CHO, UK HYUN. Ph.D. Impact of Multidimensionality on unidimensional IRT Linking and Equating Methods. (2024)
Directed by Dr. Kyung Yong Kim. 219 pp.

The present study investigates the influence of multidimensionality on linking and equating in a unidimensional IRT. Two hypothetical multidimensional scenarios are explored under a nonequivalent group common-item equating design. The first scenario examines test forms designed to measure multiple constructs, while the second scenario examines a test aimed to measure a primary latent trait but contaminated with a nuisance factor. Classification measures and equating equity properties are used to compare the baseline multidimensional IRT and unidimensional IRT under these scenarios. The findings suggest that multidimensionality is not the primary factor influencing the behavior of linking constants A and B. However, interacting factors such as mean shift, covariance structure, and linking method do have an impact. Test structure alignment is crucial for achieving quality equating results, as equating bias constitutes a substantial proportion of the total error. Classification indices demonstrate that unidimensional IRT generally outperforms the baseline MIRT, with semi-equivalent test structures showing higher performance. Equating equity properties indicate that test structure alignment and choice of linking methods significantly influence equating quality and predictability. The study highlights the importance of considering factors in achieving accurate and precise equating results. Further, Approximate Multidimensional IRT True Score (AMT) equating is proposed as a possible solution to assess the impact of multidimensionality to address the limitations of conventional equating methods in capturing dimension-specific changes in scores between test forms.

IMPACT OF MULTIDIMENSIONALITY ON UNIDIMENSIONAL IRT LINKING AND EQUATING METHODS

by

Uk Hyun Cho

A Dissertation
Submitted to
the Faculty of The Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Greensboro

2024

Approved by

Dr. Kyung Yong Kim
Committee Chair

APPROVAL PAGE

This dissertation written by Uk Hyun Cho has been approved by the following committee of the Faculty of The Graduate School at The University of North Carolina at Greensboro.

Committee Chair

Dr. Kyung Yong Kim

Committee Members

Dr. Richard Luecht

Dr. Robert Henson

Dr. Robert Furter

March 6, 2024

Date of Acceptance by Committee

March 6, 2024

Date of Final Oral Examination

ACKNOWLEDGEMENTS

I would like to express my deep gratitude to the members of my dissertation committee, namely Dr. Kyung Yong Kim, Dr. Richard Luecht, Dr. Bob Henson, and Dr. Robbie Furter. Their unwavering dedication, guidance, and assistance throughout this academic journey deserve special acknowledgement. I am particularly grateful to Dr. Kim for his exceptional patience, valuable feedback, and prompt responsiveness to my inquiries. Dr. Luecht's insightful perspectives and practical wisdom have significantly influenced my growth during my time in the program. Dr. Henson's invaluable support and encouragement have played a pivotal role in my success. Lastly, I extend profound appreciation to Dr. Furter for his leadership, research expertise, and extensive knowledge in psychometrics. Learning from their exceptional expertise has truly been an honor.

I would also like to express my heartfelt appreciation to my family, whose unwavering support has been instrumental in my achievements. Their constant love and encouragement have nurtured my personal and academic growth. I extend my deepest thanks to my wife Wonhye for your unwavering assistance and support. To my children, Youngin, Keala, and Joon, thank you for your unconditional support and patience throughout this journey. I am eagerly looking forward to what lies ahead for all of us. Finally, I would like to express my gratitude to all those individuals who have contributed significantly to my success but whom I may not have had the opportunity to acknowledge personally. Their invaluable contributions are deeply appreciated.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I: INTRODUCTION

BACKGROUND

In educational and psychological testing, the purpose of an assessment is to provide valid empirical evidence (e.g., a test score) of an examinee's latent ability with respect to one single construct of interest, mostly theoretically defined (Ozer, 2001). However, empirical response data collected from the real world are inherently multidimensional (i.e., multidimensionality) because multiple aspects of cognitive or psychological response processes of examinees to given items may differ substantially (Ansley & Forsyth, 1985).

Multidimensionality can be observed in two hypothetical testing situations. One plausible case of multidimensionality is often observed in a situation where a test, even with one construct of interest, inevitably measures multiple latent traits. For example, many professional certification or licensure exams consist of multi-faceted content domains of integrated knowledge, and its applications to real-life problems in the profession. However, the primary purpose of the assessment is to assess candidates' competency as one underlying construct of interest. As a result, the response data may be multidimensional because examinees may have differing true abilities on each sub-domain area (Luecht, 1996). Another case can be found when a test may unintentionally introduce one or more secondary or nuisance dimensions. To illustrate, suppose a reading test is built to assess a student's reading ability. However, unintended dimensions are often introduced in reading comprehension tests due to passage features and item features (Drasgow & Lissak, 1983; Lawrence & McHale, 1988; Drasgow & Lissak, 1983). Consequently, multidimensionality is likely to occur in the response data, for the sensitivity to the nuisance dimensions differs significantly between subgroups (e.g., Sawatdirakpong, 1993).

In psychometric practice, such cases of the multidimensionality bring up challenges. That is, although a test measures multiple traits, because of practical constraints, it is demanded to report a unidimensional score such that it meaningfully represents the summary of an examinee's latent abilities. Another challenge is to provide an unbiased score, free of the influence of unintended factors when a test unintentionally introduces nuisance dimensions in addition to its primary dimension of interest. Such challenges emerge clearly when multidimensional response data is analyzed with a unidimensional item response theory (UIRT) model because its strong assumption that performance on the response data depends on only one latent trait is rarely met in practice (e.g., Humphreys, 1986; Ozer, 2001). These challenges are recognized as the validity-verse-unidimensionality dilemma (Ip, 2010). That is, in practical testing situations, "there exists a dilemma between the psychometric desire for assessing a single construct versus the need for a test to function meaningfully as a valid instrument" (Strachan et al., 2022, p.348).

IRT APPROACHES

In the IRT framework, three different approaches may be recognized to address the two challenges: (1) multidimensional IRT (MIRT; Reckase & McKinley, 1983) models, (2) a locally dependent unidimensional IRT model (Ip, 2010), and (3) unidimensional IRT (UIRT; Lord, 1980; Rasch, 1980) models.

First, item response theory (IRT) can be defined as a collection of statistical models that represent the probability that an examinee obtains a particular score on a given item, based on the examinee latent ability (often denoted by θ) and item parameters (e.g., difficulty, discrimination, and pseudo-guessing). When an IRT model is correctly specified with one or more latent variables (i.e., dimensionality assumption; Lord, 1980), the probability of an item

endorsement is independent of that of another item (i.e., conditional independence assumption, Lord, 1980). The conditional independence assumption follows automatically from the dimensionality assumptions (Lord, 1982, p.19).

In MIRT, multiple latent variables are modeled such that an examinee's proficiencies are estimated for multiple constructs of interest. MIRT appears as a promising solution for the first multidimensionality case (MC-I) in which a test measures multiple dominant dimensions. However, due to its complexity, and the uncertainty about the definition of a dimension, multidimensional IRT (MIRT) has not yet been widely accepted as an operational psychometric model (Luecht & Miller, 1992).

The locally dependent unidimensional IRT model was recently proposed by Ip (2010) and can be applied to the second case of the multidimensionality (MC-II; Ip & Chen, 2012) as a feasible solution. That is, when a test consists of one primary dimension as the construct of interest and one or more secondary dimensions that are not of substantive interest, a targeted dimension can be obtained by projecting the secondary dimensions onto it (i.e., projective IRT). In other words, the targeted dimension is purified by getting rid of the influence of the nuisance factors. The projective IRT (PIRT) model have been gaining attention in the literature for its utilities (e.g., Ip & Chen, 2012; Ip et al., 2019; Kim, 2022; Strachan et al., 2020), yet it has not been commonly used in practice.

Although their applications are limited in practice, MIRT and PIRT are theoretically and empirically recognized as solutions for MC-I and MC-II, respectively. In the current study, MIRT is used as a base model and PIRT as a competing model.

As explicit in its name, UIRT requires one latent variable, leading to the unidimensionality assumption often recognized as being ideal or too strict in practice (e.g.,

Humphreys, 1986; Ozer, 2001; Yen, 1993). However, the unidimensionality assumption has been relaxed in terms of statistical and substantive perspectives (e.g., Ip, 2010). More specifically, a test with a dominant factor and one or more minor factors is considered as essentially unidimensional (Stout, 1987). In the same spirit, for the unidimensionality assumption, a dominant component or factor is required to be met to a satisfactory extent by a set of test data (Hambleton, 1989; Reckase, 1979). To illustrate unidimensionality in multidimensional data, Reckase (1990) showed two cases where a UIRT model fits well to a set of test items that measure more than one dimension. That is, when all items in a test measure the same set of skills in the same way, and when item difficulty and dimensionality are confounded. In addition to its relaxation in the statistical unidimensionality assumption, from the substantive view, the unidimensionality assumption requires that items in a test measure the same composite of abilities, rather than a single ability (Reckase et al., 1988). In other word, when a UIRT model is fitted to multidimensional data, the unidimensional latent ability represents a linear composite of the multiple dimensions present in a test (Wang, 1987; Zhang & Stout, 1999). Therefore, the unidimensionality assumption can be restated as follows: “all the items in a test are measuring the same skill or same composite of multiple skills” (Ackerman, 1995, p. 256).

For the robustness of UIRT to multidimensionality, two important findings summarized by Gibbons, Immekus, and Bock (2007) seem to be encouraging. If the data consists of a predominant general factor and major dimensions that are relatively small, the presence of multidimensionality has little effect on item parameter estimates and the associated ability estimates (i.e., essential dimensionality; Stout et al., 1996). In contrast, when response data are multidimensional with multiple dominant factors, UIRT results in parameter and ability estimates that are drawn towards the strongest factor, and the ability estimates become a

weighted composite of the measures from each individual dimension (Goldstein & Wood, 1989). However, when dimensions are uncorrelated, parameter and ability estimates are distorted by the existing degree of multidimensionality (Folk & Green, 1989, Strachan et al., 2022).

Even with extensive studies providing the important information about the impact of multidimensionality (e.g., Ackerman, 1989; Ansley, 1984; Ansley & Forsyth, 1985; De Ayala, 1994; Doody, 1985; Drasgow & Parsons, 1983; Folk & Green, 1989; Harrison, 1986; Ip, 2010; Luecht & Miller, 1992; Oshima & Miller, 1990; Reckase, 1979, 1987, 1990; Strachan et al., 2022; Way, Ansley, & Forsyth, 1986), their findings are inconclusive (Hsu & Yu, 1989).

EVALUATION

The multidimensionality issue can be concerned particularly in two areas of assessment: classification and score comparability. Scores on psychological and educational tests are widely used when making selections, diagnostic, qualification, and admission decisions. These tests are used to quantify examinees' relative standings on certain constructs of interest to classify them into performance categories (i.e., criterion-referenced tests). Thus, a primary focus of measurement precision of a IRT scale should be the degree of correctly and consistently classifying examinees into performance categories such as pass/fail (Rikli & Jones, 2013).

When a UIRT model is used to calibrate multidimensional data, problems can arise in estimating an examinee's level of ability (Walker & Beretvas, 2013). The resulting unidimensional estimate of ability is a linear combination of their ability estimates that would be obtained if a compensatory MIRT model had been used (Ackerman, 1994). Furthermore, if difficulty and dimensionality are confounded in the data, the composite of ability remains inconsistent throughout the estimated unidimensional ability scale (Reckase et al., 1986). Therefore, when the composite latent measure is used instead of multiple measures, a

classification decision on candidates' qualification may not be accurate. That is to say that when decisions are made in consideration of multiple latent abilities, the composite ability may lack ability-specific information because of its compensatory nature.

Several classification methods have been developed in UIRT framework (e.g., Lee, 2010; Lathrop & Cheng, 2014; Rudner, 2001). For example, Lathrop and Cheng (2017) claimed that their nonparametric approach outperformed Lee's approach when the ability distribution digresses from the normality assumption, but the multidimensionality issue was not fully addressed in the previous studies. Recently, Park et al. (2022) proposed the multidimensional classification procedures, and their findings assure the robustness of UIRT classification procedures on the simple multidimensional latent structure, but further investigation with different latent distributions and latent structures is suggested.

To be properly used for high-stakes decisions, test scores must be comparable across alternate forms. However, even with the best effort of psychometricians to make them parallel in terms of content and statistical specifications, alternate forms differ in difficulty in practice. Thus, it is necessary to adjust for the minor difference in difficulty (i.e., equating; Kolen & Brennon, 2004), and the quality of the equating should be evaluated to enhance the appropriateness of score interpretation (American Educational Research Association et al., 2014). Therefore, when equating is performed on two forms under a satisfactory condition of assumptions, each individual in the test population is anticipated to have the same expected score and measurement accuracy on both tests (i.e., equating equity; e.g., Morris, 1982; Tong & Kolen, 2005; Yen, 1983). However, test equity may not hold because a UIRT equating function is unable to account for dimension-specific changes in difficulty across test forms (Bolt, 1999). In this context, fitting UIRT models to multidimensional data has drawn researchers' attention to

assess the effect of multidimensionality on UIRT equating (e.g., Béguin, & Hanson, 2001; Béguin et al., 2000; Bolt, 1999; Luecht & Miller, 1992; Stocking & Eignor, 1986). Several studies indicate that the influence of multidimensionality on the quality of UIRT true score equating (TSE) seems to be of little practical importance (e.g., Bolt, 1999; Goldstein & Wood, 1989; Luecht & Miller, 1992) when the multidimensional latent structures of two forms are parallel or two forms have the same reference composite of latent traits. However, these studies mainly focus on the multidimensional content structures with homogeneous populations (Champlain, 1996). In addition, the multidimensional structure, with one primary dimension as the construct of interest and one or more nuisance dimensions (MC-II), has not attracted much attention from researchers, even with its practical significance. For example, when a test consists of items that require additional abilities for more difficult items, item parameter and ability estimates of UIRT could produce misleading and biased results regarding the targeted construct of interest, resulting in misclassification of high-stakes decisions (Ip et al., 2019). Furthermore, a few attempts have been made to assess the effect of the interaction between multidimensional test structure and heterogeneous populations (e.g., Béguin et al., 2000; Champlain, 1996; Dorans & Kingston, 1985; Stocking & Eignor, 1986). Their findings, in general, suggest that the increase in difference in abilities and multidimensionality adversely affects the accuracy of parameter estimates and results of UIRT equating procedures.

However, each study was narrowly focused because of different study conditions, purposes, and evaluation methods (e.g., Champlain, 1996; Harris & Crouse, 1993). To be specific, real data analyses provide empirical evidence (e.g., Camilli, Wang, & Fesq, 1995; Dorans & Kingston, 1985; Yen, 1984), but were limited in explanatory power. In simulation studies, it appears that the multidimensionality affects UIRT true score equating (Bolt, 1999) less

than UIRT observed score equating (OSE; Béguin et al., 2000). In the case of pre-equating, mean shift of ability distribution, the degree of multidimensionality, and combinations of both adversely affect recovery of parameters (Stocking & Eignor, 1986). These studies are inconclusive when it comes to revealing the dynamics of multidimensional test structures and different populations. Assessing the interplay of different dimensional structures and heterogeneous populations is complex and needs to be addressed more comprehensively (Skaggs, 1990).

RESEARCH PURPOSE AND QUESTIONS

The purpose of this dissertation is to evaluate the impact of multidimensionality to UIRT linking and equating when an UIRT model is applied to multidimensional response data. Two simulation experiments are designed for two testing scenarios of multidimensionality; that is, (1) a test measuring more than one trait (MC-I) and (2) a test measuring one primary trait with one or more nuisance dimensions (MC-II). In the two scenarios, the reference composite of two forms may appear to be similar or different contingent upon the interaction of test items with groups. Thus, factors to consider include different latent structures, form differences, and group differences. With these factors in mind, this study aims to address the following five questions:

1. To what extent does multidimensionality affect classification accuracy?
2. To what extent does multidimensionality affect classification consistency?
3. To what extent does multidimensionality affect first-order equity?
4. To what extent does multidimensionality affect second-order equity?
5. To what extent does multidimensionality affect the combined measure of the first- and second-order equities?

To be more specific, response data will be generated from a 2D MIRT model and a UIRT (MC-I and MC-II) and PIRT (MC-II) will be fitted to the data to seek the answers for the five questions. Note that for MIRT equating, only OSE is considered due to due to the limitation of TSE in MIRT; that is, the one-to-many relationship between an expected score and its corresponding combinations of latent abilities in the test characteristic surface makes it impossible to find the unique combination in the multidimensional latent space.

In the first simulation study for MC-I, three linking procedures are compared: (1) separate calibration (SC) with Haebara (HB) method, (2) concurrent calibration (CC), and (3) fixed parameter calibration (FPC); and two UIRT equating procedures are compared: (1) TSE and (2) OSE. The base model will be two dimensional MIRT (2DMIRT) with three linking methods: (1) SC with the extended version of HB method (Oshima et al., 2000), (2) MIRT CC and (3) MIRT FPC. But due to the limitation of TSE in MIRT, only MIRT OSE is considered. In contrast, in the second simulation study for MC-II, UIRT linking and equating procedures are the same as the first case, but 2DMIRT is a generating model and PIRT model is used as a competing model for UIRT. For evaluation, marginal indices of classification and equating properties are compared.

CHAPTER II: LITERATURE REVIEW

This chapter consists of two sections: the first section is for the fundamentals of item response theory and the second section is for the literature review on the research topic.

1 FUNDAMENTALS OF ITEM RESPONSE THEORY

OVERVIEW

Item response theory (IRT) models are widely used in scoring examinees' performance on a given test. A typical process of scoring follows obtaining item parameter estimates. It is one of the foremost benefits of the IRT models that an examinee's score is put on the same scale of item difficulty such that for test takers it is possible to understand their performance on difficulty levels. The IRT scale is arbitrarily defined with its origin and unit. For example, for the scale identification purpose, the common procedure of IRT calibration software packages is setting the distribution of the ability to be a standard normal distribution with mean zero and standard deviation with one, resulting in a scale with its origin zero and unit one (Note that Rasch models use the model-embedded scale, logit).

When two forms constructed based on even the same test specification are administered to different groups of examinees, there rise two technical issues. First, there is no medium to address the arbitrariness of two IRT scales; that is, putting two scales on one common scale. Second, because of the first issue, it is not possible to make the adjustment for the differences in form difficulty. The statistical procedure to handle the former issue is known as "Scale Linking", and the other statistical procedure to solve the latter issue is known as "Score Equating". Often common or shared items appearing in both forms are used as a medium or anchor to disentangle ability difference and test form difference. The traditional wisdom recommends that common item sets be sufficient in number and representative of the whole test form in terms of content

and statistical specifications. In the linking and equating context, this data collection design is called the “Common Item Non-equivalent Group (CINEG)” design (Kolen & Brennan, 2014) or the “Non-Equivalent groups with Anchor Test (NEAT)” design.

In general, IRT models are categorized into unidimensional IRT (UIRT) and multidimensional IRT (MIRT) by the number of latent dimensions; and, into one, two, and three-parameter IRT models by different parameterization. For instance, 3PL2DMIRT indicates the IRT model that has three item parameters (i.e., item difficulty, item discrimination, and pseudo-guessing) and two latent dimensions. MIRT can be further subdivided into “Simple Structure”, “Approximately Simple Structure”, and “Complex Structure” by latent dimensional structure; and into “Compensatory” and “Non-Compensatory” by the interaction of latent dimensions in producing the probability of item endorsement. For an approximately simple structure MIRT model, clusters of items load primarily onto one pre-specified dimension or factor with trivial cross-loadings. For a compensatory MIRT model, latent variables compensate each other for the probability of getting a given item correct. Based on the number of item categories, IRT models are divided into dichotomous models with two categories and polytomous models with more than two categories. As noticed, UIRT models are special cases of MIRT models and dichotomous models are also constrained cases of polytomous IRT models. Lastly, it is worth noting that the item response functions of the item characteristic curve (ICC) are divided into two: normal ogive and logistic function. Often 1.7 is added to the logistic function as the multiplier to match the shapes of the two ICCs.

Dimensionality is a widely used term to describe the characteristics of the latent space in the IRT framework. It is also closely related to defining a metric. In UIRT, the latent dimension is only one and its scale is defined by origin and unit, while in MIRT, two more technical

considerations are required to identify the multidimensional scale; that is, the correlation between dimensions and the rotation of the coordinate system in the multidimensional latent space. In modeling item response data, the assumption of dimensionality is essential. If the assumption is significantly violated for a fitted IRT model, the estimates of item parameters and latent abilities are biased, resulting in inaccurate linking, and equating results. As a result, the test scores might become invalid for their defensible interpretation and use.

PROPERTIES AND LIMITATIONS

Item response theory (IRT) is a popular modeling method for item response data in psychological and educational settings. Compared to the classical test theory (CTT) whose scale (e.g., the origin is zero and unit is one item) is sample-dependent, IRT provides three distinct properties. First, item difficulty and ability are on the common scale where individual items and persons are relatively located to one another. In addition, such item and ability properties do not depend on specific samples of items and examinee groups, which is called “invariance property”. In other words, the psychometric properties of items are invariant across different examinee groups and the latent ability is invariant across different test forms. The invariance property allows practitioners to improve the quality of assessment in parallel form development, quality control of test forms, and score comparison. Finally, understanding the dimensionality of the latent ability becomes more feasible. A unidimensional latent ability can be decomposed into multiple latent abilities when multiple factors are involved in the test. Such information is vital for practitioners to ensure that a test form targets the intended dimension(s).

With its convenient psychometric properties, however, IRT cannot completely be free from its limitations. Depending on the parameterization of IRT models, the estimation process is complex and time-consuming. Moreover, a large sample is required, which is often not feasible

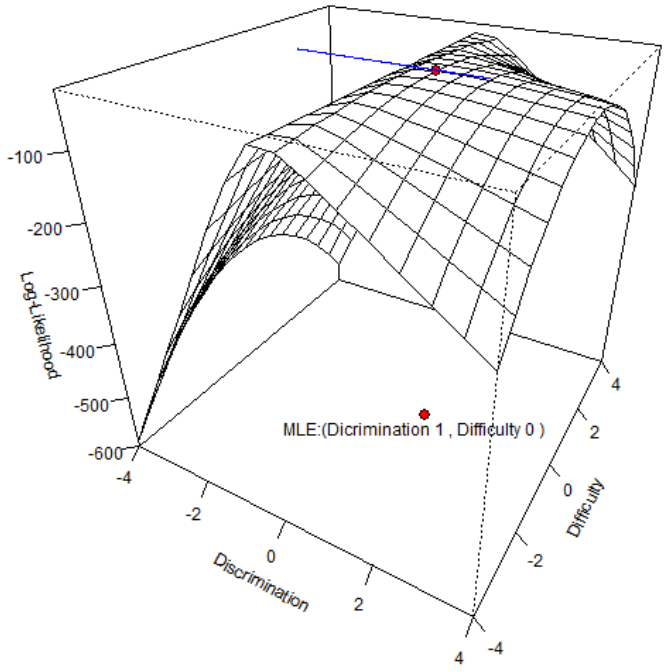
in practical testing situations. Score comparison across forms and administrations can be challenging because the IRT scale is arbitrarily defined, called “scale indeterminacy”. The common practice to resolve this issue is to treat the ability as a random variable and set its distribution into the standard normal distribution regardless of the examinees’ true ability distribution.

Such limitations become more evident in MIRT. Beyond the large sample size requirement, and complex linking and equating procedures for score comparison in MIRT, identifying the latent ability structure in the multidimensional space seems daunting. As such, in practice, UIRT is a common operational IRT model for calibration, linking, scoring, and equating despite the widely accepted notion that item response data are multidimensional.

CALIBRATION

In the educational measurement context, calibration is defined as estimating item parameters. The most widely used estimation procedure is the maximum likelihood estimation (MLE), a searching algorithm to find a set of parameter estimates that maximizes the log-likelihood of the joint probability function. Figure1 visualizes the item parameter estimates via MLE.

Figure 1. Visualization of MLE for UIRT with Item Difficulty and Discrimination



Note: The blue line indicates the gradient on the ML Estimates

To make the maximum likelihood estimation work, two assumptions for the likelihood function must be satisfied: monotonicity and local independence. The former assumption posits that the probability of a correct response increases as the latent ability level increases. The latter one postulates that responses to given items are mutually independent conditional on a given ability level. In addition, due to the unknown latent ability, typical calibration packages (e.g., FlexMIRT) treat the latent ability as a random variable that presumably follows a standard normal distribution, and then integrate the latent ability out from the likelihood function, resulting in the marginal likelihood function for item parameter estimation. The marginal likelihood function could be written as follows:

$$L(\boldsymbol{\Delta}) = P(\boldsymbol{U}\,|\boldsymbol{\Delta}) = \prod_{j=1}^{N} P\left(\boldsymbol{u_j}\,|\boldsymbol{\Delta}\right) \tag{2.1.1}$$

$$= \prod_{j=1}^{N} \int_{-\infty}^{\infty} P\left(\boldsymbol{u_j}|\boldsymbol{\Delta}, \theta_j\right) h\left(\theta_j\right) d\theta_j$$

$$= \prod_{j=1}^{N} \int_{-\infty}^{\infty} \prod_{i=1}^{n} P\left(\boldsymbol{u_{ji}}|\boldsymbol{\delta_i}, \theta_j\right) h\left(\theta_j\right) d\theta_j \,,$$

where j is the number of examinees, N; i is the number of items, n; $\boldsymbol{u_j}$ is a response vector with length n for examinee j; $\boldsymbol{\delta_i}$ is the item parameter vector with size v (the number of item parameters in the model); $\boldsymbol{\Delta}$ is an item parameter matrix with n by v dimension; $\boldsymbol{U}$ is a response matrix with N by n dimension, In the equation, $P\left(\boldsymbol{u_j}|\boldsymbol{\Delta}, \theta_j\right)$ is re-expressed as $\prod_{i=1}^{n} P\left(\boldsymbol{u_{ij}}|\boldsymbol{\delta_j}, \theta_i\right)$, a likelihood function under local independence assumption; and $h(\theta_i)$ is the probability density function of ability. Also, notice that according to the Bayes' theorem, $P\left(\boldsymbol{u_j}|\boldsymbol{\Delta}, \theta_j\right) h\left(\theta_j\right)$ is the proportionality of the posterior ability distribution denoted by $P\left(\theta_j|\boldsymbol{u_j}, \boldsymbol{\Delta}\right)$, given response data and item parameters. In other words, the goal of the marginal maximum likelihood estimation (MMLE) is to obtain the item parameter estimates that maximize the marginal likelihood function after the latent ability is integrated out from the posterior ability distribution. The posterior ability distribution is approximated by weighting empirical information from response data (i.e., likelihood) with $h\left(\theta_j\right)$, a prior distribution from a subjective information on the true ability distribution.

To obtain the posterior distribution, however, the item parameters must be known in advance, which are not before calibration. To solve this problem, the MMLE is implemented with the expectation–maximization (EM) algorithm (Bock & Aitkin, 1981; Mislevy, 1986). The

EM algorithm is an iterative procedure for finding maximum likelihood estimates of probability models in the presence of unobserved random variables, i.e., latent variable in IRT (Baker & Kim, 2004, p. 169). In the E step, conditional on the observed data and the provisional parameter estimates, the posterior expectations of Θ are computed, which consist of the expected number of attempts ($\overline{f}_{iq}$) and correct responses ($\overline{r}_{iq}$) to each item as such:

$$\overline{f}_{iq} = \sum_{j=1}^{N} \left[\frac{L(X_q)A(X_q)}{\sum_{q=1}^{Q} L(X_q)A(X_q)} \right] \tag{2.1.2}$$

and

$$\overline{r}_{iq} = \sum_{j=1}^{N} \left[\frac{u_{ij}L(X_q)A(X_q)}{\sum_{k=1}^{q} L(X_q)A(X_q)} \right], \tag{2.1.3}$$

where j is the number of examinees, N; i is the index of items n; u_{ij} is the response of an examinee j for the item i; q is the quadrature point for a particular ability level; $L(X_q) = \prod_{i=1}^{n} P_i(X_q)^{u_{ij}} Q_i(X_q)^{1-u_{ij}}$ is a likelihood with a given discrete ability value, q; $A(X_q)$ is the quadrature weight (i.e., density) for the corresponding value q in the prior distribution (Baker & Kim, 2004, p. 167).

In the M step, with those quantities, the maximum likelihood estimation algorithm finds the item parameter estimates that maximize the posterior expectation. Because finding the optimal parameters is not analytically tractable, numerical approaches (e.g., Newton-Raphson method) are employed. Note that the $A(X_q)$ can be either fixed to a chosen prior distribution or adjusted by the response data by normalizing the quantity of $\overline{f}_{iq}$ as such:

$$A^{(r+1)} = \frac{1}{N} \sum_{j=1}^{N} \left[\frac{L(X_q)A(X_q)^{(r)}}{\sum_{q=1}^{Q} L(X_q)A(X_q)^{(r)}} \right], \tag{2.1.4}$$

where r is the notation for the iteration of the EM cycle; and N is the sample size. The adjusted quadrature weights from the response data are called "empirical histogram" which can be computed with an option in calibration software packages. At the end of each M step, the updated quadrature weights are rescaled, along with the current item parameter estimates for scale identification. Then, updated quadrature weights and item parameter estimates are fed into the E step to move to the next iteration. Finally, the iteration stops when a set conversion criterion is met (for details see Baker & Kim, 2004; Kim, 2017). This MMLE-EM procedure is employed in different calibration procedures which are discussed next.

In the context of defining a metric, the MMLE is referred to as person (or norm) centering because the scale is established from the person ability distribution. In contrast, the most popular calibration package for the Rasch model, Winsteps (Linacre, 2022) estimates both item parameters and latent ability simultaneously, by setting the sum of item difficulty parameter estimates to be zero unless otherwise specified. This procedure is called item (or criterion) centering. In short, items and persons are put on the same scale regardless of either of which metric identification methods is used to define the scale.

There are three calibration procedures based on the information used: separate, concurrent, and fixed item parameter calibrations. As the name suggests, in the separate calibration (SC), with only its own response data, each test form is independently calibrated. Due to the scale indeterminacy, however, SC requires linking to put the two scales from each form on a common scale. In general, the scale of the old form is considered as a base, reference, or common scale on which the new scale will be put.

Unlike SC as a single group calibration, the current calibration (CC) is the multiple-group calibration that estimates multiple forms simultaneously by utilizing all the response data

together. In other words, the base scale is identified with the common item information from both groups. In CC, the equations 2.1.2; 2.1.3; and 2.1. 4 can be re-expressed as follows (Bock & Zimowski, 1997; Kim, 2017):

$$n_{gq}^{(r+1)} = \sum_{j=1}^{N_g} f\left(X_{gq}\middle|\boldsymbol{u_{gj}}, \boldsymbol{\Delta^{(r)}}, \pi_g^{(r)}\right) = \sum_{j=1}^{N_g} \frac{f\left(\boldsymbol{u_{gj}}\middle|X_{gq'}, \boldsymbol{\Delta}^{(r)}\right)\pi_g^{(r)}}{\sum_{q'}^{Q} f\left(\boldsymbol{u_{gj}}\middle|X_{gq'}, \boldsymbol{\Delta}^{(r)}\right)\pi_g^{(r)}} \tag{2.1.5}$$

$$r_{gq}^{(r+1)} = \sum_{j=1}^{N_g} u_{gji} f\left(X_{gq}\middle|\boldsymbol{u_{gi}}, \boldsymbol{\Delta}^{(r)}, \pi_g^{(r)}\right) = \sum_{j=1}^{N_g} u_{gji} \frac{f\left(\boldsymbol{u_{gj}}\middle|X_{gq'}, \boldsymbol{\Delta}^{(r)}\right)\pi_{gq}^{(r)}}{\sum_{q'}^{Q} f\left(\boldsymbol{u_{gj}}\middle|X_{gq'}, \boldsymbol{\Delta}^{(r)}\right)\pi_{gq'}^{(r)}} \tag{2.1.6}$$

and

$$\pi_{gq}^{(r+1)} = \frac{n_{gq}^{(r)}}{N_g} = \frac{1}{N_g}\sum_{j=1}^{N_g} \frac{f\left(\boldsymbol{u_{gj}}\middle|X_{gq'}, \boldsymbol{\Delta}^{(r)}\right)\pi_{gq}^{(r)}}{\sum_{q'}^{Q} f\left(\boldsymbol{u_{gj}}\middle|X_{gq'}, \boldsymbol{\Delta}^{(r)}\right)\pi_{gq'}^{(r)}} \ , \tag{2.1.7}$$

where N_g is the number of examinees in group g; u_{gji} is the item response of examinee j in group g to item i; $\boldsymbol{u_{gj}}$ is the item response vector of examinee j in group g to all n items; $\Delta^{(r)}$ is the vector of the provisional item parameter estimates obtained at iteration r^{-1} of the EM algorithm; Q is the number of quadrature points; X_{gq} is the q^{th} quadrature point for group g; and $\pi_{gq}^{(r)}$ is the quadrature weight for group g corresponding to X_{gq} estimated at iteration r^{-1} of the EM algorithm.

The notations show that the MMLE-EM procedure can be applied to multiple group response data (note: two groups are considered as an exemplary case here) with shared common items. In the E step, the two quantities (n_{gq} and r_{gq}) can be computed, by treating the portion of the test items not presented to each group as missing or as not administered. Because of the local

independence assumption, each item can be estimated independently; unique items use only the response data of the specific group, but the common items take advantage of all response data from both groups. In the EM cycle, the reference group ability distribution is fixed to the standard normal distribution, while that of the focal group is freely estimated. The common items continue to adjust the scale of the focal group on the reference group until the iteration completes, and the empirical quadrature weights (π_{gq}) are computed and updated in the similar fashion aforementioned. The resulting scale of the new form can be comparable to that of the base form and the underlying ability distribution of the focal group can be estimated relative to that of the reference group. In essence, the scale of CC is established with the information from both groups.

With more information used, CC may provide more accurate common item parameter estimates than SC and FPC do when the sample size is small with an appropriate model fit; and is more efficient than SC because multiple forms are calibrated in one computer run and no additional linking is required. Attributable to the unavailability of all response data and more importantly, inconsistent item parameter estimates after each calibration, however, CC may not be a practical choice as a routine calibration procedure in operation with a few exceptions such as item bank recalibration.

The fixed item parameter calibration (FPC) uses the common item parameter estimates from the previously calibrated form of which examinees are assumed to be representative of the true population because the base scale is strictly identified with only information from the reference group who took the base form. In the estimation process, the parameter estimates of common items of the base form are fixed in the new form and only unique items are estimated with existing common parameters and the response data of the new form. To make the procedure

clearer in comparison with CC, the three quantities (Kim, 2006; Kim, 2017; and Kim & Kolen, 2016) can be re-expressed as follows:

In the first iteration,

$$n_q^{(0)} = \sum_{j=1}^{N} f\left(X_q \middle| \boldsymbol{u_{ci}}, \boldsymbol{\Delta}_{ci}, \pi^0\right) = \sum_{j=1}^{N} \frac{f\left(\boldsymbol{u_{ci}} \middle| X_q, \boldsymbol{\Delta}_{ci}\right)\pi_q^{(0)}}{\sum_{q'}^{Q} f\left(\boldsymbol{u_{ci}} \middle| X_{q'}, \boldsymbol{\Delta}_{ci}\right)\pi_{q'}^{(0)}} \tag{2.1.8}$$

$$r_q^{(0)} = \sum_{j=1}^{N_g} u_{jci} f\left(X_q \middle| \boldsymbol{u_{ci}}, \boldsymbol{\Delta_{ci}}, \pi^0\right) = \sum_{j=1}^{N} u_{jci} \frac{f\left(u_{ci} \middle| X_q, \boldsymbol{\Delta_{ci}}\right)\pi_q^{(0)}}{\sum_{q'}^{Q} f\left(u_{ci} \middle| X_{q'}, \boldsymbol{\Delta_{ci}}\right)\pi_{q'}^{(0)}} \tag{2.1.9}$$

and

$$\pi_q^{(1)} = \frac{n_q^{(0)}}{N} = \frac{1}{N}\sum_{j=1}^{N} \frac{f\left(\boldsymbol{u_{ci}} \middle| X_q, \boldsymbol{\Delta_{ci}}\right)\pi_g^{(0)}}{\sum_{q'}^{Q} f\left(\boldsymbol{u_{ci}} \middle| X_{q'}, \boldsymbol{\Delta_{ci}}\right)\pi_g^{(0)}}, \tag{2.1.10}$$

Note that at the first E step, with $\pi^{(0)}$ from the standard normal distribution as a prior choice, only the common item parameters $\boldsymbol{\Delta_{ci}}$ and observed data $\boldsymbol{u_{ci}}$ for the common items are used to compute the posterior probabilities of the quadrature points. This is the necessary preparation to put the new scale onto the base scale over the subsequent iterations.

