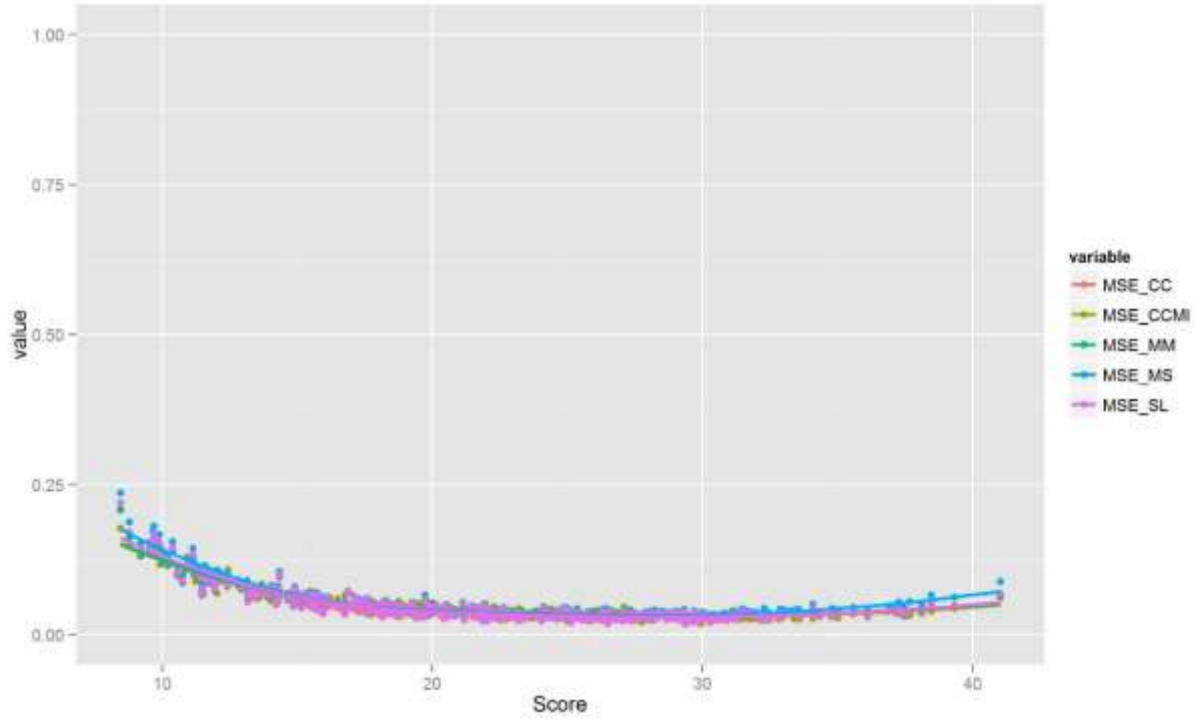

Figure 44 MSE result of condition ENS 2PL Model

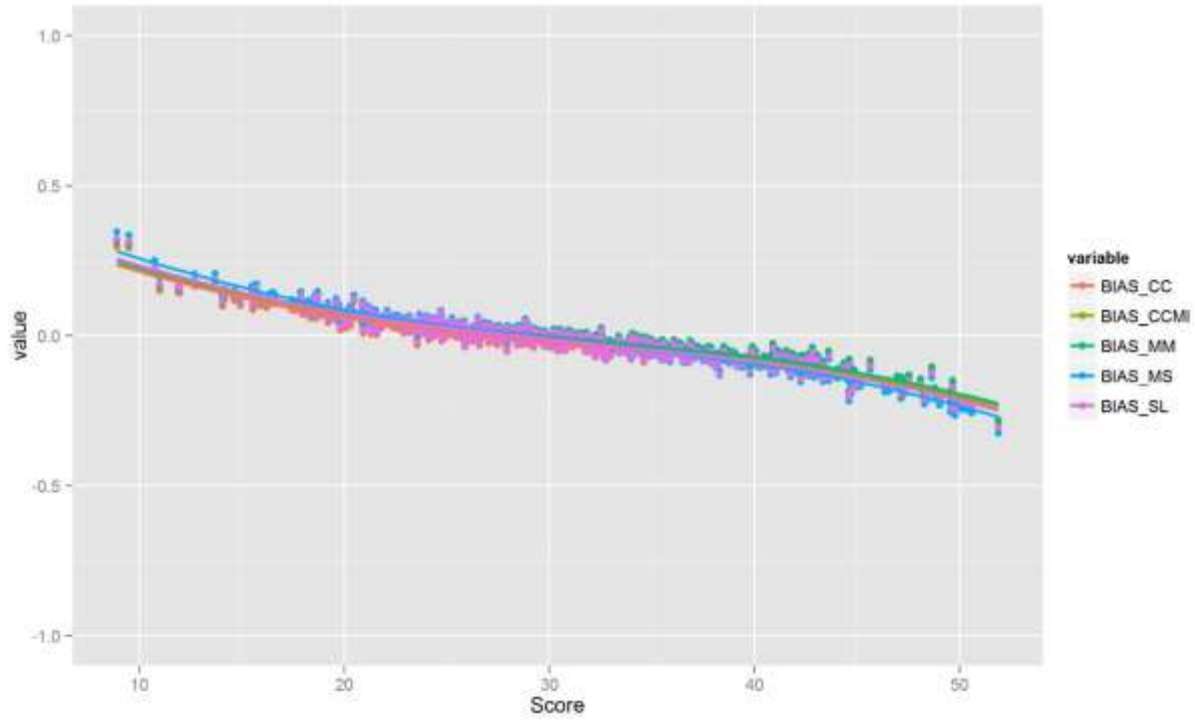

Figure 45 Bias result of condition ENL 2PL Model

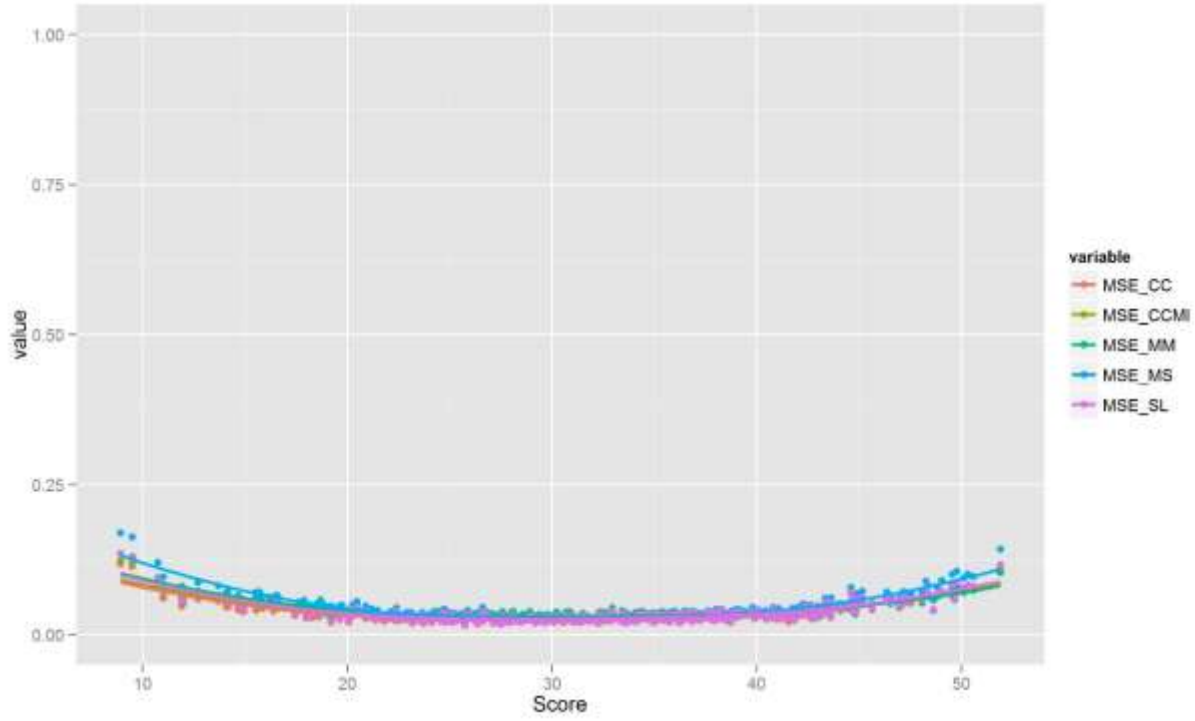

Figure 46 MSE result of condition ENL 2PL Model

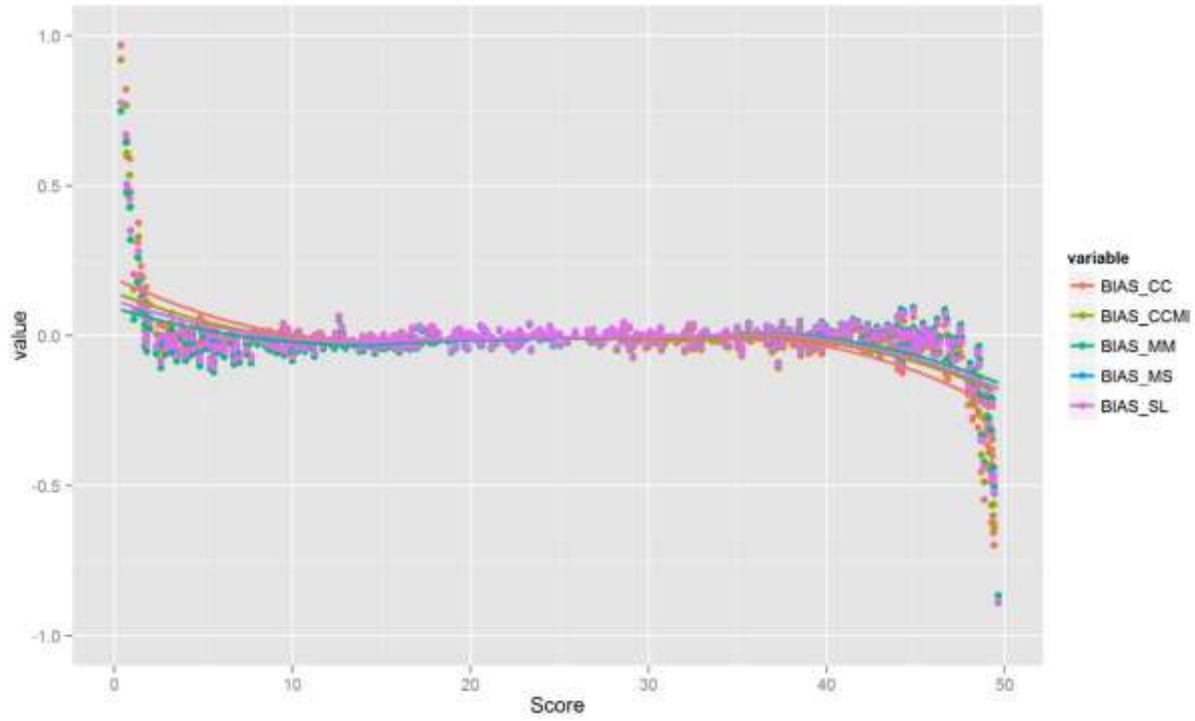

Figure 47 Bias result of condition EWS 2PL Model

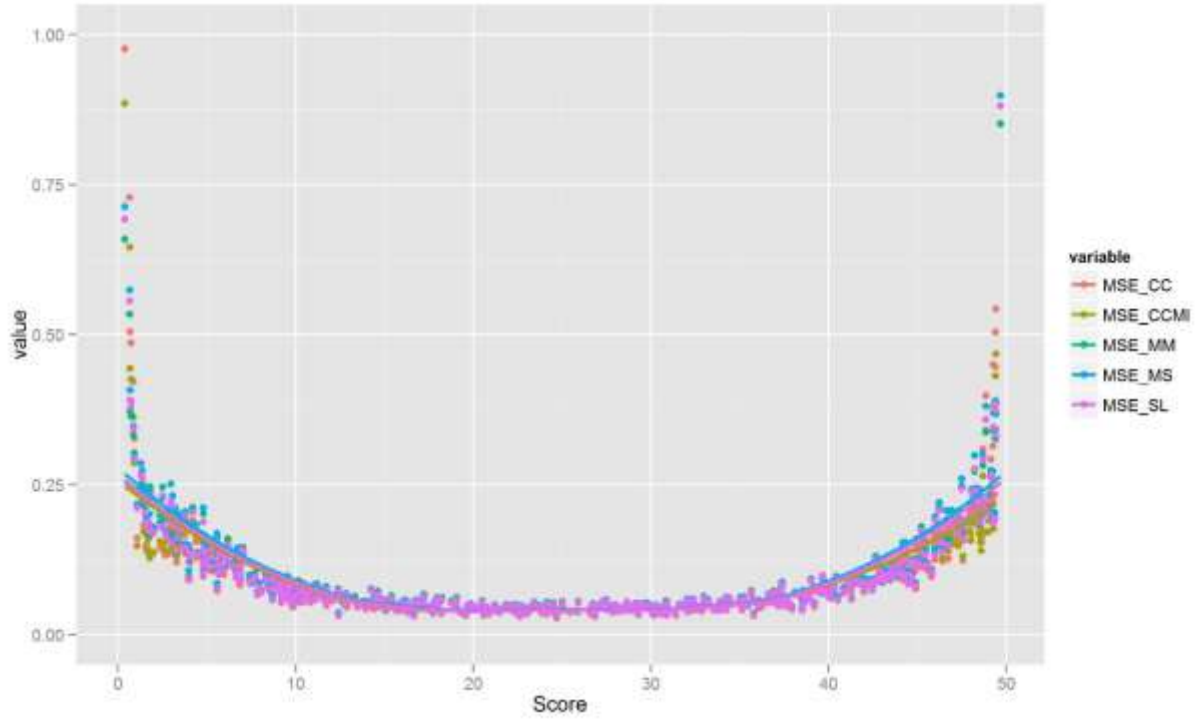

Figure 48 MSE result of condition EWS 2PL Model

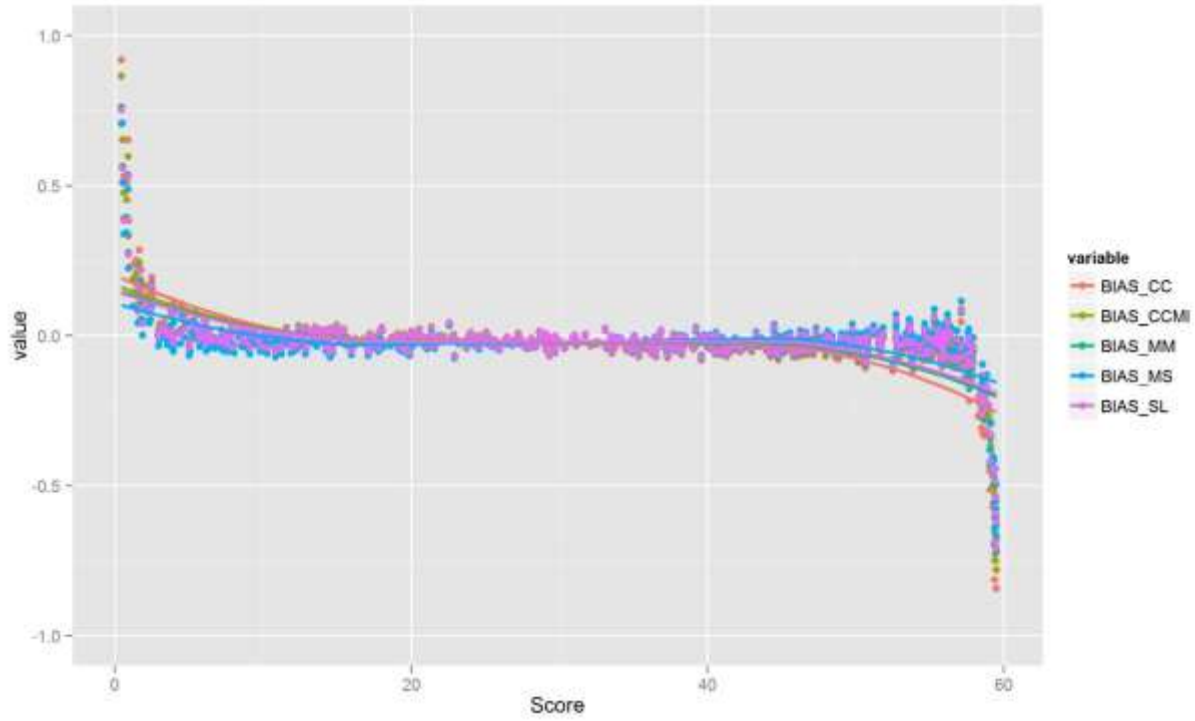