In the first M step, the unique item parameters $\boldsymbol{\Delta_{ui}^{(1)}}$ are estimated. From the second EM iteration, the fixed common items and all response data of the new form are used to obtain the final unique item parameters. The equations can be re-expressed as follows:

$$n_q^{(r+1)} = \sum_{j=1}^{N} f\left(X_q \middle| \boldsymbol{u_{ci}}, \boldsymbol{\Delta_{ci}}, \boldsymbol{\Delta_{ui}^{(r)}}, \pi^{(r)}\right)$$

$$= \sum_{j=1}^{N} \frac{f\left(\boldsymbol{u_j} \middle| X_q, \boldsymbol{\Delta_{ci}}, \boldsymbol{\Delta_{ui}^{(r)}}\right)\pi_q^{(r)}}{\sum_{q'}^{Q} f\left(\boldsymbol{u_j} \middle| X_{q'}, \boldsymbol{\Delta_{ci}}, \boldsymbol{\Delta_{ui}^{(r)}}\right)\pi_{q'}^{(r)}} \tag{2.1.11.a}$$

$$r_q^{(r+1)} = \sum_{j=1}^{N_g} \boldsymbol{u}_{ji} f\left(X_q \middle| \boldsymbol{u}_j, \boldsymbol{\Delta}_{ci}, \boldsymbol{\Delta}_{ui}^{(r)}, \pi^{(r)}\right)$$

$$= \sum_{j=1}^{N} \boldsymbol{u}_{ji} \frac{f\left(\boldsymbol{u_j}\middle|X_q, \boldsymbol{\Delta}_{ci}, \boldsymbol{\Delta}_{\boldsymbol{ui}}^{(\boldsymbol{r})}\right)\pi_q^{(r)}}{\sum_{q'}^{Q} f\left(\boldsymbol{u_j}\middle|X_{q'}, \boldsymbol{\Delta}_{ci}, \boldsymbol{\Delta}_{\boldsymbol{ui}}^{(\boldsymbol{r})}\right)\pi_{q'}^{(0)}} \qquad (2.1.11.b)$$

and

$$\pi_q^{(r+1)} = \frac{n_q^{(0)}}{N}$$

$$= \frac{1}{N}\sum_{j=1}^{N} \frac{f\left(\boldsymbol{u_j}\middle|X_q, \boldsymbol{\Delta_{ci}}, \boldsymbol{\Delta}_{\boldsymbol{ui}}^{(\boldsymbol{r})}\right)\pi_q^{(r)}}{\sum_{q'}^{Q} f\left(\boldsymbol{u}_j\middle|X_{q'}, \boldsymbol{\Delta}_{ci}, \boldsymbol{\Delta}_{\boldsymbol{ui}}^{(\boldsymbol{r})}\right)\pi_{q'}^{(r)}}, \qquad (2.1.11.c)$$

where $\boldsymbol{\Delta}_{\boldsymbol{ui}}^{(\boldsymbol{r})}$ denotes the unique item parameters in iteration r. $\boldsymbol{\Delta_{ci}}$ without the superscript r indicates that the common item parameters are fixed (or not updated) in the EM iteration. Kim (2006) compared five FPC methods and out of the five methods, the Multiple Prior Weights Updating and Multiple EM Cycles method is presented above and used in the current study because of its better performance.

FPC neither requires multiple computer-runs for calibration nor additional linking. In addition, FPC can be useful for validating the invariance property of item parameters and developing an item pool.

Because of the common scale established via the calibration process, CC and FPC are considered as an alternative to traditional IRT linking procedures. In contrast, for SC, linking is imperative. Four linking procedures are discussed in detail in the following section.

IRT LINKING

Due to the scale indeterminacy, the common procedure to identify the UIRT metric with its origin and unit is setting the examinee distribution to be a standard normal distribution with mean zero and standard deviation one regardless of the true population distribution. Therefore, when two forms are calibrated independently, the resulting two scales are not the same. Thus, the population differences must be adjusted. The statistical procedure to adjust population

differences is called linking. Linking is possible on account of the linear relationship of item parameters and the preservation of probability for item endorsement after linear transformation.

To illustrate, the linear relationship of difficulty parameters of common items on scale Y and X can be written as follows (Cook & Eignor, 1991):

$$\frac{b_{Y_j} - \mu(b_Y)}{\sigma_{b_Y}} = \frac{b_{X_j} - \mu(b_X)}{\sigma_{b_X}}, \tag{2.1.12}$$

or equivalently,

$$b_{Y_j} = Ab_{X_j} + B, \tag{2.1.13}$$

Under 3PL UIRT, the probability of item endorsement is shown as follows:

$$P\left(\theta_{Y_i}, a_{Y_j}, b_{Y_j}, c_{Y_j}\right) = c_{Y_j} + \left(1 - c_{Y_j}\right)\frac{exp\left[Da_{Y_j}\left(\theta_{Y_i} - b_{Y_j}\right)\right]}{1 + exp\left[Da_{Y_j}\left(\theta_{Y_i} - b_{Y_j}\right)\right]} \tag{2.1.14}$$

$$= c_{X_j} + \left(1 - c_{X_j}\right)\frac{exp\left\{D\frac{a_{X_j}}{A}\left[\left(A\theta_{X_i} + B\right) - \left(Ab_{X_j} + B\right)\right]\right\}}{1 + exp\left\{D\frac{a_{X_j}}{A}\left[\left(A\theta_{X_i} + B\right) - \left(Ab_{X_j} + B\right)\right]\right\}}$$

$$= c_{X_j} + \left(1 - c_{X_j}\right)\frac{exp\left[Da_{X_j}\left(\theta_{X_i} - b_{X_j}\right)\right]}{1 + exp\left[Da_{X_j}\left(\theta_{X_i} - b_{X_j}\right)\right]}$$

$$= P\left(\theta_{X_i}, a_{X_j}, b_{X_j}, c_{X_j}\right),$$

where $\theta_{Y_i} = A\theta_{X_i} + B; a_{Y_j} = \frac{a_{X_j}}{A}; b_{Y_j} = Ab_{X_j} + B; c_{Y_j} = c_{X_j}, and\ D \approx 1.7$. Notice that the guessing parameter is independent of the scale transformation. The probability of examinee i getting the item j correctly on the scale Y, denoted as $P\left(\theta_{Y_i}, a_{Y_j}, b_{Y_j}, c_{Y_j}\right)$ equals to that on the scale X, expressed as $P\left(\theta_{X_i}, a_{X_j}, b_{X_j}, c_{X_j}\right)$.

Based on the information used to find a set of the optimal linking constants, A (multiplicative constant) and B (additive constant) that minimize the difference of two sets of

common item parameters, there, in general, are two linking procedures: moment methods and characteristic curve methods.

The moment methods are mean/sigma and mean/mean methods. The mean/sigma (MS) method uses the mean and standard deviation of difficulty parameter estimates of common items from both test forms. A is the ratio of the standard deviation of the common item parameter estimates in the old form (Y) to that in the new form (X). B is the mean difference between the common item parameter estimates from the old and new forms.

$$A = \frac{\sigma(a_Y)}{\sigma(a_X)}, \tag{2.1.15}$$

and

$$B = \mu(b_Y) - A * \mu(b_X), \tag{2.1.16}$$

The mean/mean (MM) method estimates the linking coefficients using item discrimination and difficulty parameters. Thus, A is the ratio of the mean of the common item discrimination parameter estimates in the old form to that in the new form, and B is computed in the same manner as does MS.

$$A = \frac{\mu(a_Y)}{\mu(a_X)}, \tag{2.1.17}$$

The characteristic curve methods utilize all common item parameters together to find the optimal linking coefficients with which each of the item characteristic curves (ICC) or a test characteristic curve (TCC) on the new scale is transformed and matched with the counterpart on the old scale. The goal of the Haebara (HB) method is to find the A and B that minimize the

criterion function, the cumulative squared difference values of each pair of the ICCs of common items (V) over all examinees (N) between the old scale and the new transformed scale, as such:

$$Hcrit = \sum_{i}^{N}\sum_{j}^{V}\left[p_{ij}\left(\theta_{Yi};\hat{a}_{Yj},\hat{b}_{Yj},\hat{c}_{Yj}\right) - p_{ij}\left(\theta_{Yi};\frac{\hat{a}_{Xj}}{A},A\hat{b}_{Xj}+B,\hat{c}_{Xj}\right)\right]^{2}, \qquad (2.1.18)$$

The Stocking and Lord (SL) method finds the linking constants that minimize the cumulative squared difference of two TCCs (or the sum of ICCs) over all examinees between the old scale and the new transformed scale, as such:

$$SLcrit = \sum_{i}^{N}\left[\sum_{j}^{V} p_{ij}\left(\theta_{Yi};\hat{a}_{Yj},\hat{b}_{Yj},\hat{c}_{Yj}\right) - \sum_{j}^{V} p_{ij}\left(\theta_{Yi};\frac{\hat{a}_{Xj}}{A},A\hat{b}_{Xj}+B,\hat{c}_{Xj}\right)\right]^{2}, \qquad (2.1.19)$$

Its underlying concept is that the true scores from the two forms would be indistinguishable for examinees.

It is worth noting that in 1PL UIRT and Rasch model, A is 1 across all four linking methods on account of the assumption that the scale unit is unchanged; and that in the characteristic curve methods, there is no difference between 2PL UIRT and 3PL UIRT in the linking constant A and B because the lower asymptotes are not involved in the scale transformation.

Unlike the moment methods that use moments of item parameter distributions, which are outcomes of relatively simple computations, the characteristic curve methods are computationally more demanding to implement, attributable to the numerical integration of the latent variable and the iterative searching algorithm for the optimal linking constants in the multivariate space. In addition, the criterion function is non-linear in respect to the constants A and B (Kim & Lee, 2004). In the current study, the characteristics curve methods are used for

their outperformance over the moment methods (Baker and Al-Karni 1991; Hanson and Béguin 2002; Kim and Cohen 1992; Lee and Ban 2010).

In an ideal case when the two sets of common item parameters are equivalent, there is no adjustment required for the origin and unit of the two scales of the common item sets. Thus, B becomes zero and A gets one. However, if there is a discrepancy between two sets of common items, then, the discrepancy explains the population difference between the two examinee groups. The role of the linking constants is to adjust the scale of the new form to the old scale in terms of the origin and unit. Under the random group linking design with an assumption that two groups are equivalent to one another, the discrepancy is expected to be, if any, inconsequential due to sampling error and parameter estimation error. Under the CINEG design, however, the disparity is assumed to be non-trivial due to the true population difference beyond such errors, which must be addressed in a proper manner.

IRT EQUATING

After the set of item parameters of the new form is put on the common/base scale, the differences of form difficulty can be adjusted for interchangeable scores. For the valid score comparison, test forms should meet rigorous requirements (Dorans & Holland, 2000; Kolen & Brennan, 2014). That is, tests to be equated should measure the same constructs (equal-construct requirement); the equating transformation of scores should be symmetric (symmetry requirement); and it should be a matter of indifference to the examinees regardless of which form of the test is administered (equity requirement). The last requirement is generally known as Lord's equity property that is possible only in the case that two alternate forms are essentially identical. In practice, however, neither such an ideal form construction is feasible nor is equating, even so, necessary for identical forms. As an evaluation criterion in the current study, a less

restrictive version of the equity property (Morris 1982, as cited in Kolen & Brennan 2014) is used, which will be further discussed in detail later. Under a satisfactory condition of the requirements, two forms can be equated with two IRT equating procedures.

Before discussing the procedures in detail, however, it is first necessary to clarify the term "score" in IRT. To be specific, related to the examinee's ability, three values can be obtained: IRT true scores (i.e., a sum of ICCs or the expected value of the observed score regressed on ability level), observed number-correct scores, and latent ability estimates. In the equating context, both IRT true scores and observed number-correct scores are treated as observed scores which depend on the length of the test, while the latent ability estimates represent the true ability of examinees, which are invariant across forms. (Thus, for the score comparison with the latent ability estimates, no equating is required.)

The observed score equating (OSE) follows the conventional equipercentile method to find score equivalents after obtaining the observed summed-score distribution. The LW recursion formula (Lord & Wingersky, 1984) can be used to compute the conditional summed-score distribution as follows:

When r = 1

$$f_1(x = 0|\theta_i) = (1 - p_{i1}),$$

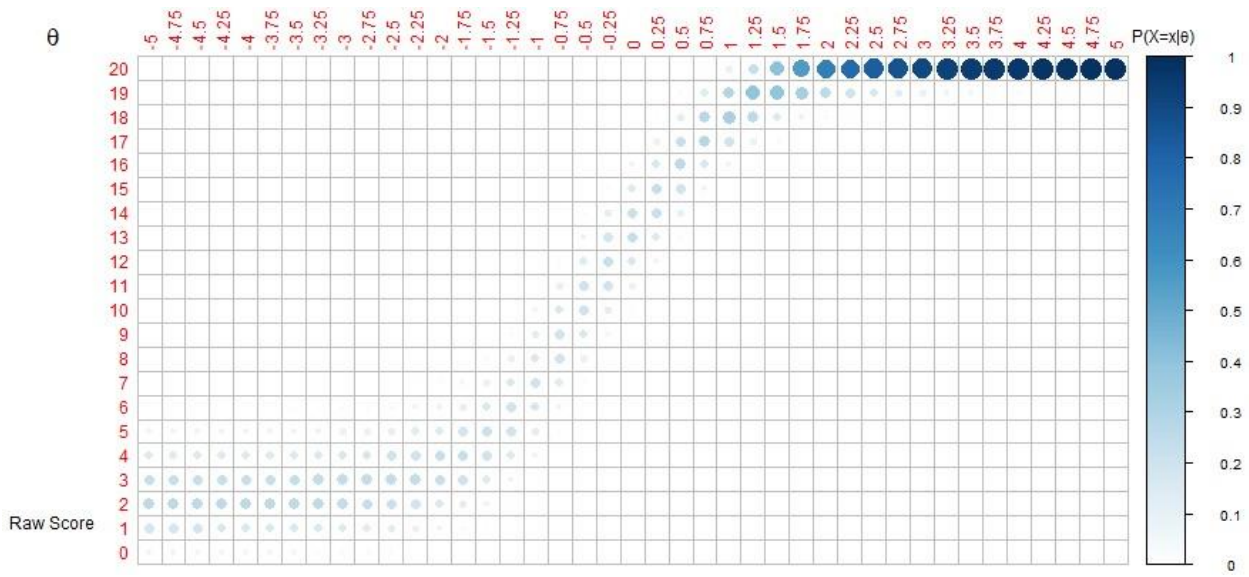
$$f_1(x = 1|\theta_i) = p_{i1},$$

When r > 1, recursive process is invoked,

$$\begin{aligned} f_r(x|\theta_i) &= f_{r-1}(x|\theta_i)(1 - p_{ir}), && x = 0 \\ &= f_{r-1}(x|\theta_i)(1 - p_{ir}) + f_{r-1}(x - 1|\theta_i)p_{ir}, && 0 < x < r \\ &= f_{r-1}(x - 1|\theta_i)p_{ir}, && x = r \end{aligned} \qquad (2.1.20)$$

where r is the number of items; x is the number-correct score; and $f_r(x|\theta_i)$ is the distribution of number-correct scores over the first r items for examinees of anility θ_i.

Figure 2. Visualization of the Conditional Summed-score Distribution Computed with LW Recursion Formula on 3PL UIRT



In Figure 2, the conditional probability of the raw scores is low in the middle of the score range where the mean of item difficulty is matched with that of the population ability and gets larger as the scores approach to the minimum and maximum scores. In this case, the conditional probabilities scatter around the low score because of pseudo-guessing.

Then, the marginal distribution is computed with the numerical integration of the conditional distribution across all latent ability space.

$$f(x) = \sum_i f(x|\theta_i)\psi(\theta_i)\,, \tag{2.1.21}$$

where $\psi(\theta_i)$ is the discrete distribution of anility θ on a finite number of equally spaced points. With the marginal distribution $f(x)$, the cumulative distribution $F(x)$ is obtained. The final step is to apply the traditional equipercentile method to the cumulative distribution:

$$e_Y(x) = F_Y^{-1}\big(F_X(x)\big), \tag{2.1.22}$$

where $e_Y(x)$ is the Form Y equivalent of score x on Form X; F_X and F_Y are the cumulative distribution functions for each scale; and $F_Y^{-1}$is the inverse function of F_Y.

Like OSE, the true score equating (TSE) is designed to map scores on a new form to those on the base form. Instead of using percentile ranks in OSE as an anchor for one-to-one score mapping, however, TSE finds score equivalents by anchoring the corresponding latent ability estimates.

DATA COLLECTION DESIGNS

With the assumption of a single population for equating, a statistical procedure to adjust the difference in difficulty of two alternate forms, individual groups taking each form should have no difference in distributional properties of examinee ability or proficiency which determines the metric of the IRT scale. To control for the differences in abilities of examinee groups, data collection methods must be implemented. Typical means used are common items, common examinees, a common ability distribution, or some combination of them. For example, the random groups design obtains the equivalence between examinee groups by assigning alternate forms randomly to examinees. In contrast, the single group design accomplishes it by administering both alternate forms to the same examinee group. The common-item nonequivalent groups (CINEG) data collection design utilizes common items as a medium to decompose the group difference and form difference. The group difference is adjusted through scale linking methods and then the form difference is adjusted through score equating procedures. Thus, the common (or anchor) items should be representative of the total test form in content and psychometric properties. IRT calibrated item pool design is similar to the CINEG design in constructing common item sets but more flexible in equating than the CINEG design

because anchor items are in common with the item pool, rather than being in common with a previous form. In the next section, the requirements of common items are explained more in detail.

The choice of data collection design can be based on the consideration of test development process, test administration, test security, sample size requirements, and statistical assumptions. In the IRT equating, the CINEG data collection design and IRT calibrated item pools require the most complex test development process due to the construction of common item sections and the strongest statistical assumptions for development of item pools. However, both approaches are easy to implement and provides greater test security because only one form needs to be administrated on a particular test data to conduct equating. However, the common item sections are not completely immune from the item security because of the repeated administrations.

The current study focuses on the CINEG data collection design which becomes more popular due to its flexibility in administration and benefits of IRT calibrated item pools. For more details of the equating data collection designs, refer to Kolen and Brennan (2014) and Petersen, Kolen, and Hoover (1989).

2. MULTIDIMENSIONAL ITEM RESPONSE THEORY

Even with the unidimensionality assumption, in the UIRT framework, response data never become unidimensional. To reflect a more realistic case, in the current study, response data are simulated via multidimensional IRT(MIRT) models, and as a frame of reference, MIRT OSE is utilized. Introducing MIRT is of importance, but its portions are discussed here, only relevant to the current study.

When two or more latent abilities are required in the item response process, the UIRT can be extended to the MIRT which provides a more accurate representation of persons and items in the multidimensional latent space. The functional form of item characteristic surface function of the 3PL MIRT is expressed as follows:

$$P\left(U_{ij}=1\middle|\boldsymbol{\theta_j},\boldsymbol{a}_i,d_i\right)=\ c_i+(1-c_i)\frac{e^{\boldsymbol{a_i\theta_j'}+d_i}}{1+e^{\boldsymbol{a_i\theta_j'}+d_i}} \tag{2.2.1}$$

or equivalently,

$$P\left(U_{ij}=1\middle|\boldsymbol{\theta}_j,\boldsymbol{a_i},b_i\right)=\ c_i+(1-c_i)\frac{e^{\boldsymbol{a_i}(\theta_j-b_i)'}}{1+e^{\boldsymbol{a_i}(\theta_j-b_i)'}}, \tag{2.2.2}$$

where $\boldsymbol{\theta_j}$ denotes the latent trait vector for examinee j; $\boldsymbol{a_i}$ is the item discrimination vector for item i; c_i is the pseudo guessing parameter for item i; d_i is the intercept for item i; $\boldsymbol{a_i\theta_j'}+d_i$ can be re-expressed as $a_{i1}\theta_{j1}+a_{i2}\theta_{j2}+\cdots+a_{im}\theta_{jm}+d_i$ or $\sum_{\boldsymbol{l=1}}^{\boldsymbol{m}}\boldsymbol{a_{il}\theta_{jl}}+\boldsymbol{d_i}$ (m is the number of latent variables).

The multidimensional difficulty of the UIRT equivalent can be obtained by the equation below:

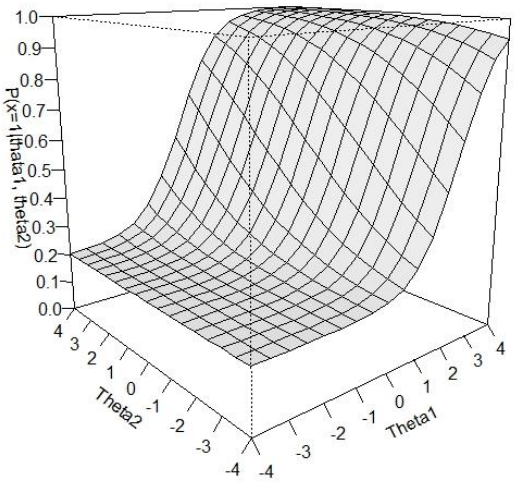
$$B_i=\frac{(-d_i)}{\sqrt{\sum_{k=1}^{m}a_{ik}^2}}=\frac{(-d_i)}{A_i}, \tag{2.2.3}$$

and the multidimensional discrimination for item i is the norm of a multidimensional discrimination vector,

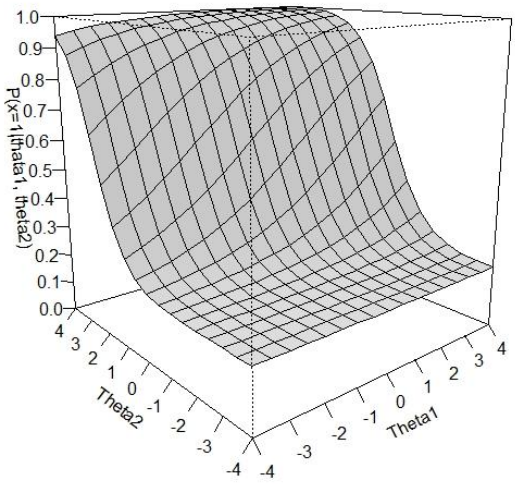
$$A_i=\sqrt{\sum_{k=1}^{m}a_{ik}^2}, \tag{2.2.4}$$

Note that when m =1, A_i becomes a_i and B_i becomes b_i by $\frac{-d_j}{a_i}$. As in UIRT, discrimination parameters are indicative of the discriminating power of a given item to examinees in the latent dimension, but in contrast, discrimination parameters in MIRT provide the information of the discriminating power to the multiple dimensions. Put differently, a discrimination parameter represents the magnitude of measurement of a given item to a specific dimension, relative to that of the other dimension(s). The MIRT pseudo-guessing parameter is directly comparable to that of UIRT.

Figure 3. Item Characteristic Surfaces of Two Items Measuring Two Latent Traits.



(a) Discrimination (1.5, 0.5), Difficulty (1, 0) for two latent dimensions



(b) Discrimination (0.5, 1.5), Difficulty (0, 1) for two latent dimensions

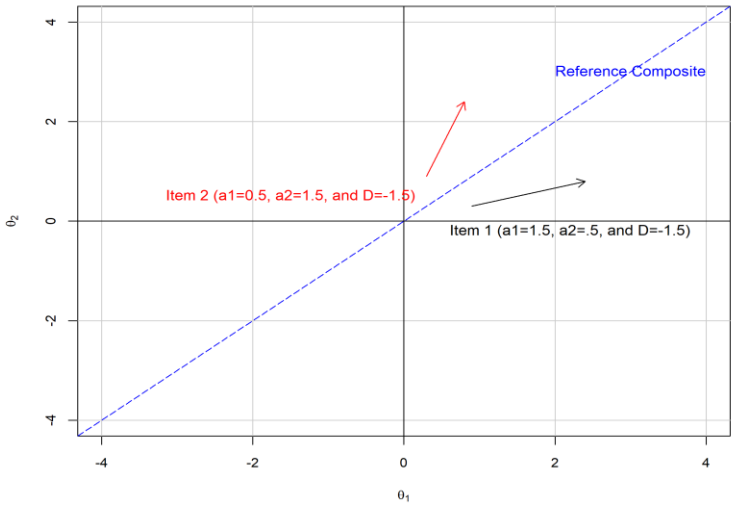
Note that two items in Figure 3 have the same A (i.e., 1.58) and B (i.e., 0.95) even though each has different discrimination values to the latent dimensions. In other words, in Figure 2.2.1, the distance from the origin of the plane to the base of two items vectors is the location of the item difficulty, 0.95 and the magnitude of the discrimination of two item vectors is the length of

the vector, 1.58. It is necessary to find the direction of the measurement, which can be expressed an angle of the item vector to a specific dimension as such:

$$\alpha_{ik} = arccos\left[\frac{a_{ik}}{\sqrt{\sum_{k=1}^{K} a_{ik}^2}}\right], \tag{2.2.5}$$

where α_{ik} denotes the angle of item i to dimension k. For instance, the first item vector has the angle (i.e., 32 degree) to the first dimension, which is the same angle of the second item vector to the second dimension, as shown in Figure 2.2.2 below. In addition, the probability of item endorsement of both items for an examinee with the latent ability (1, 1) is the same 0.70. It is worth noting that the two items measure the two latent traits with the same difficulty level, magnitude of discrimination, and probability of item endorsement, but the first item measures dominantly the first latent trait, while the second item primarily does the second latent trait, which is visually illustrated in Figure 4.

Figure 4. Item Characteristic Surfaces of Two Items Measuring Two Latent Traits.



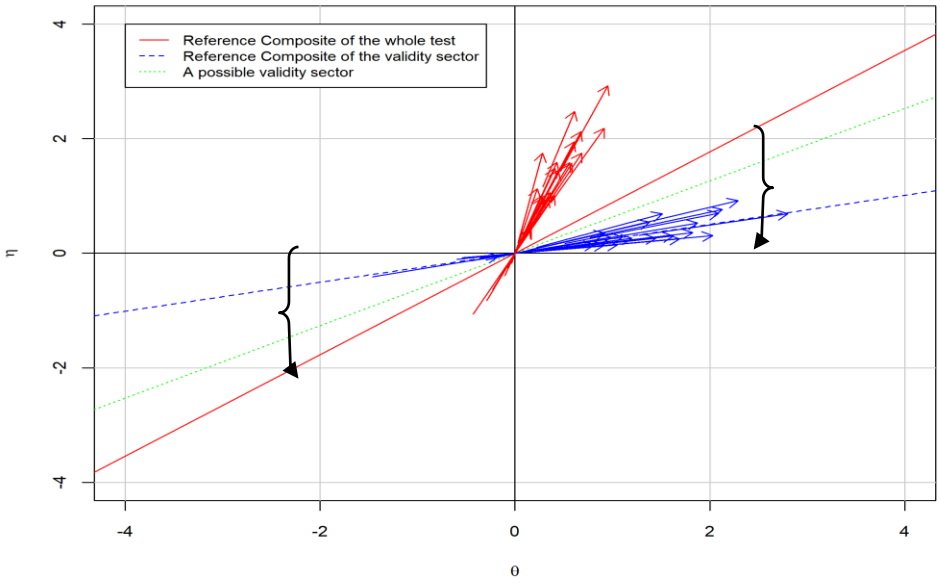
Note that the angles to the first-dimension axes are $\alpha_1 = 32°$ for the itme 1, $\alpha_2 = 58°$ for the item 2, and $\theta_\alpha = 45°$ for the refrence composite, the blue dotted line, which is the linear combination of the two items.

Unidimensional Approximation

Wang (1987) conjectured the relationship between the multidimensional latent space and its unidimensional projection which is a fitted UIRT to multinational response data. She defined the approximate unidimensional scale as the "reference composite", which represents a linear composite of the dimensions present in the test, or the mean of the directional cosines for a given item cluster (Luecht & Miller, 1992). That is, a linear composite indicates the direction of the first eigenvector of the matrix $\boldsymbol{A^T A}$, where $\boldsymbol{A}$ is the matrix of item discrimination parameters. The line of the reference composite shown in Figure 5 passed the origin with the slope 1 because the two elements of the first eigen vector is same, 0.71.

Akerman (1992) proposed a validity sector, the narrow angle between the primary latent dimension and the reference composite of the item set that measure the purported trait. In the case where the purported trait is the first latent dimension, and the second dimension is considered as a nuisance factor, items out of the validity sector are invalid. He also suggested the construct validity index (CVI) which is obtained by $cos^2(\alpha_i - \theta_\alpha)$ which ranges from 0 (i.e., completely invalid) to 1 (i.e., totally valid). The CVI of both items is .95, indicating that the two items are valid, measuring the same combinations of skills as the reference composite. However, the two items should be interpreted differently. One caveat of the reference composite is that the reference composite of the test with enough invalid items can be pulled out of the validity sector's reference composite (Ackerman, 1992, p. 74), as shown in Figure 2.2.3. With the given two items as an example, the second item may be incorrectly considered as a valid item when the construct of interest is the first dimension.

Figure 5. Visualization of Reference Composites and Validity Sector



Similar in the concept, but different in the mathematical formulation, Zhang (1996), Zhang and Stout (1999a and 1999b) derived the unidimensional linear composite referred to as "the direction of best measurement", (see Reckase, 2009 for the concise explication) and Zhang and Wang (1998) derived corresponding item parameter estimates, expressed as such:

$$\theta_\alpha = \boldsymbol{\alpha^t \theta} \tag{2.2.6}$$

$$P_i\,(Y\ =\ 1|\theta_\alpha\,)\ =\ c_i + (1 - c_i)\frac{1}{1\ +\ exp[\,-\boldsymbol{a_i^\star}\,\boldsymbol{\theta_\alpha} - d_i^*\,]} \tag{2.2.7}$$

$$a_i^\star\ =\ \left(1\ +\ \sigma_i^{*2}\right)^{-1/2}\boldsymbol{a_i^T}\sum\alpha \tag{2.2.8}$$

$$d_i^*\ =\ \left(1\ +\ \sigma_i^{*2}\right)^{-1/2} d_i \tag{2.2.9}$$

and

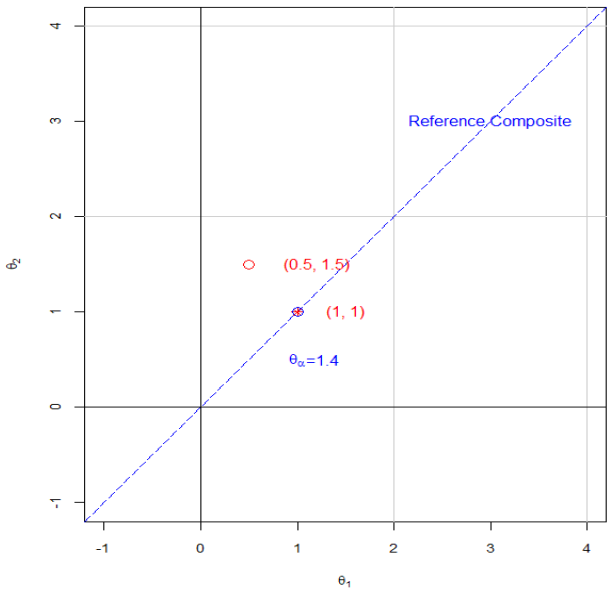
$$\sigma_i^*\ =\ \boldsymbol{a_i^T}\ \sum a - \left(\boldsymbol{a_i^T}\ \sum a\right)^2, \tag{2.2.10}$$

where $\boldsymbol{\alpha}$ is a vector of weights with constraint $\boldsymbol{\alpha^T}\sum\boldsymbol{\alpha} = 1$; $\sum$ is a correlation matrix; θ_α represent the composite latent dimension; $a_i^\star$, and d_i^* denote the discrimination and intercept of

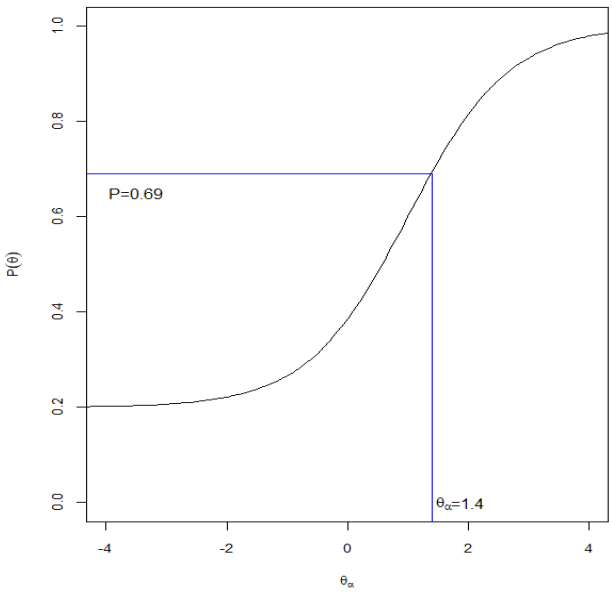
the composite latent dimension; $P_i\,(Y\ =\ 1|\theta_\alpha\,)$ is the probability of item endorsement. Strachan et.al. (2022) investigated the validity of the linear composite conjecture. Their simulation study demonstrated that the fitted UIRT model sufficiently approximates the linear composite direction in a multidimensional space.

In the given example above, $\boldsymbol{A} = \begin{bmatrix}1.5 & 0.5\\ 0.5 & 1.5\end{bmatrix}$, and the eigenvector of the first eigne value of eigen decomposition of $\boldsymbol{A^T A}$ is $\upsilon_{\lambda_i} = \binom{0.71}{0.71}$. The first weight α_1 becomes 0.71 obtained by $\frac{\upsilon_1}{\|\upsilon_{\lambda_i}\|}$ and α_2 is obtained by $\sqrt{1-\alpha_1^2}$. For the sake of simplicity, set the correlation zero, and with $\theta = \binom{1}{1}$, $\alpha = \binom{0.71}{0.71}$ and $c = 0.2$, the resulting properties of the two items in the composite latent space are identical. That is, $\alpha^\star$ =1.2, $d^\star = -1.2$, and $p(x = 1|\boldsymbol{\theta}])$ = 0.69 for both items and the composite latent ability θ_α is 1.4. Furthermore, holding all properties for the items equal, another examinee with $\theta = \binom{0.5}{1.5}$ gets $\theta_\alpha = 1.4$, and 0.68 probability, which is same as the examinee with $\theta = \binom{1}{1}$. In this simplified case, two points can be noticed; two items that measure two latent traits with different compositions of item property can become identical in the composite dimension, which is considered unidimensional; in contrast, two examinees with different combinations of proficiency levels on latent traits can become identical at the composite dimension. A graphical representation for the two item vectors and latent abilities are visualized in Figure 6 below.

Figure 6. Unidimensional Approximation (aka, Reference Composite)



(a) Both latent ability vectors: (0.5, 1.5) and (1, 1) are projected onto the reference composite of the value 1.4



(b) The probability of an item endorsement for $\boldsymbol{\theta_{\alpha} = 1.4, \alpha^{\star} = 1.2, , d^{\star} = -1.2, and\ c = 0.2}$

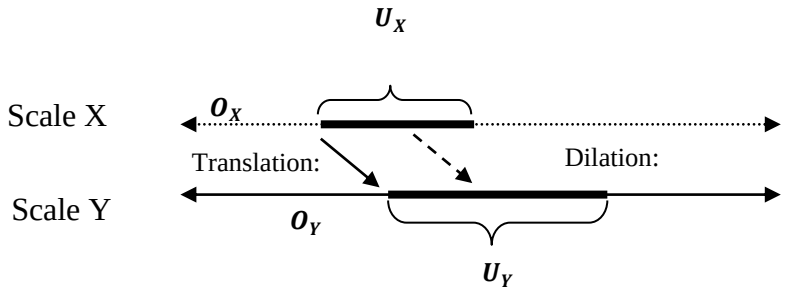
Previous studies (e.g., Ackerman, 1992; Luecht & Miller, 1992; Wang, 1987; and Zhang & Wang, 1998) illuminated the relationship between the UIRT and the MIIRT in terms of latent dimensions and item parameters and provided evidence about the validity and robustness of UIRT fitted to the multidimensional response data.

In the operational setting, for licensing and certification organizations, it is one of the major concerns to corroborate the validity of test scores which should be a valid representation of examinees' true ability on the target latent dimension. Thus, the current study takes the practical concerns further and deeper in investigating the impact of differential dimensional structures on results of linking and equating. The following is a brief introduction of the linking and equating of MIRT.

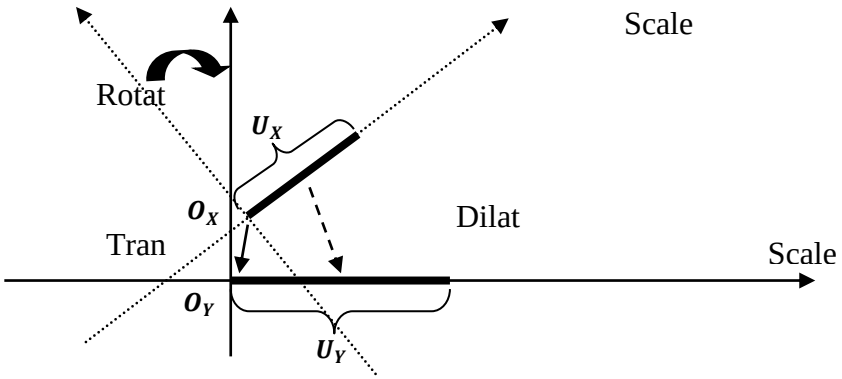
MIRT LINKING

In addition to the translation of the origin and dilation of the unit in UIRT, the scale linking in MIRT requires the consideration of one more indeterminacy, rotation of scale. The pictorial comparison of linking scales between UIRT and MIRT was illustrated in Figure 7.

Figure 7. UIRT and MIRT Linking Components



(a) UIRT Scale Linking



(b) (2D) MIRT Scale Linking

Note: O is the location of origin, and U is the length of unit, and $X \rightarrow Y$ denotes metric transformation from scale X to scale Y. (Modified from Min, 2007, p. 43). Note that for the scale identification in MIRT, a multivariate standard normal distribution is used with a mean vector with zero and a diagonal matrix with 1 for the variance/covariance matrix (i.e., $\boldsymbol{\theta} \sim MVN(\mathbf{0}, \boldsymbol{I})$).[1]

[1] $\boldsymbol{MVN}(\mathbf{0}, \boldsymbol{I})$ is often a preferred choice for the reference group, even though real traits are likely to be correlated at some degree, because it consists of an orthogonal rotation, a translation transformation, and a single dilation or contraction (Li & Lissitz, 2000).

As an extension of the 3PLUIRT linking, MIRT linking procedures obtain the linking coefficients in a matrix and a vector form as follows:

$$\boldsymbol{\theta_{Yi}} = \boldsymbol{T^{-1}\theta_{Xi}} + \boldsymbol{\beta} \tag{2.2.11}$$

$$\boldsymbol{a_{Yi}^{t}} = \boldsymbol{a_{Xi}^{t}T} \tag{2.2.12}$$

$$d_{Yi} = d_{Xi} - \boldsymbol{a_{Yi}^{t}\beta} \tag{2.2.13}$$

and

$$c_{Yi} = c_{Xi} \tag{2.2.14}$$

where $\boldsymbol{\theta}$ is a vector of multidimensional ability estimates, $\boldsymbol{T}$ is a transformation matrix to account for rotational indeterminacy (i.e., off-diagonal elements for correlation between dimensions) and dilation indeterminacy (i.e., diagonal elements for the unit of measurement for each dimension), $\boldsymbol{\beta}$ represents the translation vector for the translation indeterminacy, $\boldsymbol{a}$, d, and c are the slope parameter vector, intercept parameter and pseudo-guessing parameter, respectively, for the base (i.e., Y) form and the new (i.e., X) form.

Of a few MIRT linking procedures to estimate $\boldsymbol{T}$ and $\boldsymbol{\beta}$ (Thompson et al., 1997; Hirsch, 1989; Davey et al, 1996; Li & Lissitz, 2000; Min, 2003; and Yon, 2006), in the current study, for the methodological consistency, test characteristic curve method is chosen from the two methods proposed by Oshima et al. (2000). Two characteristic curve methods are direct extensions of the UIRT linking methods: Stocking and Lord (1983) and Haebara (1980) methods. In both procedures, rescaled parameters are obtained by minimizing the cumulative squared difference between TCCs over items and ICCs for each item for examinees of a particular ability. For instance, for the 2D MIRT, the SL method can be expressed as:

$$\sum_{q1=1}^{Q1}\sum_{q2=2}^{Q2} w_{q1q2}\left[T_X\left(\theta_{q1},\theta_{q2}\right) - T_Y^{\star}\left(\theta_{q1},\theta_{q2}\right)\right]^2 \tag{2.2.15}$$

and HB method can also be shown as:

$$\sum_{j=1}^{V}\sum_{q1=1}^{Q1}\sum_{q2=2}^{Q2} w_{q1q2}\left[ICC_X(\theta_{q1},\theta_{q2}) - ICC_Y^{\star}(\theta_{q1},\theta_{q2})\right]^2 \quad (2.2.16)$$

where $Q1$ and $Q2$ are the number of quadrature points for the first and second dimensions, respectively; w_{q1q2} is the weights for corresponding quadrature points. $T_X(\theta_{q1},\theta_{q2})$ denotes the test characteristic surface (TCS) for Form Y, and $T_Y^{\star}(\theta_{q1},\theta_{q2})$ indicates the transformed TCS of Form X; and $ICC_X(\theta_{q1},\theta_{q2})$ is the ICC for Form Y and $ICC_Y^{\star}(\theta_{q1},\theta_{q2})$ is the transformed ICC for Form X.

With a rotation matrix, $\boldsymbol{A}$ and a translation vector, $\boldsymbol{\beta}$, obtained, the equality of the probability for item endorsement can be expressed as follows:

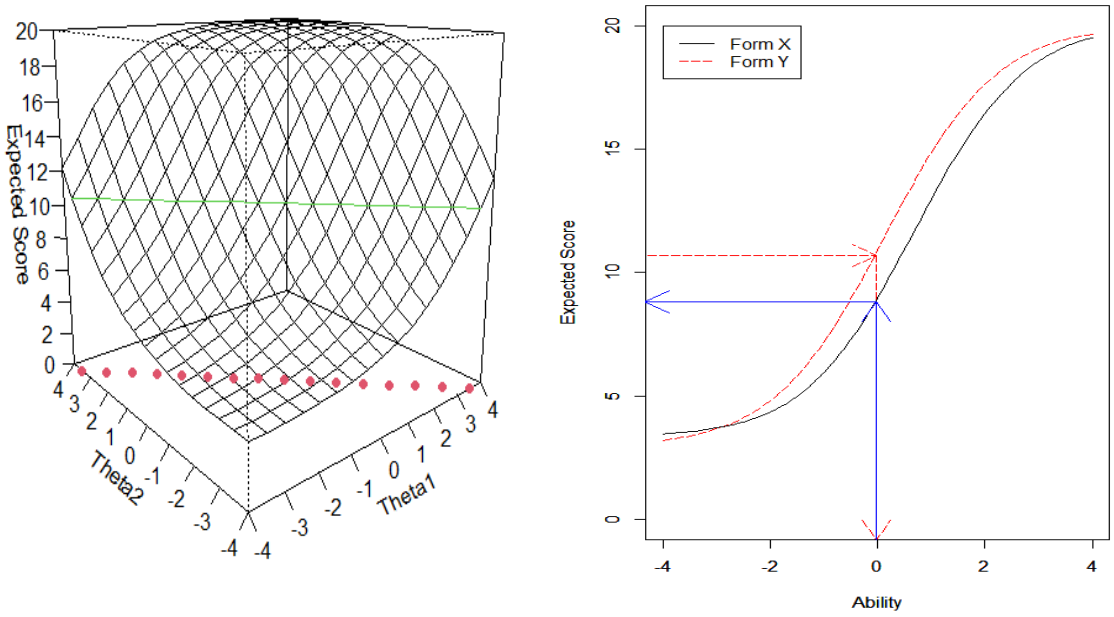
$$\boldsymbol{a_i^{T^*}\theta_j^*} + d_i^* = \left(\boldsymbol{a_i^T A^{-1}}\right)\left(\boldsymbol{A\theta_j + \beta}\right) + \left(d_i - \boldsymbol{a_i^T A^{-1}\beta}\right) = \boldsymbol{a_i^T\theta_j} + d_i \quad (2.2.17)$$

where * indicates metric transformation onto the base scale.

MIRT EQUATING

After the scale linking performed to put the new scale to the base scale for score comparability by adjusting its rotation, unit, and origin, various equating methods can be employed to find the equivalents of scores in the base scale. Unlike the case in UIRT, however, TSE is not a feasible option for the full MIRT due to the one-to-many relation between an expected score and its corresponding combinations of latent abilities in TCS, which makes it impossible to find the unique combination in the multidimensional latent space shown in Figure 8 below.

Figure 8. Ture Score Equating in IRT



(a) MIRT TCS (b) TSE in UIRT

In contrast, MIRT OSE does not suffer the one-to-many relationship between an observed score and its combinations of latent variables. Instead, as the direct extension of UIRT OSE, MIRT OSE obtains the marginal distribution $f(x)$ of observed scores by summing out the latent variables from the conditional summed-score distribution $f(x|\boldsymbol{\theta_i})$.

$$f(x) = \sum_i f(x|\boldsymbol{\theta_i})\psi(\boldsymbol{\theta_i}), \qquad (2.2.18)$$

One convenient choice for the multivariate ability density $\psi(\boldsymbol{\theta_i})$ can be MVN (**0, I**), where **0** is a mean vector with zeros and **I** is an identity matrix with ones (i.e., variance) in diagonal entries and zeros (i.e., covariance/correlation between latent variables) in the off-

diagonal entries. That is, all latent variables have the same origin and unit of measurement with the orthogonal angle between axes in the multidimensional coordinate system.

3. *REVIEW OF RELEVANT LITERATURE*

CALIBRATION

The concurrent and fixed item parameter calibration methods are often chosen as the convenient alternative for linking, and the effectiveness of those methods are evaluated in comparison with the separate calibration method.

Kim and Cohen (1998) examined the performance of two linking procedures: SC using BILOG-MG (Mislevy & Bock, 1990) and SL linking performed by EQUATE (Baker, 1993), and CC using BILOG-MG with MMAPE (marginal maximum a posteriori estimation) and MULTILOG (Thissen, 1991) with MMLE. Their study was conducted on simulated response data generated with 2PL UIRT, with factors: different lengths of common items (5, 10, 30, and 50 items) in the test of 50 items, group difference, and sample size (500 examinees). Based on the evaluation criteria, root mean square difference (RMSE) and the mean Euclidean distance (MED), the authors found that with a large number of common items (more than 5 in a test of 50 items), three linking methods performed similarly, while with a small number of common items, SC performed better.

In contrast, based on the equating results, Petersen et al. (1983) concluded that CC did not perform better, leaving the further investigation; they expected that CC would produce more stable equating results because it does not make assumptions about the relationship between the item parameter scales in the case of SC (i.e., indifference of probability of item endorsement between the base scale and the transformed scale onto the base scale).

Hanson and B*é*guin (2002) used MULTILOG and BILOG-MG (Zimowski et al., 1996) for both CC and SC methods to remove the confounding effect of difference between computer programs found in the study by Kim and Cohen (1998). In their simulation study, five factors were incorporated: group difference, calibration method, estimation program, sample size, and length of common items. Based on the evaluation criteria: MSE based on the true score and MSE based on the weighted and unweighted ICC, the authors found that CC provided more accurate results than SC with SL except that when groups were non-equivalent and the number of common items were small, but as expected, with a larger sample size and common items, and equivalent group ability, both programs performed similarly and produced smaller MSE.

Kim and Cohen (2002) used simulated graded response data to examine the performance of SC with SL linking and CC. The conditions of the study included sample size (300 for both base/target group, 1000/1000, and 1000/300), group ability (N (1, 1) for base group and N (0, 1), N (1, 1) for target group), and length of the common item set (5, 10, and 30 items) for a 30-item test. With all program default options, MULTILOG was used to estimate item parameters, and both sets of item parameters from SL liking and CC were put on the metric of generating item parameters. With evaluation criteria: root mean square difference (RMSD), and mean distance measure (MDM) for item parameter recovery and RMSD for ability recovery, they concluded that CC outperformed SC with linking in all conditions even at a small degree.

Supporting the findings of Kim and Cohen (2002), the study by Kim and Kolen (2007) found that CC outperformed SL and HB linking methods in different ability distributions, and that CC seems to be less sensitive to scale shrinkage or expansion in non-equivalent group combinations. In contrast, Lee and Ban (2010) found that SC procedures outperformed CC. This is inconsistent with the results from some previous studies (Hanson & Béguin, 2002; Kim &

Kolen, 2006). The authors conjectured that the lack of common items and examinees between forms might be a source of potential bias for CC.

Keller and Keller (2011) investigated the accuracy of examinee classification and the long-term sustainability of IRT scaling methods between linear transformation methods, and FPC across six administrations of a test. Under the 3PL UIRT model, 5000 examinees were simulated on the conditions of the changes of ability distributions: baseline case N (0, 1), mean-shift case with the mean increased by 0.15 between each administration from the baseline case, and skew-shift case with the skewness increased by -0.15, resulting the medians become 0, 0.2, 0.46, 0.64, 0.93, and 1.08. PARSCALE (Muraki, 1992) was used with Posterior option, which allows the prior distributions to be updated after both the E and M stages of the EM cycles. With the evaluation criteria: RMSE, bias of the latent ability estimates and classification accuracy with threshold 2% proposed by Keller, Wells, and Keller (2010), the authors found that the characteristic curve methods produced the most accurate results in the case of the mean shift case, whereas FPC performed best in the skew shift case. The difference of classification accuracy was less than the threshold 2 % which can be essentially ignored for all methods.

Kang and Petersen (2012) conducted a comparison study of three item calibration methods: SC with SL linear scale transformation, CC, and FPC. The methods were compared using summations based on actual testing program data. The response data were simulated for a test of 50 items with common items (10, 20 and 40) under the 3PLUIRT model. The group ability distributions were fixed for the base group to N (0, 1) and manipulated such that the target group had N(0, 1), N(0.25, 1.1^2), or N(0.5, 1.2^2) with the sample size of 500 and 2,000. BILOG-MG was used for SC, and BILOG-MG without the prior update and PARSCALE with the prior update were used for FPC. The accuracy or performance of the four UIRT linking procedures

were evaluated on the recovery of the underlying ability distributions, ICC criterion by Hanson and Beguin (2002), and TCC criterion. The authors concluded that three calibration methods produced similarly accurate results, while FPC without the prior update performed poorly.

A concise summary can be found in Kolen and Brennan (2014) that when the data fit the UIRT models and assumptions are met, CC performs better than SC with linking because it uses all available information from the data. However, SC is more robust to violations of the IRT assumptions than CC due to the nature of the data collection design. When SC is implemented properly, FPC can be an efficient alternative to SC. Sample size, length of a common item set, and group equivalency are common factors to be considered in choosing a calibration procedure.

ANCHOR/COMMON ITEMS

Along with the increasing popularity of the CINEG design in linking and equating, the property of a common-item set has become of great interest to test developers and researchers. Under the CINEG equating design, a set of common items in both forms is used to adjust for group difference or to minimize equating error resulting from differences between two forms in group ability (Cook & Petersen, 1987). The anchor item set has been assumed to be a parallel miniature or "mini" version of the operational forms being equated with respect to both content and statistical characteristics (e.g., Angoff, 1968; Kolen & Brennan, 2014). Klein and Jarjoura (1985) investigated the effect on linear equating procedures by manipulating the content of common items and concluded that the failure of the content representativeness may lead to substantial equating error.

The content representativeness of an anchor item set is justifiable from a content validity standpoint. To put it differently, the inconsistent content composition with the operational forms indicates measuring different constructs. Consequently, the validity of score interpretation may

be questionable. Such a content requirement becomes eminently important in criterion-referenced assessments, of which purpose is to classify examinees on an established criterion.

The statistical assumption of anchor items is not as straightforward as for the content requirement. The conventional wisdom (e.g., Kolen & Brennan, 2014; and Livingston, 2014) states that the statistical property of the anchor items should reflect the full range of the operational forms to be equated such that the distribution of item difficulty of the anchor items is equivalent to that of the overall test. In test construction, such a statistical constraint can be alternatively relaxed with the "miditest" consisting of moderate difficulty items proposed by Sinharay and Holland (2006a, 2006b, and 2007). Benefits of such flexibility of the miditest in test construction were reassured by Cho, Wall, Lee, and Harris (2010), Fitzpatrick and Skorupski (2016), Liu, Sinharay, Holland, Curley, and Feigenbaum (2011a), Liu, Sinharay, Holland, Feigenbaum, and Curley (2011b), Yi (2009), and Sinharay (2018).

Unlike the two anchor item choices that are based on the information of item difficulty distribution of the overall test, the test response function (TRF) was used as an additional option for the anchor test to evaluate the multidimensional linking procedures (Yao, 2011). Constructed with a matched TRF to that of the whole form, an anchor test retains the information of the full test characteristics including discrimination, difficulty, and pseudo-guessing. In her simulation study on item parameter recovery, the extended version of the unidimensional SL linking method outperformed the multidimensional version of Mean/Sigma and Mean/Mean methods.

The statistical representativeness also includes correlation, reliability, and length of an anchor item set. Budescu (1985) argued that the anchor-test correlation is the most critical determinant of the efficiency of the equating process and that the correlation functionally depends on two factors: the reliability of the total test and the relative length of an anchor set.

The higher correlation between the anchor item set and the form being equated produces better equating results (Angoff, 1971; Budescu, 1985; Petersen, Kolen, & Hoover, 1989; Sinharay & Holland, 2006a).

Equal reliability is one of five fundamental requirements to test equating (Angoff, 1971; Budescu, 1985; Dorans & Holland, 2000; Kolen & Brennan, 2014; Lord, 1980). Moses and Kim (2007) evaluated the impact of unequal reliability on test equating methods and found that unequal reliability inflates equating function variability. With fewer items, however, the anchor item set is more likely to have lower observed reliability than the overall test.

The length of an anchor item set is also non-trivial to ensure successful equating due to the significance of the influence of the number of anchor items on linking stability. Even with no absolute agreement on the length of the anchor item set due to the characteristics of the set of anchor items, the purpose of testing, and the nature of the test specification, a rule of thumb found in the literature (e.g., Angoff, 1971; and Kolen & Brennan, 2004) is that the number of common items should be at least 30 items or 20 % of a full-length test containing forty or more.

In the current study, the parallel miniature (or "mini" version) of the operational forms is selected as the anchor choice for the CINEG equating design. The mini-version anchor set is viewed as the best representative of the content and statistical pretties of the whole form and its wide acceptance in practice.

DIMENSIONALITY

DEFINITION

From a substantive standpoint, test dimensionality can be understood as a minimum number of latent factors that adequately account for the underlying examinees' performance. In this context, a dimension of interest can be interpreted as a construct which is a theoretical

representation of the underlying trait, concept, attribute, process, and/or structure that a test is designed to measure (Messick, 1989; AERA, APA, & NCME, 2014). In an operational setting, the construct is further interpreted as a realized manifestation of such conceptual entities, which are identified, measured, and quantified into a score as the object of inference for its interpretation and use (Kane, 2006).

From a statistical perspective, test dimensionality can be defined as the number of latent factors that account for the correlations among item responses in a test form to achieve local independence and monotonicity. The local independence assumption implies that item responses are unrelated conditional on the latent variable(s). That is, off-diagonal entries in the item variance-covariance matrix become close to zero. Outlining two forms to be strong and weak forms in local independence, McDonald (1981) stated that the weak form, commonly used, only requires the partial correlations of the test items zero when the latent traits are partialled out yet ignoring moments beyond the second order. To put it in the context of inference, locally dependent items (e.g., items in an item bundle) are redundant because they provide similar information; thus, they do not contribute to making an accurate inference about an examinee's ability (Wainer, 1995).

UIRT assumes that one latent ability is required for a test taker to get items correct in a test form. Such a strict assumption is relaxed both in the substantive and the statistical dimensionality. More specifically, Reckase et al. (1988) argued that the unidimensionality assumption requires that items in a test measure the same composite of abilities, rather than a single ability. Stout (1987) coined the term, "essential unidimensionality"; that is, with a major factor and one or more minor factors, a test is essentially unidimensional. In a similar vein, Hambleton (1989), and Reckase (1979) provided a reassurance that for the assumption of

unidimensionality, a dominant component or factor is required to be met to a satisfactory extent by a set of test data. These studies collectively lead to a conclusion that the test can be claimed to be unidimensional in terms of both substantive and statistical dimensionality, provided that the same compound traits are identified as a dominant factor in a test. Even if the unidimensionality assumption requires satisfactory properties from both substantive and mathematical unidimensionality, nevertheless, this does not prove the logical connection or agreement between such entities (McDonald, 1981) without confirmatory validity evidence (AERA, APA, & NCME, 2014).

In an operational test setting, many factors may cause multidimensionality, such as content specification, item type, and administration condition. In addition, two or more cognitive traits due to different backgrounds and experiences may influence an examinee's responses to an item. For example, language skills may be required to get a math item correct. A test with mixed-format items such as multiple-choice and constructed-response items can lead to two distinct dimensions (e.g., Manhard, 1996; Sykes et al., 2002). Due to many factors involved, items rarely measure a single trait (Reckase, 2009).

Such multiple factors can be either international or unintentional. The adverse influence created by unintentional factors on local independence needs to be identified and controlled to avoid overestimates of information and underestimates of the standard error of the ability estimates (Sireci et al., 1991; Wainer, 1995; Wainer & Wang, 2000; Yen, 1993). For example, Hoskens and Boeck (1997) provided a conceptual framework of modeling item associations beyond those explained by a target latent variable, while Glas and Suarwz Falcon (2003) proposed statistical tests for the 3PL UIRT model. Chen and Thissen (1997) detailed two models that can induce local dependence: surface local dependence (SLD) and underlying local

dependence (ULD). The ULD model implies that an unmodeled underlying latent variable causes the local dependent set of items. Unlike the ULD model, the SLD model does not assume the latent variable as an underlying cause, but item similarity either in content or location.

Further, Camilli, Wang, and Fesq (1995) argued that statistical dimensionality is necessary, but not sufficient without defining functional dimensionality to provide a complete conceptualization of dimensionality; functional dimensionality depends on the testing situation and the use of test scores; in contrast, statistical dimensionality is a requirement for item local independence. Their theoretical grounds lie in Hattie (1985) and Messick (1989). Hattie defined dimensionality as a joint property of the item set and a particular sample of examinees from its underlying population. Messick stated that from a content viewpoint, conceptualizing content validity should be in the judgment of experts about domain relevance and representativeness, but not in the test.

Lastly, Henning (1992) argued that psychometric unidimensionality should be distinguished from psychological unidimensionality. Even with the commonality of measuring some primary dimension or trait, psychometric dimensionality and psychological dimensionality need not agree. For instance, in the test where a kind of psychological unidimensionality is present, certain fluctuations in the distribution of item difficulty or ability patterns can lead to psychometric multidimensionality or vice versa.

It is concluded that in the item response theory, test dimensionality means the psychometric dimensionality which is mainly determined by the characteristics of content, item, and population; and that its unidimensional assumption can be relaxed in the statistical and substantive perspective, provided that it offers a meaningful interpretation of factor structure and meets the statistical assumptions (e.g., local independence and monotonicity).

ASSESSMENT

The common purpose of dimensionality analysis is to identify the simple structure for meaningful interpretation. Ackerman et al. (2003) suggested that substantive judgment in consideration of test specification, content analysis, and psychological analysis should guide dimensionality assessment. In the context of dimensionality assessment, the conventional item factor methods are exploratory in nature, utilizing tetrachoric or polychoric correlations, because Pearson product-moment correlations are not applicable in the categorical item factor analysis (see Mislevy, 1986 for details). However, this approach is not perfect without limitations: nonlinear relationship of item performance and the underlying latent ability (Hattie, 1984); lack of mathematical requirement (e.g., positive definite) of the correlation matrix, which leads to Heywood cases; and no standard rule for deciding the number of interpretable factors (Mislevy, 1986).

Procedures may be divided into parametric and non-parametric. Parametric methods require a functional form that is specified for the dependence of items on the dimensions. Mplus (Muthen & Muthen, 1998-2017), TestsFACT (Bock et al., 1999), NOHARM (Fraser & McDonld, 2003), and Parallel analysis (Horn, 1965) leverage exploratory item factor analysis on an inter-item correlation matrix to extract meaningful factors. In contrast, non-parametric methods aim to identify dimensionally homogeneous clusters of items. For example, DETECT (Stout et al., 1999; Stout et al. 2001) is designed to achieve approximate simple structure by maximizing the difference between within-cluster and between-cluster conditional covariances on examinees' scores. Two computer programs, CCPROX and HCA developed by Roussos (1992) employ the hierarchical cluster technique to create distinct item clusters on the proximity matrix. (see Svetina & Levy, 2014 for details).

Hattie (1985) conducted an extensive methodology review on assessing the unidimensionality of tests and items. The methodology of quantifying the extent of unidimensionality with indices was grouped into five different approaches based on the answer patterns, reliability, principal components, factor analysis, and latent trait models. A concise summary of the underlying assumptions and criticisms of each approach is as follows:

- The answer pattern approach is based on the idea that a perfectly unidimensional test is a function of the amount by which a set of item responses deviates from the ideal scale pattern (Guttman, 1944), but the ideal or perfect scale is not realistic, and no method exists that enables to distinguish a test of just one trait from a test composed of an equally weighted composite of abilities.
- The reliability approach is based on Cronbach's claim (Cronbach, 1951) that alpha estimates the proportion of the test variance due to all common factors among the items provided that the inter-item correlation matrix becomes of unit rank. Novick and Lewis (1967) showed that there is no systematic relationship between the rank of a set of variables and alpha (or internal consistency measure) which is not a monotonic function of unidimensionality.
- The principal component approach is based on the notion that the maximum variance, expressed as the percentage of the total variance, is explained by the first component (i.e., factor) of eigen decomposition of a tetrachoric correlation matrix. The maximum variance to be considered unidimensional is arbitrary; 20% (Reckase, 1979) or 40% (Carmines & Zeller, 1979). In addition, there is no proven rule for choosing the number of components. For instance, Kaiser rule (1970) is one of many.

- The factor analysis approach is based on the normality assumption of latent variables, which is strong, and nonlinearity with binary data, which may cause a spurious factor. In a second-order factor structure with one high order factor and several specific factors, variance in specific factors, each of which is unidimensional, may not be clear; that is, the item intercorrelation matrix will be of unit rank, but items are not measuring the same thing (Lumsden, 1957). In addition, conceptualizing the second-order factor is an independent task.
- The latent variable approach is based on the notion that responses to items can be accounted for by latent traits, the characteristics of the examinees, which are monotone nondecreasing functions (Rosenbaum, 1984). The fundamental assumption is local independence, which should not be taken as unidimensionality. Unidimensionality is strictly defined as the existence of one latent trait underlying the set of items. Model fit tests (e.g., Chi-squared test) are the typical measures of unidimensionality. It is supported by studies (e.g., Von den Wollenberg, 1982a, 1982b) that the Rasch model is robust to the violation of unidimensionality.

He concluded that an index must be viewed as a critical part of the evidence to determine the degree to which a test is unidimensional, but that even with an index, judgment must be used when interpreting it. It is also worth noting that in his previous study (Hattie, 1984), he contended that a unidimensional test is not necessarily reliable, internally consistent, or homogeneous, but may rather be factorially complex in terms of the linear common-factor model.

Almost two decades later, Tate (2003) conducted a comprehensive study on dimensionality based on simulated data and real data. The real data was collected from the

reading test with 8 reading passages and 62 items, administered on two testing days. The results of various dimensionality analysis methods confirmed the two-factor solution with testing day effect, different from the author's initial assumption of being essentially unidimensional or multidimensional with passage dependencies. Simulated data were generated on UIRT and MIRT models factoring in presence of guessing, extreme values of slope and difficulty parameters, local item dependency, and different factor structures.

The results showed that all methods performed well for the unidimensional and multidimensional cases without guessing. In general, extreme difficulty and discrimination parameters became problematic in parametric methods. For the simulated test with weak multidimensionality, none of the methods was able to detect a single locally dependent item pair. When factor complexity increased away from the simple structure, the performance of dimensionality recovery was poor across most methods.

The author concluded that the parametric methods (e.g., the factor analysis or MIRT) would correctly recover the underlying true structure when the assumed model is correct; when the model parameters are not extreme; and when multidimensionality is considerable. In case of the violation of such strong requirements for the parametric methods, nonparametric methods (e.g., HCA/CCPROX, DIMTEST, and DETECT) should be considered.

More recently, Svetina and Levy (2013) provided a framework for the dimensionality assessment. The authors analyzed the 1996 NEAT Science Assessment data, which has 3 content areas and 16 total items, consisting of 8 multiple-choice and 8 constructed-response items. The data were analyzed with four major approaches: confirmatory and exploratory with parametric and nonparametric methods, respectively. The results showed that conclusive evidence was not found to support the theoretical three-dimensional model based on three content areas. They

concluded that the assessment procedure should be decided in consideration of data in terms of missingness, scoring method, presence of a lower asymptote, and distributional assumption.

Dorans and Lawrence (1999) suggested that the unit of dimensionality analysis should be defined based on the purpose of the analysis. The authors proposed the relative dimensionality principle, which states with stress on the unit of analysis:

> …[T]he dimensions extracted from data depend on the number of each type of measure entered into the analysis, the metric of the analysis, the methods used for analysis, and the unit of analysis (p, 7).

The authors stated that the main purpose of the item-level (micro) analysis is to assess unidimensionality assumption (e.g., DIF), while that of the test score-level (macro) analysis is to assess what different tests measure and how they relate to each other (e.g., equating). They analyzed SAT verbal data at the two different levels as the unit of analysis: item-level and test score-level with the scores of item parcels (Dorans & Lawrence, 1987), small collections of non-overlapping items thought to measure the same underlying dimension or dimensions. Item parceling is a way of linearization to circumvent nonlinearity and item difficulty differences (see McDonald & Ahlawat, 1974). The results showed that the dimensions from the two different levels did not agree with each other in terms of the number, structure, and interpretation of factors. The item-level analysis found that a speed factor was consistently identified in the two- and the three-factor solutions across the content domains, while the test score-level analysis found the three-factor solution and the four-factor solution with a general second-order factor based on combinations of content domains.

Lastly, it is worth mentioning that due to the interaction between a sample of items and a sample of examinees, the dimensionality of a data matrix is the lesser of the number of

dimensions between that the items are sensitive to and that examinees vary on (Reckase, 2009). To put it differently, as the outcome of the interactions between examinees and items, item response data will become multidimensional when test items are designed to measure multiple abilities or when examinees' mastery level varies on multiple skills. Even with items measuring multiple skills, however, a test produces unidimensional response data if examinees' proficiency varies in only one of their skills. Reversely, response data will be unidimensional on the condition that items measure only one of the skills on which examinees vary.

In short, dimensionality cannot be completely invariant either across populations or test forms, but rather it is determined by the characteristics of a test and a sample of the population taking the test, and the degree of the interaction between the two under a given set of testing conditions.

STRUCTURE OF MULTIDIMENSIONALITY

Adams, Wilson, and Wang (1997) recognized MIRT models into two categories based on the number of constructs that an item measures: between-item MIRT models and within-item MIRT models. In between-item models, subsets of items are mutually exclusive and measure different latent variables. That is, a test of the between-item models consists of multiple subscales that measure distinct latent dimensions, and items of each subset are loaded on a specific latent dimension. This type of test structure is also known as "simple structure." In a more realistic setting, the simple structure can be recognized as an approximately simple structure.

In contrast, within-item models are known as "complex structures" because each item is designed to measure multiple latent dimensions. Within-item models are appropriate for modeling interactions between multiple latent abilities and task requirements.

From a substantive ground, the simple structure can be recognized as a confirmatory approach because of the prior knowledge of item loading structure, while the complex structure is more related to the exploratory approach because it does not impose any restriction on the item loading structure. The bifactor model (Gibbons & Hedeker, 1992) can be viewed as an instance of a combined structure. That is, the model is a complex structure because the probability of correct responses can be modeled as a function of a combination of general and specific dimensions for a given item. In addition, the bifactor model can also be considered a confirmatory approach because its item loading structure should be prespecified. When multiple specific factors are correlated, a second-order factor can logically be introduced to account for the correlation.

This second-order factor structure model can be recognized as a restricted bifactor model when a multivariate normal distribution is assumed for the latent factors (Rijmen, 2009). See Li et al. (2006) and Rijmen (2009) for further details on the equivalent relationship between high-order model, bifactor model, and testlet model (Bradlow, Wainer, & Wang, 1999).

UNIDIMENSIONAL APPROXIMATION

For the unidimensional approximation of multidimensional latent structures, three statistical approaches are recognized: direction of best measurement (Zhang & Stout, 1999a and 1999b), reference composite (Wang, 1985), and projective IRT (IP, 2010; Ip & Chen, 2012). Zhang and Stout proved that the unidimensional latent composite indicates the direction that the test measures best, which is essentially unidimensional with a major factor and one or more minor factors (Stout, 1987). Wang showed that when a unidimensional IRT model is fitted to multidimensional response data, unidimensional item parameters are unidimensional

projections of test items with respect to the test composite. That is, the unidimensional latent ability is a weighted linear combination of multiple abilities.

Brossman (2010) and Reckase (2009) summarized the first two methods succinctly. It can be agreeably posited by the two researchers that those methods are similar in the concept of the linear composite of abilities in the multidimensional latent space but are mathematically different. That is, the reference composite is the orientation of the unidimensional line in the multidimensional latent space given by the first eigenvector with the largest eigenvalues from the eigen decomposition of $\boldsymbol{A'A}$ where $\boldsymbol{A}$ is the item slope matrix. In contrast, the direction of best measurement is the direction corresponding to the average of multidimensional information function evaluated in all directions. (see Reckase, 2009; Zhang & Stout, 1999a and 1999b for details)

Unlike the previous methods, the projective IRT method (PIRT) reduces the multidimensional latent space onto a specific target dimension or a "purified" dimension presumed to be the dimension of interest. Thus, PIRT aims to remove the contamination caused by the multiple minor nuisance dimensions (see Ip et al.,2019 for details; and Kim & Cho, 2020 for application).

Brossman and Lee (2013) conducted, with real data sets under equivalent group equating design, a comparison study of 5 different equating procedures with equipercentile equating as a benchmark: UIRT OSE and TSE, full MIRT OSE, and unidimensional approximation of MIRT OSE and TSE. Their findings suggested that psychometric frameworks (i.e., choice of UIRT or MIRT) is an influential factor on equating results. That is, UIRT TSE and OSE are close to each other, while MIRT OSE and two UIRT approximation procedures have a similar equating pattern and result.

Recently, MIRT has gained much attention from researchers and practitioners for its various operational applications in calibration, scoring, and computer adaptive testing. Nevertheless, the applications of MIRT are very limited in the field of linking and equating due to the indeterminacy of MIRT scale, as explicated in the previous section. Consequently, it is no surprise that UIRT linking and equating procedures are a de facto standard in operation. In the following section, selected studies on the impact of multidimensionality on UIRT linking and equating are reviewed.

4. RELEVANT STUDIES

Roussos and Stout (1996) provided the conceptual ground for multidimensionality-based differential item functioning (DIF). They argued that DIF manifests itself through differences in the marginalized IRF, written as follows:

$$P_G(\theta) = \int P(\theta, \eta) f_G(\eta|\theta) d\eta \tag{2.3.1}$$

where G is the group indicator; and the probability of getting an item correct as a function of θ is obtained by averaging the response function $P(\theta, \eta)$ over the distribution of η for each fixed value of θ. For reference and focal groups, the item endorsement probability $P(\theta)$ cannot be equal when the conditional probability $f(\eta|\theta)$ are not equal. The conditional probability $f(\eta|\theta)$ can be different for groups of examinees when the item is sensitive to both the primary construct θ and some secondary construct η; when the conditional distribution $p(\eta|\theta)$ is different to two subgroups; and when the interaction of the two exists (Shealy & Stout, 1993, as cited in Roussos & Stout, 1996).

The expected difference in the means of η conditional on θ for the two groups can be expressed as such:

$$E_R(\eta|\theta) - E_F(\eta|\theta) = \left(\mu_{\eta_R} - \mu_{\eta_F}\right) + \theta\left(\rho_R \frac{\sigma_{\eta_R}}{\sigma_{\theta_R}} - \rho_F \frac{\sigma_{\eta_F}}{\sigma_{\theta_F}}\right) + \left(\mu_R \rho_R \frac{\sigma_{\eta_R}}{\sigma_{\theta_R}} - \mu_F \rho_F \frac{\sigma_{\eta_F}}{\sigma_{\theta_F}}\right), \quad (2.3.2)$$

And equation 2.3.2 can be simplified when both groups have the same standard deviations ($\sigma_{\theta_R} = \sigma_{\theta_F}$ and $\sigma_{\eta_R} = \sigma_{\eta_F}$) and correlation ($\rho_R = \rho_F$) and where $\sigma_\eta = \sigma_\theta$, as such:

$$E_R(\eta|\theta) - E_F(\eta|\theta) = \left(\mu_{\eta_R} - \mu_{\eta_F}\right) + \rho\left(\mu_{\theta_R} - \mu_{\theta_F}\right), \quad (2.3.3)$$

where E is the expectation operator, and ρ denotes correlation between η and θ. The expected difference in the means of η conditional on θ for the two groups can occur when $\mu_{\eta_R} \neq \mu_{\eta_F}$, but less likely to occur when both $\mu_{\eta_R} - \mu_{\eta_F}$ and $\mu_{\theta_R} - \mu_{\theta_F}$ have the same sign even though $\mu_{\eta_R} \neq \mu_{\eta_F}$. Even in the case where $\mu_{\eta_R} \approx \mu_{\eta_F}$, DIF can occur when $\rho \neq 0$ and $\mu_{\theta_R} \neq \mu_{\theta_F}$. In the equation 2.3.3, three factors should be considered: the difference of the secondary constructs, the difference of the construct of interest between groups, and the correlation between the two constructs between the groups.

They laid out the theoretical foundation of DIF in the context of multidimensionality, which, by the author of this study, can be viewed as an instance of manifested differential dimensionality. The more details are found in the study by Ackerman (1992).

Spence (1996) examined the effect of multidimensionality on unidimensional equating under equivalent and non-equivalent groups. The generating model was 2PL 2D compensatory and non-compensatory MIRT with alpha angles from 0 to 64 as a violation of the unidimensionality. Responses of 1,000 simulees generated from a bivariate standard normal distribution were analyzed. The scale indeterminacy issue of the two forms was handled through CC, and SC with Mean/Sigma method, and SL method. Evaluation criteria were based on the comparison between unidimensional item parameters approximated with Wang's equations

(1985) and those obtained by UIRT calibration, i.e., the comparison between the linear combinations of the true multidimensional abilities and the unidimensional estimates of the parameters. The results showed that with randomly equivalent groups, there was little difference attributable to the unidimensional equating procedures. In contrast, the large mean differences were displayed in the concurrent calibration of nonequivalent groups.

Bolt (1999) investigated the impacts of multidimensionality on the equity of the first two moments of the conditional equated score distributions for two forms: first-order equity and second-order equity. The study was based on both simulated data and the Law School Admissions Test (LSAT) with traditional equating methods (i.e., linear and equipercentile equating) and the IRT TSE. The dimensionality of the LSAT data was checked with an exploratory two-dimensional NOHARM analysis to verify a two-dimensional MIRT model. On the real data, linking was not involved under the assumption that two groups are equivalent. For the simulation study, the generating model was 2PL 2D MIRT model with different correlations between constructs. The results showed that for the high correlations (e.g., 0.7, or larger), the IRT TSE method performed slightly better than conventional linear and equipercentile equating.

In his study, the first and second-order equity criteria proposed by Thomasson (1993) are worthy of being recognized. First, the conditional bias $(i.e., d_1(\boldsymbol{\theta}))$ of the equating is the difference between the conditional means of scores X and the conditional means of equated scores $x(Y)$, expressed as:

$$d_1(\boldsymbol{\theta}) = E_X[X|\boldsymbol{\theta}] - E_Y[x(Y)|\boldsymbol{\theta}], \tag{2.3.4}$$

where E is the expectation operator and θ is an ability vector. $d_1(\theta)$ is an indicator of how well the equating function has matched expected scores for examinees having an ability. Often, it is necessary to compute single-valued marginal measures that provide a condensed

summary of information across the range of abilities given some density, although such measures lose detailed information about particular ability ranges of interest. By integrating $d_1(\theta)$ over, a weighted average difference of the first conditional moments (wad_1) is obtained, and to eliminate the cancellation of bias, a weighted average absolute difference of the first conditional moments ($waad_1$). Both equations for the case of 2D MIRT can be expressed as:

$$wad_1 = \int_{\theta_1} \int_{\theta_2} d_1(\boldsymbol{\theta}) f(\boldsymbol{\theta}) d\theta, \tag{2.3.5}$$

and

$$waad_1 = \int_{\theta_1} \int_{\theta_2} |d_1(\boldsymbol{\theta})| f(\boldsymbol{\theta}) d\theta, \tag{2.3.6}$$

where $\int$ is the integral and $f(\boldsymbol{\theta})$ is the bivariate density function of $\boldsymbol{\theta}$. The $d_1(\boldsymbol{\theta})$ and wad_1 are indicators of the first conditional moments at the examinee and the test level, respectively. The $waad_1$ indicates the expected magnitude of the conditional bias of equating at the test level.