Figure 49 Bias result of condition EWL 2PL Model

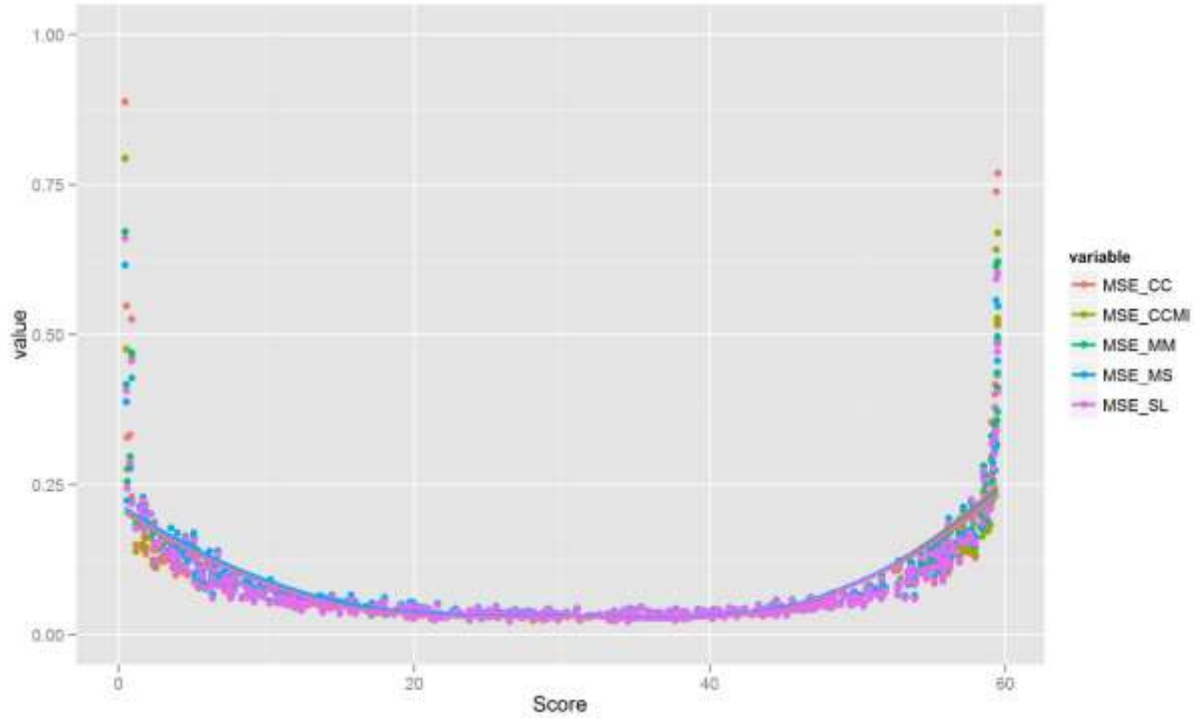

Figure 50 MSE result of condition EWL 2PL Model

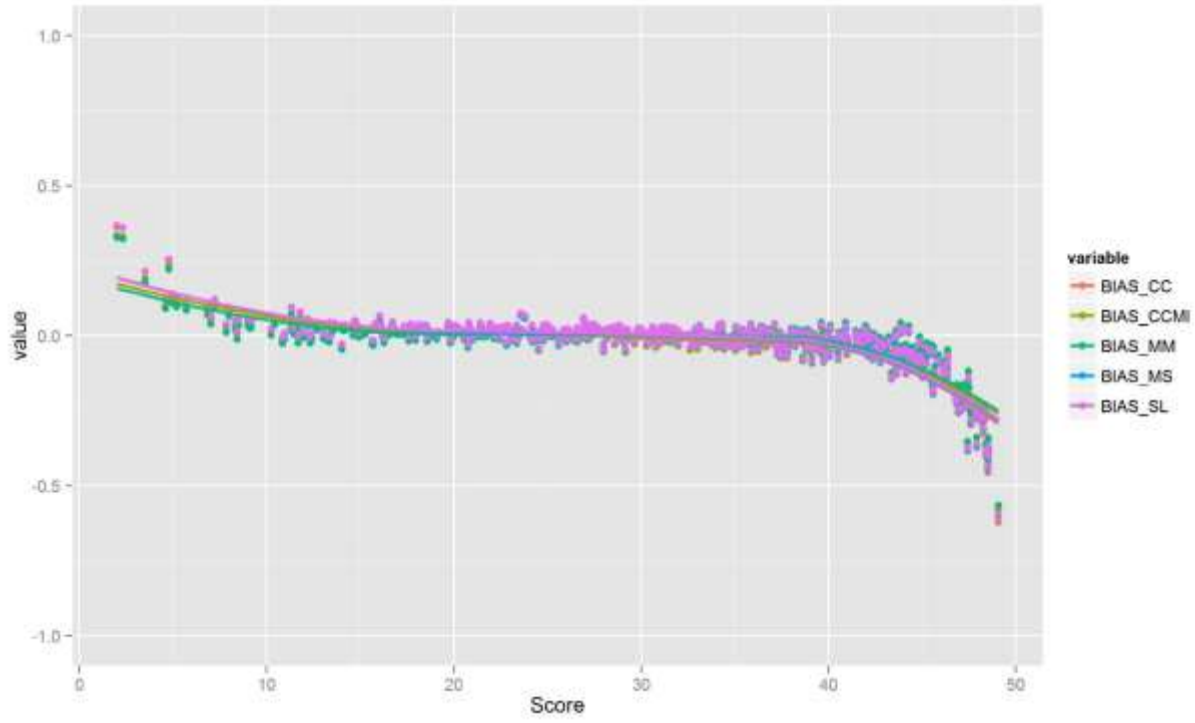

Figure 51 Bias result of condition DIS 2PL Model

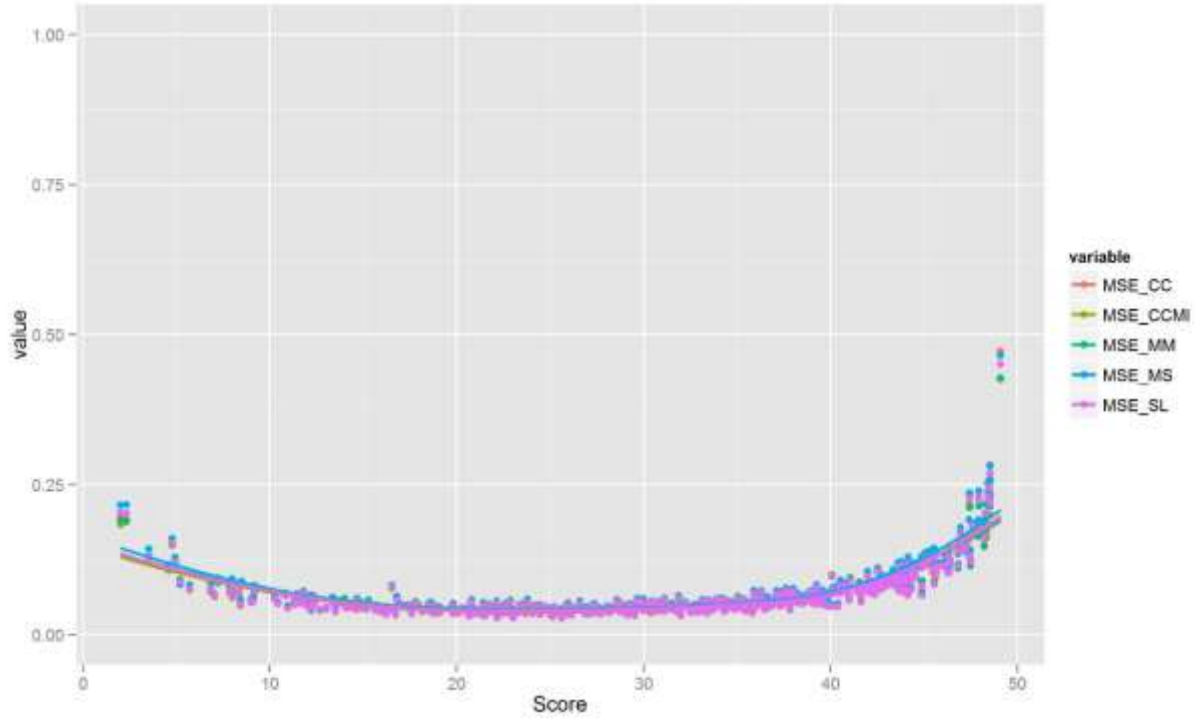

Figure 52 MSE result of condition DIS 2PL Model

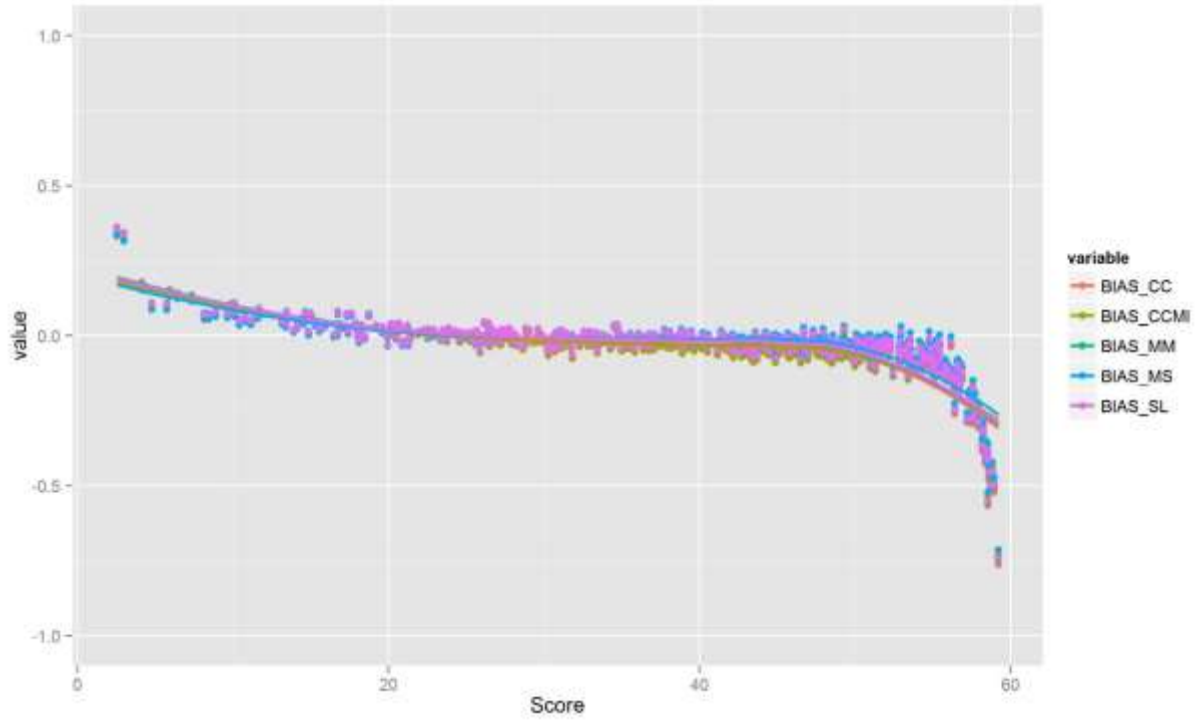

Figure 53 Bias result of condition DIL 2PL Model

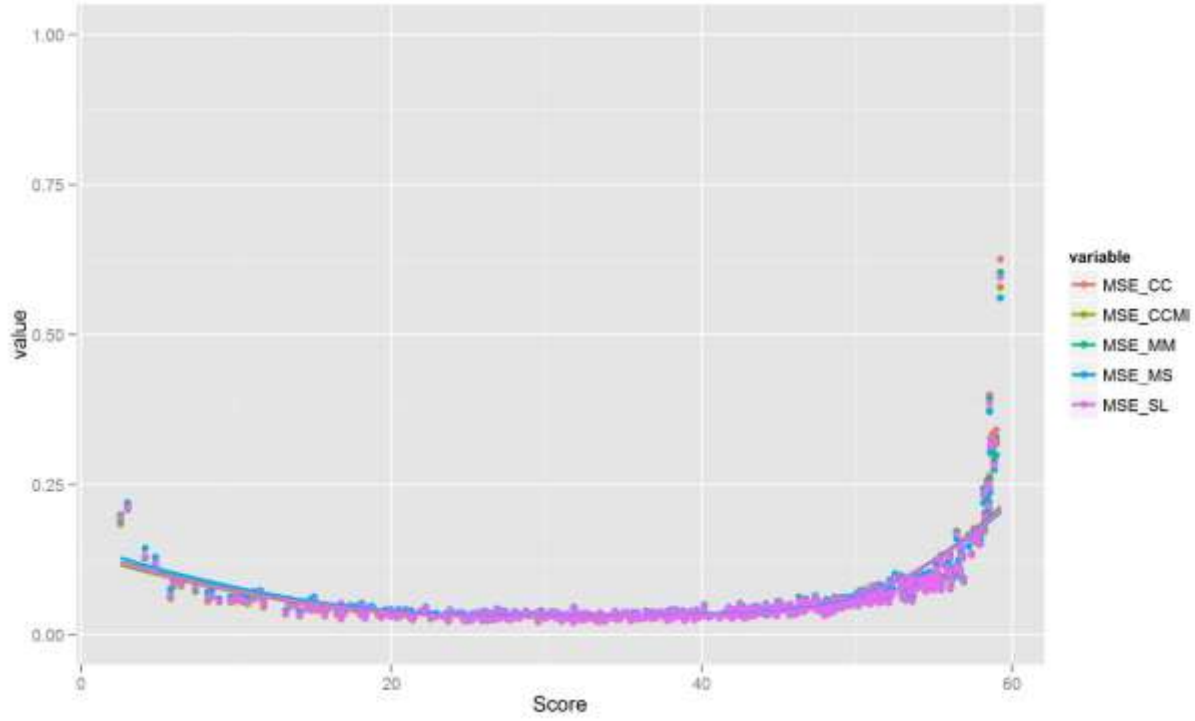

Figure 54 MSE result of condition DIL 2PL Model