The second-order equity, equal conditional variance, is not meaningful unless the first-order equity holds. For the discrepancies of the first and second moment together between the conditional score distributions of the two tests being equated, Thomasson (1993) computed the total conditional variance ($tcv(\boldsymbol{\theta})$), which was expressed in a different notation in Bolt (1999) as such:

$$tcv(\boldsymbol{\theta}) = E_Y[x(Y) - E_X(X)|\boldsymbol{\theta}]^2$$

$$= d_1^2(\boldsymbol{\theta}) + Var_Y[x(Y)|\boldsymbol{\theta}]. \tag{2.3.7}$$

Total conditional variance indicates the precision with which the equating transformation predicts an examinee's expected score on test X given the score on test Y. To incorporate $Var_x(X|\boldsymbol{\theta})$, the author constructed an index $d_{1,2}(\boldsymbol{\theta})$ as a combined first – and second – order equity criterion as such:

$$d_{1,2}(\boldsymbol{\theta}) = E_Y[x(Y) - E_X(X)|\boldsymbol{\theta}]^2 - E_X[X - E_X(X)|\boldsymbol{\theta}]^2$$

$$= tcv(\boldsymbol{\theta}) - Var_X(X|\boldsymbol{\theta}). \quad (2.3.8)$$

This index represents the accuracy of the equating transformation in terms of how well Y predicts expected score on X as compared to an actual administration of X. In other words, consistent with the definition of equity given earlier, the equating function is evaluated by comparing tests X and Y with respect to their relative capacities to predict expected performance on test X. As a result, this index perhaps serves as a better extension from conditional bias to one of conditional variance. Thomasson called the combined measures (e.g., tcv) a "global" index, while calling the single measures (i.e., $d_1(\boldsymbol{\theta}), wad_1, and\ waad_1$) a "local" index of equating performance.

Béguin, Hanson, and Glas (2000) compared the effect of multidimensionality on unidimensional IRT equating based on SC and CC. In the simulation design, the item parameter estimates of the 2PL 2D MIRT model from the real data were used to simulate data, and conditions were constructed varying in mean proficiency level of new forms and covariance of both forms. Parameter estimates were obtained using BILOG-MG and MCMC with the Gibbs-sampler. For SC, the SL linking method was utilized to resolve the scale issue. IRT OSE was used for both unidimensional and multidimensional models. They found that in non-equivalent group conditions with an increase in the covariance and variance of the second proficiency

dimension, the error for unidimensional equating methods was substantial compared to that of multidimensional equating.

Béguin and Hanson (2001) conducted a simulation experiment as a derivative of their previous study (Béguin and Hanson, 2000) to examine the effects of multiplicative or non-compensatory multidimensionality on unidimensional IRT equating based on SC and CC. They found that the result was consistent with the previous study.

Béguin (2002) investigated the robustness of equating to violations of the representativeness of the set of common items. Various conditions were manipulated, on the length of the subset, the correlation between the constructs, and the difference between the proficiency of the populations. They concluded that in general the unidimensional equating procedure was found to be robust to violation of the assumption of representativeness of the common item set. However, cases with large not-represented common items and a low correlation (less than .7) between the dimensions showed a clear increase in the error of the estimated score distribution.

The simulation study by Lin and Dorans (2010) was summarized well in Lin, Dorans, and Weeks (2016); that is, when the two tests are nonparallel in content structure but the groups are equivalent, both the anchor type and the extent of multidimensionality did not show evident impact on the equating results from most of the traditional linear equating (e.g., the Levine and the Tucker method) and chained equipercentile methods, with an exception that IRT true-score method was sensitive to both anchor types and multidimensionality.

As an extension of the previous study to the nonequivalent groups with anchor test design (NEAT), Lin, Dorans, and Weeks (2016) investigated the impact of content representativeness and length of anchor test on linking when the two tests are multidimensional and nonparallel in

content structure. With the 1PL 2D MIRT model as a generating model, the ability level of subgroups, the correlation between two content areas, the length of anchor test, and the proportional mix of content specifications were factored into the simulation design. The results showed that under the single group equating as criteria, in all cases, the Levine method performs better than the Tucker or chained methods, except for the 10-item anchor because the Levine method tends to be the least sensitive to the length and representativeness of the anchor item set. Also, the results suggested that equating the tests with different content structures should be avoided due to additional bias introduced by an inadequate anchor test.

One distinct stream of study on the multidimensionality to test score equating is about the multidimensionality caused by different item formats in a test, which typically consists of multiple-choice (MC) items and constructed-response (CR) or free-response (FR) items. Kim, Lee, and Kolen (2020) proposed a theoretical and conceptual framework for true-score equating using a simple-structure MIRT model. Under the multidimensional IRT framework, unlike the observed score equating, the true score equating is limited because of the one-to-many relation between a composite score and many corresponding combinations of multiple latent values on the test characteristic surface. The authors addressed the limitation well by taking advantage of the unique property of the simple structure and then introducing weights to compute composite scores. Even with the satisfactory results of their study, however, the proposed method is not completely free from limitations such as a limited number of factors, factor structure, and long process time for its operational use.

From the literature review, the performance of linking and equating under unidimensional item response models can be affected by the multidimensionality of the latent structure, the content and statistical representativeness of the common-item set, and the choice of the common-

item set between the mini and midi test. In addition, the theoretical foundation of differential dimensionality (Sawatdirakpong, 1993) can be borrowed from the conceptualization of DIF in the context of multidimensionality. Furthermore, the interpretation of the results was not clear with the term, multidimensionality.

However, the impact of multidimensionality in the UIRT linking and equating under CINEG design has not been fully examined. In other words, Spencer (1996) created two forms with different levels of multidimensionality by carefully designing multidimensional items that measures the second dimension at different levels but fixed the abilities of the two examinee groups to the standard bivariate normal distribution. Bolt (1999) factored in the multidimensionality caused by the correlation of the latent variables and corresponding sets of discrimination parameters under the equivalent group design. That is, two forms have the same multidimensional item structures at the different degrees of multidimensionality. In contrast, Beguin and Glas (2000) examined the impact of multidimensionality under the condition of different latent distributions without considering the influence of the different item compositions in the multidimensional latent space.

As stated by Reckase (2009), and Roussos and Stout (1996), however, the dimensionality of the test form is determined by the interaction of items and examinee groups who took the form. It is necessary to take both players into consideration. Thus, the goal of the current study is to investigate the impact of multidimensionality caused by items and examinee groups by designing two simulation experiments, which lays out in the following chapter.

CHAPTER III: METHODS

This chapter describes the design of a simulation experiment to investigate factors of interest to address the research questions in the first chapter. The benefit of a simulation experiment is that with known truth about the examinee ability and item properties, study conditions can be manipulated and evaluated, which is less likely to be feasible with real data. The simulation experiment consists of three sections: data generation and calibration, linking and equating procedures, and evaluation criteria.

SIMULATION DESIGN

DATA GENERATION AND CALIBRATION

Response data for two forms X (new form) and Y (base form) will be simulated on combinations of factors: mean, variance, and covariance/correlation of latent abilities, and test dimensional structures under the compensatory 2-dimensional extension of the two-parameter logistic model (Reckase, 1997) with a guessing parameter (i.e., 3PL 2D MIRT) which is given by

$$P\left(X_{ij}=1\middle|\boldsymbol{\theta}_j,\boldsymbol{a_i},d_i,g_i\right)=g_i+(1-g_i)*\left(\frac{exp[D(\boldsymbol{a_i^T\theta_j}+d_i)]}{1+exp[\,D(\boldsymbol{a_i^T\theta_j}+d_i)]}\right), \tag{3.1}$$

where $\boldsymbol{\theta_j}$ is a 1×2 vector of person latent traits; that is,

$$\boldsymbol{\theta_j}=\left(\theta_{j1},\theta_{j2}\right)\sim MVN(\boldsymbol{\mu},\boldsymbol{\Sigma}),$$

where

$$\boldsymbol{\Sigma}=\begin{pmatrix}\sigma_1^2 & \sigma_{12}\\ \sigma_{12} & \sigma_2^2\end{pmatrix},$$

is the variance and covariance matrix with σ_m^2 $(m = 1, 2)$ denoting the variance of dimension m and σ_{12} denoting the covariance between dimensions 1 and 2. The parameter a_i is a 1×2 vector of item discrimination parameters for the ith item; d_i is the intercept for the ith item; and D is the scaling constant that is set to 1.7 for this study. To solve the rotational indeterminacy of MIRT models, m*(m-1)/2 constraints can be imposed on the item discrimination parameters for an m-dimensional model (McDonald, 1997). For example, in flexMIRT (Cai, 2017), the loading of the second discrimination parameter of the first item is fixed to zero for the 3PL-2D MIRT model.

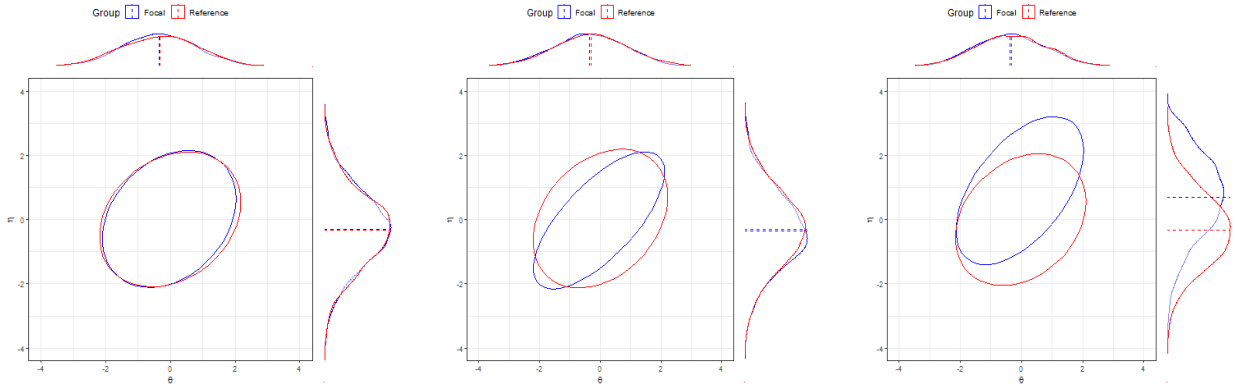
The test lengths of both Forms X and Y are set to 50 with 40 unique and 10 common items (i.e., 20% of the test form). Group differences in means, variances, and covariance of the latent variables are considered in the simulation study to investigate the sensitivity of populations to different dimensions. The covariance/correlation between the two latent variables is also studied to vary the degree of multidimensionality in populations; that is, when ρ is close to 0, the multidimensionality emerges to the population clearly, whereas when ρ is close to 1, the unidimensionality emerges to the population (e.g., Bolt, 1999). For the reference and target groups, abilities for 2,000 examinees are generated from a bivariate normal distribution with parameters (i.e., BN (mean vector (μ_θ, μ_η) covariance matrix $(\sigma_\theta^2, cov_{\theta,\eta}, \sigma_\eta^2)$) provided in Table 1. Three representative group differences by mean and/or covariance shift are visualized in Figure 9. The ability distributions for the two groups are simulated based on the modified version of the study of Beguin and Glas (2000) in which responses were generated based on the information obtained from empirical data.

Table 1. Overview of the Latent Ability Distributions

Group	$\boldsymbol{\mu}$	$\boldsymbol{\Sigma}$	Sample size

Reference	(0, 0)	$\begin{pmatrix} 1 & \\ 0.3 & 1 \end{pmatrix}$	
Target	(0, 0), (0.1), (1, 0), (1, 1)	$\begin{pmatrix} 1 & \\ 0.3 & 1 \end{pmatrix}, \begin{pmatrix} 1 & \\ 0.7 & 1 \end{pmatrix}, \begin{pmatrix} 1 & \\ 0.32 & 1.14 \end{pmatrix}, \begin{pmatrix} 1 & \\ 0.9 & 1.65 \end{pmatrix}$	2,000

Figure 9. Visual Illustration of Latent Ability (θ and η) Distributions with Mean-shift, and Var/Covariance-shift. Reference Group in Red and Target Group in Blue.



Ref and Foc:
BN ((0,0), (1, 0.3,1)*)

Foc:
BN ((0,0) (1, 0.7,1))

Foc:
BN ((0,1) (1, 0.7,1.49))

For MC-I, all 50 items in the base form measure both traits, and the angle of the reference composite (RC) to the first dimension is 45 degrees as shown in Figure 10. (1). In contrast, the RC for the first new form has the same angle as the base form but the first item cluster (item 1-25) measures dominantly the first trait θ, whereas the second item cluster (item 26-50) measures mainly the second trait η, with the angle of 30 degrees between the two clusters as shown in

* For the sake of brevity, the covariance matrix is symmetric and presents three entries: the first and third entries represent the variances of the first and second ability distributions, respectively, while the second entry represents the covariance. The remaining notation follows the same pattern.

Figure 10. (2). The RC for the second new form has an angle of 51 degrees to the first dimension, which is shifted from the RC of the base form by 6 degrees to the second dimension, but the angle between the clusters remains the same, 30 degrees as shown in Figure 10. (3).

To be more specific, for the base form, both slope parameters are sampled from a uniform distribution between 0.57 and 1.14 (denoted as $Unif(0.57, 1.14)$) such that the alpha of the items are all 45 degrees. For the first new form with the same RC as the base form, the first slope parameters of the first 25 items and the second slope parameters for the last 25 sampled from a $Unif\ (0.57, 1.14)$. To create two item clusters (c_1 and c_2) symmetric against the RC, the second slope parameters (a_2) of c_1 and the first slope parameters (a_1) of c_2 are obtained from (a_1 of c_1) $*$ 1.73 and (a_2 of c_2) $*$ 1.73. That is, the resulting angles for $c_1$and c_2 are 30 and 60 degrees, respectively, from the first dimension. For the second new form with the different RC from the base form, a_1 of c_1 are sampled from a $Unif(0.285, 0.57)$ and a_2 of c_1 and a_1 of c_2 are obtained from (a_1 of c_1) $*$ 2.75. Finally, a_1 of c_2 are obtained from (a_1 of c_2) $*$ 0.84. The resulting angles for $c_1$and c_2 are 40 and 70 degrees, respectively, from the first dimension. The intercept parameters are generated from uniform distribution between -1.5 and 1.5 for the base form Y, while for the new form X, the intercept values are generated between -.95 and 1.5 to create form difference in difficulty (e.g., Bolt, 1999). For all form. the guessing parameters are generated from a beta distribution with shape parameters 5 and 17 (i.e., the default values of BILOG-MG). The item parameter generation scheme is summarized in Table 2.

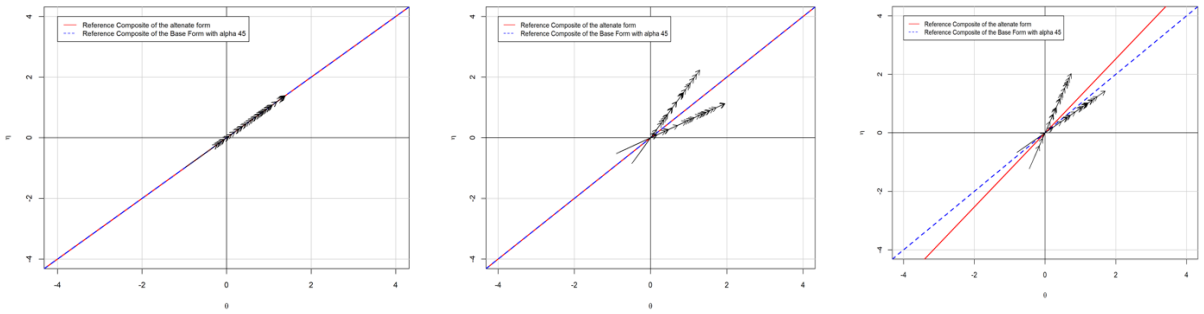
The intention of this design is to investigate following three conditions of multidimensionality. The first condition as a criterion demonstrates that two forms are parallel both multidimensionally and unidimensionally. In contrast, the second condition is designed to demonstrate the case when two forms are parallel unidimensionally but not multidimensionally.

To be specific, the base form and the new form have the same direction of RC, but the new form of this case has two item clusters, each of which measures dominantly the first and second traits, respectively. The last condition demonstrates the case when the base form and the new form are neither parallel multidimensionally nor unidimensionally.

Table 2. Overview of Item Parameter Generation Scheme for MC-I

Form		Item parameters					
	Item cluster	a_1	a_2	d	g	RC_angle	Angle_btw Clusters
Base	1-50:	0.57:1.14	0.57:1.14	$Unif(-1.5, 1.5)$	$Beta(5, 17)$	45	0
Same_RC	1-25	0.57:1.14	0.98:1.95	$Unif(-0.95, 1.5)$		45	30
	26-50	0.98:1.95	0.57:1.14				
Diff_RC	1-25:	0.26:0.57	0.78:1.57			51	30
	26-50:	0.78:1.57	0.66:1.31				

Figure 10. Schematic Illustration of Two Test Structures in Two-Dimensional Latent Space (θ and η) of MC-I



(1) Base form (2) Same reference composite (3) Different reference composite

For MC-II, all 50 items in the base form measure the first trait as the construct of interest with the angle between an item and the axis of the first dimension (i.e., alpha denoted as α) within 20 degrees, which is referred to as the validity sector in the literature (Ackerman, 1992) , as shown in Figure 11.(1). However, the new form has the first 25 items that measure exclusively the first dimension, but the other 25 items that additionally measure the second dimension in an increasing pattern of the angle from alpha 0 to 60 degrees such that the RC (29.4 degree) is out

of the validity sector, as shown in Figure 11.(2). The new form is more difficult than the base form to accommodate the condition from the previous studies (e.g., Bolt, 1999).

The first 25 item slope parameters of the base form are generated from a uniform distribution ranging from 0.94 to 1.0, and the corresponding second slope parameters are generated with $a_1 * \frac{\sqrt{1-cos(x)^2}}{cos(x)}$ (Ruecht & Miller, 1984), where x is cosign values ranging from 0.94 to 1 such that the alphas of the first set of items are within the validity sector (i.e., 20 degrees from the first axis). The second set of 25 items measure the primary dimension only, resulting in an alpha value of 0. For the new form, the slope parameters are generated in the same manner except for the cosign values ranging from 0.5 to 1 such that the alpha angles spread from 0 to 60 degrees, resulting in the RC of the new form to be located out of the validity sector. The intercept parameters are generated from uniform distribution between -1.5 and 1.5 for the base form Y, while for the new form X , the intercept values are generated between -.95 and 1.5 to create form difference in difficulty (Bolt, 1999). The lower asymptotes are generated in the same manner as in MC-I.

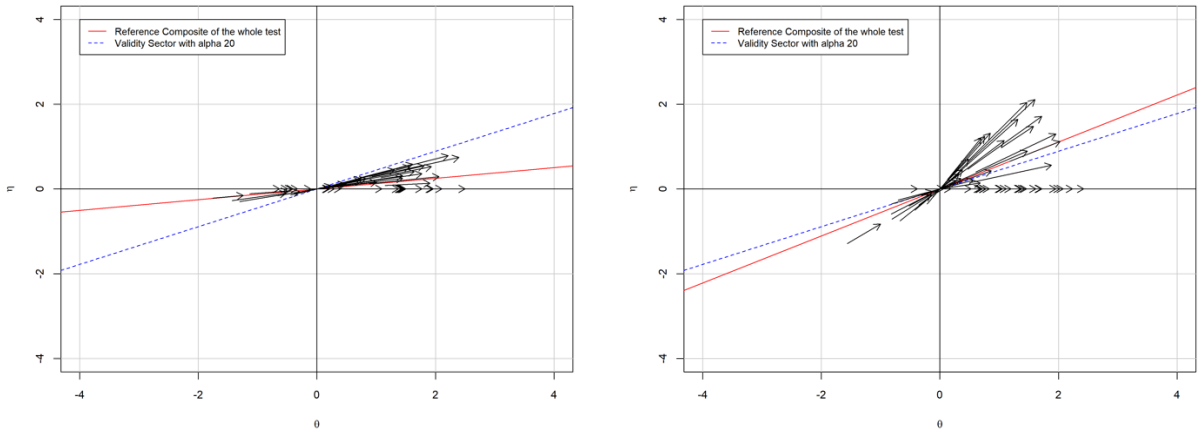
For MC-II, the underlying assumption is that the first dimension is the construct of interest, and the second dimension is viewed as a nuisance dimension, and that all items are crafted to measure primarily the first dimension, even with some items measuring the secondary dimension only at the trivial level. The validity sector proposed by Ackerman (1992) is set to the angle at 20 degrees from the axes of the first dimension of the base form. The purpose of the different test structure is to examine the impact of possible multidimensionality by varying the cosine angles of items away from the validity sector. The item parameter generation schedule is summarized in Table 3.

With given group and item information, the probability of item endorsement for each person is computed with equation 3.1 and converted to dichotomous responses by comparing it with randomly generated values between 0 and 1. If the probability of getting the item correct is equal or greater than the random value, then its response gets one otherwise zero. The response data generation is implemented in R (R Core Team, 2022).

Table 3. Overview of Item Parameter Generation Scheme for MC-II

Form	Item parameters							
	Item cluster	a_1	a_2	alpha (α)	d	g	RC_ angle	Validity sector
Base	1-25	0.5-1.5	0-0.46	0-20	$Unif$ (-1.5, 1.5)	Beta (5, 17)	6.95	20
	26-50	1	0	0				
New form	1-25	0.5-1.5	0-2.37	0-60	$Unif$ (-1.5, 0.95)		29.4	
	26-50	1	0	0	$Unif$ (-1.5, 1.5)			

Figure 11. Schematic Illustration of Two Test Structures in Two-Dimensional Latent Space (θ and η) of MC-II

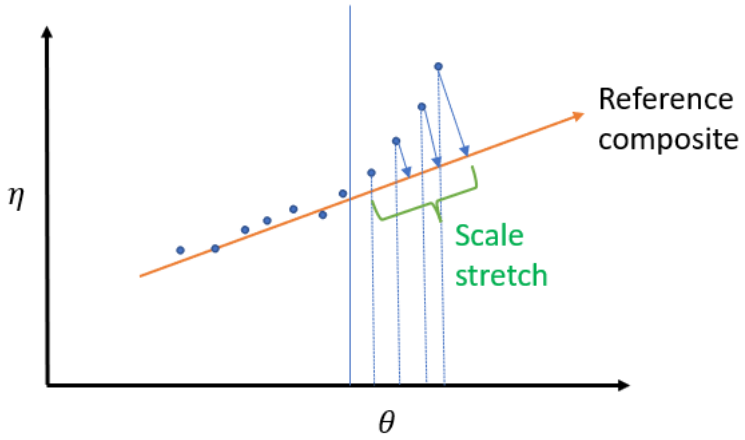


(1) Base form with items within the validity sector

(2) New form with some items out of the validity sector

It is important to note that the true model is PIRT in MC-II. The primary interest of MC-II is to evaluate the robustness of UIRT against PIRT when the primary dimension is contaminated with a nuisance factor. Specifically, a unidimensional test score is likely to overestimate the primary ability when a secondary ability is more required for difficult items (i.e., nonproportional abilities requirement; Ip et al., 2019). That is, unlike MC-I where the true model is the generating model, 3PL-2D MIRT, the generating model is used as an instrument in MC-II to introduce the influence of a nuisance factor to the estimation of the primary factor in UIRT. Test form structures were visualized in the vector plot Figure 11. Figure 12 visually illustrates the case that the UIRT scale stretches in the high ability when an item heavily loaded on the secondary dimension, the case of the new form.

Figure 12. Visualization of the Scale Stretch on RC



Note: A dot represents an observed sum score. (Modified from Ip et al., 2019, p. 150).

LINKING AND EQUATING PROCEDURES

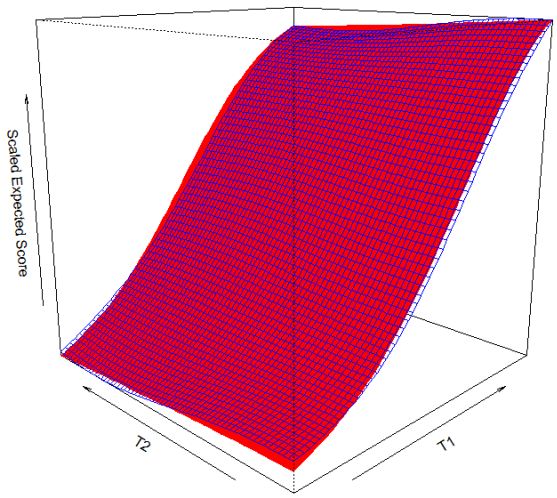
To establish a common scale, CC, FPC, and SC with SL linking procedure are performed to investigate different approaches to scale establishment. For CC, the common scale is determined with the response information from both forms, whereas for FPC and SC, the common scale is the same as the scale of the base form. For MC-I, the extended version of the univariate SL linking procedure (Oshima et al., 2000) is used to resolve the scale indeterminacy in the two-dimensional latent space. Plink R package (Weeks, 2022) is used for IRT scale linking for UIRT and MIRT.

flexMIRT (Cai, 2017) is used to calibrate response data. As a base model, 3PL 2D MIRT is fitted to the response data. For MC-II, as the items in the second set are designed to measure only the first trait, the loadings of the items are fixed to the first dimension. In addition, the prior distribution for slope parameters is set to a lognormal distribution with a mean of 0 and a variance of 0.5, and a beta distribution with shape parameters 5 and 7 is used for the prior of the

lower asymptote. The same response data are also fitted using the unidimensional three-parameter logistic (3PL) IRT model with the same prior configurations. For both unidimensional and multidimensional IRT models, calibration is performed first to estimate item parameters and then with the estimated parameters, scoring is conducted with *Score = SSC* option in the <*Options*> section to obtain the marginal probability distribution of the number correct scores for OSE. *EmpHist = Yes* option in the <*Groups*> section is used to handle the distribution of latent variables that are away from a standard normal distribution (Kim, 2019).

To adjust population difference, a total of 10 items are carefully selected as an anchor set such that the statistical property of the anchor set is close to that of the whole test by minimizing the absolute difference between the TCS of the anchor set and the TCS of the whole test (Yao, 2011), presented in Figure 13. 10 items are 20 percent of the 50-item test, which is practically sufficient (e.g., Angoff, 1971; Kolen & Brennan, 2004).

Figure 13. An illustrative Example of the Comparison between the Scaled TCS (in red) of the Base Form Y and the Scaled TCS (in blue) of the Common Item Set.



Note: The mean of absolute difference is 0.034. T1 and T2 indicate the first and second traits, respectively.

UIRT TSE and OSE, and MIRT OSE are performed with Plink R package (Weeks, 2022) (Note: weights for MIRT OSE is separately computed by the author). For OSE, the weighting scheme for the synthetic population is 0 for reference group and 1 for the target group to create a direct comparison of how the new group performed on the new form to how the test takers in the new group would have performed had they taken the base form (Brennan & Kolen, 1987; Kolen & Brennan, 2014). The quadrature values and weights of the marginal probability distribution of the number correct scores to be used for OSE are extracted from the flexMIRT output files.

For MC-I, only IRT OSE equating is performed due to the infeasibility of TSE in MIRT, the one-to-many relationship between one true score and many combinations of latent values. For MC-II, both IRT TSE and OSE are performed. For IRT TSE, the projective IRT (PIRT) or locally dependent unidimensional IRT model (Ip, 2010; Ip & Chen, 2012) is used to obtain the equivalent projected UIRT item parameters, put on the base scale with HB linking procedure after separate calibration. To be specific, PIRT is to obtain transformed item parameters of the primary dimension by projecting items in the two-dimensional latent space onto the primary unidimensional space. Because of the projection, the item response model is locally dependent. The item response function of the locally dependent unidimensional IRT model is given by

$$P\left(X_{ij} = 1 \middle| \theta_j, a_i^*, d_i^*, g_i^*\right) = g_i^* + (1 - g_i^*) * \frac{exp\left(a_i^* \theta_j + d_i^*\right)}{1 + exp\left(a_i^* \theta_j + d_i^*\right)}, \quad (3.2)$$

where a_i^* and d_i^* are the projected item parameters of the PIRT and can be computed as follows:

$$a_i^* = \lambda_{logit(i)}\left(a_{i1} + \frac{a_{i2}\rho\sigma_2}{\sigma_1}\right),$$

$$d_i^* = \lambda_{logit(i)} * d_i,$$

$$g_i^* = g_i$$

where σ_1 and σ_2 are the standard deviations of θ_1 and θ_2, respectively, ρ represents the population correlation between θ_1 and θ_2, and a_{i1} and a_{i2} are discrimination parameters for the i^{th} item. The scalars $\lambda_{logit(i)}$ and k are expressed as $\lambda_{logit(i)} = \frac{1}{\sqrt{1+k^2 a_{i2}^2 (1-\rho^2)\sigma_2^2}}$, and $k = \frac{16\sqrt{3}}{15\pi} = .588$. Note that g_i^* is same as g_i. Finally, the obtained item parameters are used for evaluation of multidimensionality against the baseline with respect to classification and equating equity. It is worth noting that the generating model 2D MIRT is used as a baseline condition for evaluation since item parameters obtained with the PIRT procedure are not locally independent (see Stucky, 2011 for more details). That is, the independence of item parameters is required for LW formula (Lord & Wingersky, 1984).

EVALUATION CRITERIA

Classification Consistency and Accuracy

Classification accuracy and consistency are pivotal validity evidence of psychometric quality for high-stakes assessments, such as admission, certification, and licensure. Classification consistency (CC) is defined as the degree to which examinees are classified into the same categories over replications of the same measurement procedure, whereas classification accuracy (CA) refers to the extent to which actual classifications using observed cut scores agree with true classifications based on known true scores (Lee, 2010, p. 1). CC and CA can conceptually be viewed as reliability and validity of classification, respectively (Lee, Hanson, & Brennan, 2000). Methods for establishing the accuracy and consistency of classification decisions are well-established. Following is a brief introduction of five methods from Diao and Sireci (2018).

Rudner's (2001, 2005) method computes CA and CC based on the IRT scale assuming that θ follows a normal distribution. Relaxing the normality assumption for θ and discretizing the θ scale with equal distances, Guo's (2006) method computes CA and CC based on the likelihood with given all examinees' θ points. Lee's (2010) approach computes CA and CC with a given cut score on the observed-raw score scale based on the summed-score distribution conditioned on θ. Both the Rudner approach and the Lee approach were developed under IRT, but the main difference of the two is the scale used to place the cut score on. Hambleton and Han's method (in Bourque, et al., 2004) is based on simulated examinees' responses on the given IRT item parameters (see Deng, 2011 for details). Lathrop and Cheng method (Lathrop & Cheng, 2014) is a non-parametric version of Rudder's method and Lee's method, without imposing the normality assumption on the θ scale.

The current study adopts Lee's (2010) method to compute CA and CC because it is developed under the IRT framework and uses the number correct score as a cut score, which is commonly used in practice. In Lee's method, CC is computed with the conditional summed-score distribution obtained using the Lord and Wingersky (1984) recursion formula and the prespecified cut scores. The conditional category probability can be computed by summing the conditional summed-score probabilities for all summed scores (x) that belong to category h; that is,

$$P_\theta(h) = \sum_{x=x_{(h-1)}}^{x_h-1} Pr(X = x|\theta)\,. \tag{3.3}$$

where h = 1, 2, ..., K. The conditional classification consistency index, ϕ_θ, is defined as the probability that an examinee having θ is classified into the same category on independent administrations of two parallel forms of a test (Lee et al., 2002). With the assumption of the independence of two testing administrations, the probability of passing/failing from the first

testing occasion is expected to be the same as the probability of passing/failing for the second testing occasion. Thus, ϕ_θ can be computed based on a single form as follows:

$$\phi_\theta = \sum_{h=1}^{K} \left[\sum_{x=x_{(h-1)}}^{x_h - 1} \Pr(X = x|\theta) \right]^2 = \sum_{h=1}^{K} [P_\theta(h)]^2. \quad (3.4)$$

The conditional classification consistency index quantifies classification consistency for different levels of θ. In case of two categories (i.e., pass or fail) that is widely used in licensure and certification assessments, the conditional consistent classification index sums the squared probability of passing and the squared probability of failing at a given θ point. The possible maximum value of the conditional consistent classification probability is one, which typically occurs at the extreme θ points. The possible minimum value of the conditional consistent classification probability is usually located near cut scores because classification decisions are most uncertain near cut scores. As a single scalar value, the marginal classification consistency index,ϕ, can be obtained by integrating the conditional classification consistency index over all quadrature points with corresponding weights, as follows:

$$\phi = \int_{-\infty}^{\infty} \phi_\theta g(\theta)\, d\theta \approx \sum_\theta \phi_\theta g(\theta). \quad (3.5)$$

where θ is a discretized quadrature points (which plays as a proxy for unobserved true scores) that span across certain range (e.g., -3 to 3 with 49 points); and $g(\theta)$ is the density of θ. Dependent on the choice of the θ for integration for computing marginal results in equation 3.5, Lee method has D method with quadrature points (aka distribution approach) and P method with individual examinee's θ values (aka individual approach). D method is the choice of the evaluation criterion in this study with the main interest in group-level statistics (Lee, 2010). The conditional and marginal classification consistency indices are evaluated against the true classification consistency indices obtained using the generating item parameters.

With the conditional probabilities (observed classification), $p_\theta(h)$, known, the conditional classification accuracy index, γ_θ, can be obtained by

$$\gamma_\theta = p_\theta(\eta) = \sum_{x=\tau_{(\eta-1)}}^{\tau_\eta - 1} Pr(X = x|\theta), \; for\; \theta \in \eta. \tag{3.6}$$

where η $(= 1, 2, \ldots, K)$ is the true categorical status of an examinee. The true category η can be determined by comparing the expected summed score for θ (computed from the test characteristic function; TCC) with the true cut scores on the summed score metric, $\tau_1, \tau_2, \ldots, \tau_{K-1}$. When the true cut score in TCC is assumed to be the same as the observed cut score, the marginal classification accuracy index, γ, is given by

$$\gamma = \int_{-\infty}^{\infty} \gamma_\theta g(\theta)\, d\theta \approx \sum_\theta \gamma_\theta g(\theta). \tag{3.7}$$

Classification indices for MIRT

Consistent with the D method in UIRT, its extended version is used for the true model, 3PL 2D MIRT in MC-I. The computation is identical to that of UIRT except for the latent trait (θ) become a latent trait vector with two elements $(\boldsymbol{\theta} = (\theta_1, \theta_2))$. Given the conditional distribution of summed-score (x) and the cut scores, the conditional probability of scoring in each performance category can be computed by summing up the conditional probabilities of all total summed-score x values that belong to category h, as follows:

$$P_{\boldsymbol{\theta}}(h) = \sum_{x=x_{(h-1)}}^{x_h - 1} Pr(X = x|\boldsymbol{\theta}). \tag{3.8}$$

Then, the conditional classification consistency index ϕ_θ is defined by

$$\phi_{\boldsymbol{\theta}} = \sum_{h=1}^{K} \left[\sum_{x=x_{(h-1)}}^{x_h - 1} Pr(X = x|\boldsymbol{\theta})\right]^2 = \sum_{h=1}^{K} [p_{\boldsymbol{\theta}}(h)]^2, \tag{3.9}$$

and the marginal classification consistency index ϕ is given by

$$\phi = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \phi_{\boldsymbol{\theta}} g(\boldsymbol{\theta}) d\theta_1 d\theta_2 \approx \sum_{\theta_1} \sum_{\theta_2} \phi_{\boldsymbol{\theta}} g(\boldsymbol{\theta}). \tag{3.10}$$

With a true cut score on the summed-score metric, the conditional classification accuracy index $\gamma(\boldsymbol{\theta})$ is given by

$$\gamma(\boldsymbol{\theta}) = P_{\boldsymbol{\theta}}(\eta), \tag{3.11}$$

and the marginal classification accuracy index, γ, is expressed as

$$\phi = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \gamma(\boldsymbol{\theta}) g(\boldsymbol{\theta}) d\theta_1 d\theta_2 \approx \sum_{\boldsymbol{\theta}_1} \sum_{\boldsymbol{\theta}_2} \gamma(\boldsymbol{\theta}) g(\boldsymbol{\theta}). \tag{3.12}$$

For MC-II in which items are designed primarily to measure the first dimension, which is the construct of interest, contaminated with a nuisance dimension, PIRT is used to obtain transformed item parameters of the primary dimension by projecting items in the two-dimensional latent space onto the primary unidimensional space. With the obtained PRIT item parameters, the Lee's D method is applied to compute baseline classification indices. To compute the baseline criterion, the marginalized conditional summed-score distribution is applied after integrating out the nuisance dimension.

This study employs summed-scores to establish cut-scores because the summed-score metric is consistent across UIRT and MIRT. In testing organizations for high-stakes examinations such as licensure and certification, a cut score is typically established by criterion-reference methods between 70 and 90% of pass rate for the first-time takers (e.g., Lineberry et al., 2020). The current study uses 80% of the test's maximum raw score. This corresponds to a total score of 40 in a 50-item test.

The performance of each linking method's marginal classification indices is compared to the corresponding true classification criteria, which are obtained using the generating item parameters for MC-I and MC-II. Results are evaluated using bias, standard error (SE), and root mean squared error (RMSE) as follows:

$$Bias = \frac{1}{R}\sum_{r=1}^{R} \widehat{ci}_r - CI\,, \tag{3.13}$$

$$SE = \sqrt{\frac{1}{R}\sum_{r=1}^{R} (\widehat{CI}_r - \overline{\widehat{CI}})^2}, \tag{3.14}$$

$$RMSE = \sqrt{Bias^2 + SE^2}, \tag{3.14}$$

whereby R is the number of replications (i.e., 100); $\widehat{CI}_r$ is an estimate of the marginal classification index (consistency and accuracy) at a given replication r; CI is the true marginal classification index; and $\overline{\widehat{CI}}$ takes the average standard deviation from all replications of the marginal classification index.

EQUATING ACCURACY and EQUITY

The goal of equating is to establish the most accurate equating relationship so that the scores of examinees can be interchangeable across forms. This goal is achieved by minimizing systematic error (Bias) and random error (e.g., RMSE, or SEE). Systematic error may occur from the violation of assumptions of the same test structure and group equivalence in ability. For instance, assumptions are violated when test structures of two parallel forms are away from agreement in dimensionality (i.e., MC-I and MC-II; e.g., Bolt, 1999; Ip et al., 2019); when latent distributions of two populations are different in the multidimensional latent space (i.e., anchor item parameter drift; Roussos & Stout, 1996) or when such cases are compounded. Random error may result from sampling examinees randomly as opposed to using the entire population. Thus, the random error is expected to decrease as sample size increases, while the systematic error may remain regardless of the sample size.

All equating procedures are vulnerable to these errors (Kolen & Brennan, 2004). The impact of multidimensionality is evaluated under the following three IRT equating procedures:

(1) unidimensional IRT true score equating, (2) unidimensional IRT observed score equating, and (3) multidimensional IRT observed score equating. The new form to be equated to the base form is assumed to cause errors to the equating results due to the different dimensional structure of the new form in MC-I and MC-II and its interaction with the ability distributions. In the context of the multidimensionality, to compare the accuracy and precision of the equating relationship between the equated scores obtained from the three equating methods and those obtained from the generating item parameters for the baseline condition, three criteria are used: bias, SE, and RMSE as follows:

$$Bias(x) = R^{-1}\sum_{r=1}^{R}\hat{e}_Y^{(r)}(x) - e_Y(x), \tag{3.15}$$

$$SE(x) = \sqrt{R^{-1}\sum_{r=1}^{R}[\hat{e}_Y^{(r)}(x) - \bar{\hat{e}}_Y^{(r)}(x)]^2}, \tag{3.16}$$

$$RMSE(x) = \sqrt{Bias(x)^2 + SE(x)^2}, \tag{3.17}$$

whereby R is the number of replications ($i.e.,\ 100$).; x is a particular score point; $\hat{e}_Y^{(r)}(x)$ is the estimated base form equivalent of score x obtained from the r^{th} replication; $e_Y(x)$ is the base form equivalent of score x obtained using the generating item parameters for MC-I and transformed item parameters with PIRT for MC-II; and $\bar{\hat{e}}_Y^{(r)}(x)$ is the mean estimated base form equivalent of score x over R replications ($i.e.,\ R^{-1}\sum_{r=1}^{R}\hat{e}_Y^{(r)}(x)$).

The bias of a particular score x is a measure of accuracy - how close, on average, the estimated base form equivalent of score x is to the base form equivalent of score x from the generating item parameters, viewed as a true score equivalent. The standard error measures the precision of the equating results. The root mean squared error considers both systematic (i.e.,

bias) and random error (i.e., SE) together due to the possible trade-off between the bias and SE. These criteria can be useful for practitioners to choose an optimal equating procedure for the purpose of assessment.

The equity property of equating proposed by Lord (1980) holds only if for examinees with a given true score, the distribution of the equated scores on the new form is identical to the distribution of the score on the old form. Lord's equity property will not hold unless the two forms are identical in which case equating is unnecessary (Kolen & Brennan, 2004). However, building identical parallel forms is not feasible. For this reason, two practical properties were suggested to evaluate the quality of equating: the first-order equity property (FOE; Divgi, 1981; Morris, 1982; Yen, 1983), and the second-order equity property (SOE; Tong & Kolen, 2005). FOE implies that the expected conditional means are equal for the alternate forms after equating. With FOE being satisfied, SOE can be meaningfully assessed, which holds if the conditional standard error of measurement (CSEM) is equal for the alternate forms after equating.

The first and second order equity properties are evaluated with the methods by Tong and Kolen (2005). With the conditional observed score probability $f(X = j|\theta_i)$, the expected equated score, $\widehat{eq}_y(x)$, at a given θ_i can be obtained as

$$E\left(\widehat{eq}_y(x)\middle|\theta_i\right) = \sum_{j=0}^{n} \widehat{eq}_y(x_j) f(X = j|\theta_i). \quad (3.18)$$

Similarly, the standard error of measurement (SEM) conditional on θ_i can be calculated:

$$SEM|\theta_i = \sqrt{var[(x)|\theta_i]} = \sqrt{\sum_{j=0}^{n}\left[\widehat{eq}_y(x_j) - E\left(\widehat{eq}_y(x)\middle|\theta_i\right)\right]^2 f(X = j|\theta_i)}\ . \quad (3.19)$$

After the expected equated scores and conditional SEMs of the new form were obtained using equations 3.18 and 3.19, the equity properties could be evaluated. Index D_1 is computed to empirically assess the adequacy of preserving the first-order equity property, which is

$$D_1 = \frac{\sqrt{\sum_i q_i \left\{ E[Y|\theta_i] - E\left[\widehat{eq}_y(x)\middle|\theta_i\right] \right\}^2}}{SD_y}, \tag{3.20}$$

where $E[Y|\theta_i]$ is the base form conditional mean for a given proficiency θ_i, $E\left[\widehat{eq}_y(x)\middle|\theta_i\right]$ is the conditional mean of an equated score for a given proficiency θ_i , q_i is the quadrature weight at θ_i , and SD_Y is the standard deviation of Form Y (base form). The quadrature points θ_i and the corresponding quadrature weight q_i can either come from the examinees taking the old form or those taking the new form, depending on the definition of the synthetic population. In this study, because the synthetic group was conceptualized as the examinees taking the new form, the quadrature points which had been transformed in the scale transformation step and the corresponding quadrature weights for the examinees taking the new form were used. The smaller the D_1 value is, the better the first-order equity is preserved. The denominator SD_Y in the equation is used to standardize the index so that D_1 indices from different tests can be compared.

Second-order equity was evaluated using the index D_2, which is calculated as follows:

$$D_2 = \frac{\sqrt{\sum_i q_i \left\{ SEM_Y|\theta_i - SEM_{\widehat{eq}_y(x)}|\theta_i \right\}^2}}{SD_y}, \tag{3.21}$$

where $SEM_Y|\theta_i$ is the conditional SEM for the base form for examinees with proficiency θ_i , and $SEM_{\widehat{eq}_y(x)}|\theta_i$ is the conditional SEM for the equated new form for examinees with proficiency θ_i. Similar to the D_1 index, the quadrature points θ_i and weights q_i from the examinees taking the new form were used in this study. A large D_2 value suggests that the second-order equity property is not sufficiently preserved. The denominator SD_Y in the equation is used to standardize the index so that D_2 indices from different tests can be compared.

The final equity evaluation criterion is the index $d_{1,2}(\theta)$, which is used in Bolt (1999) to evaluate the first- and second-order equities together in equation 2.3.8 in Chapter 2. This index provides more comprehensive evaluation on equating transformation, considering both equity properties together, in terms of the prediction of the expected score on the old form with the score of the new form in comparison to the actual score of the old form. All evaluation criteria are computed by the R code written by the author.

In summary, this simulation experiment is to evaluate the robustness of UIRT under the CINEG equating design in terms of classification indices and equity property indices. Two major cases are considered with different latent structure. In the first case (MC-I), the test forms are built to measure two latent dimensions of which the latent structures vary between forms. In the other case (MC-II), the test forms have one primary dimension and one secondary dimension. Thus. items are assumed to measure only the primary latent dimension as the construct of interest.

CHAPTER IV: RESULTS

This chapter is structured into two sections that are dedicated to the presentation of findings from the MC-I and MC-II simulation experiments, respectively. To assess the influence of multidimensionality, four linking methods were employed: separate calibration utilizing the SL and HB methods (SC-HB and SC-SL), as well as two calibration approaches employing the CC and FPC methods. To compute classification indices, the item parameters of 3PL 2D MIRT were employed as the baseline, and classification indices were computed for the new forms to facilitate comparison. The computation of these indices involved utilizing Lee's D method (2010) for UIRT and its extended version for MIRT.

In the context of equating procedures, the MIRT OSE served as the baseline for establishing mean absolute equating bias (MAB) and root mean square error (RMSE). This involved utilizing the generating item parameters. Additionally, to obtain the marginal distribution of the observed scores for the baseline, weights were derived from a bivariate normal distribution characterized by a mean vector of (0, 0) and a covariance matrix of (1, 0.3, 1). For comparison, UIRT OSE and TSE were employed. For SC-HB and SC-SL, weights were derived from a normal distribution, with the linking constant B as a mean and A as standard deviation, obtained with plink R package (Weeks, 2022). For CC and FPC, the mean and variance freely estimated by flexMIRT were used for weight to obtain the marginal distribution. As described in Chapter 3, OSE utilizes a weighting scheme wherein the synthetic population is assigned a weight of zero for the reference group and a weight of one for the focal group.

In evaluating the quality of an equating method, equating equity property indices (D1 for first order equity and D2 for second order equity) were computed using equated scores. The establishment of the baseline for these equity property indices involved employing an extended

version of Tong and Kolen's (2005) methodology to MIRT. A combined index referred to as "D12" was used to indicate weighted average absolute difference of the first and second conditional moments, following equation 2.3.8 in Chapter 2. To facilitate meaningful comparisons, the index was computed in absolute value, thereby avoiding cancellation effects. The purpose of this index is to assess the accuracy of the equating transformation by evaluating how well scores on one form can predict equated scores on another form, relative to an actual administration of X Base Form (Bolt, 1999). As a reference point, the baseline was established using MIRT OSE with the generating parameters. Furthermore, a bivariate standard normal distribution was utilized to derive a marginal index. It is important to note that the evaluation encompassed the complete set of outcomes, without dividing them according to the equating method. This approach was adopted since both equating methods employed in the study, under the given conditions, did not indicate any significant differentiation.

MC-I : COMPLEX STRUCTURE MEASURING TWO CONSTRUCTS OF INTEREST

MC-I aims to demonstrate the calibration of forms designed to measure multiple constructs of interest using UIRT. The X Base and Y Base share an equivalent test structure featuring a linear reference composite (RC) at a 45-degree angle, intended to measure the two constructs equally. Conversely, the Y Same RC consists of two distinct sets of items with different loadings on each dimension, although their linear composite remains consistent with that of X Base. In contrast, the Y Different RC incorporates two split item sets with varying loadings on each dimension, accompanied by a linear composite that differs from that of X Base. The common items were carefully chosen to minimize the discrepancy with the test characteristic surfaces of the whole X Base Form. A comprehensive overview of the process involved in generating the test forms is presented in Table 4. Varying mean and covariance

matrices were considered to assess the interaction between test structures and latent distributions (see Table 1 in Chapter 2). The characteristics of the common items and unique items of the base and new forms are summarized in Table 4.

Table 4. Descriptive Statistics for Generating Item Parameters of MC-I

	Common Items for all forms		Unique Items							
			X Base		Y Base		Y Same RC		Y Diff RC	
	Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd
Slope 1	0.82	0.18	0.86	0.17	0.86	0.17	1.18	0.42	0.81	0.41
Slope 2	0.82	0.18	0.86	0.17	0.86	0.17	1.18	0.49	1.09	0.24
Intercept	0.32	0.58	0.33	0.68	0.31	0.73	0.43	0.77	0.39	0.69
Guessing	0.16	0.03	0.14	0.03	0.14	0.04	0.15	0.04	0.15	0.04
Item Angle	45	0	45	0	45	0	45	15	55	15.19
MDISC	1.17	0.25	1.21	0.24	1.21	0.24	1.73	0.34	1.41	0.33
MID	-0.31	0.51	-0.29	0.59	-0.25	0.65	-0.27	0.48	-0.28	0.56
RC Angle	45	NA	45	NA	45	NA	45	NA	51.68	NA

Notes: Item angle denotes the measurement direction of an item; MDISC represents item discrimination in the multidimensional latent space, similar to UIRT; MID indicates item difficulty, also analogous to UIRT; RC Angle (Reference Composite Angle) signifies the measurement direction of the test form.

Linking Constants. Figure 14 visually depicts the mean values of A and B, categorized by test structure, mean shift and covariance structure of the focal group, and linking method.

Based on the observation, the linking constant A displays consistent patterns in its estimates across different test structures and mean shifts. However, the constant A exhibits variation depending on the covariance structure and linking method used. Specifically, as the covariance increases, the estimates of the constant A also increase. The linking method CC yields the highest estimated values, followed by FPC, while both SC linking methods produce equivalently lower estimates. For instance, within the base mean vector (0,0) and under the covariance structure (1, 0.3, 1), the mean estimates obtained are 1.28 using the CC method, 1.16 using the FPC method, 1.00 via the SC-HB method, and 0.99 through the SC-SL method.

Conversely, when considering a different covariance structure (1, 0.9, 1.65), the CC method yields an estimate of 1.53, the FPC method produces an estimate of 1.42, while both the SC-HB and SC-SL methods yield an estimate of 1.20.

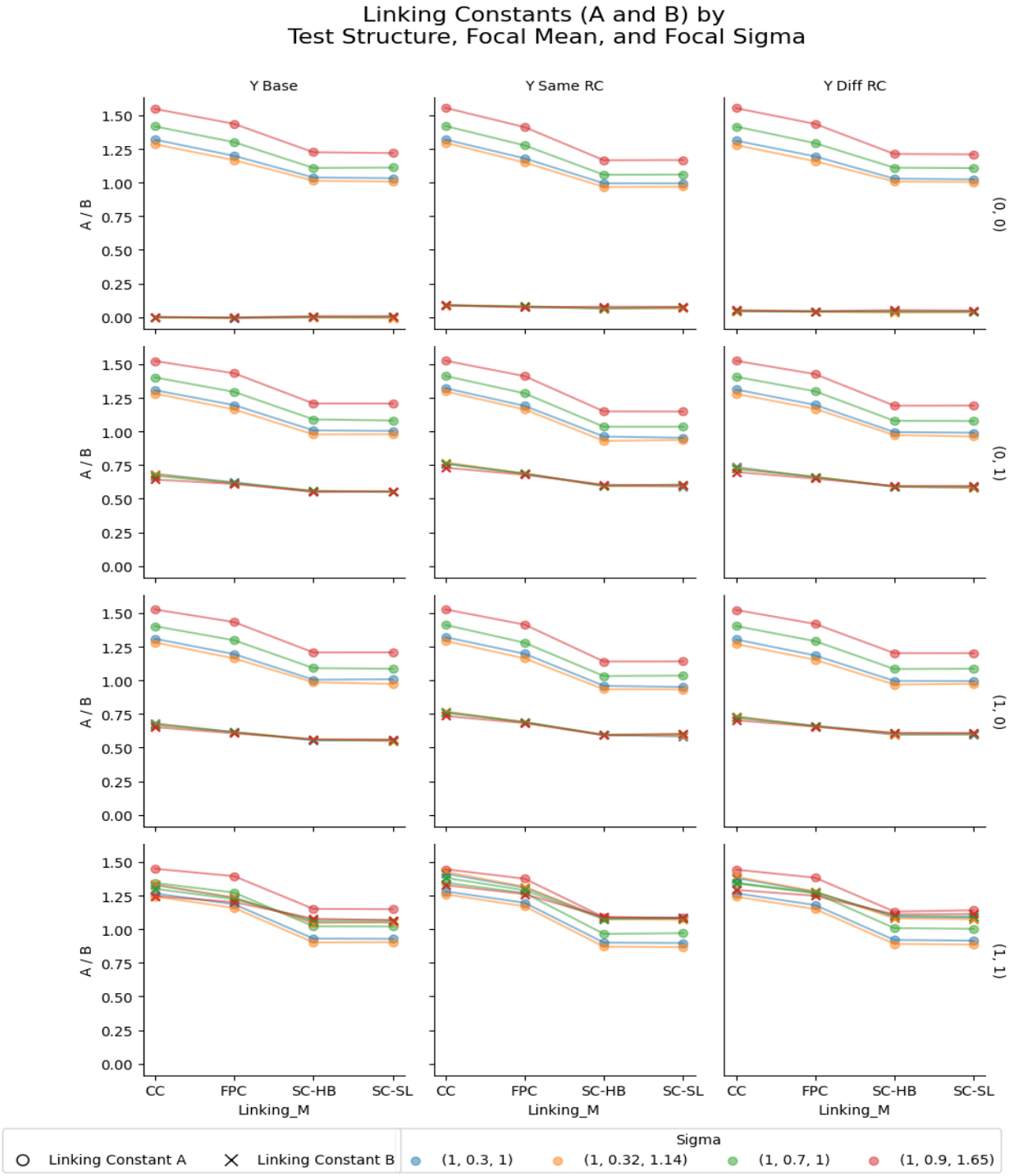
One potential hypothesis to account for the disparities in A and B estimates by the linking methods is related to the information derived from response data and the methodologies employed within each linking method. To be specific, the CC method estimates linking constants by utilizing response data from both groups, while the FPC method utilizes response data solely from the new group. On the other hand, the SC-HB and SC-SL methods only employ response data from the common items within the new group. Moreover, both SC methods utilize characteristic curves, whereas CC and FPC rely on calibration approaches.

The linking constants reported in this study, despite having slightly different interpretations, are referred to as "linking constants" for comparison purposes. The linking constants obtained from SC linking methods were based solely on common items. However, for CC and FPC, the "linking constants" were estimated using the mean and standard deviation derived from the focal group ability distribution, which included all items from the exam, not just the common items. As a result, the comparison between the linking methods is not entirely unbiased.

To investigate the factors affecting the variability of linking constants, a multivariate multiple linear regression analysis was conducted to account for potential correlations between variables A and B, summarized in Table 5. Also, this regression analysis was selected due to its ability to provide the magnitudes of factors for the purpose of comparison. The observed variability can be separated into two components: "Within" and "Between" variation. "Within" variation refers to the variation within a test structure, attributable to factors such as mean shift,

covariance structure, and linking method. On the other hand, "Between" variation pertains to the variation observed between different test structures.

Figure 14. Mean Linking Constants by Test Structure, Mean, Sigma, and Linking Methods



The linking constants A and B were treated as dependent variables, while predictors included the test structure, mean shift, covariance structure, and linking method. The reference point was set at a mean of (0, 0), covariance structure of (1, 0.3, 1), the base test structure (Y Base Form that is equivalent to X Base Form), and CC linking method. The results indicate that, for linking constant A and B, Y Diff RC with a different test structure did not exhibit statistical significance. For Y Same RC with a same reference composite but with a different test structure, despite the presence of statistical significance, the magnitudes of the effects for the linking constant A and B are relatively weak compared to the impact of covariance structure on the linking constant A and the influence of mean shift on constant B. In other words, after accounting for other factors, the variation in linking constant A and B was not influenced by the different test structures (i.e., no significant indication between-variability by test structure).

Regarding A, as the mean shift moves further away from the base mean vector, it leads to a greater reduction in estimates on a smaller scale. Additionally, the covariance structure exhibits an increasing pattern of coefficients as it deviates from the base covariance structure. On the other hand, the covariance structure does not reach statistical significance for constant B. Putting differently, the variation of the linking constant A and B estimates appears within covariance structure and mean shift.

In summary, the results from both observation and regression analyses consistently indicate that mean shift, covariance structure, and linking methods significantly influence linking constants A and B. In other words, the impact of test structures on the estimation of linking constants A and B is found to be limited, as the primary aim of scale linking is to account for population differences.

Table 5. Regression Results of Linking Constants

	Linking Constant A Estimate (SE)	Linking Constant B Estimate (SE)
(Intercept)	1.314 (0.005)***	0.087 (0.011)***
Test Structure (Y Same RC)	-0.025 (0.003)***	0.061 (0.008)***
Test Structure (Y Diff RC)	NA	0.044 (0.008)***
Focal Mean (0, 1)	-0.016 (0.004)***	0.594 (0.009)***
Focal Mean (1, 0)	-0.016 (0.004)***	0.596 (0.009)***
Focal Mean (1, 1)	-0.068 (0.004)***	1.154 (0.009)***
Focal Sigma (1, 0.32, 1.14)	0.027 (0.004)***	NA
Focal Sigma (1, 0.7, 1)	0.114 (0.004)***	NA
Focal Sigma (1,0.9,1.65)	0.234 (0.004)***	NA
Linking Method (FPC)	-0.108 (0.004)***	-0.054 (0.009)***
Linking Method (SC-HB)	-0.332 (0.004)***	-0.135 (0.009)***
Linking Method (SC-SL)	-0.333 (0.004)***	-0.136 (0.009)***

Note. Standard errors are shown in parentheses.
Significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.
NA: Not presented as it is not statistically significant.

Equating. For each equating method, the results of 192 simulation conditions (3 test structures x 4 different means x 4 different covariance structures x 4 linking methods) are summarized below. The equating outcomes, as measured by MAB and RMSE, were computed for evaluation purposes. Furthermore, the practical implications of equating results were examined using the Difference That Matters (DTM) threshold (Dorans & Feigenbaum, 1994). DTM is defined as the absolute value of 0.5, representing the point at which a score would warrant rounding up to the next integer on the observed score scale.

UIRT True Score Equating

Based on the observation depicted in Figure 15, the variation in test structures results in differences in equating performance measured by MAB and RMSE, and also the DTM threshold is visualized in red as a reference. The disparity between MAB and RMSE values is minimal, suggesting that the majority of the total error is attributable to bias.

Regarding the case when the test structures of two test forms are equivalent both unidimensionality and multidimensionally (i.e., Y Base), it was found that the two SC linking methods consistently exhibit superior performance across different mean shifts and covariance structures, while the CC method performed the poorest, with the FPC method falling between CC and the SC methods in terms of performance. The variation in performance becomes more noticeable as the mean shift (1, 1) deviates from the X Base (0, 0). For practical consideration, the CC method consistently surpasses the DTM threshold across all conditions. On the other hand, the FPC method partially surpasses the DTM method when the mean vector is set to (0,1) or (1, 0) and entirely surpasses it when the mean vector is (1, 1). Additionally, both SC linking methods fall within the threshold established by the DTM method.

As to the case when the test structures of two test forms are equivalent both unidimensionality and multidimensionally (i.e., Y Same RC), contrasting the findings of Y Base, the CC linking method exhibited the best performance, followed by the FPC and two SC linking methods, regardless of mean shift and covariance structure. Unlike Y Base, the variation in performance attributed to covariance structures becomes more pronounced even when utilizing the base mean shift. Overall, as the covariance increases between two latent variables, MAB increases, except for CC and FPC in the mean vector (1, 1), showing the opposite pattern. Regarding DTM, in contrast to the result observed in the Y Base case, two SC linking methods consistently surpass the DTM threshold across all conditions. However, the CC linking method falls below the DTM threshold when the mean vectors are set to (0, 1) and (1, 0), and the covariance is less than 0.9. Similarly, the FPC method also falls below the DTM threshold when the mean vectors are (0, 1) and (1, 0), and the covariance is equal to or less than 0.32.

In the case of the test structures of the two test forms were not equivalent either unidimensionality or multidimensionally (i.e., Y Diff RC), the results were most complex. To be specific, it was found that in general, all linking methods yielded the highest performance with the mean shifts (0,0) and (0, 1). In contrast, two SC linking methods outperformed CC and FPC with the mean shift (0, 1). Conversely, when the mean vector shifted in both abilities (1,1), CC and FPC outperformed SC-HB and SC-SL with the highest covariance (0.9). It is worth noting that as the covariance structures exhibited higher covariance values, the overall performance improved with the mean vector (0, 1), while when the mean vector was shifted to the second dimension (i.e., (0, 1)), an opposite performance pattern was observed. Y Diff RC shows mixed results regarding DTM. Specifically, when there is an unbalanced shift in mean vectors, such as (0.1) and (1,0), none of the linking methods fall below the DTM threshold. However, under the base mean vector of (0,0), all linking methods except for two SC linking methods with a covariance of 0.9 fall within DTM. Furthermore, when considering the mean vector of (1,1) and a covariance of 0.9, both the CC and FPC linking methods fall under the DTM threshold.

Table 6 summarizes the result of the regression analysis on MAB and RMSE. Test structure, mean shift, and linking method were considered as predictors and the covariance structure was dropped due to its low contribution to the model and clear interpretation of the interaction effects. The model fits the data well and approximately 94 % of the variability in the dependent variables (i.e., MAB and RMSE) can be explained the independent variables included in the model. The baseline model with Y Base with CC produces 0.57 MAB and 0.62 RSME. In other words, on average, the predicted values from the model using Y Base and CC are off by 0.57 equated score units in terms of absolute bias. Additionally, the model's predictions deviate

from the actual observed values by an average of 0.62 equated score units, as indicated by the RMSE value.

In comparison to the baseline model in Table 6, three noteworthy observations can be made. Firstly, in the case of Y Same RC with a mean shift of (1, 0), the CC method demonstrated the best performance (MAB:-0.42 and RMSE:-0.41). Secondly, when considering Y Diff RC with a mean shift of (0, 1), the CC method produced the poorest performance (MAB:0.89 and RMSE:0.86). It is important to note that the MAB is slightly larger in comparison to the RMSE due to the fact that these represents the predicted values generated by the model. Third, there are cases of interaction between factors that improve the equality quality: two-way interactions between Y Same RC and mean vectors (0,1) and (1,0), and three-way interactions among Y Diff RC, mean vector (0, 1), and two SC linking methods.

Figure 15. MAB and RMSE by Test Structure, Mean, Sigma, and Linking Methods

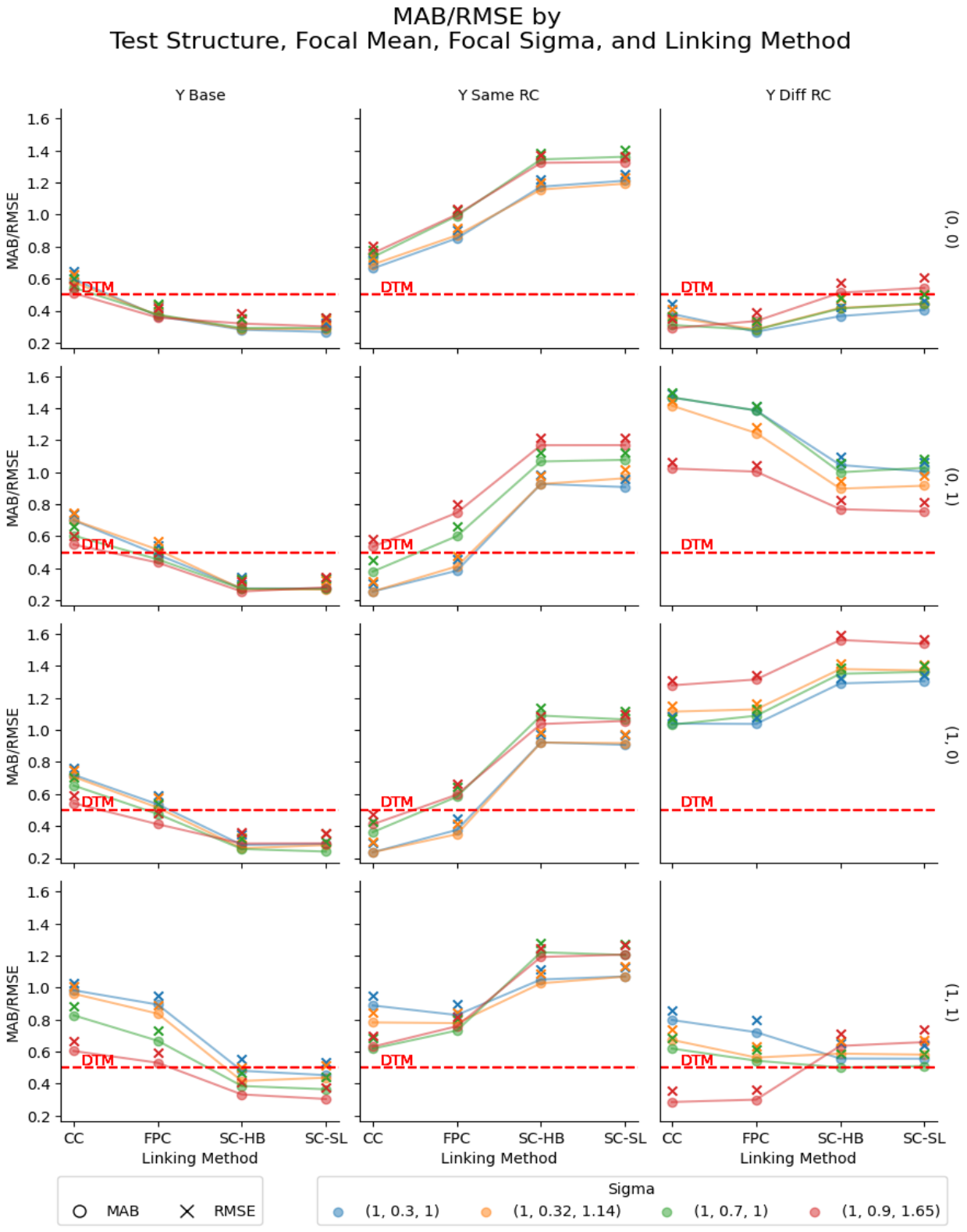


Table 6. Regression Results of UIRT TSE

	MAB	RMSE
	Estimate (SE)	Estimate (SE)
(Intercept)	0.574 (0.044) ***	0.623 (0.043) ***
Y Same RC	0.125 (0.062) *	0.125 (0.061) *
Y Diff RC	-0.236 (0.062) ***	-0.227 (0.061) ***
Mean(1, 1)	0.203 (0.062) **	0.205 (0.061) **
FPC	-0.167 (0.062) **	-0.152 (0.061) *
SC-HB	-0.274 (0.062) ***	-0.259 (0.061) ***
SC-SL	-0.273 (0.062) ***	-0.262 (0.061) ***
(Y Same RC)*(0, 1)	-0.384 (0.088) ***	-0.376 (0.086) ***
(Y Diff RC)*(0, 1)	0.891 (0.088) ***	0.862 (0.086) ***
(Y Same RC)*(1, 0)	-0.420 (0.088) ***	-0.405 (0.086) ***
(Y Diff RC)*(1, 0)	0.711 (0.088) ***	0.695 (0.086) ***
(Y Same RC)*(1, 1)	-0.194 (0.088) *	-0.182 (0.086) *
(Y Same RC)*(FPC)	0.354 (0.088) ***	0.336 (0.086) ***
(Y Same RC)*(SC-HB)	0.795 (0.088) ***	0.773 (0.086) ***
(Y Diff RC)*(SC-HB)	0.368 (0.088) ***	0.360 (0.086) ***
(Y Same RC)*(SC-SL)	0.793 (0.088) ***	0.777 (0.086) ***
(Y Diff RC)*(SC-SL)	0.370 (0.088) ***	0.363 (0.086) ***
(Y Diff RC)*(0, 1)*(SC-HB)	-0.383 (0.125) **	-0.368 (0.122) **
(Y Diff RC)*(1, 0)*(SC-HB)	0.262 (0.125) *	0.249 (0.122) *
(Y Diff RC)*(0, 1)*(SC-SL)	-0.402 (0.125) **	-0.382 (0.122) **

Note. Standard errors are shown in parentheses.
Significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.
Adjusted R-squared: 0.94 both for MAB and RMSE

UIRT Observed Score Equating

The results obtained from the OSE approach closely resemble those from the TSE approach. Thus, in order to gain further insights into the visual interpretation of Figure 16, it is advised to consult the visual examination of Figure 15 in the TSE.

Summarized in Table 6, the results of the regression analysis in OSE were also comparable to those in TSE. However, the coefficient of the intercept (MAB: 0.557 and RMSE:

0.607) was slightly smaller than that of TSE (MAB: 0.574 and RMSE: 0.623). In contrast, the standard errors (MAB: 0.052 and RMSE: 0.051) in OSE were slightly larger than those in TSE (MAB: 0.044 and RMSE: 0.043). Finally, TSE exhibited a slightly better model fit (Adjusted R-squared: 0.94) compared to OSE (Adjusted R-squared: 0.92). This indicates that TSE estimates coefficients with slightly higher accuracy than OSE.

Compared with TSE, OSE produce consistent results, but with higher coefficients and standard errors. For instance, when comparing Y Same RC to the base condition, the best-performing case with a mean vector of (1, 0) exhibits average reductions of -4.99 in MAB (with a standard error of 0.103) and -0.482 in RMSE (with a standard error of 0.101). On the other hand, the worst-performing case of Y Diff RC with a mean vector of (0, 1) shows average increases of 0.972 in MAB (with a standard error of 0.103) and 0.902 in RMSE (with a standard error of 0.101).

Figure 16. MAB and RMSE by Test Structure, Mean, Sigma, and Linking Methods

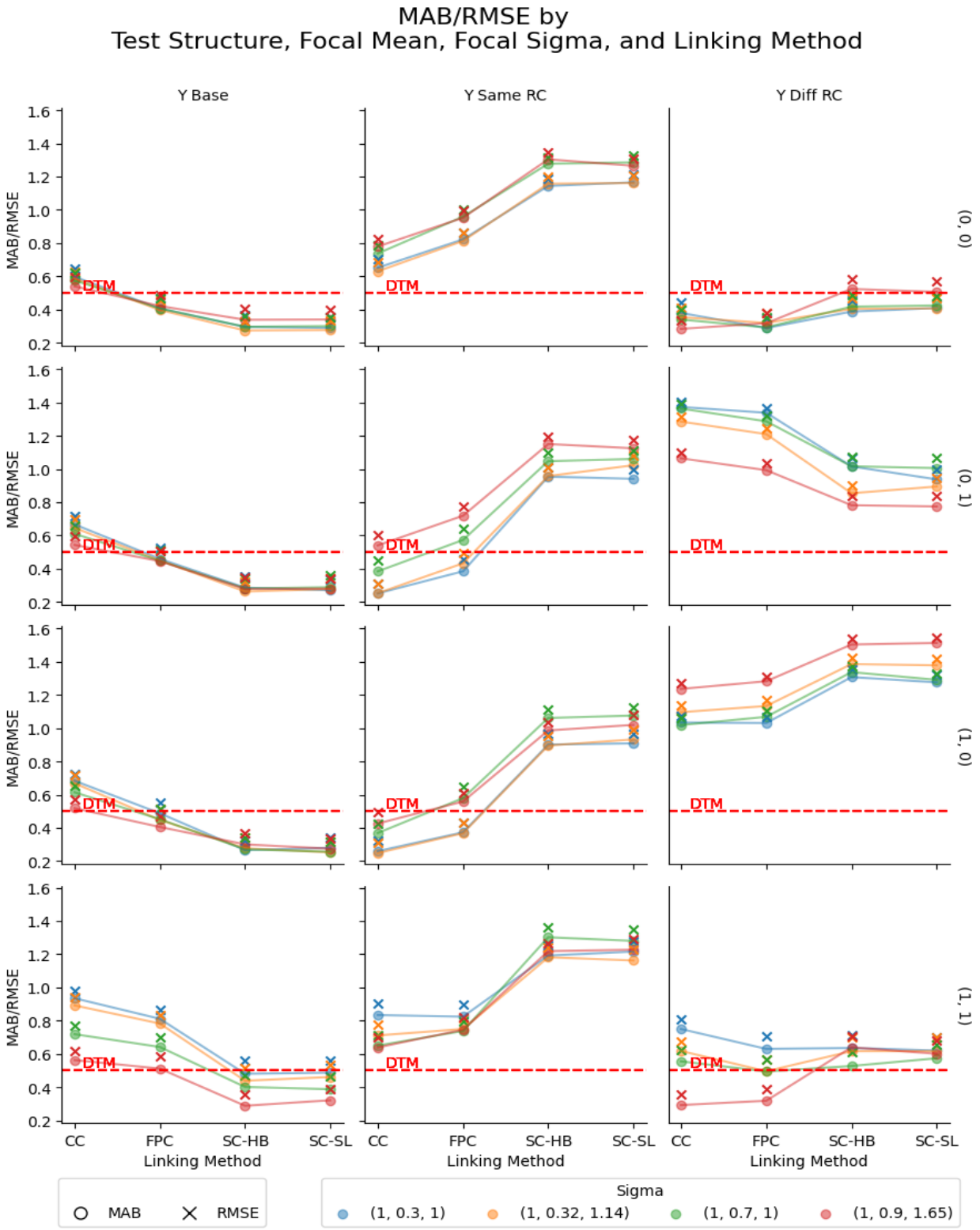


Table 7. Regression Results for UIRT OSE

	MAB Estimate (SE)	RMSE Estimate (SE)
(Intercept)	0.557 (0.052) ***	0.607 (0.051) ***
Y Same RC	0.155 (0.073) *	0.152 (0.071) *
Y Diff RC	-0.222 (0.073) **	-0.214 (0.071) **
Mean(1, 1)	0.287 (0.073) ***	0.288 (0.071) ***
FPC	-0.189 (0.073) *	-0.178 (0.071) *
SC-HB	-0.262 (0.073) ***	-0.250 (0.071) ***
SC-SL	-0.270 (0.073) ***	-0.260 (0.071) ***
(Y Same RC)*(0, 1)	-0.437 (0.103) ***	-0.424 (0.101) ***
(Y Diff RC)*(0, 1)	0.927 (0.103) ***	0.902 (0.101) ***
(Y Same RC)*(1, 0)	-0.499 (0.103) ***	-0.482 (0.101) ***
(Y Diff RC)*(1, 0)	0.683 (0.103) ***	0.665 (0.101) ***
(Y Same RC)*(1, 1)	-0.269 (0.103) *	-0.252 (0.101) *
(Y Same RC)*(FPC)	0.407 (0.103) ***	0.390 (0.101) ***
(Y Same RC)*(SC-HB)	0.801 (0.103) ***	0.783 (0.101) ***
(Y Diff RC)*(SC-HB)	0.355 (0.103) ***	0.345 (0.101) ***
(Y Same RC)*(SC-SL)	0.833 (0.103) ***	0.814 (0.101) ***
(Y Diff RC)*(SC-SL)	0.394 (0.103) ***	0.386 (0.101) ***
(Y Diff RC)*(0, 1)*(SC-HB)	-0.400 (0.146) **	-0.384 (0.143) **
(Y Diff RC)*(1, 0)*(SC-HB)	0.307 (0.146) *	0.296 (0.143) *
(Y Diff RC)*(0, 1)*(SC-SL)	-0.446 (0.146) **	-0.426 (0.143) **

Note. Standard errors are shown in parentheses.
Significance levels: *** p < 0.001, ** p < 0.01, * p < 0.05.
Adjusted R-squared : 0.92 for both MAB and RMSE

In short, the results of the present study indicate that the quality of equating is primarily influenced by the alignment of test structures between test forms. Additionally, it is important to ensure equivalence of populations among groups and carefully select appropriate linking methods. These factors should be considered when conducting equating.