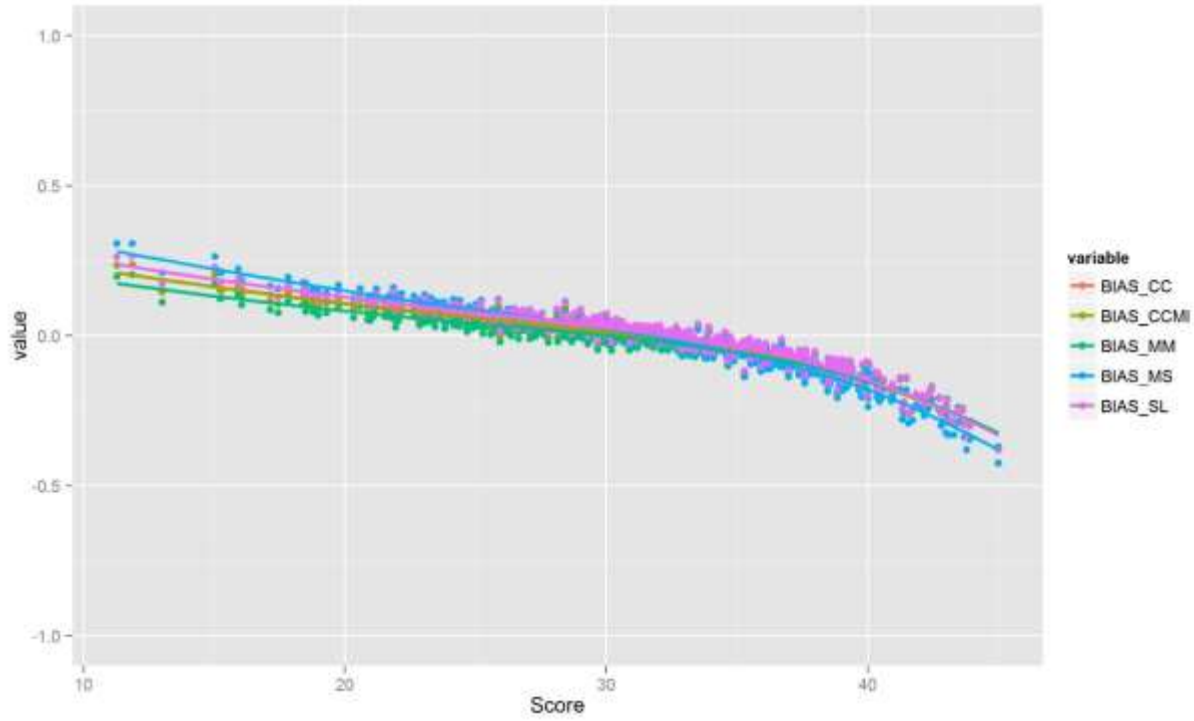

Figure 55 Bias result of condition DNS 2PL Model

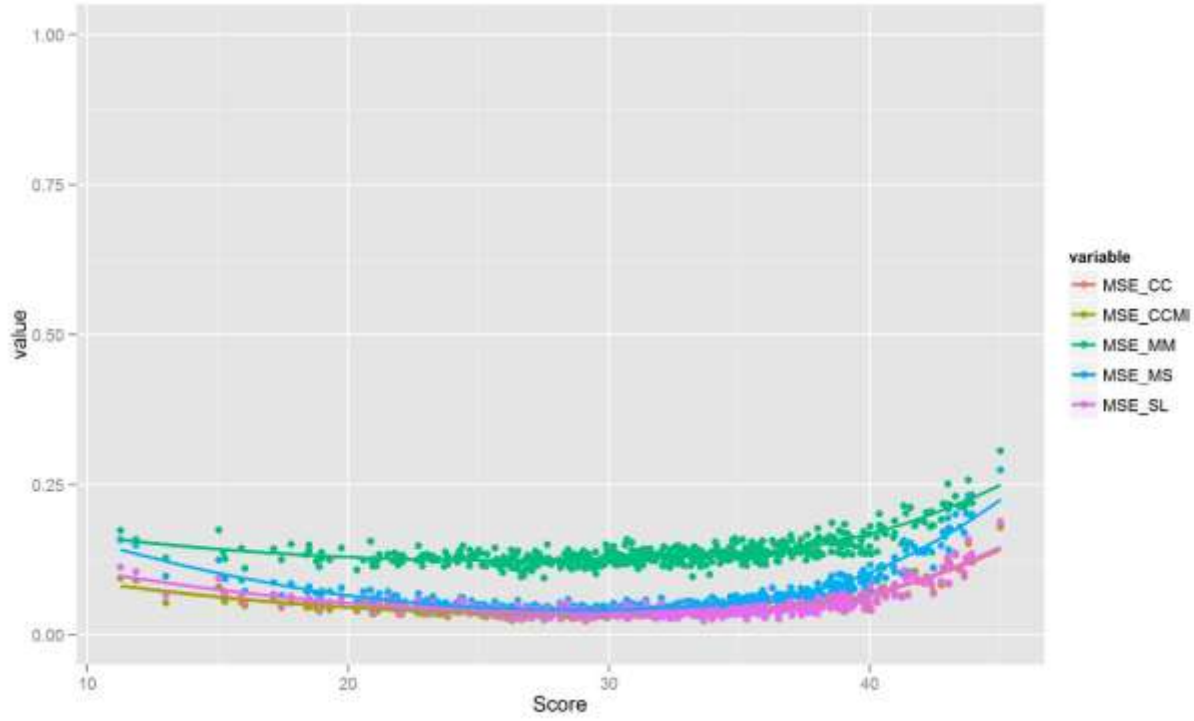

Figure 56 MSE result of condition DNS 2PL Model

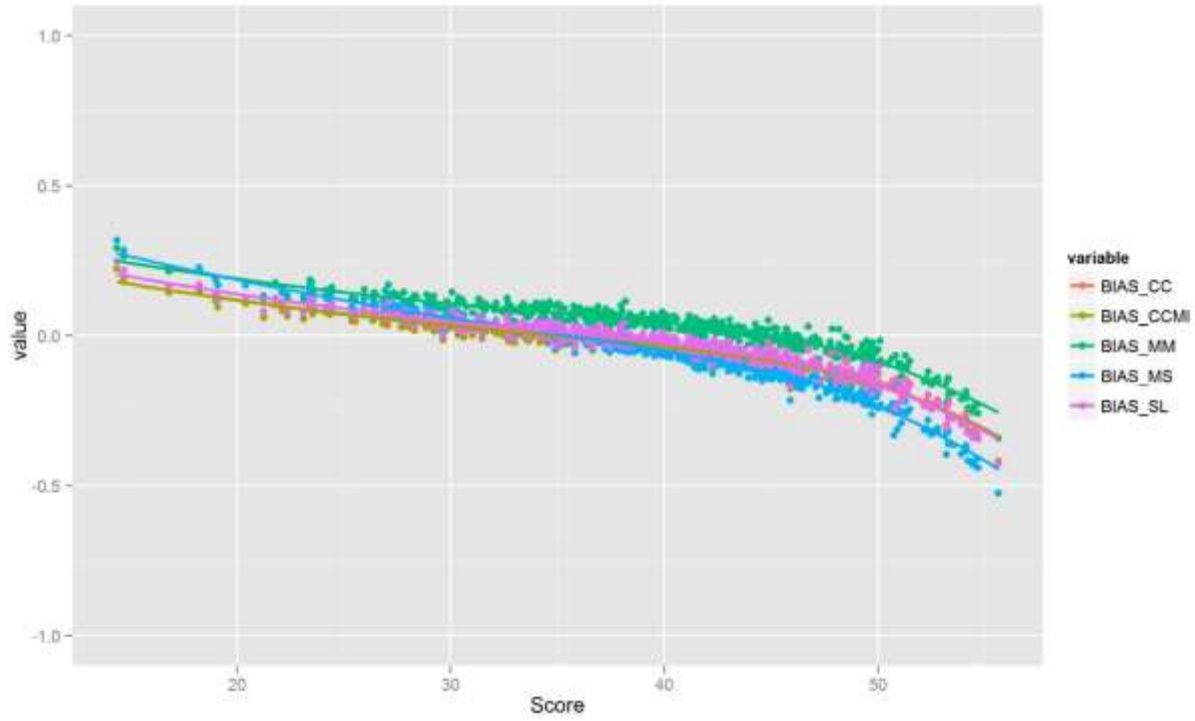

Figure 57 Bias result of condition DNL 2PL Model

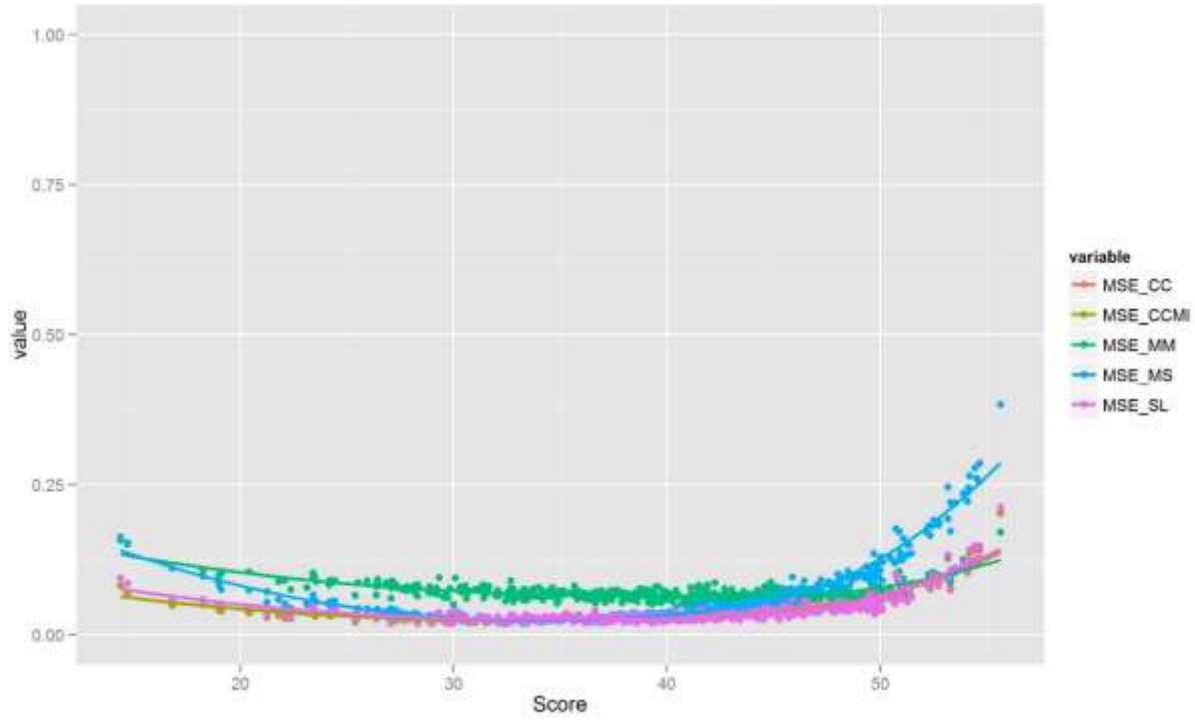

Figure 58 MSE result of condition DNL 2PL Model

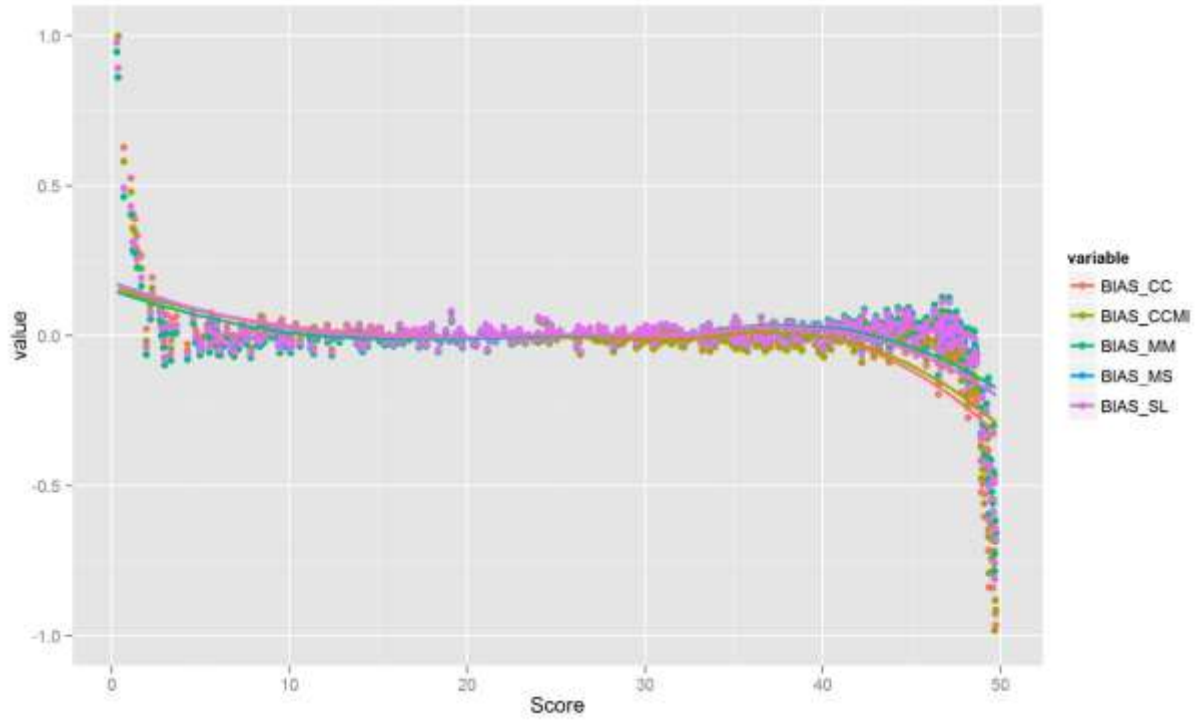

Figure 59 Bias result of condition DWS 2PL Model

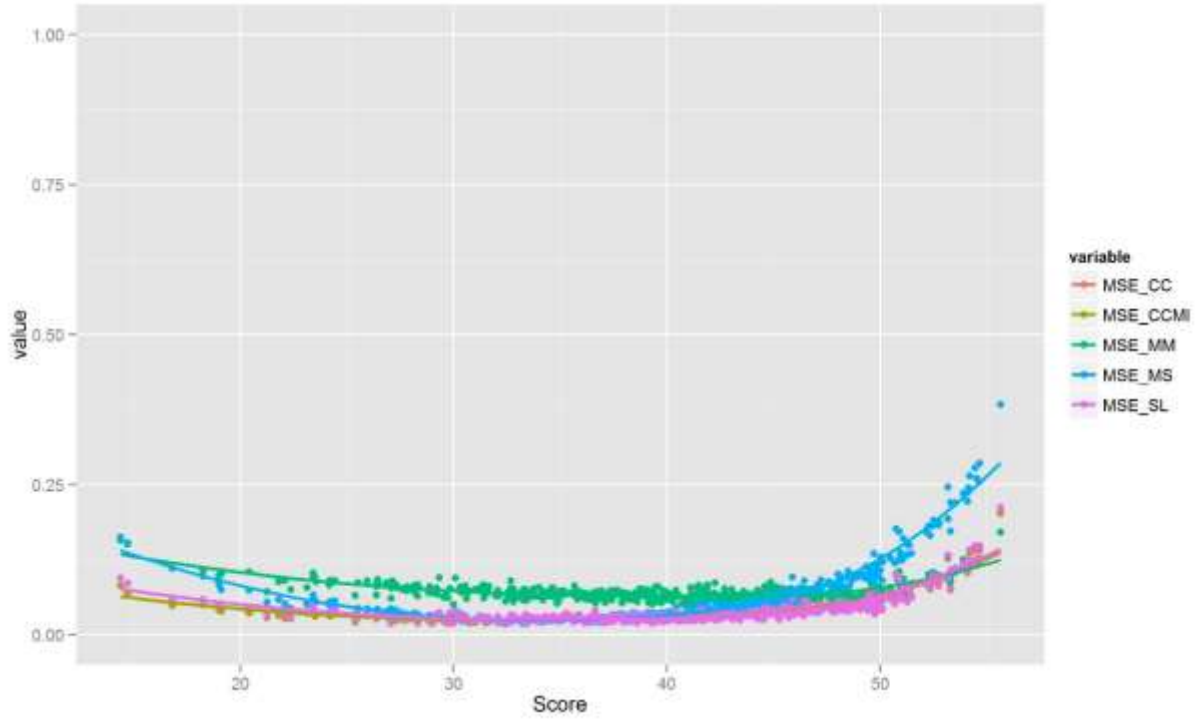

Figure 60 MSE result of condition DWS 2PL Model

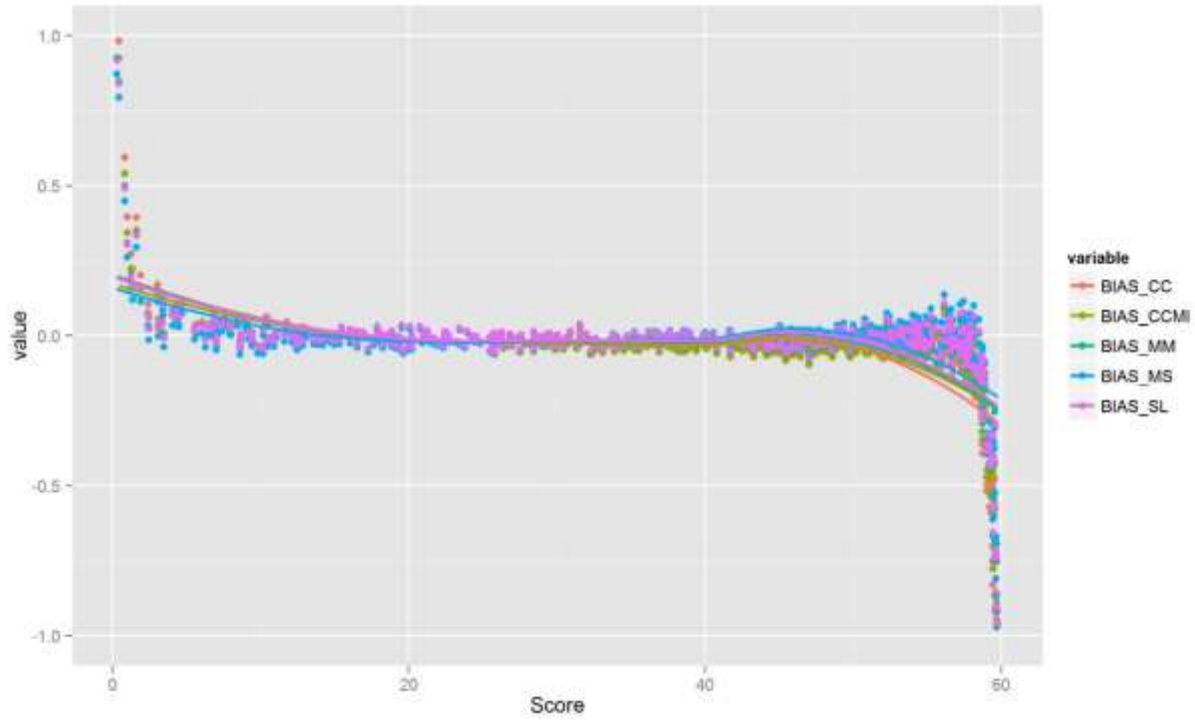

Figure 61 Bias result of condition DWL 2PL Model

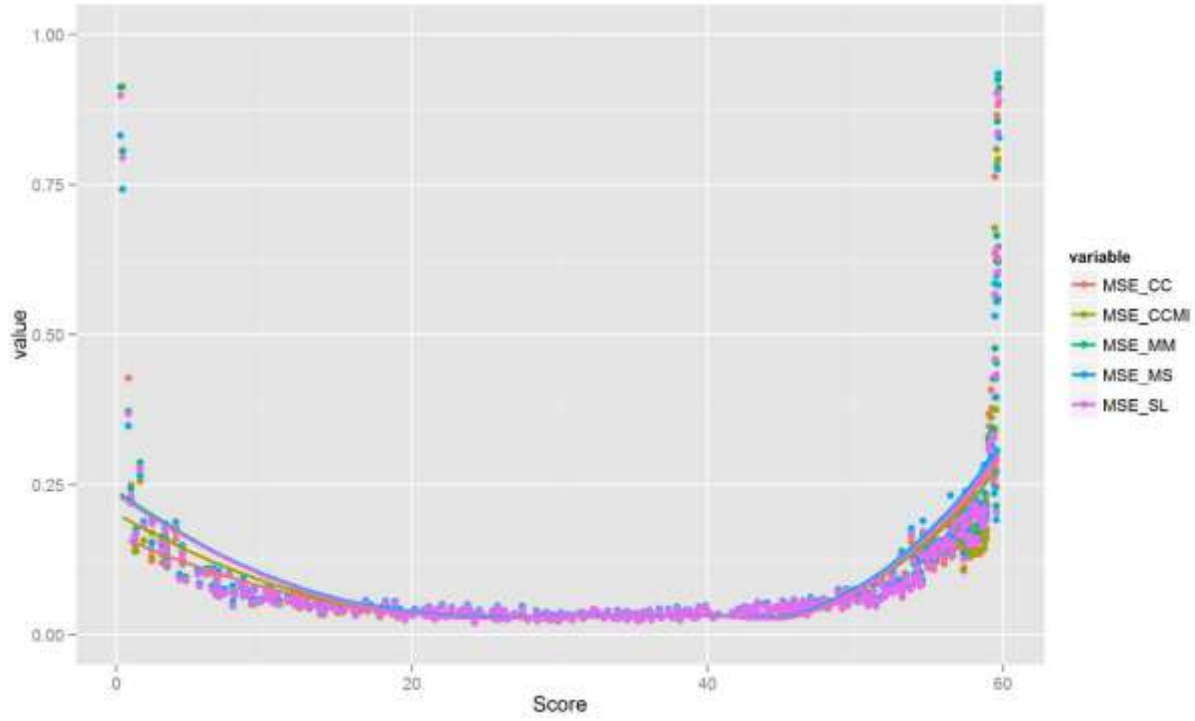

Figure 62 MSE result of condition DWL 2PL Model

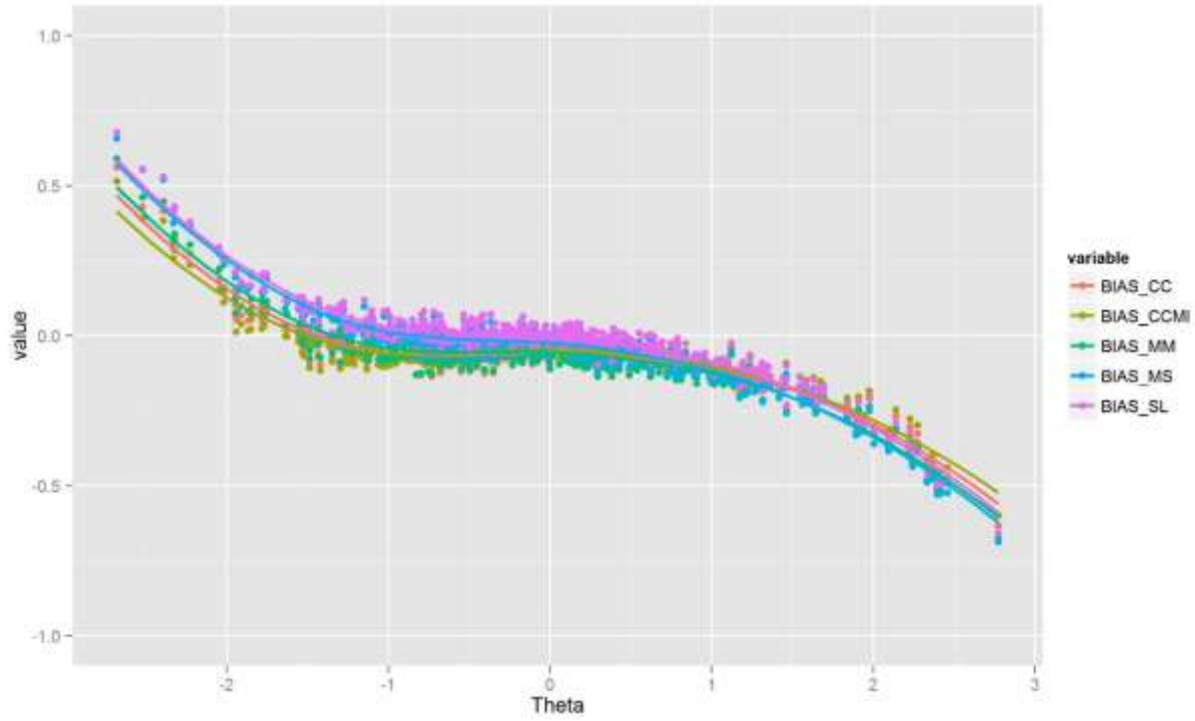

Figure 63 Bias result of condition EIS 3PL Model

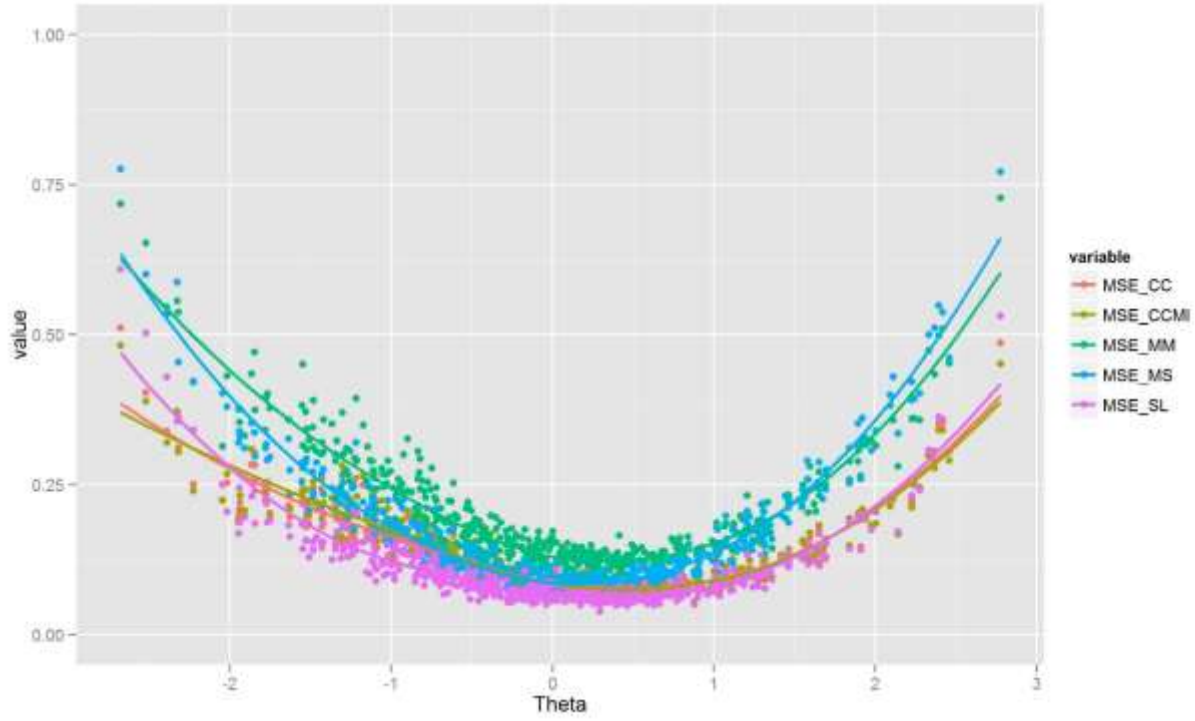

Figure 64 MSE result of condition EIS 3PL Model

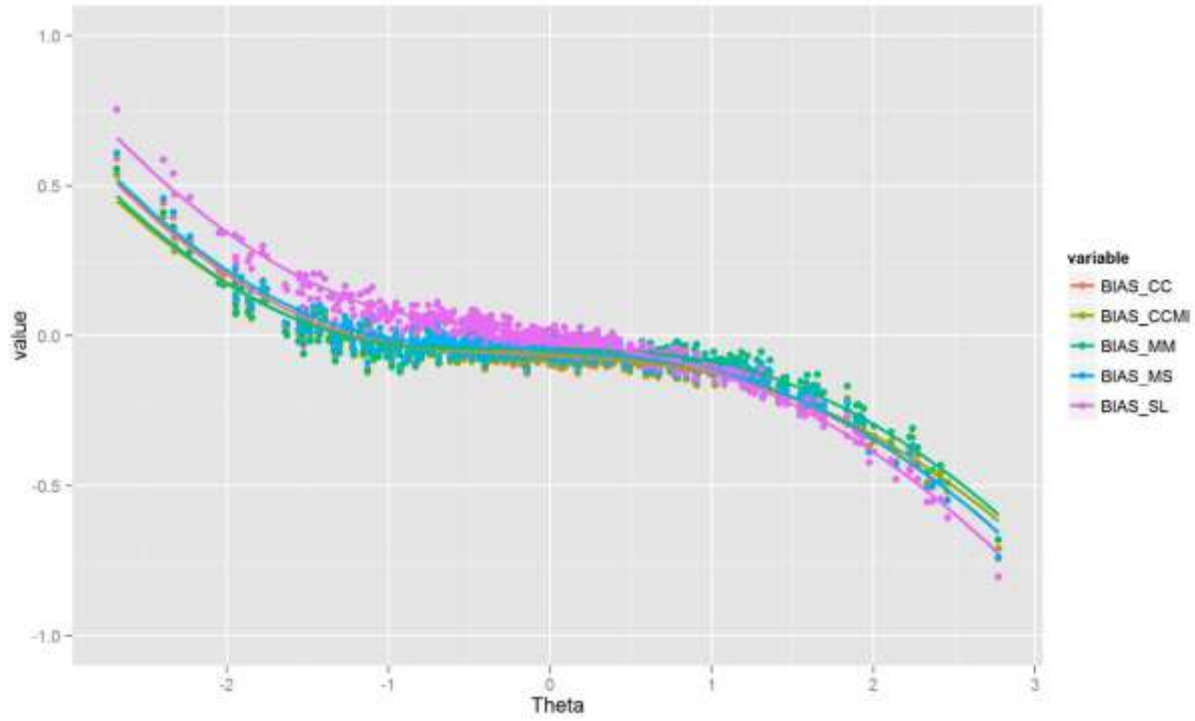