Evaluation

In order to assess the influence of multidimensionality on UIRT linking and equating, classification indices were employed to compare the linking results, while equity indices were used to evaluate the equating outcomes. These comparisons were conducted with consideration for factors such as test structure, mean shift, and covariance structure. Due to the absence of any significant differences in the results pertaining to equating equity properties and their interpretation across different equating methods, the findings from the combined data are presented herein to highlight the main conclusions.

Classification

The results of the classification consistency (CC) and classification accuracy (CA) analyses are presented visually in Figures 17, considering the intended factors aforementioned. The findings indicate that, overall, the classification indices for the new forms, assessed using the UIRT approach, are higher than the baseline with the generating MIRT. Furthermore, regardless of the conditions, the CA indices consistently outperform the CC indices. In essence, CA represents the correlation between true and observed scores, whereas CC represents the correlation between observed scores. Naturally, CC will be attenuated due to the presence of measurement error in the observed scores.

Based on the findings depicted in Figure 17, overall, the pattern of classification indices demonstrates similarity between test structures, but Y Same RC exhibits the highest

classification indices (i.e., 0.95 in CC and 0.96 in CA) across all conditions. The CC linking method demonstrates higher performance in both classification indices (i.e., 0.95 in CC and 0.95 in CA), compared to other linking methods, with the exception of FPC in the case of Y Diff RC and mean vectors ((0, 1), and (1, 0)). Following CC, the FPC method and two SC methods come next, while the two SC linking methods demonstrates the lowest performance.

Regarding mean shift, it is observed that, unlike other mean vectors, when the mean vector (1, 1) deviates from the base in both latent variables, the variation in performance is the smallest. Additionally, as the covariance between latent variables increases, the indices of CC and CA increase across all conditions. This suggests that when multidimensional test structures callops into a unidimensional latent space, the classification indices increase.

A summary of the regression analysis in Table 8 with only significant coefficients at a significant level at 0.05 examined the main effects of test structure, mean shift, and linking methods, excluding the influence of covariance structure due to its minimum contribution and clear interpretation. The model was able to explain approximately 80% of the variability in the classification indices. The findings indicate that, in comparison to the baseline with a mean of (0, 0), a covariance structure of (1, 0.3, 1), and CC as the linking method, the most influential factors in classification indices were Y Same RC, the CC linking method, and a mean shift of (1, 1). After controlling for other factors, Y Same RC exhibited, on average, higher values by 0.009 in CC and 0.005 in CA compared to Y Base, which had values of 0.939 in CC and 0.956 in CA. Among factors, the mean shift (1, 1) demonstrates the highest coefficients, specifically 0.016 in CC and 0.012 in CA. Moreover, a clear trend of increasing coefficients (i.e., from 0.001 to 0.004 and 0.008) is observed as the covariance within the covariance structure increases from 0.32, to 0.7, and finally to 0.9.

Figure 17. CC and CA by Test Structure, Mean, Sigma, and Linking Methods

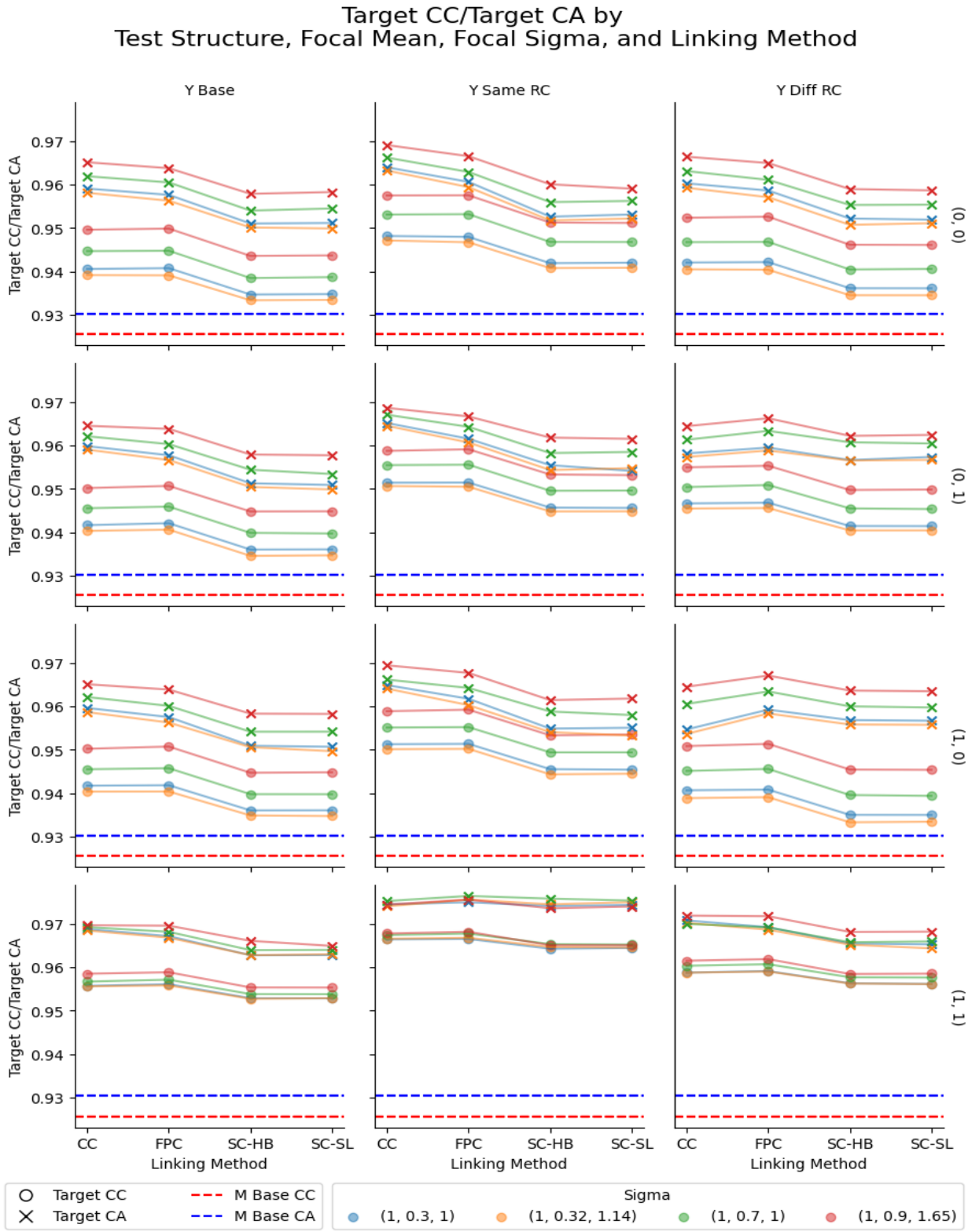


Table 8. Regression Results of CC and CA

	CC Estimate (SE)	CA Estimate (SE)
(Intercept)	0.939 (0.000)***	0.956 (0.001)***
Test Structure (Y Same RC)	0.009 (0.000)***	0.005 (0.000)***
Test Structure (Y Diff RC	0.002 (0.000***	0.002 (0.000)***
Focal Mean (0, 1)	0.003 (0.000)***	0.001 (0.000)**
Focal Mean (1, 0)	0.001 (0.000)*	0.001 (0.000)*
Focal Mean (1, 1)	0.016 (0.000)***	0.012 (0.000)***
Focal Sigma (1, 0.32, 1.14)	0.001 (0.000)*	0.003 (0.000)***
Focal Sigma (1, 0.7, 1)	0.004 (0.000)***	0.006 (0.000)***
Focal Sigma (1,0.9,1.65)	0.008 (0.000)***	0.006 (0.000)***
Linking Method (FPC)	-0.005 (0.000)***	-0.006 (0.000)***
Linking Method (SC-HB)	-0.005 (0.000)***	-0.006 (0.000)***
Linking Method (SC-SL)	N/A	-0.001 (0.000)*

Note. Standard errors are shown in parentheses.
Significance levels: *** p < 0.001, ** p < 0.01, * p < 0.05.
NA: Not presented as it is not statistically significant.
Adjusted R-squared: 0.80 both for MAB and RMSE

Equity Properties

The equating equity properties, the first-order equity (FOE) and second-order equity (SOE) indices (D1 and D2, respectively), are visually represented in Figure 18, revealing differences between the UIRT equating approach and the generating MIRT models (i.e., X Base and Y Base). The findings suggested that, in general, the FOE and SOE indices derived from UIRT equating were higher, indicating poorer bias and precision compared to the baseline MIRT case. Overall, the equivalent test structure demonstrated superior performance compared to the other two test structures. Across all conditions, two linking methods with separate calibration (i.e., SC-HB and SC-SL) demonstrated virtually identical performance.

Regarding the equivalent test structure (i.e., Y Base), among the examined linking methods, CC displays the poorest performance, followed by FPC and two SC linking methods

perform best with the lowest equity indices (i.e., 0.05 in D1 and 0.01 in D2). Furthermore, as the covariance increases, the equality indices show improvement.

In contrast, in the case of the semi-equivalent test structure (i.e., Y Same RC), the opposite pattern of performance was observed for the linking methods. Specifically, the CC linking method demonstrates superior performance across all targeted factors, achieving average values of 0.08 in D1 and 0.01 in D2. Moreover, with an increase in covariance, in general, the equality indices exhibit a decline in performance.

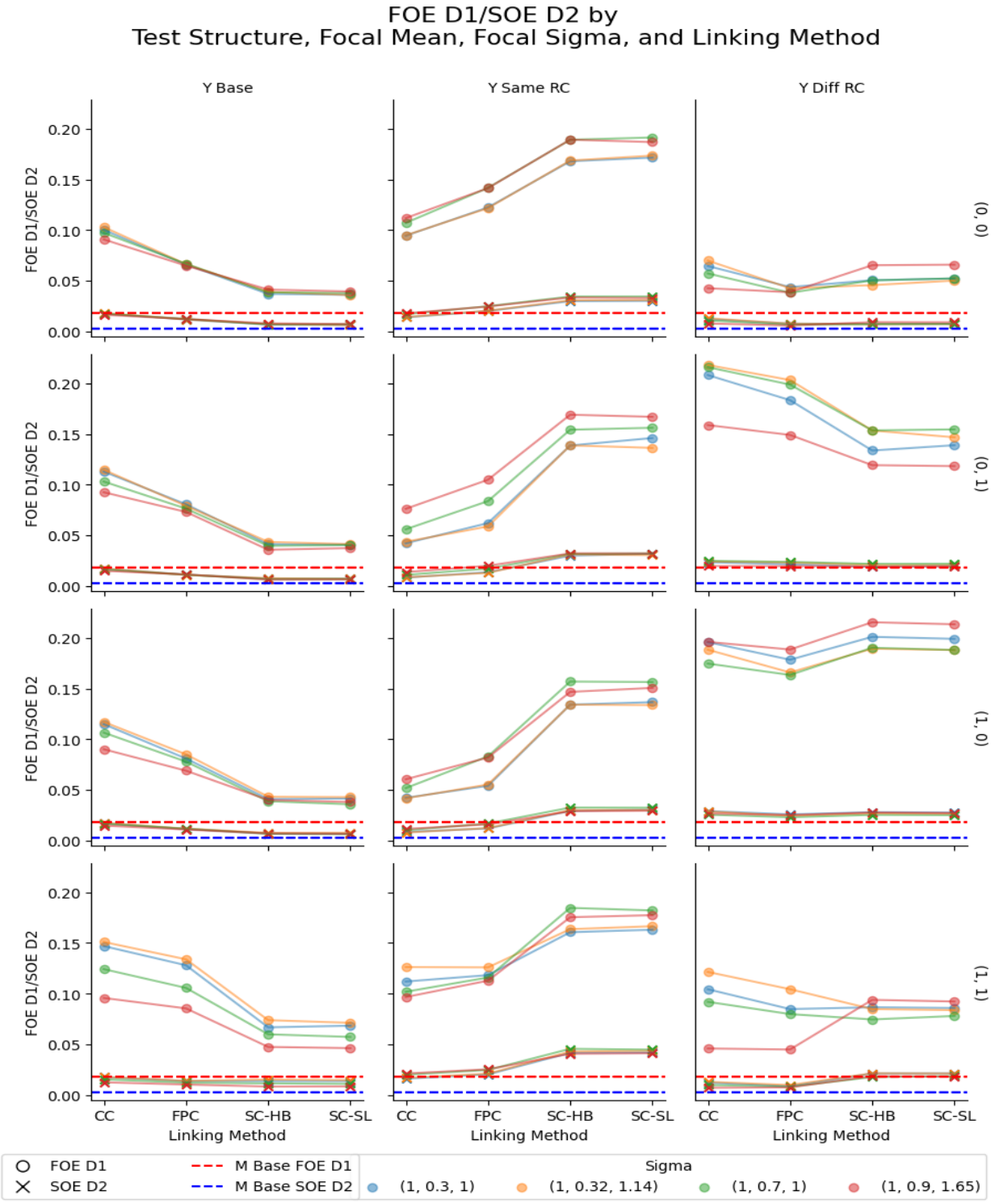
The non-equivalent test structure (i.e., Y Diff RC) yielded mixed results due to the compounded interaction effects with other factors. Among different mean vectors, the base mean vector demonstrates the lowest FOE and SOE indices. Additionally, irrespective of the covariance structure within the base mean vector (0,0), FPC consistently exhibited superior performance compared to other linking methods. However, when the mean vector was (0, 1), two SC linking methods outperformed CC and FPC. Additionally, as the covariance increased, there was an improvement in the FOE and SOE indices. Conversely, when considering the mean vector of (1, 0), the performance order was reversed, and as the covariance increased, the indices demonstrated a decline in performance. Within the mean vector (1, 1), CC and FPC performed better with the highest covariance structure.

Summarized in Table 9, a regression analysis was performed to examine the effects of test structure, mean shift, and linking method, with the covariance structure excluded due to its minimal contribution. The results of the regression analysis align with the observations made through visual examination in Figure 18.

For Y Same RC, in comparison to the baseline, the two-way interaction with the mean vector (1, 0) showed an average decrease of -0.062 in D1 and -0.005 in D2. Similarly, the two-

way interaction with the mean vector (0, 1) resulted in an average decrease of -0.056 in D1 and -0.004 in D2. In contrast, when considering the interactions involving FPC, SC-HB, and SC-SL, it is observed that the quality of equating diminishes in terms of D1 and D2 (0.062, 0.135, and 0.139 in D1; and 0.012, 0.027, and 0.027 in D2, respectively).

Figure 18. FOE and SOE by Test Structure, Mean, Sigma, and Linking Methods



For Y Diff RC , the two-way interactions involving mean vectors (0, 1), and (1, 0); and with two SC linking methods result in increased values for both D1 and D2. On average, D1 increases by 0.134, 0.121, 0.053, and 0.057, while D2 increases by 0.013, 0.018, 0.007, and 0.007, respectively. However, the three-way interaction with mean vector (0, 1), and SC-HB reduces D1 on average by -0.048 , whereas the three-way interaction with mean vector (1, 1) and SC-SL increases D2 on average by 0.006.

Table 9. Regression Results of FOE and SOE

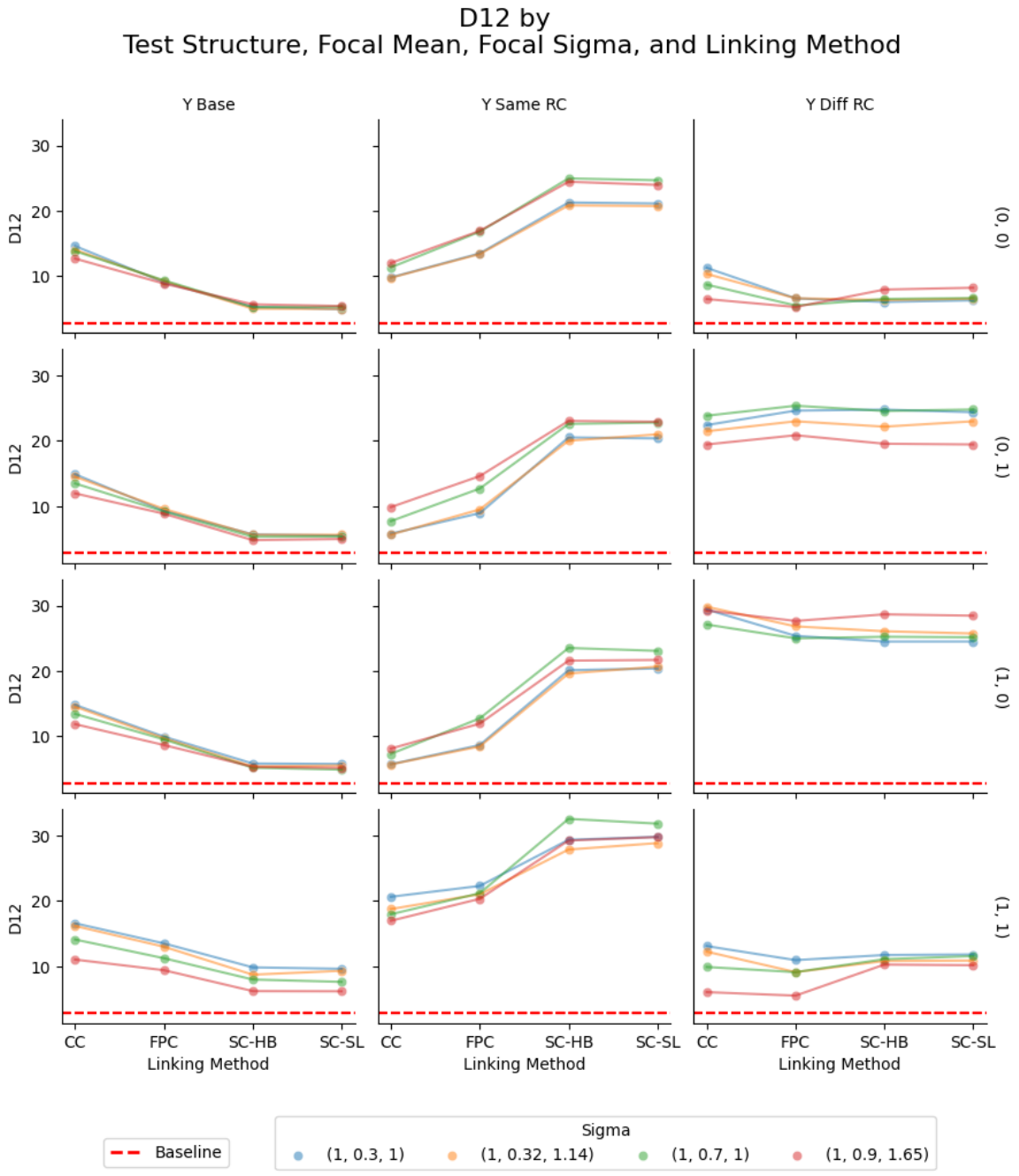
	FOE D1	SOE D2
	Estimate (SE)	Estimate (SE)
(Intercept)	0.098 (0.007) ***	0.017 (0.001) ***
Y Diff RC	-0.039 (0.010) ***	-0.006 (0.001) ***
(1, 1)	0.032 (0.010) **	-0.002 (0.001)
FPC	-0.032 (0.010) **	-0.005 (0.001) ***
SC-HB	-0.059 (0.010) ***	-0.010 (0.001) ***
SC-SL	-0.060 (0.010) ***	-0.011 (0.001) ***
(Y Diff RC) * (0, 1)	0.134 (0.014) ***	0.013 (0.002) ***
(Y Same RC) * (0, 1)	-0.056 (0.014) ***	-0.004 (0.002) *
(Y Diff RC) * (1, 0)	0.121 (0.014) ***	0.018 (0.002) ***
(Y Same RC) * (1, 0)	-0.062 (0.014) ***	-0.005 (0.002) **
(Y Same RC) * (FPC)	0.062 (0.014) ***	0.012 (0.002) ***
(Y Diff RC) * (SC-HB)	0.053 (0.014) ***	0.007 (0.002) ***
(Y Same RC) * (SC-HB)	0.135 (0.014) ***	0.027 (0.002) ***
(Y Diff RC) * (SC-SL)	0.057 (0.014) ***	0.007 (0.002) ***
(Y Same RC) * (SC-SL)	0.139 (0.014) ***	0.027 (0.002) ***
(1, 1)*(SC-HB)	NA	0.007 (0.002) ***
(1, 1)*(SC-SL)	NA	0.007 (0.002) ***
(Y Diff RC) * (0, 1) * (SC-HB)	-0.048 (0.019) *	NA
(Y Same RC) * (0, 1) * (SC-SL)	-0.052 (0.019) **	NA
(Y Diff RC) * (1, 1) * (SC-HB)	NA	0.006 (0.002) *
(Y Same RC) * (1, 1) * (SC-SL)	NA	0.006 (0.002) *

Note. Standard errors are shown in parentheses.
Significance levels: *** p < 0.001, ** p < 0.01, * p < 0.05.
NA: Not presented as it is not statistically significant.
Adjusted R-squared: 0.94 for D1 and 0.97 for D2

Figure 19 visually depicted the combined index, D12. The overall pattern of D12 closely resembled that of the FOE D1 index, presumably due to the predominant contribution of D1. Any slight discrepancies between the two can be attributed to variations in the computation

approach employed. Specifically, the weights for D1 and D2 were derived from the standard normal distribution, whereas for D12, the weights were determined using a bivariate normal distribution with a mean of (0, 0) and a sigma of (1, 0.3, 1). Notably, despite the numerous similarities, it is worth highlighting that in Y Diff RC, with mean vectors of (0, 1) and (1, 0), the performance gap attributed to the linking method was reduced.

Figure 19. D12 by Test Structure, Mean, Sigma, and Linking Methods



MC-II : ONE CONSTRUCT OF INTEREST WITH A NUISANCE FACTOR

MC- II aims to exemplify the calibration of forms specifically designed for the assessment of a single construct of interest accompanied by a nuisance factor through the application of unidimensional item response model. The base form X (i.e., X Base), comprising 50 items, primarily targets measurement of the first dimension, exhibiting a 20-degree inclination from the axis of the primary dimension. Within this paradigm, two forms were introduced, characterized by the employment of both an equivalent test structure (i.e., Y Base) and a modified test structure (i.e., Y 60) where the targeted measurement of items deviated by 60 degrees from its axis, thereby highlighting the noticeable influence of the nuisance dimension. A detailed exposition of the test form generation process is presented in Table 10. To comprehensively evaluate the interplay between test structures and latent distributions, the same variations of latent distribution were applied as in MC-I.

Table 10. Descriptive Statistics for Generating Item Parameters of MC-II

	Common Items of X Base		Unique Items					
			X Base		Y Base		Y 60	
	Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd
Slope 1	0.97	0.18	0.96	0.19	0.98	0.21	0.99	0.22
Slope 2	0.12	0.15	0.11	0.13	0.11	0.14	0.42	0.57
Intercept	-0.1	0.81	-0.19	0.82	-0.18	0.80	-0.31	0.71
Pseudo-guessing	0.14	0.03	0.15	0.04	0.15	0.04	0.14	0.04
Item Angle	7.59	8.8	6.62	7.62	6.62	7.62	19.39	22.9
MDISC	0.99	0.18	0.98	0.19	0.99	0.22	1.17	0.8
MID	0.19	0.81	0.22	0.86	-0.16	0.88	0.27	0.70
CV	0.99	0.01	0.98	0.01	0.98	0.01	0.83	0.20
RC Angle	6.95	0	6.95	0.00	7.56	0.28	29.40	0.00

Notes: Item angle denotes the measurement direction of an item; MDISC represents item discrimination in the multidimensional latent space, similar to UIRT; MID indicates item difficulty, also analogous to UIRT; RC Angle (Reference Composite Angle) signifies the measurement direction of the test form.

Linking. The results of the estimates for the linking constants A and B are presented in Figure 20, considering factors such as test structure, mean shift, covariance structure, and linking method. Similar to the analysis conducted for MC-I, regression analysis was performed to examine the impact of these intended factors on A and B, and the results are summarized in Table 11.

Figure 20 was visually examined, revealing distinct sensitivities of the linking constants A and B. The linking constant A was found to be sensitive to variations in covariance structure, particularly when multidimensionality tends to collapse onto unidimensionality, leading to a marginal increase in the estimates of A due to the fact that A is responsible for adjusting the unit of a scale in IRT, representing the standard deviation of the population distribution. The linking constant A demonstrated the greatest sensitivity to different linking methods as shown in MC-I. To clarify, the CC linking method yields the highest value for A, followed by FPC which closely aligns with CC. On the other hand, the two SC linking methods consistently generate the lowest value, regardless of test structures, mean shifts, and covariance structures.

On the other hand, across all conditions, the linking constant B consistently exhibits lower values for Y 60 than for Y Base. The linking constant B shows greater responsiveness to variations in mean shift, indicating its role in adjusting the center of a scale. When the mean vector deviates from the baseline in the primary dimension, the variation in B estimates increases by linking method. However, when the mean vector shifts in the nuisance dimension, the estimates of the linking constant B appear to be consistent. Regarding linking methods, when there is a positive shift in the primary latent ability such as mean vectors (1, 0) and (1, 1), B estimates demonstrate a decreasing trend from CC, FPC, to two SC linking methods for both Y Base and Y 60. However, when considering mean vectors of (0, 0) and (0, 1), B estimates remain

constant across linking methods in Y Base. In contrast, in Y 60, there exists an increasing pattern of B estimates observed from CC, FPC, to two SC linking methods.

The regression analysis results, presented in Table 11, confirm the findings from visual inspections and provide statistical evidence. The test structure showed statistical significance for both A and B, with a larger effect on B (-0.22) than on A (0.01). Mean shift and covariance structure were found to be significant for A, with the highest effect (0.04) observed for the covariance structure, while (1, 0.9, 1.65) was the only significant factor for covariance structure. The results indicate that the choice of linking method had the strongest effect, with the highest effect observed for SC- BH (-0.27) followed by SC-SL (-0.26), and FPC (-0.10). In contrast, for B, covariance structure did not show statistical significance, while mean shift exhibited the most influential effect, with an increasing pattern of coefficients as the mean vector deviated from the baseline (i.e., (0, 1): 0.09; (1, 0): 0.77; and (1, 1): 0.86). Notably, the mean vector's deviation (1, 0) from the primary dimension had a greater impact than its deviation (0, 1) from the nuisance dimension due to the fact that the test was designed to measure dominantly the primary dimension.

Figure 20. Linking Constants by Test Structure, Mean, Sigma, and Linking Methods

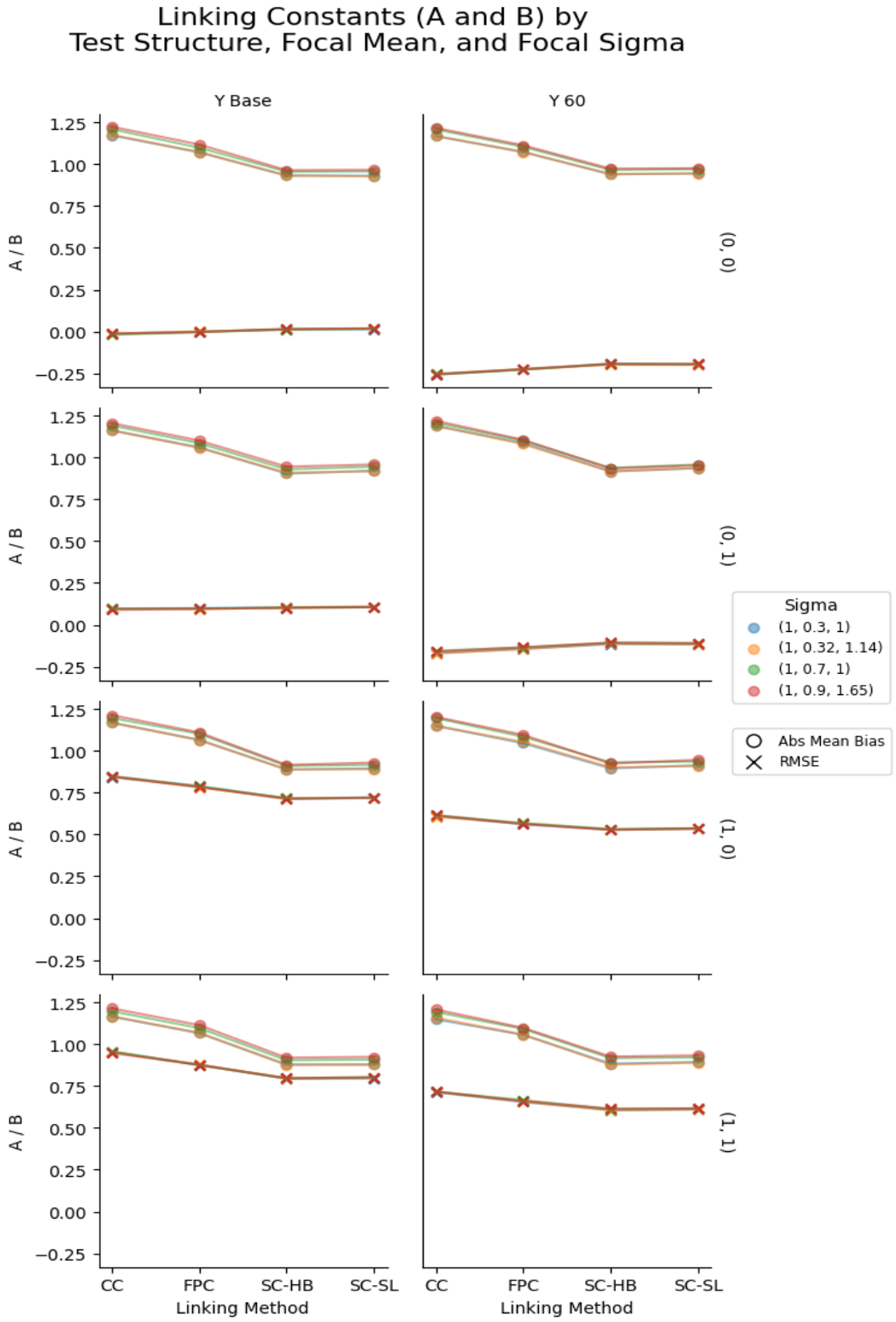


Table 11. Regression Results of Linking Constants

	Linking Constant A Estimate (SE)	Linking Constant B Estimate (SE)
Intercept	1.184 (0.003) ***	0.025 (0.011) *
Test Structure (Y 60)	0.005 (0.002) *	-0.217 (0.006) ***
Focal Mean (0, 1)	-0.011 (0.003) ***	0.092 (0.009) ***
Focal Mean (1, 0)	-0.026 (0.003) ***	0.771 (0.009) ***
Focal Mean (1, 1)	-0.031 (0.003) ***	0.860 (0.009) ***
Focal Sigma (1,0.9,1.65)	0.039 (0.003) ***	NA
Linking Method (FPC)	-0.103 (0.003) ***	-0.022 (0.009) *
Linking Method (SC-HB)	-0.267 (0.003) ***	-0.041 (0.009) ***
Linking Method (SC-S)	-0.258 (0.003) ***	-0.038 (0.009) ***

Note. Standard errors are shown in parentheses.
Significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.
NA: Not presented as it is not statistically significant.

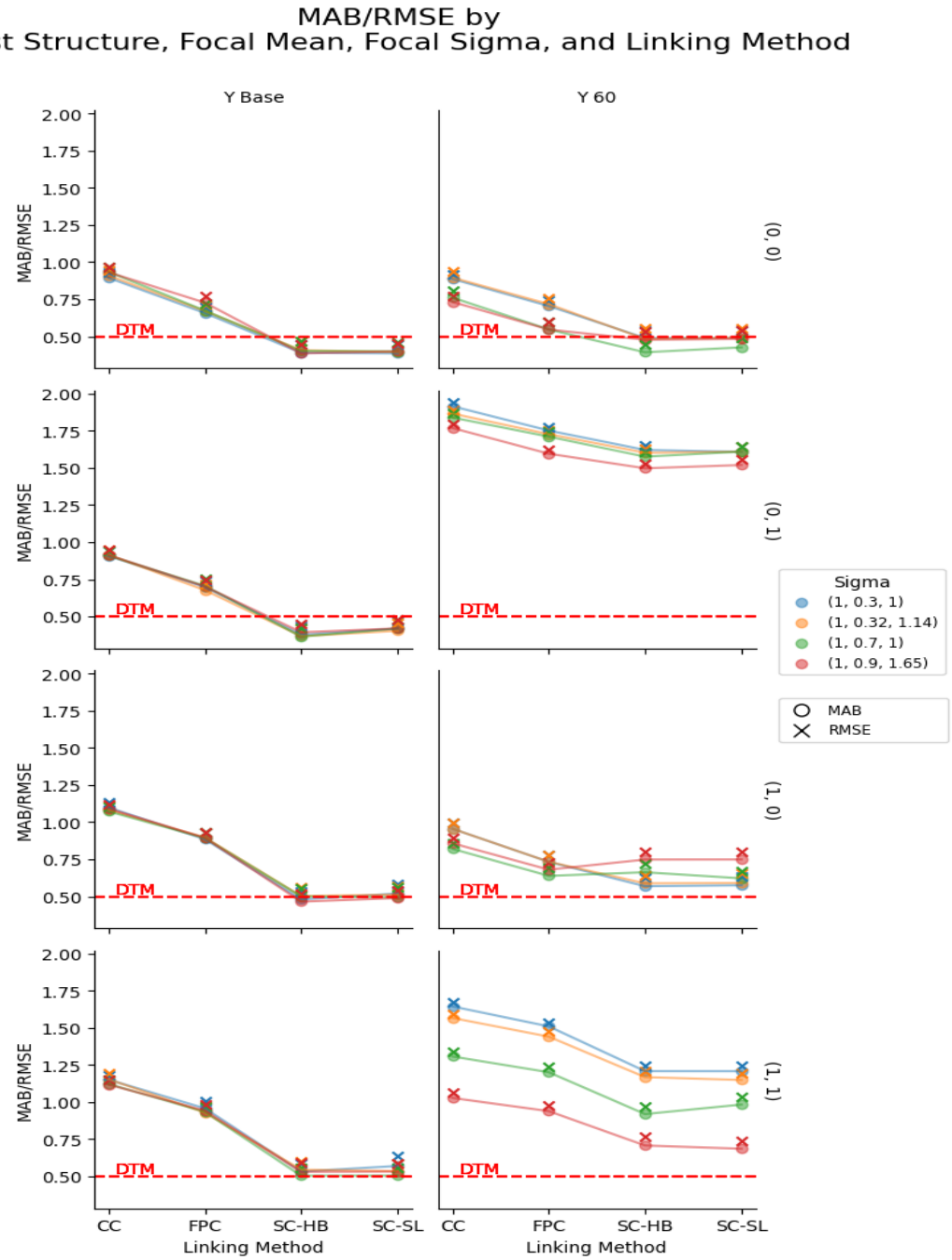
Equating. For each equating method, the results of 128 simulation conditions (2 test structures * 4 different means * 4 different covariance structures * 4 linking methods) are summarized below.

UIRT True Score Equating

The visual comparison presented in Figure 21 offers an intuitive understanding of how intended factors influence equating results, specifically MAB and RMSE. In general, the major source of error in equating is primarily attributed to MAB due to the fact that RMSE has slightly higher than MAB. In the equivalent test structure (i.e., Y Base), the equating performance remains consistent across different covariance and mean structures. However, when the mean vectors are shifted, particularly in the primary dimension, the equating results deteriorate. Notably, the choice of linking method emerges as the most influential factor affecting the equality of equating results, with SC-HB and SC-SL performing similarly and yielding the best outcomes across various conditions. CC performs the poorest, and FPC follows in terms of equating quality. In relation to the DTM, only two SC linking methods fall below the DTM

threshold when mean vectors are set to (0, 0) and (0, 1). Additionally, these two SC linking methods are close to the DTM threshold when the mean vectors are (1, 0) and (1, 1).

Figure 21. MAB and RMSE by Test Structure, Mean, Sigma, and Linking Methods



On the other hand, in the test structure influenced by the nuisance factor (i.e., Y 60), there is a significant variance in equating performance due to study factors. For instance, the equating quality is optimal when the mean vector matches the baseline (0, 0), while a mean vector shifts only on the nuisance dimension (0, 1) results in the poorest performance. When the mean vector is shifted solely in the primary dimension (1, 0), the performance improves. However, when the mean vector deviates in both dimensions (1, 1), the performance variation increases with varying covariance structures, and higher covariance values lead to improved performance. With the baseline mean vector (0, 0) and covariance structure (0, 0.7, 1), two SC linking methods fall below the DTM threshold.

To examine the influence of test structure, mean shift, and linking method on equating, a regression analysis was carried out and the findings are summarized in Table 12. The results align with the observations from the visual inspection. A noticeable negative impact was observed for the mean vector (0, 1), with the largest magnitude effects recorded for both MAB (1.03) and RMSE (1.02). The interaction between Y 60 and the mean vector (1, 1) ranked second in terms of its effect on both MAB (0.35) and RMSE (0.34). Furthermore, it was found that choosing linking methods other than CC led to an improvement in equating quality.

Table 12. Regression Results of UIRT TSE

	MAB Estimate (SE)	RMSE Estimate (SE)
(Intercept)	0.918 (0.049)***	0.952 (0.048)***
Focal Mean (1, 0)	0.168 (0.070)*	0.165 (0.068)*
Focal Mean (1, 1)	0.218 (0.070)**	0.216 (0.068)**
Linking Method (FPC)	-0.235 (0.070)**	-0.225 (0.068)**
Linking Method (SC-HB)	-0.521 (0.070)***	-0.496 (0.068)***
Linking Method (SC-SL)	-0.522 (0.070)***	-0.498 (0.068)***
(Y_60)*(0, 1)	1.034 (0.098)***	1.021 (0.096)***
(Y_60)*(1, 1)	0.349 (0.098)***	0.342 (0.096)***

Note. Standard errors are shown in parentheses.
Significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Adjusted R-squared:0.94 for both MAB and RMSE

UIRT Observed Score Equating

The results obtained from 128 different conditions of UIRT OSE demonstrate a similar pattern as observed in UIRT TSE. Figure 22 visually represents the results, with Table 13 providing a summary of the regression analysis on MAB and RMSE. These analyses demonstrate consistent findings with those obtained from TSE, indicating comparable results. However, the coefficients of OSE in absolute value are larger than those of TSE. For instance, the SC-SL linking method that improved equating equality most reduced MAB and RMSE by, on average, -0.56 and -0.53, respectively in OSE, but by, on average, -0.52 and -0.50, respectively in TSE. Further interpretation of the results can be referred to that of TSE.

Figure 22. MAB and RMSE by Test Structure, Mean, Sigma, and Linking Methods

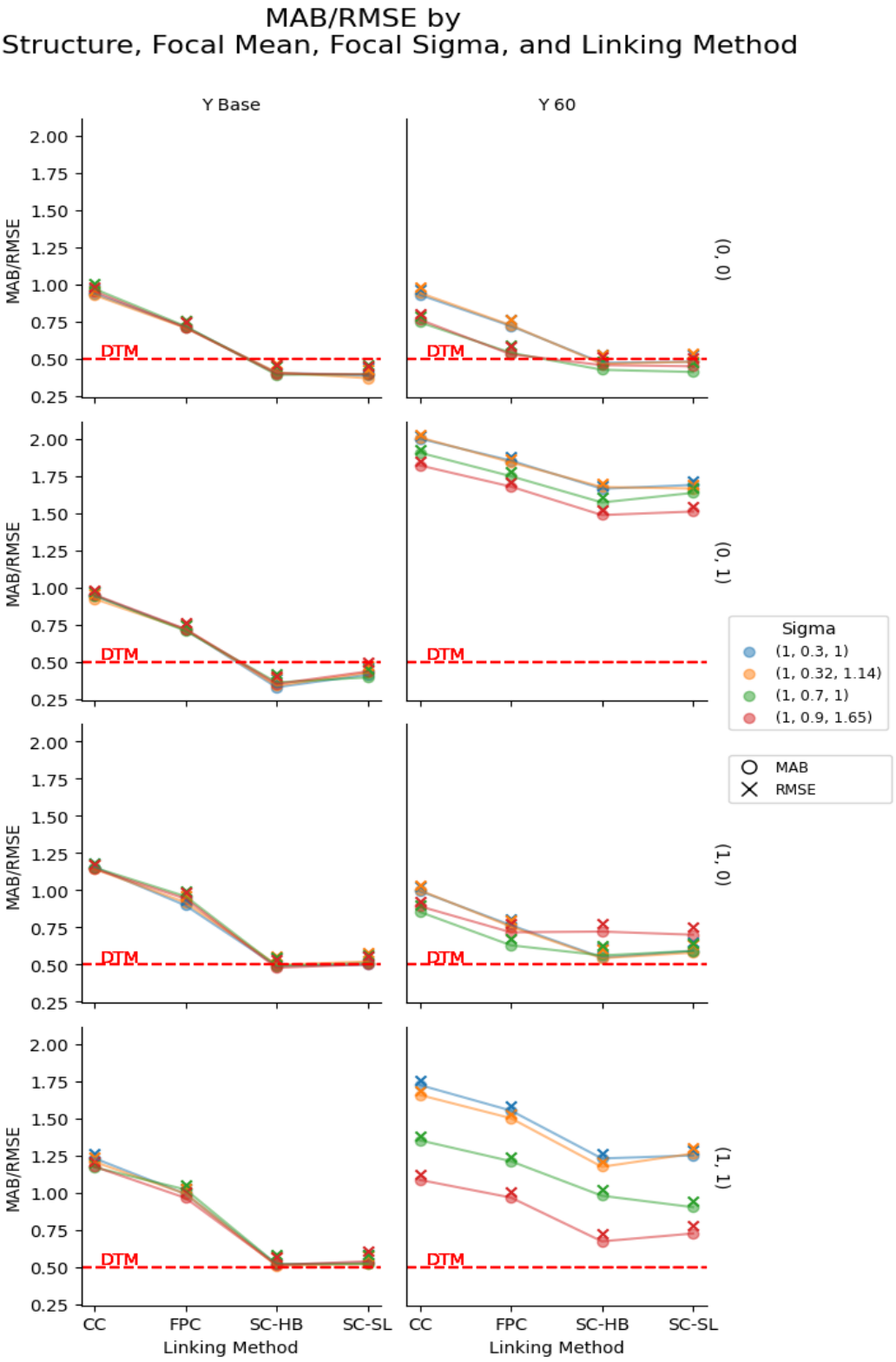


Table 13. Regression Results of UIRT OSE

	MAB Estimate (SE)	RMSE Estimate (SE)
(Intercept)	0.948 (0.054)***	0.980 (0.053)***
Focal Mean (1, 0)	0.199 (0.076)*	0.199 (0.074)**
Focal Mean (1, 1)	0.248 (0.076)**	0.248 (0.074)**
Linking Method (FPC)	-0.235 (0.076)**	-0.226 (0.074)**
Linking Method (SC-HB)	-0.545 (0.076)***	-0.517 (0.074)***
Linking Method (SC-SL)	-0.560 (0.076)***	-0.533 (0.074)***
(Y_60)*(0, 1)	1.095 (0.108)***	1.080 (0.105)***
(Y_60)*(1, 1)	0.361 (0.108)***	0.350 (0.105)***

Note. Standard errors are shown in parentheses.
Significance levels: *** p < 0.001, ** p < 0.01, * p < 0.05.
Adjusted R-squared:0.94 for both MAB and RMSE

EVALUATION

Classification

Classification indices across different factors are depicted visually in Figure 23, with the corresponding regression analysis results summarized in Table 14.

Consistent with the findings from MC-I and visual assessments in Figure 23, the CA indices for UIRT surpass those for the generating MIRT model, irrespective of the considered factors. However, the CC indices exhibit higher values only at the means of (0, 0) and (0, 1). Specifically, when there is a positive shift in ability within the primary dimension, CC indices align closely with the baseline when considering CC and FPC linking methods. Conversely, employing SC-HB and SC-SL results in lower CC indices than the baseline. It is further observed that the presence of a mean shift in the primary dimension leads to increased variation in CC indices influenced by the covariance structure. Notably, this variation is more pronounced in Y 60. More precisely, as covariance becomes higher, there is an improvement in CC indices, across all linking methods.

In Y Base, the CC indices are generally smaller than those for Y 60 when mean vectors are (0, 0) and (0, 1). When mean vectors are (1, 0) and (1, 1). the CC indices fall below the baseline when using two SC linking methods. On the other hand, Y 60 shows the similar trend of CC on the same condition but with a large variation by covariance structure. In the presence of the mean vector (1, 0), there is a contrasting trend in CA indices across the linking methods between Y Base and Y 60. Specifically, in Y Base, there is an increasing pattern of CA indices observed from CC, FPC, and two SC linking methods. However, in Y 60, a decreasing pattern of CA indices is observed. Generally, both CC and CA indices exhibit improvement with an increase in covariance between latent variables. However, an exception occurs when employing two linking methods with the mean vector (0, 1) in Y 60. In this case, the higher covariance value diminishes the classification indices.

The outcomes of a regression analysis on CC and CA indicate that, on average, Y 60 exhibits slightly higher classification indices (by 0.006 in CC, and 0.004 in CA) compared to Y Base. In CC, there are two-way interactions involving test structure and mean shift with (1, 0) and (1, 1). These interactions lead to average reductions of CC indices by -0.06 and -0.008, respectively. Meanwhile, CA demonstrates three-way interactions involving test structure, mean shift with (1, 0), and linking method. That is, Y 60, mean vector (1, 0), and FPC interaction results in an average CC index decrease of -0.01. On the other hand, the three-way interactions involving Y 60, mean vector (1, 0), and SC-HB lead to an average CA index decrease of -0.12, and SC-SL leads to an average CA index decrease of -0.13.

Figure 23. CC and CA by Test Structure, Mean, Sigma, and Linking Methods

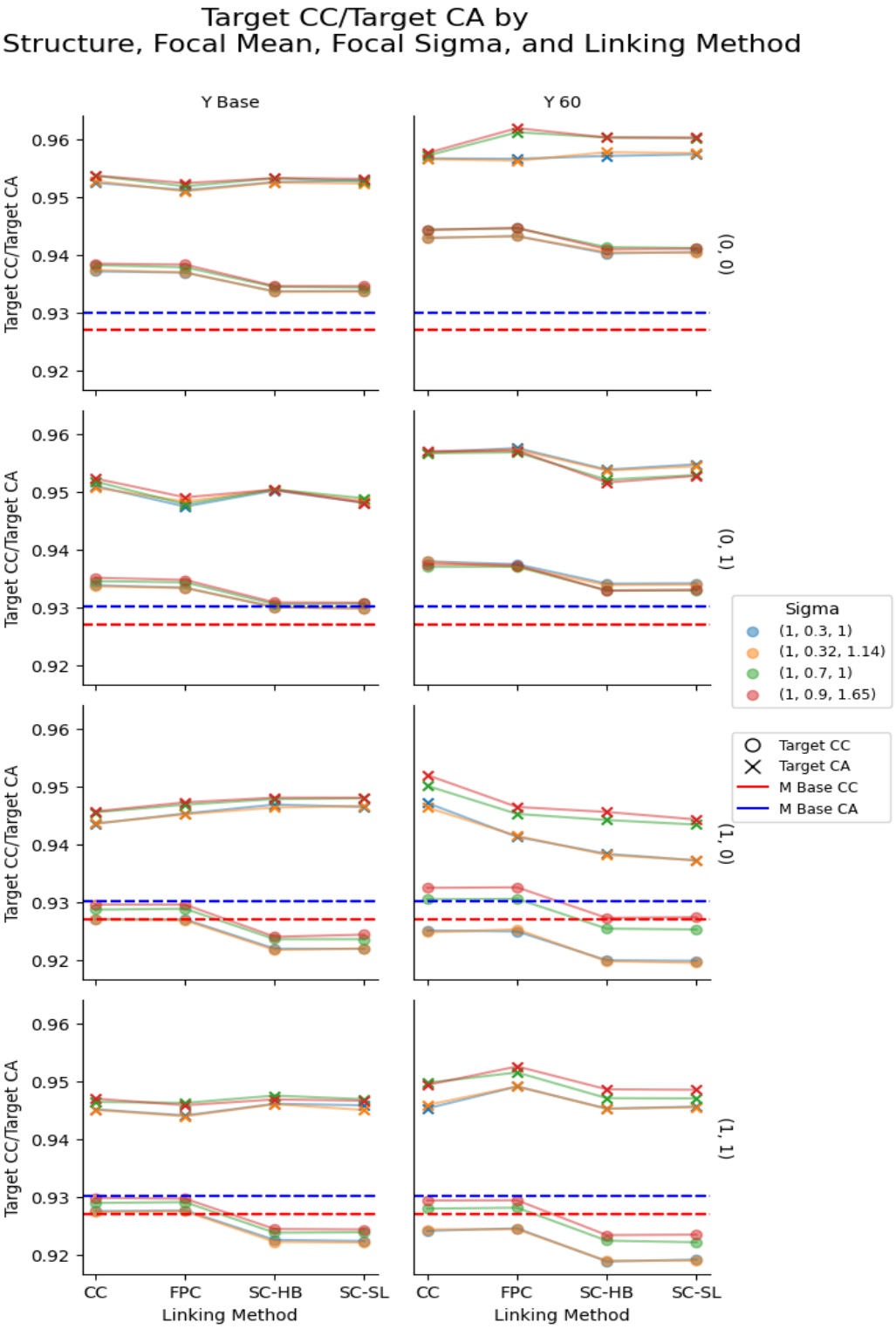


Table 14. Regression Results of CC and CA

	CC Estimate (SE)	CA Estimate (SE)
Intercept	0.938 (0.001)***	0.953 (0.001)***
Test Structure (Y 60)	0.006 (0.001)***	0.004 (0.001)**
Focal Mean (0, 1)	-0.004 (0.001**	N/A
Focal Mean (1, 0)	-0.010 (0.001)***	-0.009 (0.001)***
Focal Mean (1, 1)	-0.009 (0.001)***	-0.007 (0.001)***
Linking Method (SC-HB)	-0.004 (0.001)**	N/A
Linking Method (SC-SL)	-0.004 (0.001)**	N/A
Y 60) *(1, 0)	-0.006 (0.002)**	N/A
(Y 60) *(1, 1)	-0.008 (0.002)***	N/A
(Y 60) *(FPC)	N/A	0.004 (0.002)*
(1, 0)*(SC-HB)	N/A	0.003 (0.002)*
(1, 0)*(SC-SL)	N/A	0.003 (0.002)*
(Y 60) *(1, 0)*(FPC)	N/A	-0.010 (0.002)***
(Y 60) *(1, 0)*(SC-HB	N/A	-0.012 (0.002)***
(Y 60) *(1, 0)*(SC-SL	N/A	-0.013 (0.002)***

Note. Standard errors are shown in parentheses.
Significance levels: *** p < 0.001, ** p < 0.01, * p < 0.05.
NA: Not presented as it is not statistically significant.
Adjusted R-squared: 0.93 for CC and 0.90 for CA

Equity Properties

The comparison of equating equity indices for D1 and D2 is visually presented in Figure 24 with respect to test structure, mean shift, covariance structure, and linking method. The impact of these factors on the indices is further analyzed through a regression analysis, as summarized in Table 15.

Overall, the CC linking method consistently yields higher D1 values compared to the base model across all conditions. For Y Base, when the mean vectors (0, 0) and (0, 1) are present, the FPC and two other linking methods produce D1 indices that are either equal to or lower than the baseline. In contrast, when the mean vectors (1, 0) and (1, 1) are present, only two SC linking methods get close to the baseline. For D2 indices, URIT demonstrates similar or slightly higher values than the baseline. For Y 60, when considering the base mean vector (0, 0),

FPC with high covariances and two SC linking methods with low covariances exhibit values below the baseline. With the mean vector of (0, 1), all linking methods yields the least favorable outcome in both D1 and D2. When employing the mean vector (1, 0) and (1, 1), the impact of covariance structures becomes more noticeable. In particular, when the mean vector is fixed at (1, 0), an increase in covariance contributes to a discernible decrease in both D1 and D2 indices, especially for two SC linking methods. Conversely, when the mean vector is (1, 1), an increase in covariance brings about a noticeable improvement in both D1 and D2 indices for all linking methods.

The results of the regression analysis, as outlined in Table 15, corroborate the findings derived from visual observations. Specifically, when Y 60 interacts with the mean shift of (0, 1), the equating bias increases on average by 0.218 after accounting for other factors. In contrast, the equating precision experiences an average increase by -0.006, suggesting an improvement in the precision of equating quality. However, it is important to note that the interpretation of D2 is only meaningful when D1 is meaningful. Additionally, the regression analysis did not reveal covariance effect, resulting in a lack of explanation for the variation observed in D1 when considering Y 60 with a mean shift of (1, 1).

Figure 24 visually represented the global index, D12, which closely resembled the pattern observed in the FOE D1 index, mainly due to the predominant influence of D1. Any minor discrepancies between the two can be attributed to variations in the computational approach used as explained previously in MC-I. Importantly, despite the numerous similarities, it is noteworthy that in Y 60, with mean vectors of (0, 1) and (1, 1), the performance gap resulting from the choice of linking method was considerably diminished. In the case of Y Base, two SC linking methods demonstrated superior or at least comparable performance to the baseline in D12.

Figure 24. FOE D1 and SOE D2 by Test Structure, Mean, Sigma, and Linking Methods

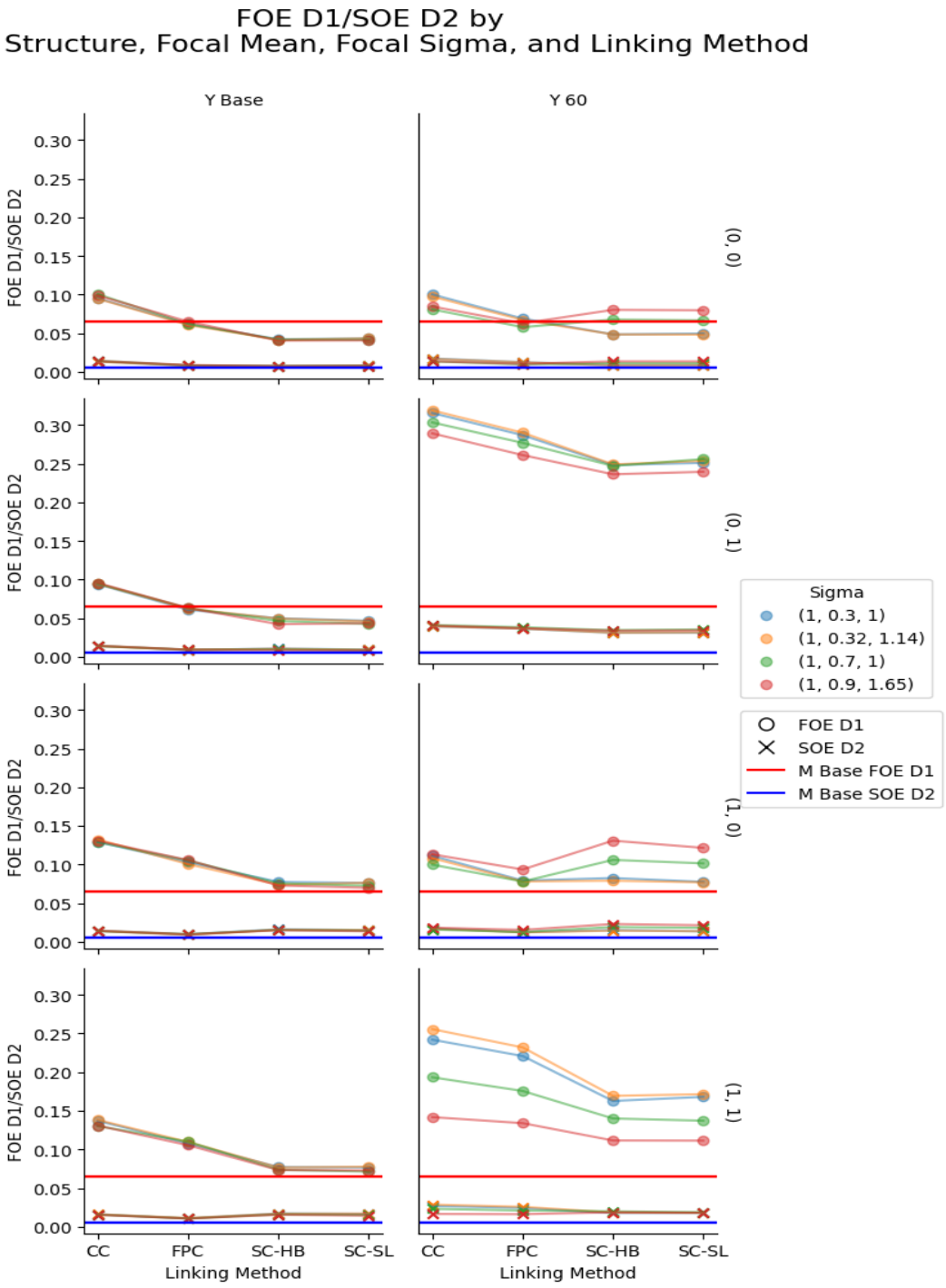
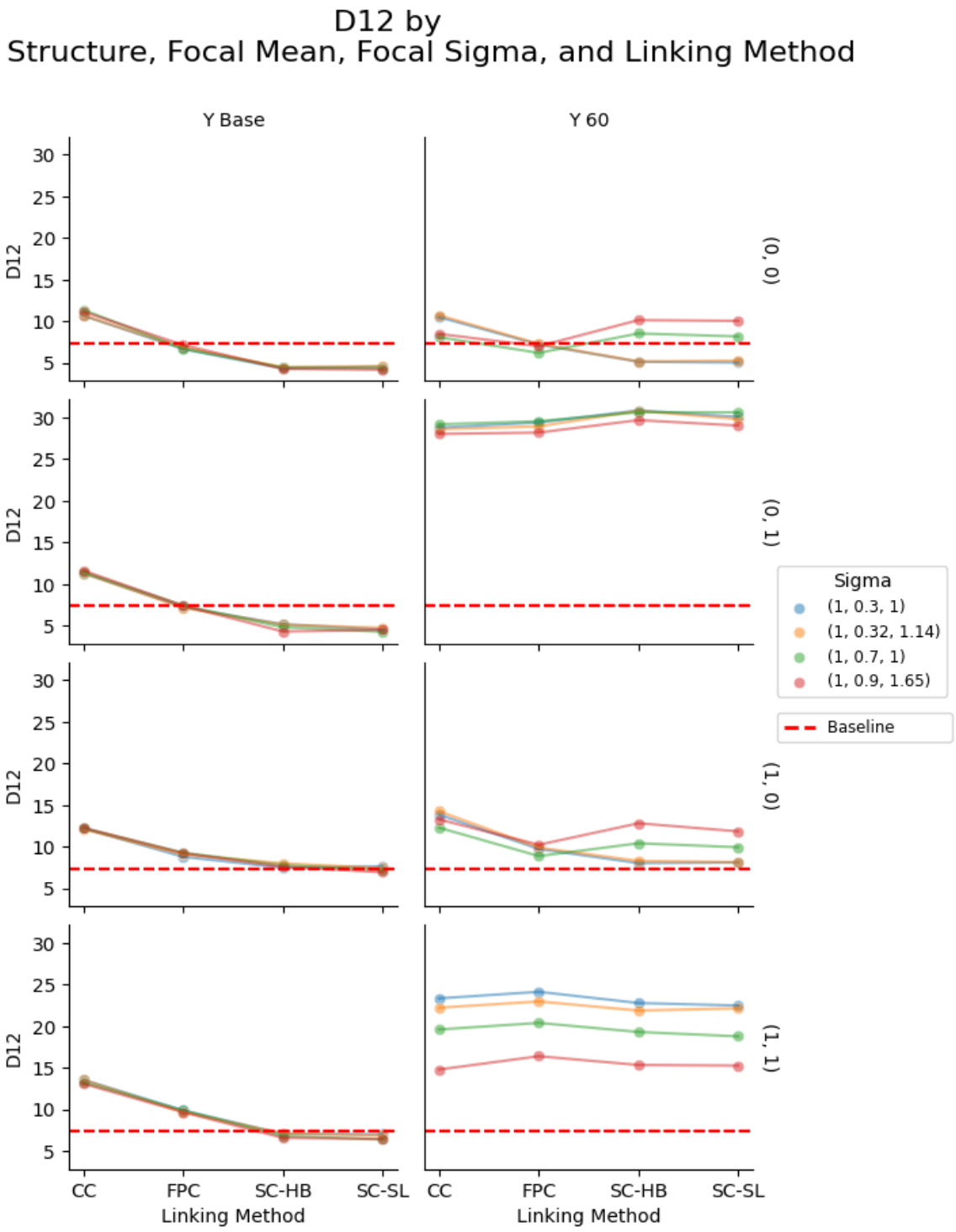


Table 15. Regression Results of FOE and SOE

	FOE D1 Estimate (SE)	SOE D2 Estimate (SE)
(Intercept)	0.097 (0.008) ***	0.014 (0.001) ***
(1, 0)	0.033 (0.011) **	NA
(1, 1)	0.037 (0.011) **	NA
FPC	-0.035 (0.011) **	-0.005 (0.001) ***
SC-HB	-0.056 (0.011) ***	-0.006 (0.001) ***
SC-SL	-0.055 (0.011) ***	-0.005 (0.001) ***
(Y 60) *(0,1)	0.218 (0.016) ***	-0.006 (0.001) ***
(1, 1) *(SC-HB)	0.081 (0.016) ***	-0.006 (0.001) ***
(1, 1) *(SC-SL)	NA	0.007 (0.002) ***
(1, 1) *(SC-HB)	NA	0.006 (0.002) **
(Y 60) *(0,1) * (SC-HB)	NA	-0.007 (0.003) **
(Y 60) *(0,1) *(SC-SL)	NA	-0.007 (0.003) *

Note. Standard errors are shown in parentheses.
Significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.
NA: Not presented as it is not statistically significant.
Adjusted R-squared: 0.95 for FOE and 0.95 for SOE

Figure 25. D12 by Test Structure, Mean, Sigma, and Linking Methods



CHAPTER V: DISSCUSION

The current study investigated the influence of multidimensionality on unidimensional item response model linking and equating through two hypothetical multidimensional scenarios within a nonequivalent group common-item equating design.

In the first scenario, the focus was on a test designed to measure multiple constructs of interest. Three new test forms were utilized to explore the effects of different test structures compared to a base form. One new form had an equivalent test structure to the base form, aligning both unidimensionally and multidimensionally. The second new form represented a case of a semi-equivalent test structure, sharing the same reference composite direction but differing in the multidimensional latent space. The final new form had a non-equivalent test structure, lacking equivalence both unidimensionally and multidimensionally.

The second scenario involved a test intended to measure a primary construct of interest but contaminated with a nuisance factor. Two new forms were examined alongside the base form. The first form and the base form contained items and a reference composite located within a chosen validity sector (Ackerman, 1992), while the other form included contaminated items, heavily influenced by the nuisance factor, causing the reference composite to deviate away from the validity sector toward the nuisance dimension.

Under these two hypothetical scenarios, the study utilized classification measures and equating equity properties to compare their marginal indices between the baseline MIRT and UIRT. The following section provides a summary of the study's findings.

The results concerning Research Questions (RQ) related to MC-I is summarized as follows:

RQ 1 and 2: impact on classification consistency and accuracy

It was found that for MC1, the classification indices of UIRT were higher compared to the baseline generated model across all conditions. Y Same RC outperformed Y Base and Y Diff RC in both classification indices. The classification indices increased as the covariance values increased. CC linking method demonstrated superior performance compared to other linking methods while controlling for other conditions. When the mean vector increased in both latent dimensions, the variation in performance reduced due to covariance structure and linking method was minimal.

RQ 3 and 4: impact on equating equity properties (first order: D1 and second order:D2)

It was observed that the equity indices of the baseline model performed better than most UIRT cases across all conditions. D1 demonstrated greater variation and divergence from the baseline across different conditions, whereas D2 displayed a more consistent pattern and was in closer alignment with the baseline. However, when the test structure of the new form aligned with the base form (i.e., Y Base), with mean vectors not deviating significantly and using two linking methods (i.e., HB and SL) with separate calibration, the D2 indices of UIRT were closer to the baseline. When the test structure of the new form was aligned unidimensionally only (i.e., Y Same RC), the CC linking method performed best in both D1 and D2 across different mean shifts and covariance structures. For the non-equivalent test structure (i.e., Y Diff RC), mixed results were observed, with CC outperforming except for the mean vector (0, 1), where SC-BH and SC-SL methods performed best. The meaningfulness of SOE is contingent upon the quality of FOE, which is an important point to consider (Thomasson, 1993).

RQ5: impact on the combined equating equity property (D12)

It was observed that the overall pattern of D12 closely resembled that of D1, despite having distinct interpretations. The D12 index served as an indicator of the predictability of expected performance on another test (Bolt, 1999). A smaller value of the index indicates that Form Y could predict equated scores on Form X with reduced bias and increased precision. The current study demonstrates that the predictability of equating performance was primarily influenced by the test structure. When the test structures of two test forms were equivalent both unidimensionally and multidimensionally, the predictability was most accurate and precise. However, in cases where the test structures differed in a multidimensional latent space but were equivalent in a composite unidimensional latent space, the accuracy of the predictability was diminished. Additionally, the choice of linking method significantly impacted the value of the D12 index. In situations where the test structures were not comparable in either a unidimensional or multidimensional sense, mixed results were observed. However, it was evident that a balanced mean vector shift yielded better results, and the choice of linking methods played a critical role. Specifically, in contrast to cases with the equivalent test structure where SC-HB and SC-SL demonstrated superior performance, the CC linking method outperformed the separate calibration linking methods in cases involving the semi-equivalent structure.

The results concerning Research Questions (RQ) related to MC-II is summarized as follows:

RQ 1 and 2: impact on classification consistency and accuracy

It was observed that for MC2, the classification accuracy indices of UIRT were consistently higher than those of the baseline model across all conditions. However, it was noted that higher classification accuracy indices (than the baseline) were predominantly observed when the mean vector remained the same as the baseline or only varied in the nuisance dimension.

Conversely, when the mean shift occurred in the primary dimension, the classification consistency decreased below the baseline level, particularly when utilizing two SC linking methods. Additionally, when the dimensionality expanded into two dimensions as the covariance of the two latent variables decreases, the classification consistency indices decrease to their lowest values. In contrast, under the same conditions, CC and FPC produced higher or comparable classification consistency indices compared to the baseline model.

RQ 3 and 4: impact on equating equity properties (first order: D1 and second order:D2)

Unlike the case of MC-I where the equity indices of the baseline model demonstrated superior performance compared to most UIRT cases across all conditions, in MC-II, when the test structures of two test forms were equivalent, UIRT equating resulted in FOE indices that were either better or equivalent to those of the baseline, particularly when utilizing two SC equating methods. Additionally, even in a new test form heavily influenced by a nuisance factor but with the same mean vector as the baseline, UIRT equating produced FOE indices that were equivalent to the baseline. However, when the mean vector changed in the nuisance dimension in Y 60, the interaction effect caused the largest discrepancy in both FOE and SOE compared to the baseline. It is also worth noting that as the dimensionality collapsed into a single dimension as the covariance of the two latent variables increases, the FOE indices exhibited improvement.

RQ5: impact on the combined equating equity property (D12)

When examining the overall pattern of D12 and comparing it to D1, it was observed that they closely resembled each other, despite having different interpretations. When considering both D1 and D2 simultaneously, the two SC linking methods demonstrated superior predictability compared to the base MIRT model, in cases where the test structures were equivalent. However, when the primary dimension was impacted by the presence of a nuisance

factor, the UIRT equating method exhibited the poorest predictability of the equating results, regardless of the linking methods and covariance structures employed.

Implications to Operational Setting

In practical settings, MIRT equating procedure is not a preferable choice due to the limit in its practical utilities. Multidimensionality can be viewed from two perspectives: (expected) multidimensionality by design and idiosyncratic multidimensionality. The first multidimensionality is the case when multidimensionality is inevitably emerged on test forms by design, while the latter case is when multidimensionality appears due to an unintended nuisance factor(s). In the current study, two cases of multidimensionality are designed into MC-I for the first case and MC-II for the second case of multidimensionality. The followings are the implications of the current study.

When designing tests to measure one construct of interest but inevitably involve multiple domains, such as medical licensure and certification examinations, or tests with integrated items to assess multiple subject areas, it is crucial to establish equivalent test structures between two forms for equating purposes in order to achieve optimal equating quality. However, when the test structures are unknown, it is recommended to employ different linking methods, as they may yield different equating results based on the dimensional structures of the forms. The findings suggest that separate calibration using SL and HB methods outperforms when the test structures of two forms are equivalent. However, in cases where the dimensional structures of forms are not comparable in the multidimensional latent space but are equivalent in the unidimensional latent space, the CC method emerges as the preferred choice. Additionally, it should be noted that controlling for equivalent abilities between populations is crucial when the dimensional structures of forms are not equivalent.

Regarding classification consistency and accuracy, the UIRT approach exhibits higher values compared to the multidimensional item response theory approach. Additionally, as a multidimensional structure collapses into a unidimensional latent space, the classification consistency and accuracy tend to increase. Across all conditions, the concurrent calibration method generally produces the best outcomes in terms of classification accuracy and consistency. However, it is important to note that classification results should be interpreted with caution. The Lee's D method (2010) chosen for classification indices is a single form procedure based its underlying assumption of parallelism between test forms. For example, in Figure 4.4, it is evident that the classification indices are highest for Y Same RC as a whole. However, it is important to recognize that the dimensional structure of Y Same RC is not equivalent to the base form, X Base. Therefore, it is crucial to ensure that the prerequisite of form parallelism is met before interpreting the indices accurately. If there are variations in classification indices between forms, it may indicate a change between those forms.

When a test aims to measure a primary latent trait but is contaminated with a nuisance factor, it is essential to minimize the influence of the nuisance factor to obtain desirable equating results. The findings indicate that SL and HB methods are recommended regardless of the test structures. However, if the forms have test structures that are not comparable to each other, it becomes crucial to minimize the ability disparity of the nuisance dimension between populations and to increase the correlation between dimensions.

Regarding classification consistency and accuracy, the findings suggest that minimizing the shift in the nuisance ability is of utmost importance. Overall, the concurrent calibration method is the preferred choice for achieving optimal classification consistency and accuracy.

Limitations and Directions for Future Research

In this section, the limitations of the current study are discussed. It should first be acknowledged that the item parameters might not accurately represent real-world conditions. To be specific, operational items may not behave as being hypothesized in the simulation conditions, which were aimed to exemplify distinct test structures in linking and equating scenarios. For example, all items may not measure two latent traits evenly as shown in X Base Form in MC-I. Additionally, for MC-II, different angles can be applied other than 20 degree for the validity sector and 60 degree for the item spread as an influence of the nuisance factor. Importantly, incorporating real data analysis within this context could enhance the research design and provide additional support.

Second, along with the first limitation, the base model employed a 2-dimensional MIRT model with a complex test structure. The complex MIRT model is commonly employed for exploratory purposes; however, in operational testing settings, a confirmatory MIRT model is defined in advance based on specific criteria. These criteria may include using different item formats, such as multiple-choice items and construct response items, for a simple structured MIRT model; item bundles for a testlet model; or a test consisting of a general factor with multiple specific factors for a bifactor model. Furthermore, a complex MIRT model is not identifiable which prevents calibration of the simulated data to obtain empirical information for its comparison with UIRT. This limitation restricts the evaluation of equating equity indices in comparison with results from calibration of generated response data.

Third, the present study employed a non-equivalent group anchor item design for equating purposes. This design utilizes a set of common items to account for the impact of population differences and variations in form difficulty. However, it is suggested that a random

group equating design may be more effective in investigating the influence of test structures, as it reduces the complexity of the study design by controlling for the impact of population differences. Furthermore, the utilization of a common item set could potentially distort the original dimensional structures being examined. One consideration is that applying the random group design in operational testing scenarios may pose a challenge due to the potential inconsistency in the test dimensionality over time under different internal and external testing conditions.

Fourth, setting the cut score should rely on maximizing the information of parallel forms and ensuring same information to be shared across parallel forms. Regrettably, this pivotal aspect was overlooked in the present study, which concentrated on variations in test structures instead. It would be prudent to include this consideration when assessing classification indices.

Fifth, unlike equating equity indices D1 and D2 (Tong & Kolen, 2005) which are standardized, the D12 index (Bolt, 1999; Thomasson, 1993) employed in the study is not a standardized measure, indicating its dependency on the specific test being examined. Thus, the interpretation of results primarily relies on relative comparisons, lacking an objective criterion for assessment.

Sixth, in future research, it would be valuable to quantify multidimensionality. Within practical contexts, exploring the threshold of multidimensionality within the UIRT framework holds importance for assessing its impact. One potential method involves assessing the consistency of pass rates by adjusting the degree of reference composites between two forms.

The last but not least, with regards to OSE in both MIRT and UIRT, when MIRT is compensatory and monotonically increasing on latent variables, the equating results between MIRT and its UIRT equivalent model produce similar outcomes, as demonstrated later.

However, it is important to note that these equating procedures are unable to account for dimension-specific changes in item difficulty across test forms (Bolt, 1999). When multidimensional response data is calibrated using UIRT, the multidimensional scale is collapsed into a unidimensional scale, resulting in the loss of dimension-specific information.

Although MIRT true score equating has not been proposed due to the one-to-many relationship between an integer score and infinite coordinates of latent variables that give the specific score on the test characteristic function, it is possible to approximate equated scores by averaging a finite number of the coordinates, which are carefully chosen. This introduces the possibility of future research on approximate MIRT true score equating, which is elucidated in the following section.

Approximate MIRT TSE

In ATM (Approximate Multidimensional Item Response Model True Score Equating), equated scores can be obtained by an approximation process with a small tolerance level, or precision. Its procedure is similar to that of the UIRT TSE, but involves several technical specifics. That is, the AMT equating procedure involves (first) specifying integer scores on a new form, (second) finding θ coordinates corresponding to the score with an appropriate precision level (e.g., 0.1), (third) determining score equivalents through the TCS of the old form, and (last) finally obtaining the equated scores by averaging score equivalents through three weighting schemes.

The first weighting method is no-weighting or equal weighting approach. The second option is integer-value weighting method, which uses the conditional summed score probability distribution obtained through the LW recursive formula (Lord & Wingersky, 1984). This method involves rounding the score equivalents on the old form to integers in order to determine the

conditional probabilities corresponding to score equivalents as weights. However, this rounding process can introduce significant rounding errors. To mitigate this issue, the real-value weighting scheme can be employed. This approach aims to minimize the impact of rounding errors by computing a conditional real-value summed score probability distribution using the generalized LW algorithm (Kim, 2013). It is worth noting that the real-value weighting scheme may require higher computational cost.

A hypothetical example of the no-weighting AMT and the integer-weighting AMT is illustrated in Table 16. The real-value weighting AMT follows a similar computation approach to the integer-weighting AMT method. Figure 26 illustrates the discrete test characteristics surface of Form X with theta coordinates for each integer score, indicating the inconsistency of accuracy in approximation due to the different numbers of coordinates and coverage of latent trait space. The contour plot shown in Figure 27 depicts ten theta coordinates evenly spaced along the optimal line, representing a score of 5 on a 10-item test. By employing the equal number of equidistant theta coordinates on the optimal line for each score, the problem of inconsistency is expected to be effectively resolved. In Figure 28, the conditional observed score probabilities of the X Base form are visually represented. These probabilities serve as weights for corresponding score equivalents.