Figure 65 Bias result of condition EIL 3PL Model

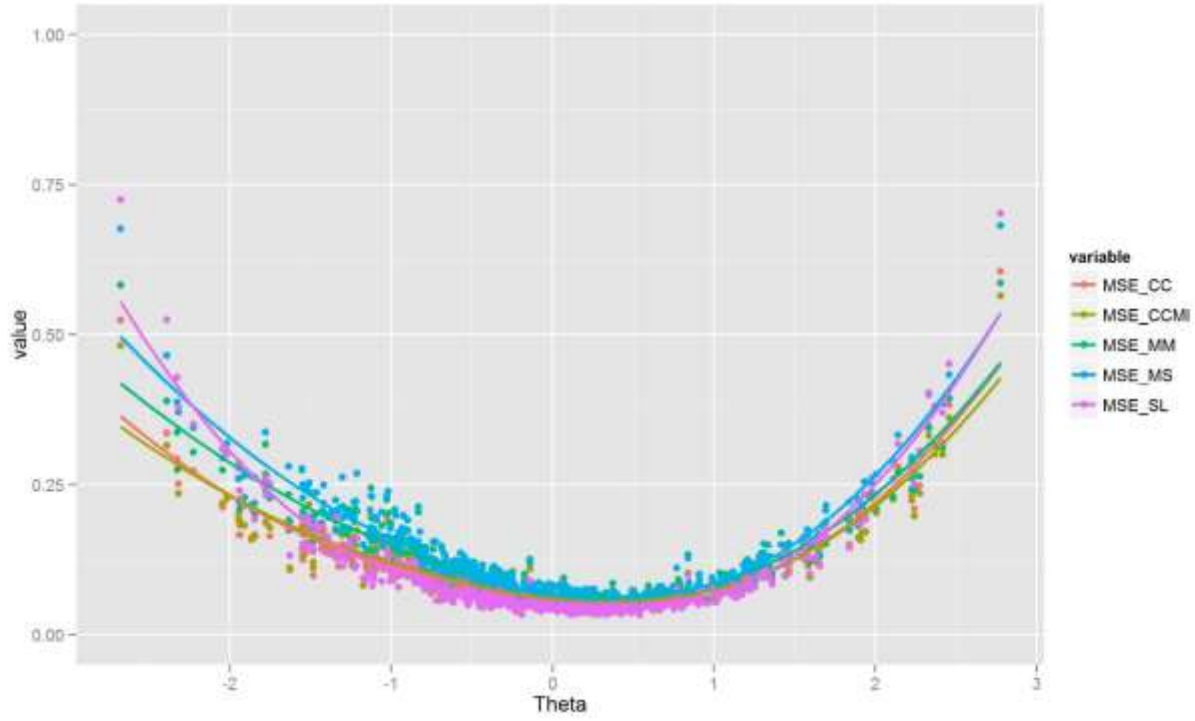

Figure 66 MSE result of condition EIL 3PL Model

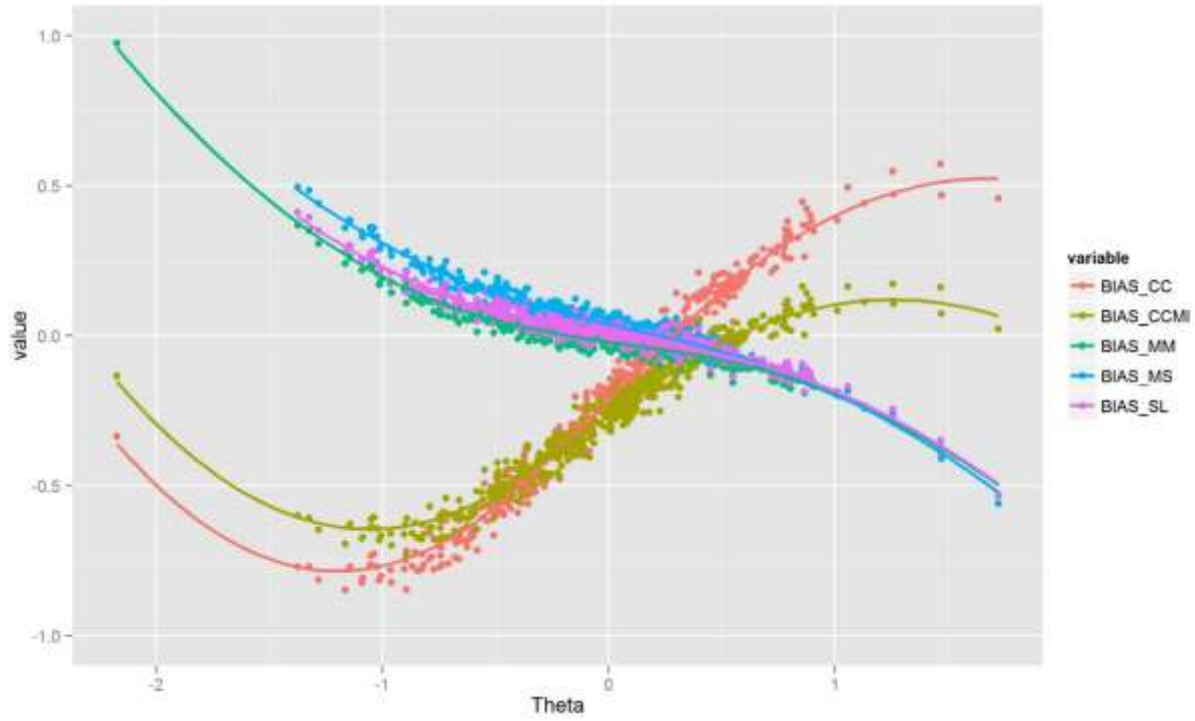

Figure 67 Bias result of condition ENS 3PL Model

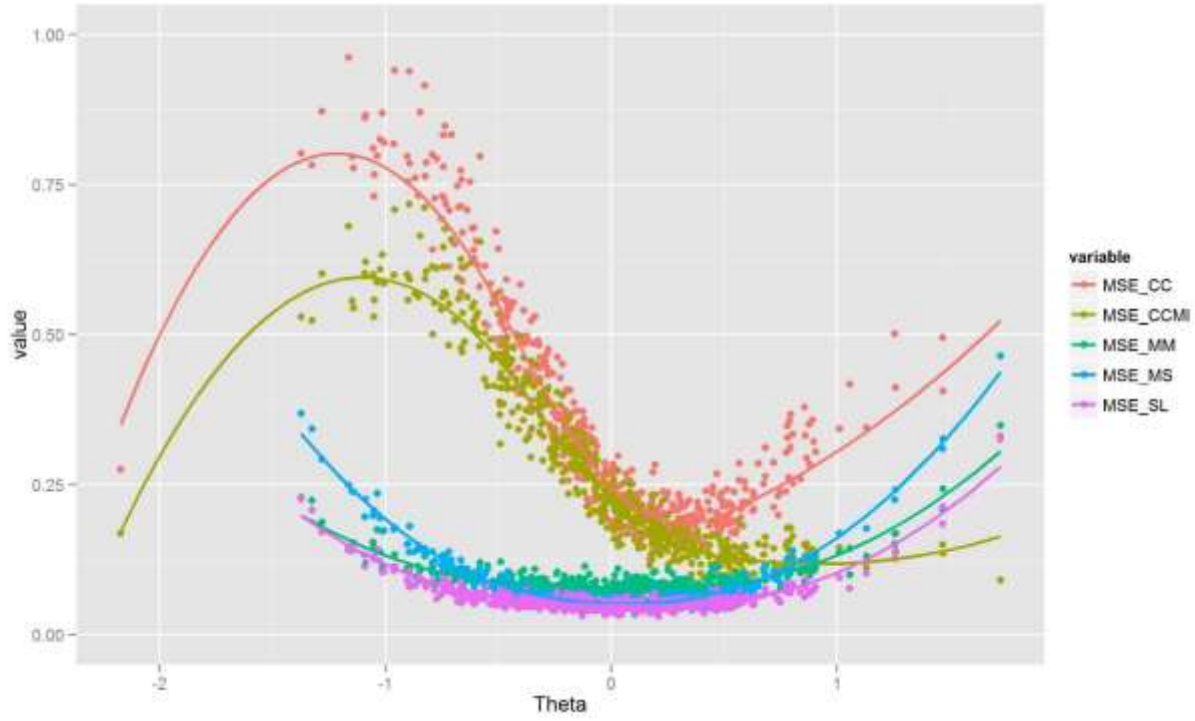

Figure 68 MSE result of condition ENS 3PL Model

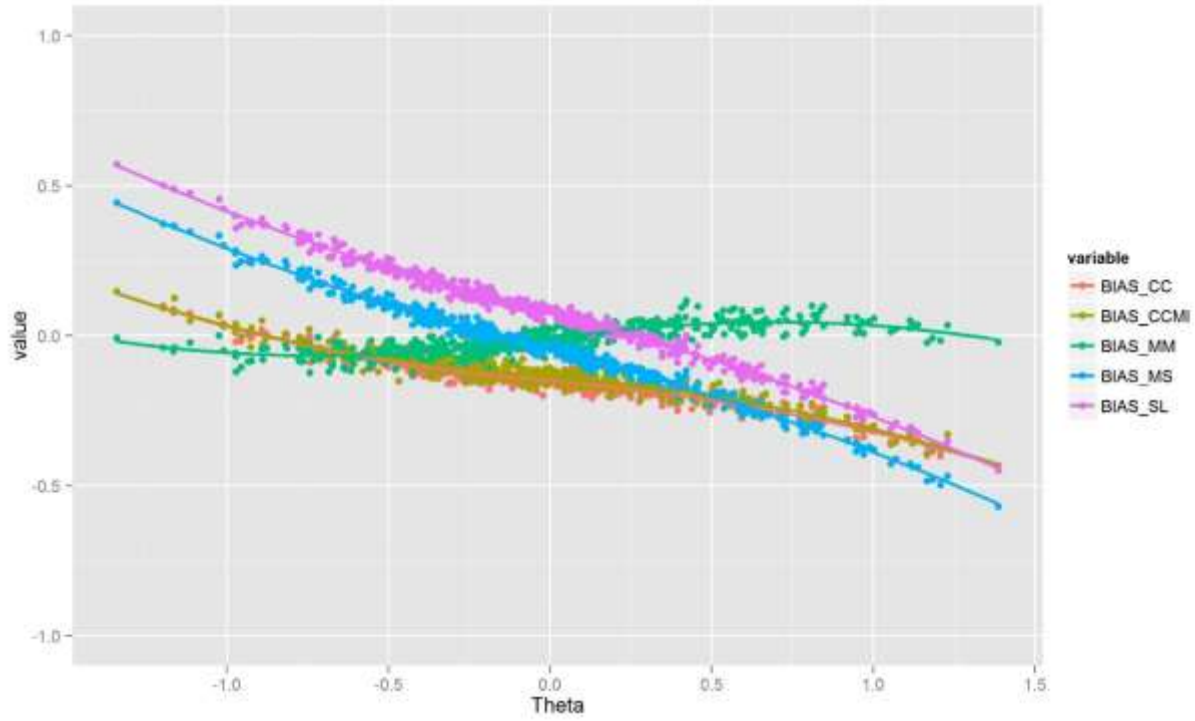

Figure 69 Bias result of condition ENL 3PL Model

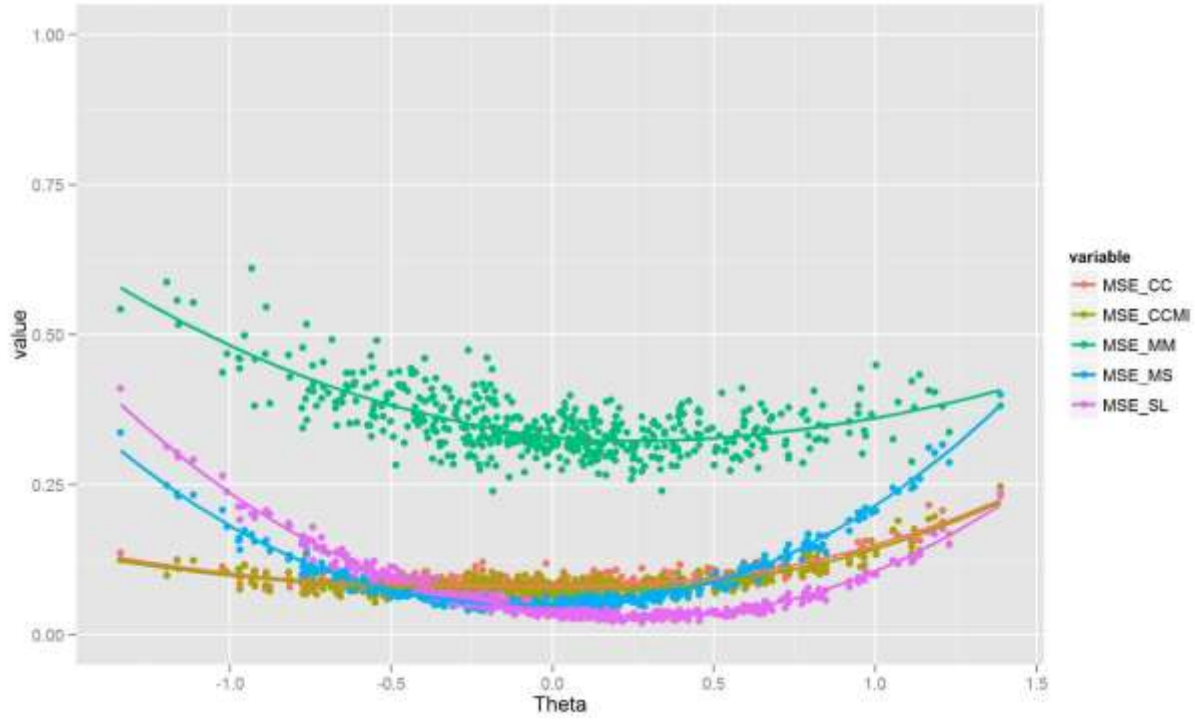

Figure 70 MSE result of condition ENL 3PL Model

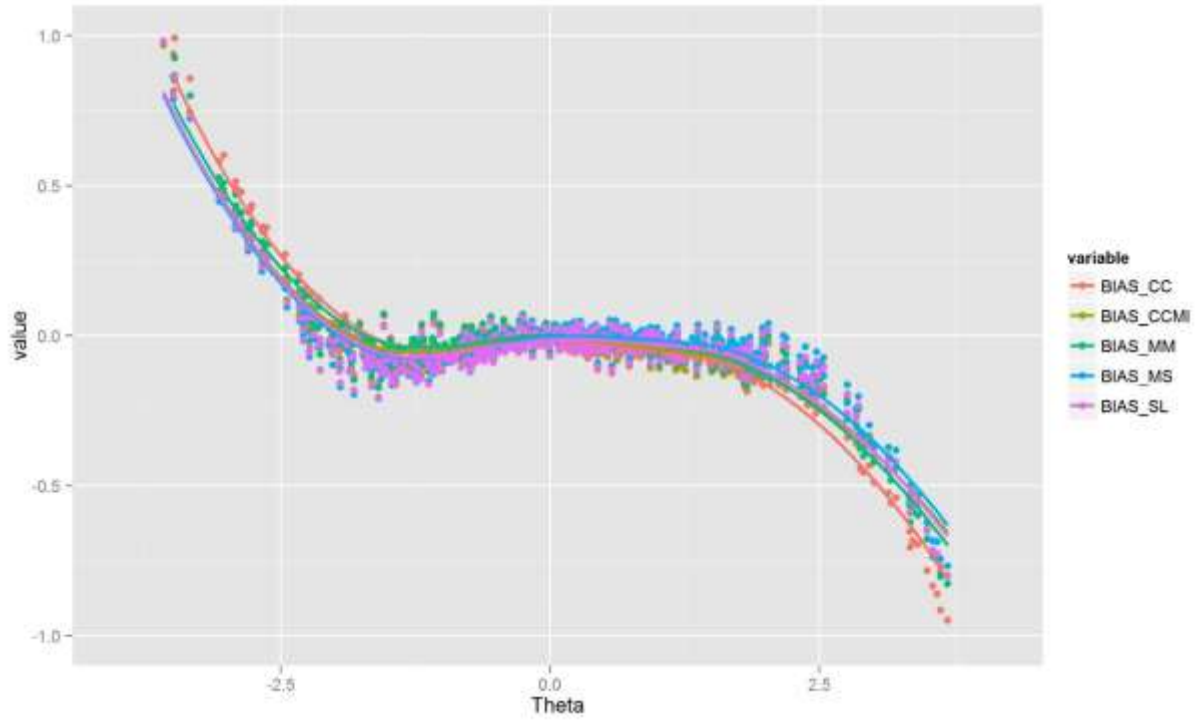

Figure 71 Bias result of condition EWS 3PL Model

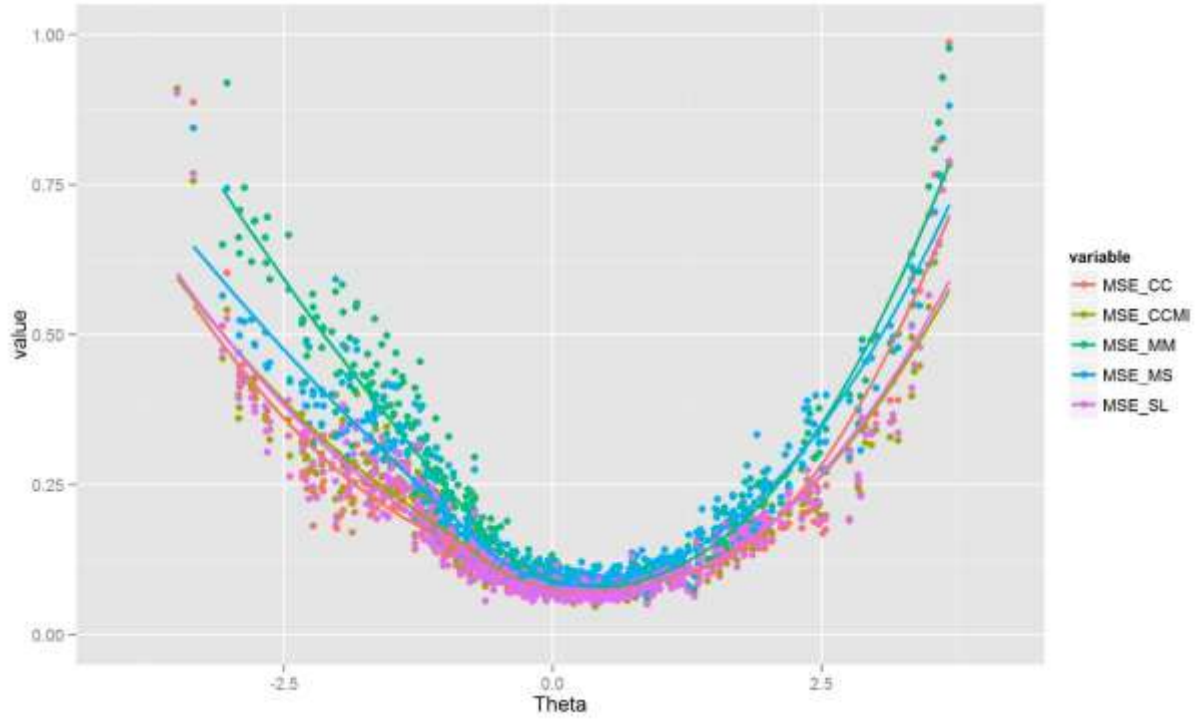

Figure 72 MSE result of condition EWS 3PL Model

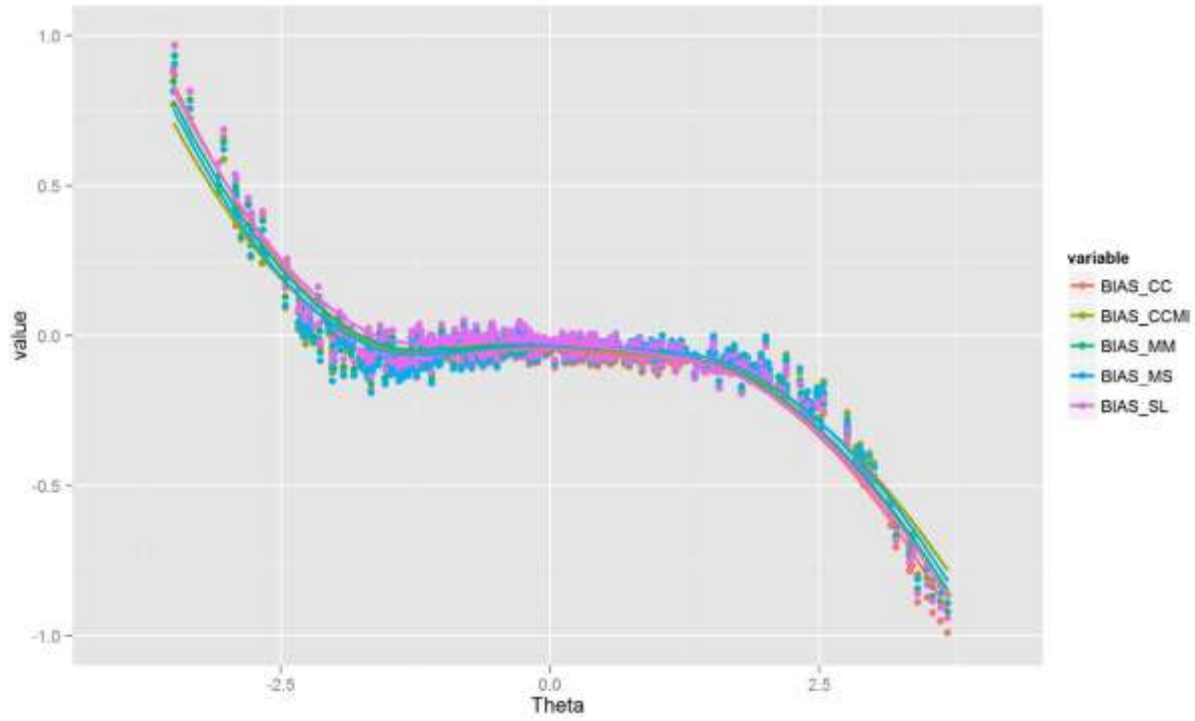

Figure 73 Bias result of condition EWL 3PL Model

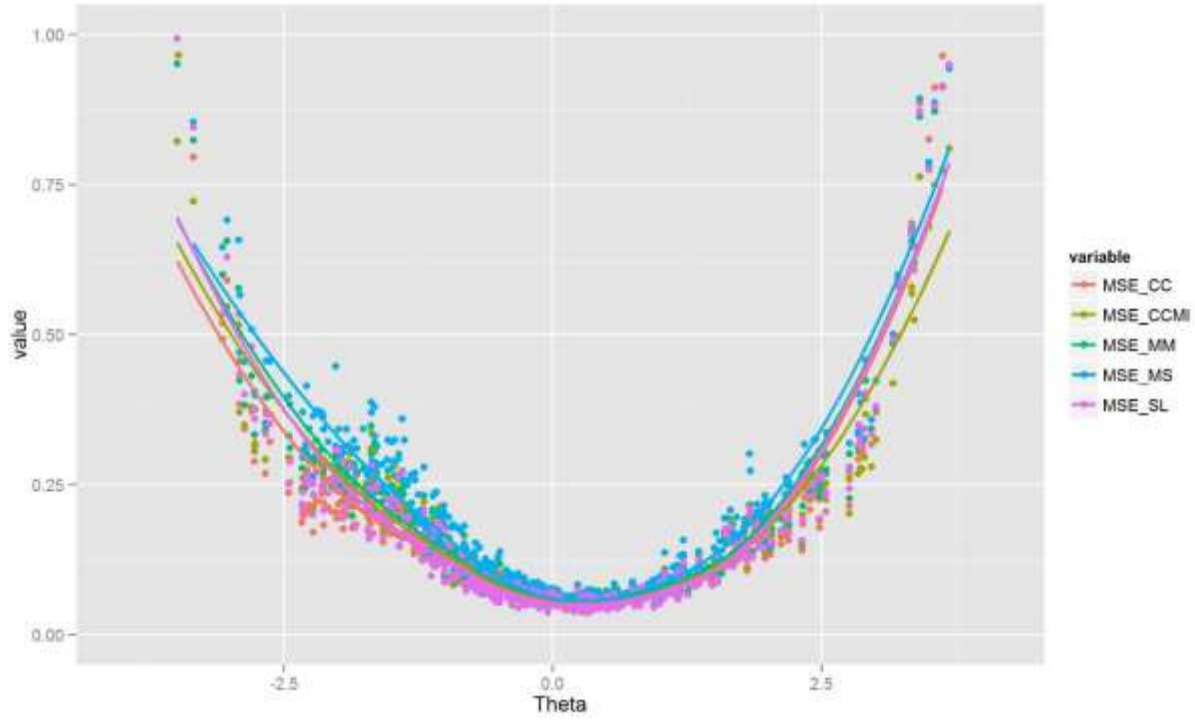

Figure 74 MSE result of condition EWL 3PL Model

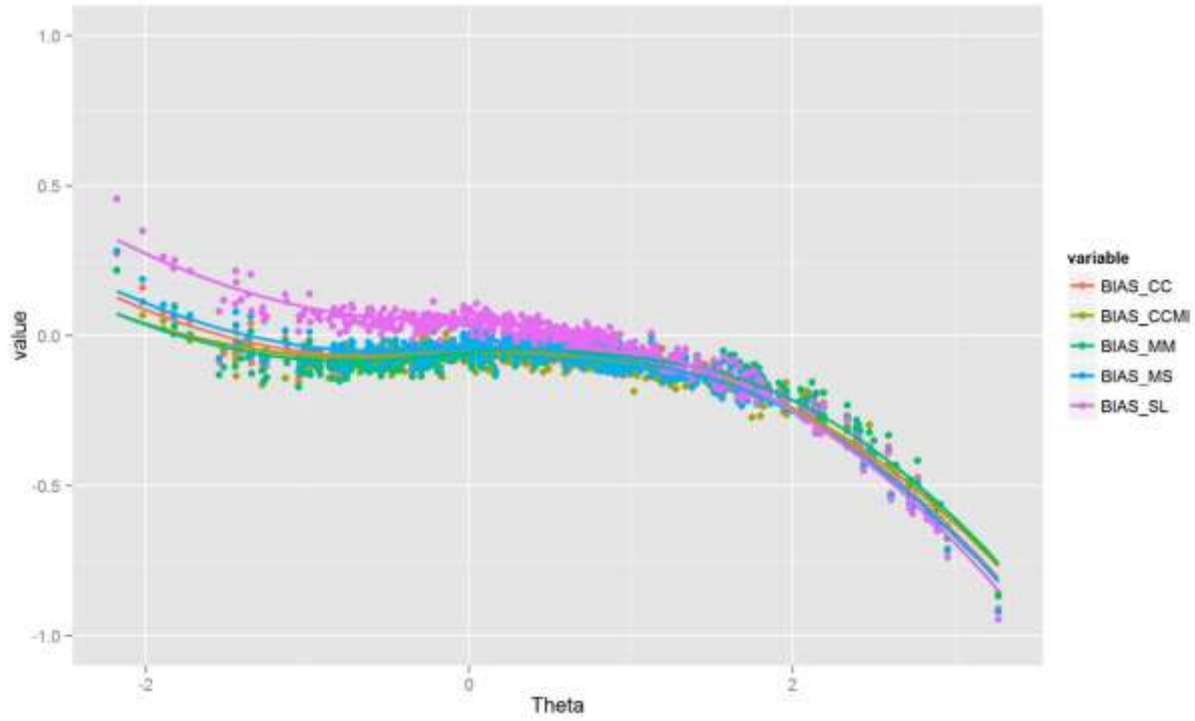

Figure 75 Bias result of condition DIS 3PL Model

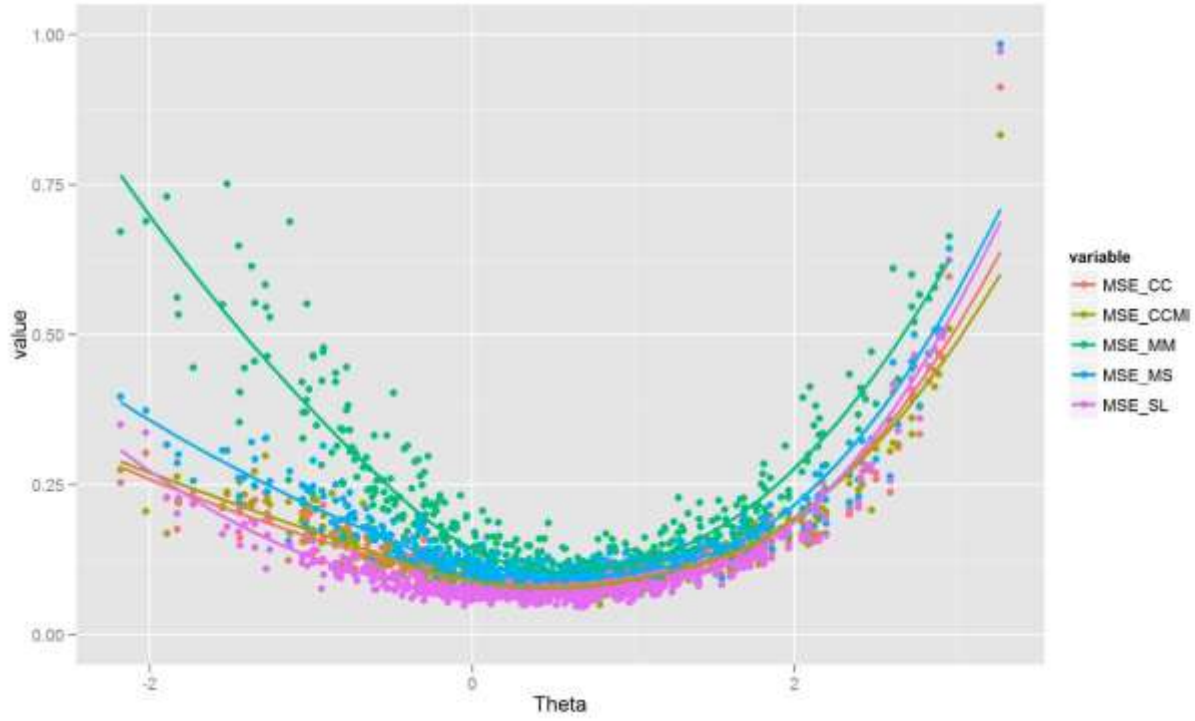

Figure 76 MSE result of condition DIS 3PL Model

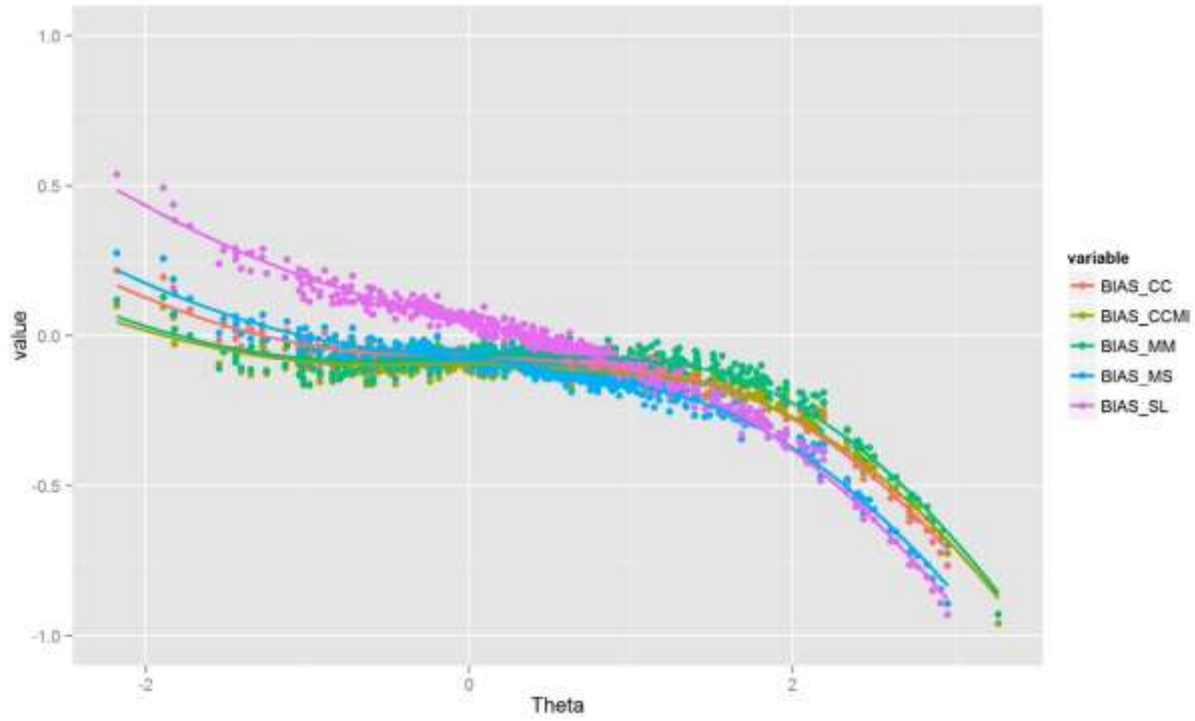

Figure 77 Bias result of condition DIL 3PL Model

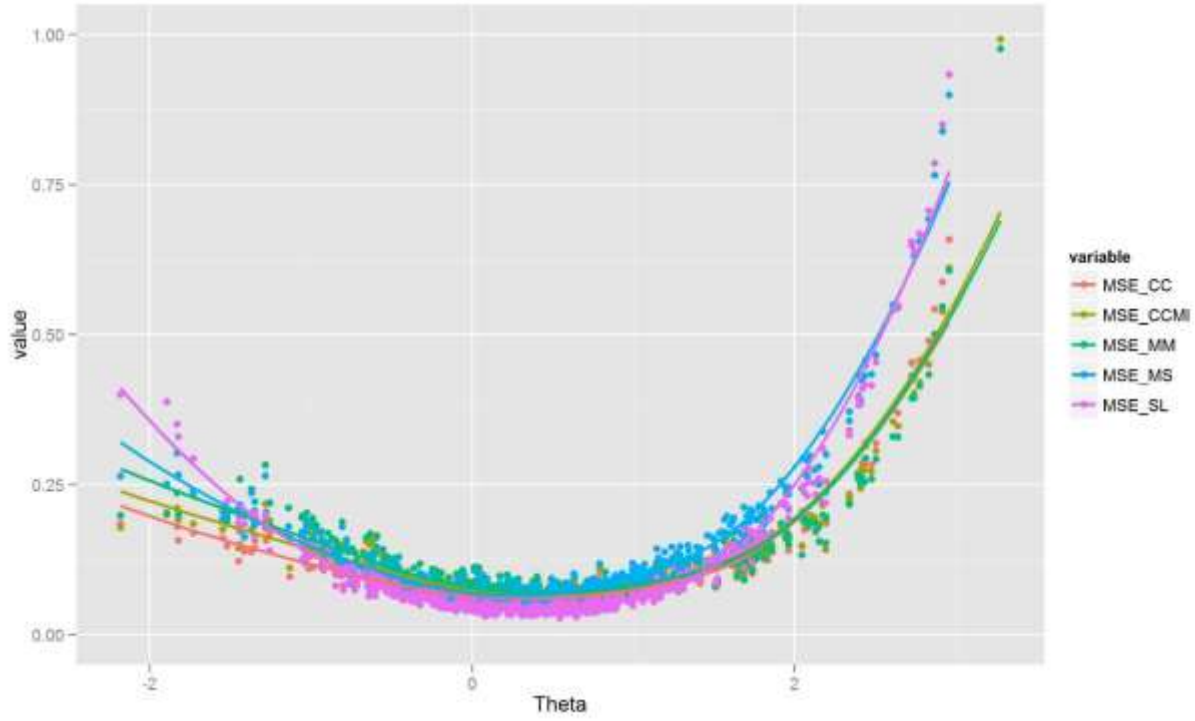

Figure 78 MSE result of condition DIL 3PL Model

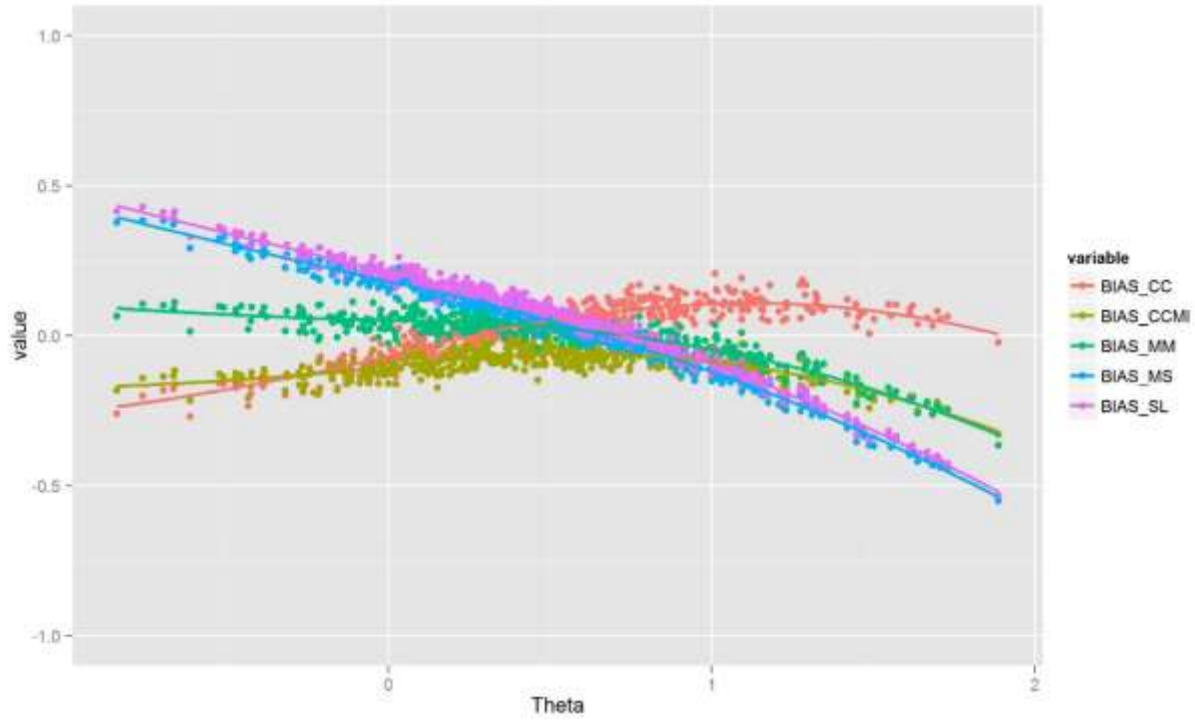

Figure 79 Bias result of condition DNS 3PL Model

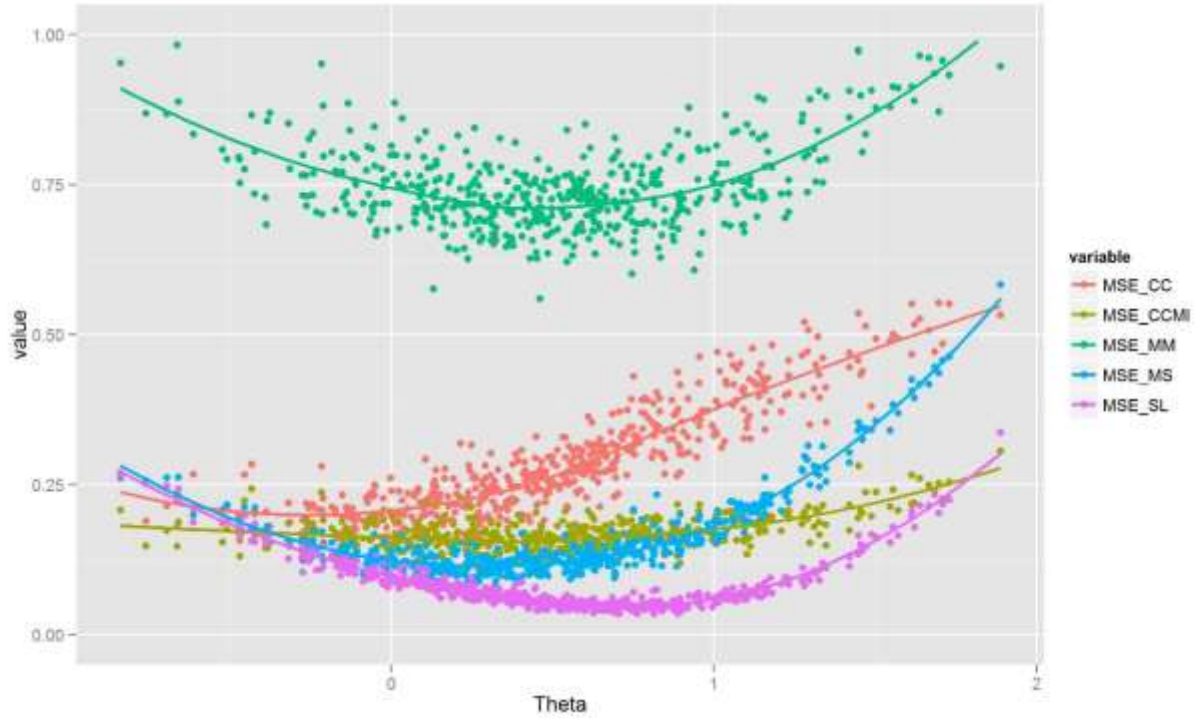

Figure 80 MSE result of condition DNS 3PL Model

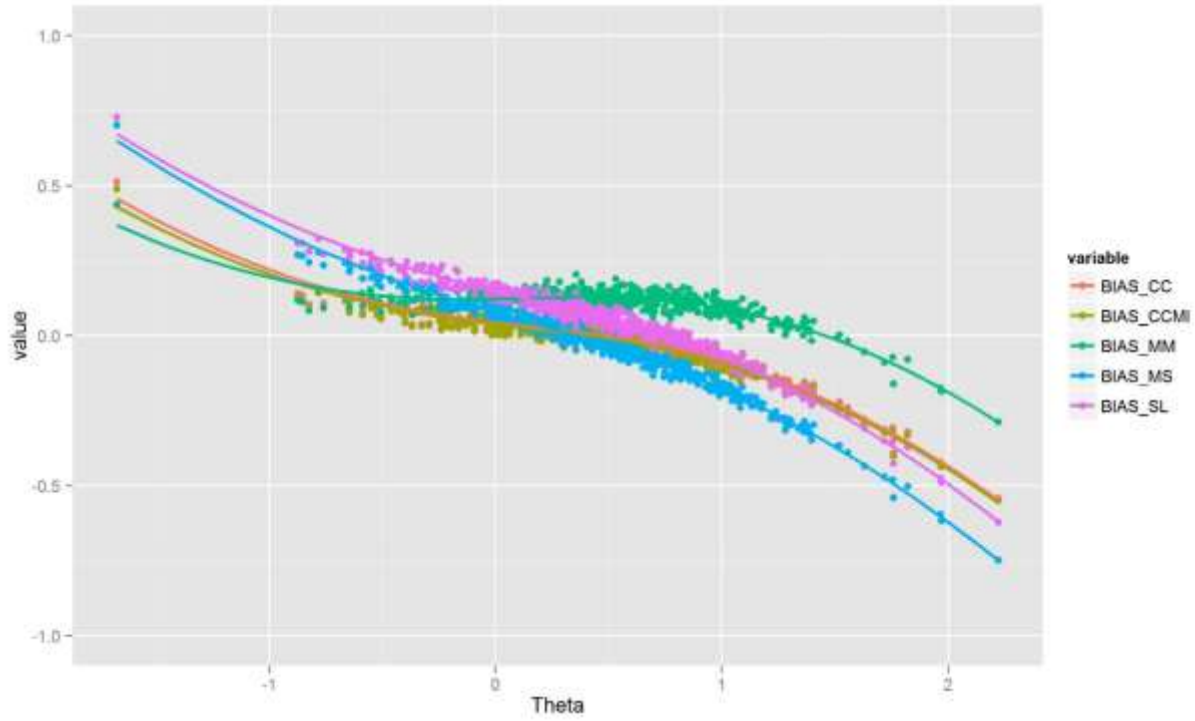

Figure 81 Bias result of condition DNL 3PL Model

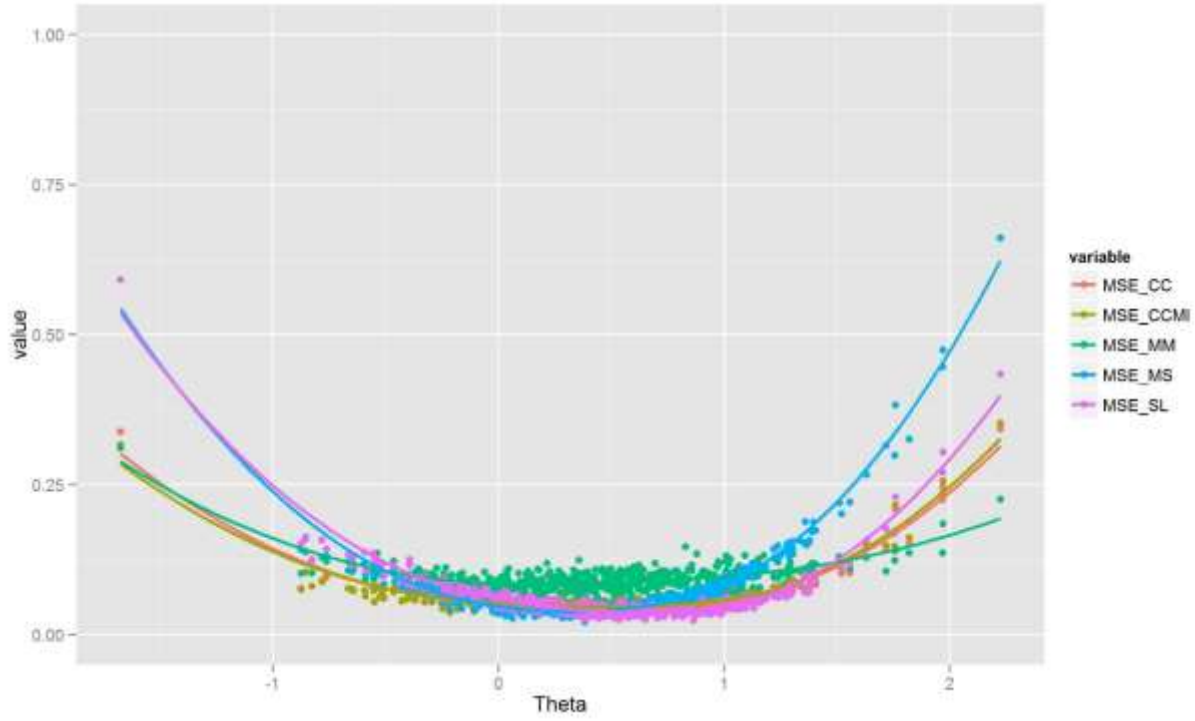

Figure 82 MSE result of condition DNL 3PL Model

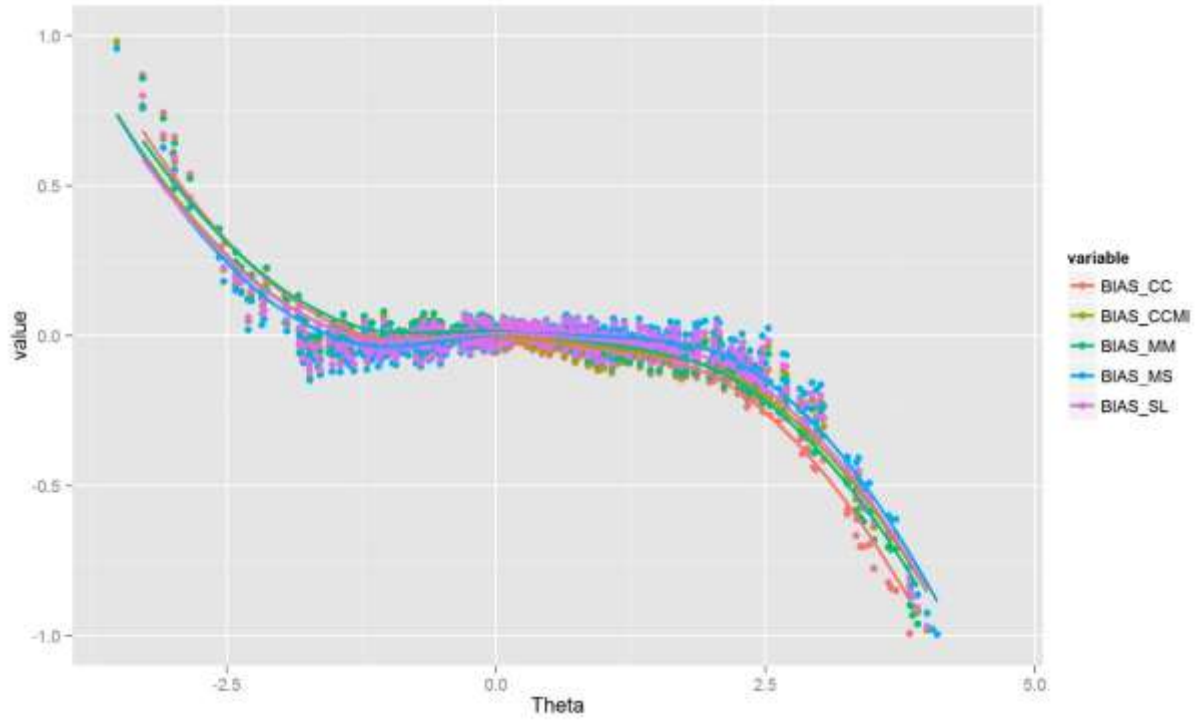

Figure 83 Bias result of condition DWS 3PL Model

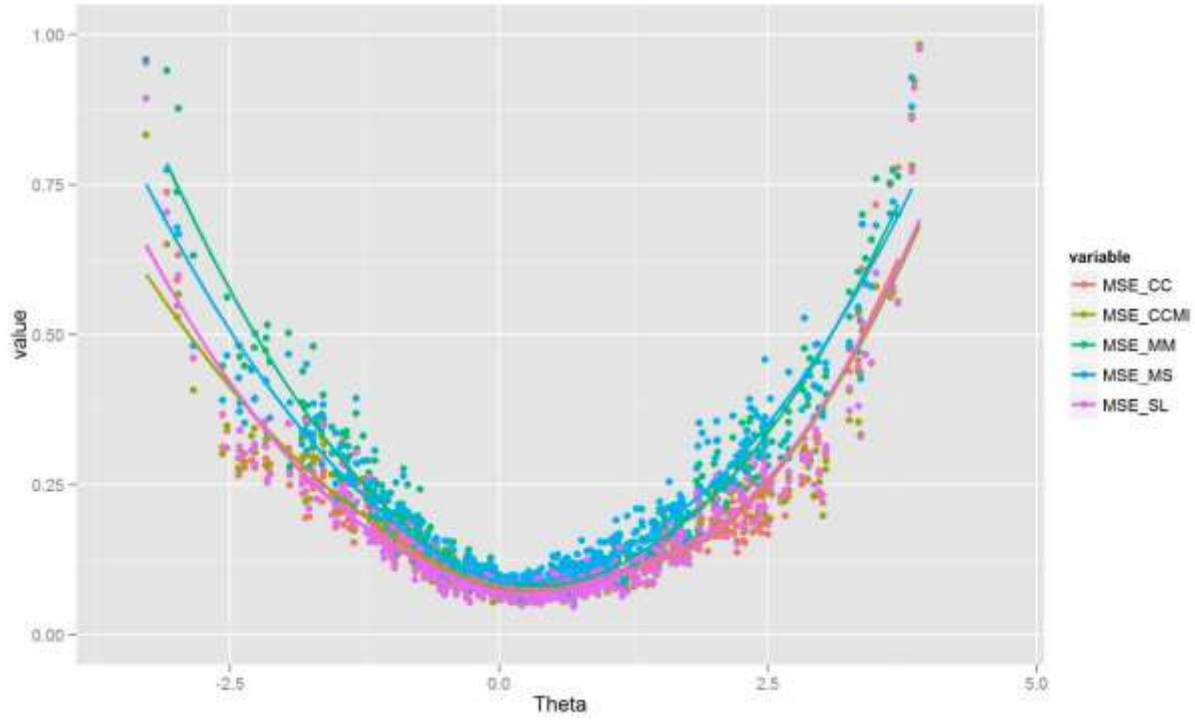

Figure 84 MSE result of condition DWS 3PL Model

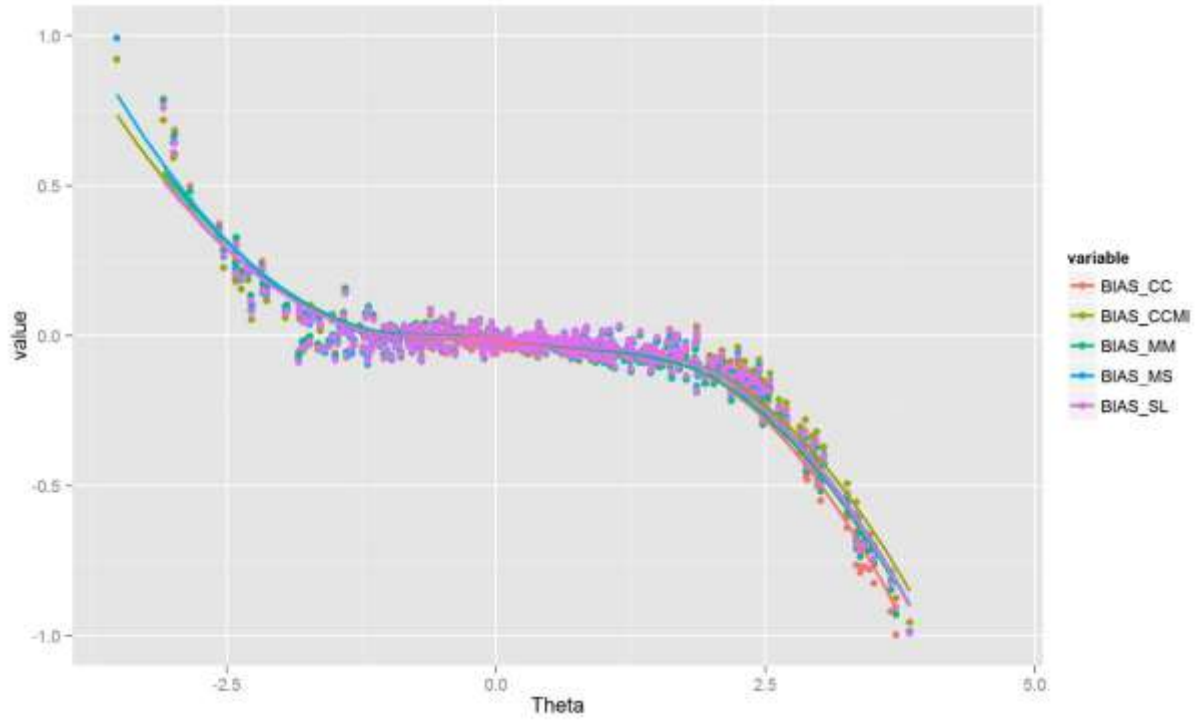

Figure 85 Bias result of condition DWL 3PL Model

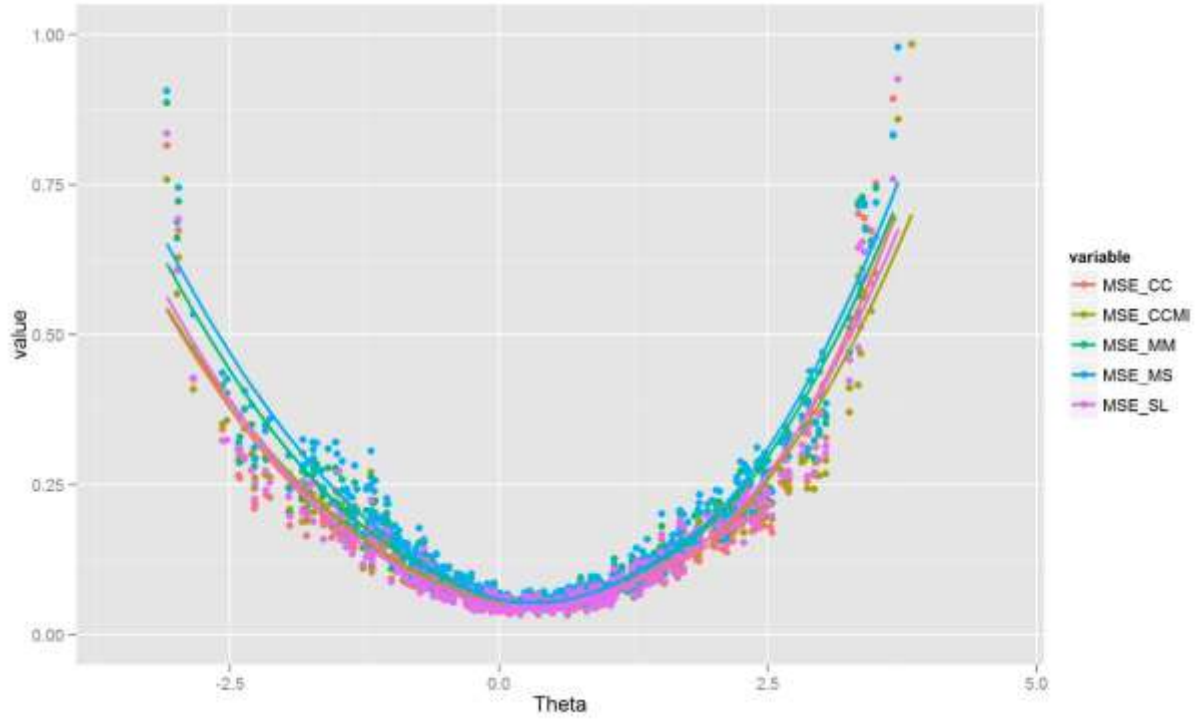

Figure 86 MSE result of condition DWL 3PL Model