Figure 29 illustrates the difference between the Y form scores and their equated scores in MC-I. The generating MIRT has items that maintain the test structures of the whole test without the anchor item set (to maintain the intended dimensional structures), which can be considered as a random group equating design. The linear composite UIRT obtained the item parameters from the MIRT item parameters using the formula (Zhang & Wang, 1988). The baseline, referred to as

X Base, acts as a reference point obtained from identity equating to assess the effectiveness of the AMP procedure in equating.

When the test structures of forms are similar, equating results between MIRT OSE, UIRT OSE, and TSE closely align, visualized in Figure 31 for MC-I and Figure 32 for MC-II. However, the score differences observed in ATM differ from those of the three conventional equating methods due to ATM's ability to consider dimensions-specific score change between forms. Figure 29 visualizes the TCSs of X Base and Y Base, while Figure 30 displays the TCSs of X Base and Y Same RC, providing an illustration of the score difference depicted in Figure 30. For example, the equated scores of Y Base exceed those of X Base, indicating that Y Base is generally more challenging than X Base except for two extreme score regions in Figure 29. The plot on the bottom right of Figure 31 shows the score difference clearly with symbols in red. In order to compare the weighting schemes, two different approaches are represented: no-weighting using * and integer-weighting using ** in Figure 31 and 32.

In short, AMT provides richer interpretation of the equating result because it absorbs the information of the dimension-specific changes in (expected) scores on TCSs between forms, unlike the conventional IRT equating procedures (see Bolt, 1999) . To be specific, for successful AMT equating, two test forms to be equated are required to have not only monotonicity of a test characteristic function for TSE and a cumulative probability function for OSE but also a congruent dimensional structure between forms. Thus, AMT could add a validity value to the current definition of equating, which has been viewed as a "statistical" procedure, but from the AMT perspective, equating could be used as a validity tool to assess and validate the degree of alignment in the test structures of two forms in the equating context.

Table 16. An Illustrative example of AMT

No Weight				Integer Value Weight					
New Form Y		Old Form X	$Eq_X(Y)$ by $AMT*$	Old Form X					$Eq_X(Y)$ by $AMT**$
$x \in \mathbb{Z}_{\geq 0}$	(θ_1, θ_2)	$x_i \in \mathbb{R}_{\geq 0}$	$\frac{\sum_i x_i}{N}$	$x_i \in \mathbb{Z}_{\geq 0}$	$P(X = \chi_i \mid \boldsymbol{\theta})$	$P_x(X = \chi \mid \boldsymbol{\theta}) = \sum_{k=1}^{n} P_x(X = \chi_k \mid \boldsymbol{\theta})$	$P_i(X = \chi_i \mid \boldsymbol{\theta}) = \frac{P_i}{\sum_{i=1}^{n} P_i}$	$x_i * P_i$	$\sum_i x_i * P_i$
Score on Y	Theta Coordinates	Equivalent Scores on X	Equated Score Y on X	Rounded Scores on X	Corresponding Conditional Probabilities	Summed Conditional Probabilities	Normalized Weights	Expected Equivalent Scores	Equated Score Y on X
30 out of 50	(-1.5, 1.5)	29.3	30.50	29	0.003	0.003	0.08	2.35	30.49
	(-1.0, 1.0)	29.6		30	0.008	0.016	0.43	12.97	
	(-0.5, 0.5)	30.0		30	0.008				
	(0, 0)	30.8		31	0.005				
	(0.5, -0.5)	30.7		31	0.005	0.015	0.41	12.57	
	(1.0, -1.0)	31.4		31	0.005				
	(1.5, -1.5)	31.6		32	0.003	0.003	0.08	2.59	

Figure 26. Test Characteristic Surface of X Base with Theta Coordinates in Red

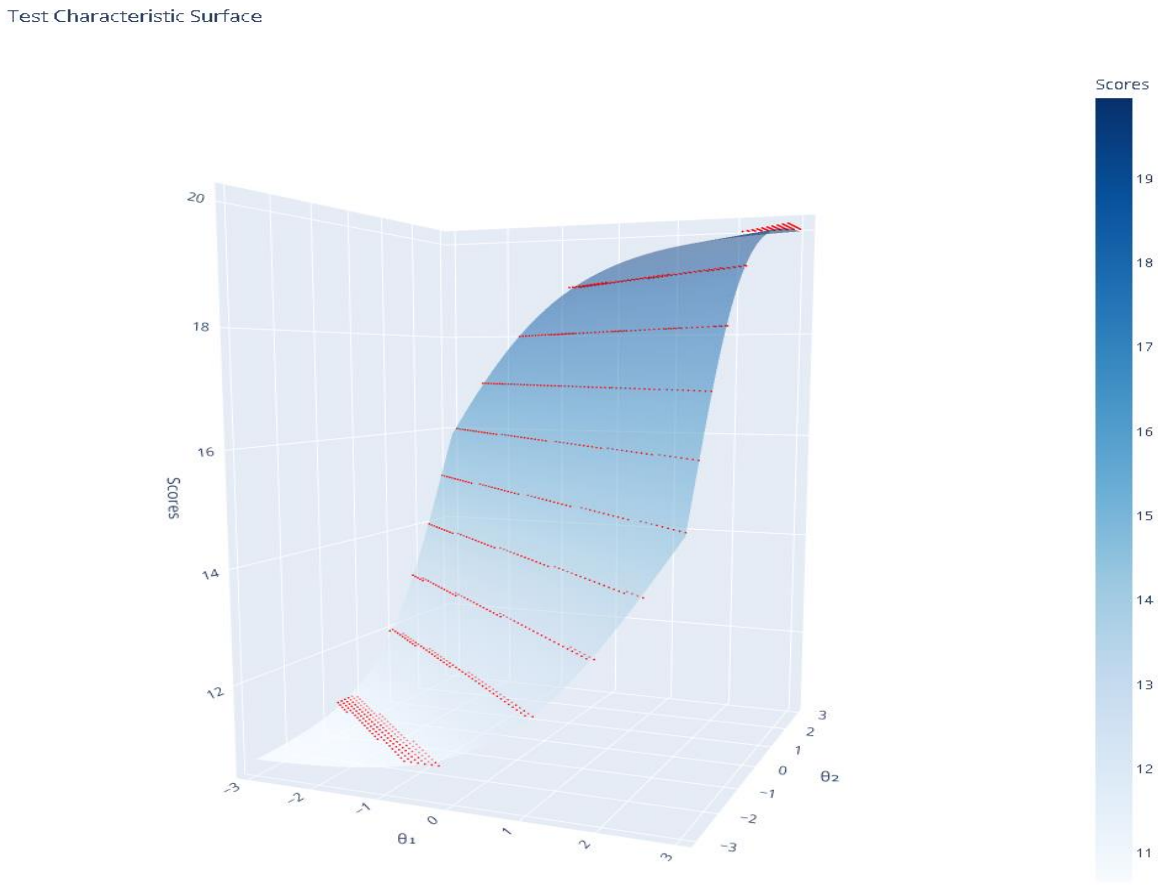


Figure 27. Ten Equidistant Theta Coordinates Located on the Optimal Line that Represents a Score of 5 on a 10-item Test in a Contour Plot

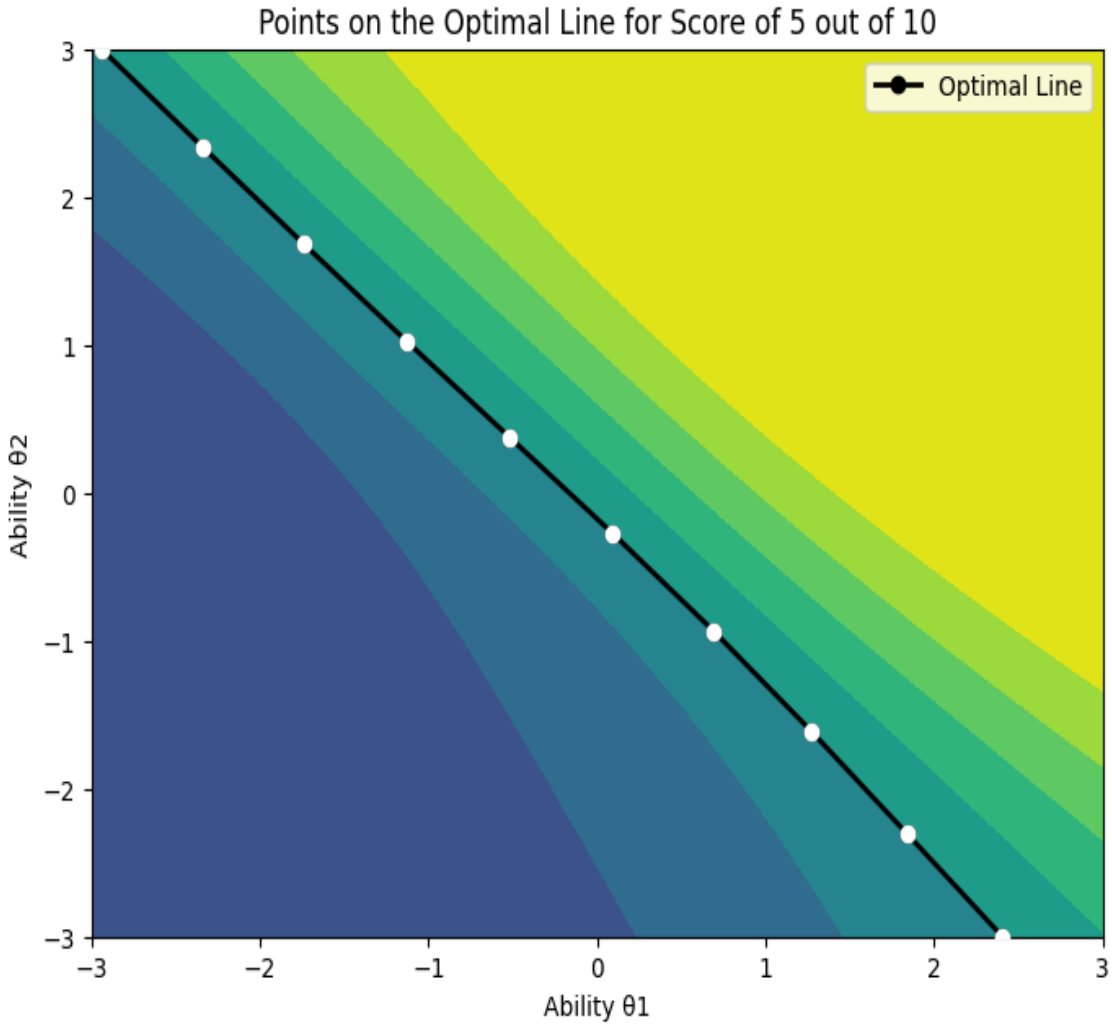


Figure 28. Conditional Observed Score Probability of X Base

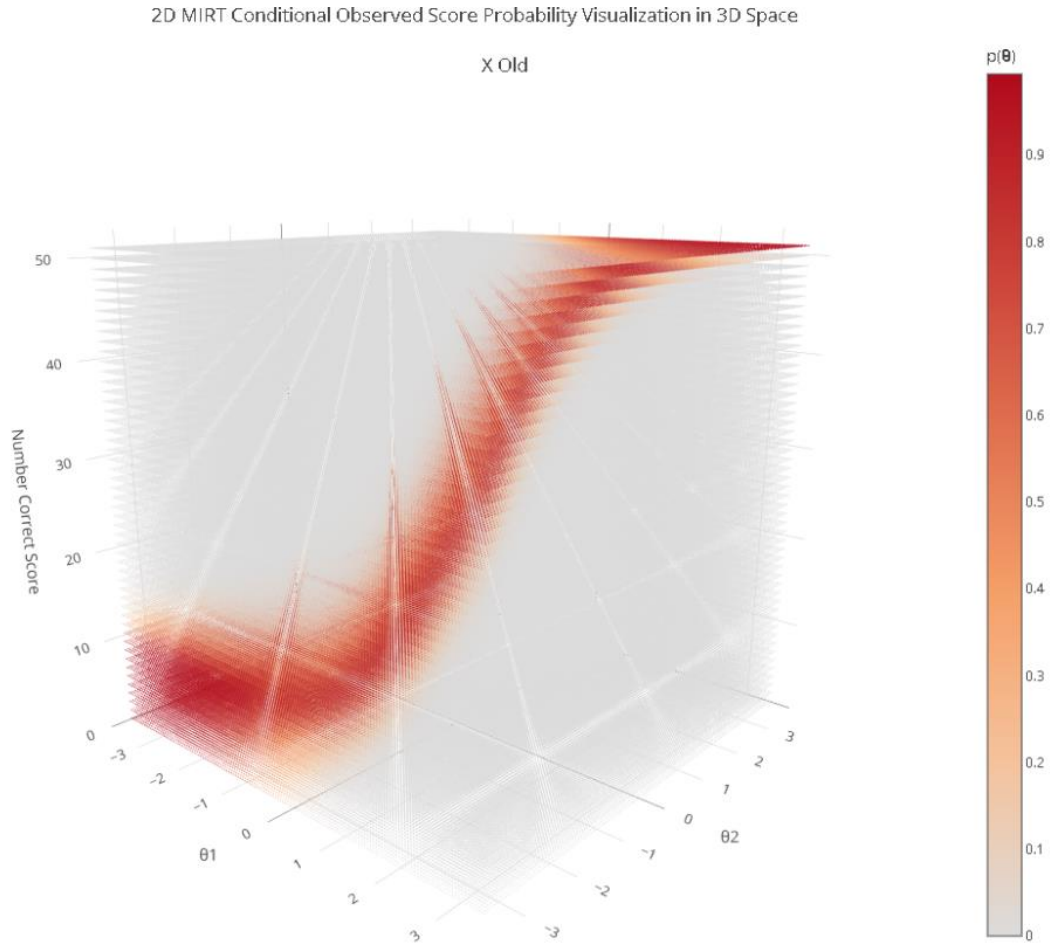


Figure 29. Test Characteristic Surface of X Base in Red and Y Base in Blue

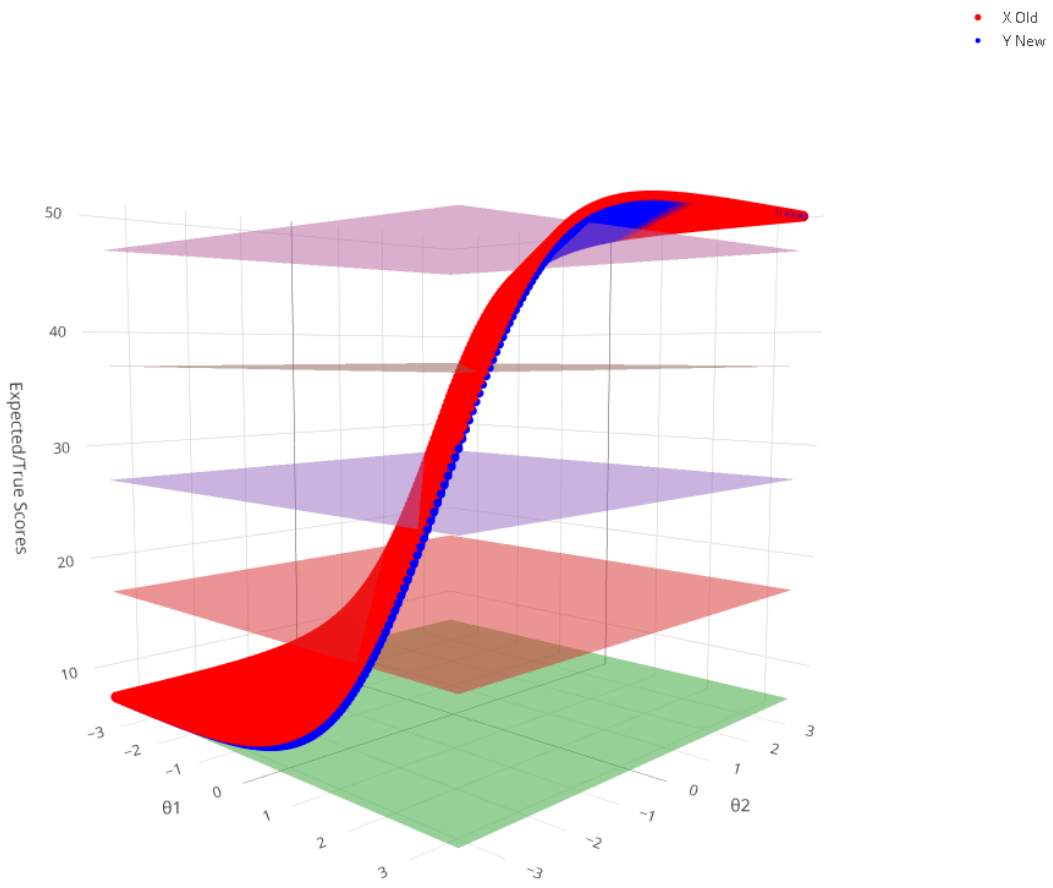


Figure 30. Test Characteristic Surface of X Base in Red and Y Same RC in Blue

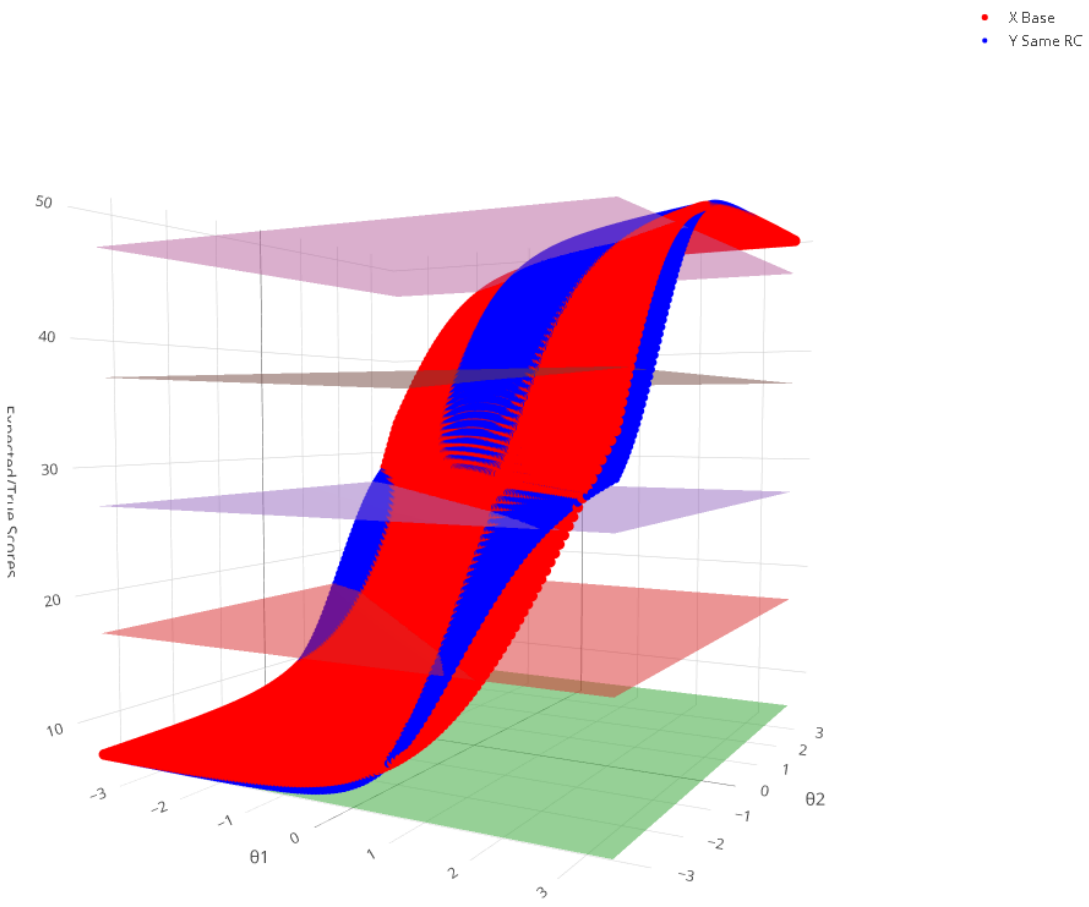
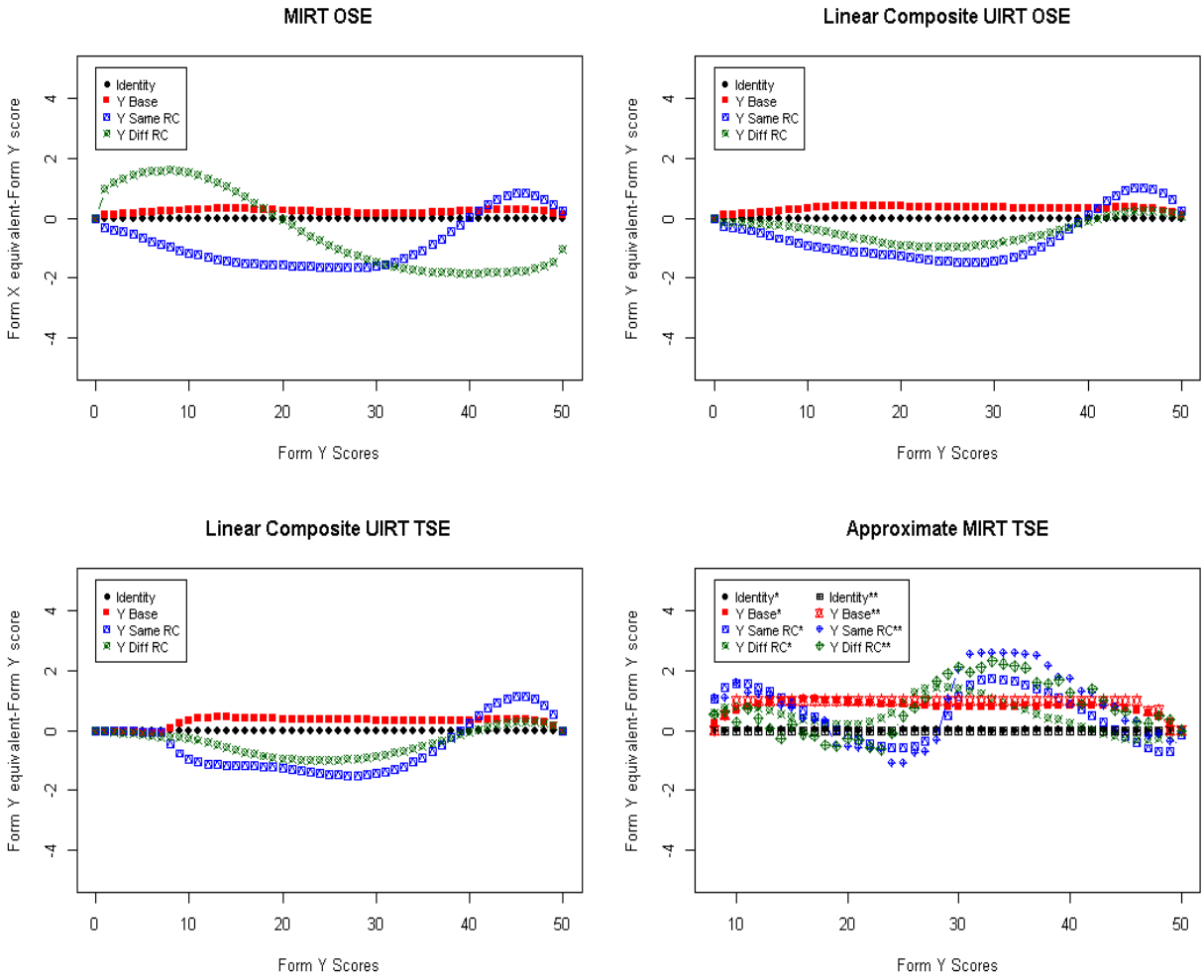
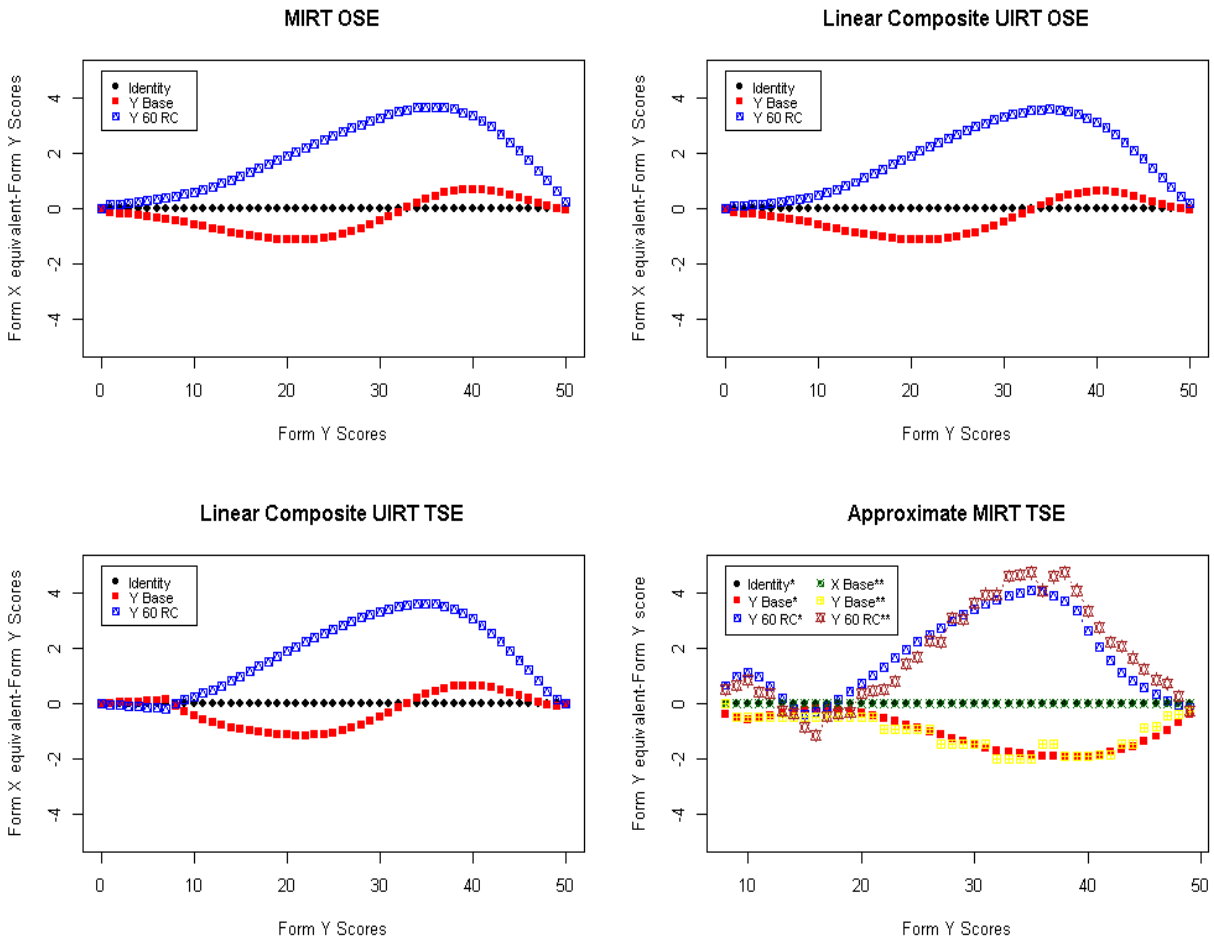


Figure 31. Comparison of Equating Score Difference in MC-I



Equated scores were truncated at 8 which is larger than the sum of guessings in AMT equating.

Figure 32. Comparison of Equating Score Difference in MC-II



Equated scores were truncated at 8 which is larger than the sum of guessings in AMT equating.

References

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. Journal of educational measurement, 29(1), 67-91.

Ackerman, T. A., Gierl, M. J., & Walker, C. M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. Educational Measurement: Issues and Practice, 22(3), 37-51.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing. (2014). Standards for educational and psychological testing. American Educational Research Association.

Babcock, B., & Albano, A. D. (2012). Rasch scale stability in the presence of item parameter and trait drift. Applied Psychological Measurement, 36(7), 565-580.

Baker, F. B., & Kim, S. H. (Eds.). (2004). Item response theory: Parameter estimation techniques. CRC Press.

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

Bastari, B. (2000). Linking multiple-choice and constructed-response items to a common proficiency scale. Unpublished doctoral dissertation, University of Massachusetts.

Béguin, A. A., Hanson, B. A., & Glas, C. A. (2000, April). Effect of multidimensionality on separate and concurrent estimation in IRT equating. In Proceedings of the Annual Meeting of the National Council on Measurement in Education (Vol. 2000, pp. 25-27).

Bock, R. D., & Zimowski, M. F. (1997). Multiple group IRT. In Handbook of modern item response theory (pp. 433-448). Springer New York.

Bolt, D. M. (1999). Evaluating the effects of multidimensionality on IRT true-score equating. Applied Measurement in education, 12(4), 383-407.

Bourque, M. L., Goodman, D., Hambleton, R. K., & Han, N. (2004). Reliability estimates for the ABTE tests in elementary education, professional teaching knowledge, secondary mathematics and English/language arts (Final Report). Leesburg, VA: Mid-Atlantic Psychometric Services.

Brossman, B. G., & Lee, W. C. (2013). Observed score and true score equating procedures for multidimensional item response theory. Applied Psychological Measurement, 37(6), 460-481.

Budescu, D. (1985). Efficiency of linear equating as a function of the length of the anchor test. Journal of Educational Measurement, 22(1), 13-20.

Cai, L. (2015). Lord–Wingersky algorithm version 2.0 for hierarchical item factor models with applications in test scoring, scale alignment, and model fit testing. Psychometrika, 80(2), 535-559.

Cai, L. (2017). flexMIRT R version 3.51: Flexible multilevel multidimensional item analysis and test scoring [Computer software]. Chapel Hill, NC: Vector Psychometric Group.

Camilli, G., Wang, M. M., & Fesq, J. (1995). The effects of dimensionality on equating the Law School Admission Test. Journal of Educational Measurement, 32(1), 79-96.

Carlson, J. E. (2017). Unidimensional vertical scaling in multidimensional space (ETS RR17-29). Princeton, NJ: Educational Testing Service.

Chalmers, R. P., Counsell, A., & Flora, D. B. (2016). It might not make a big DIF: Improved differential test functioning statistics that account for sampling variability. Educational and Psychological Measurement, 76(1), 114-140.

Cho, Y. W., Wall, N. L., Lee, W., & Harris, D. J. (2010, April). The effects of common item selection on equipercentile equating for mixed format tests. Paper presented at the meeting of the National Council on Measurement in Education, Denver, CO.

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

Deng, N. (2011). Evaluating IRT-and CTT-based methods of estimating classification consistency and accuracy indices from single administrations. University of Massachusetts Amherst.

Diao, H., & Sireci, S. G. (2018). Item response theory-based methods for estimating classification accuracy and consistency. Journal of Applied Testing Technology, 19(1), 20-25.

Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equatability of tests: Basic theory and the linear case. Journal of educational measurement, 37(4), 281-306.

Dorans, N. J., & Kingston, N. M. (1985). The effects of violations of unidimensionality on the estimation of item and ability parameters and on item response theory equating of the GRE verbal scale. Journal of Educational Measurement, 22(4), 249-262.

Dorans, N. J., & Lawrence, I. M. (1987). THE INTERNAL CONSTRUCT VALIDITY OF THE SAT 1. ETS Research Report Series, 1987(2), i-101.

Dorans, N. J., & Lawrence, I. M. (1999). The role of the unit of analysis in dimensionality assessment. ETS Research Report Series, 1999(2), i-39.

Fitzpatrick, J., & Skorupski, W. P. (2016). Equating with miditests using IRT. Journal of Educational Measurement, 53, 172–189.

Guo, F. (2006). Expected classification accuracy using the latent distribution. Practical Assessment, Research, and Evaluation, 11(1), 6.

Hattie, J. (1984). An empirical study of various indices for determining unidimensionality. Multivariate behavioral research, 19(1), 49-78.

Hattie, J. (1985). Methodology review: assessing unidimensionality of tests and ltenls. Applied psychological measurement, 9(2), 139-164.

Hanson, B. A., & Béguin, 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.

Henning, G. (1992). Dimensionality and construct validity of language tests. Language Testing, 9(1), 1-11.

Ip, E. H. S., & Chen, S. H. (2012). Projective item response model for test-independent measurement. Applied Psychological Measurement, 36(7), 581-601.

Ip, Edward H., Tyler Strachan, Yanyan Fu, Alexandra Lay, John T. Willse, Shyh-Huei Chen, Leslie Rutkowski, and Terry Ackerman. "Bias and bias correction method for nonproportional abilities requirement (NPAR) tests." Journal of Educational Measurement 56, no. 1 (2019): 147-168.

Jonathan P. Weeks (2010). plink: An R Package for Linking Mixed-Format Tests Using IRT-Based Methods. Journal of Statistical Software, 35(12), 1-33. URL http://www.jstatsoft.org/v35/i12/.

Kang, T., & Petersen, N. S. (2012). Linking item parameters to a base scale. Asia Pacific Education Review, 13(2), 311-321.

Kane, M. (2006). Validation. In R. Brennan (Ed.), Educational measurement (4th ed., pp. 17–64), Westport, CT: American Council on Education & Praeger.

Keller, L. A., & Keller, R. R. (2011). The long-term sustainability of different item response theory scaling methods. Educational and Psychological Measurement, 71(2), 362-379.

Kim, K. Y. (2017). IRT linking methods for the bifactor model: a special case of the two-tier item factor analysis model. The University of Iowa.

Kim, S. (2006). A Comparative Study of IRT fixed parameter calibration methods. Journal of Educational Measurement, 43, 355-381.

Kim, S. H., & Cohen, A. S. (1992). Effects of linking methods on detection of DIF. Journal of educational measurement, 29(1), 51-66.

Kim, S. H., & 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. (2007). Effects on scale linking of different definitions of criterion functions for the IRT characteristic curve methods. Journal of Educational and Behavioral Statistics, 32(4), 371-397.

Kim, S., & Kolen, M. J. (2016). Multiple group IRT fixed-parameter estimation for maintaining an established ability scale. Center for Advanced Studies in Measurement and Assessment Report, 49.

Kim, S., & Lee, W. C. (2004). IRT scale linking methods for mixed-format tests. ACT, Incorporated.

Kim, S. Y., Lee, W. C., & Kolen, M. J. (2020). Simple-structure multidimensional item response theory equating for multidimensional tests. Educational and psychological measurement, 80(1), 91-125.

Klein, L. W., & Jarjoura, D. (1985). The importance of content representation for common-item equating with non-random groups. Journal of educational Meusurement, 22,197-206.

Kolen, M. J., Brennan, R. L., & Kolen, M. J. (2004). Test equating, scaling, and linking: Methods and practices (pp. 177-180). New York: Springer.

LaFond, Lee James. Decision consistency and accuracy indices for the bifactor and testlet response theory models. The University of Iowa, 2014.

Lathrop, Q. N., & Cheng, Y. (2014). A nonparametric approach to estimate classification accuracy and consistency. Journal of Educational Measurement, 51(3), 318-334.

Lawrence, I. M., Curley, W. E., & McHale, F. J. (1988). Differential item functioning for males and females on sat®-verbal reading subscore items. ETS Research Report Series, 1988(1), i-55.

Lee, W. C. (2010). Classification consistency and accuracy for complex assessments using item response theory. Journal of Educational Measurement, 47(1), 1-17.

Lee, W. C., & Ban, J. C. (2009). A comparison of IRT linking procedures. Applied Measurement in Education, 23(1), 23-48.

Lee, W., Hanson, B. A., & Brennan, R. L. (2002). Estimating consistency and accuracy indices for multiple classifications. Applied Psychological Measurement, 26, 412–432.

Li, Y. H., & Lissitz, R. W. (2000). An evaluation of the accuracy of multidimensional IRT linking. Applied Psychological Measurement, 24(2), 115-138.

Linacre, J.M. (2022). Winsteps® (Version 5.2.3) [Computer Software]. Portland, Oregon: Winsteps.com. Available from https://www.winsteps.com/

Lineberry, M., Park, Y. S., Hennessy, S. A., & Ritter, E. M. (2020). The Fundamentals of Endoscopic Surgery (FES) skills test: factors associated with first-attempt scores and pass rate. Surgical endoscopy, 34(8), 3633-3643.

Liu, J., Harris, D. J., & Schmidt, A. (2007). Statistical procedures used in college admissions testing. In C. R. Rao & S. Sinharay (Eds.), Handbook of statistics, Vol. 26: Psychometrics (pp. 1057–1091). Amsterdam, The Netherlands: Elsevier.

Liu, Y., Magnus, B., O'Connor, H., & Thissen, D. (2018). Multidimensional item response theory.

Liu, J., Sinharay, S., Holland, P., Curley, E., & Feigenbaum, M. (2011a). Test score equating using a mini-version anchor and a midi anchor: A case study using SAT data. Journal of Educational Measurement, 48, 361–379.

Liu, J., Sinharay, S., Holland, P. W., Feigenbaum, M., & Curley, E. (2011b). Observed score equating using a mini-version anchor and an anchor with less spread of difficulty: A comparison study. Educational and Psychological Measurement, 71, 346–361.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score" equatings". Applied Psychological Measurement, 8(4), 453-461.

Luecht, R. M. (1996). Multidimensional computerized adaptive testing in a certification or licensure context. Applied psychological measurement, 20(4), 389-404.

Luecht, R. M., & Miller, T. R. (1992). Unidimensional calibrations and interpretations of composite traits for multidimensional tests. Applied Psychological Measurement, 16(3), 279-293.

Lumsden, J. (1957). A factorial approach to unidimensionality. Australian Journal of Psychology, 9(2), 105-111.

Manhart, J. J. (1996). Factor Analytic Methods for Determining Whether Multiple-Choice and Constructed-Response Tests Measure the Same Construct.

Martineau, J. A. (2006). Distorting value added: The use of longitudinal, vertically scaled student achievement data for growth-based, value-added accountability. Journal of Educational and Behavioral Statistics, 31(1), 35–62

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

McDonald, R. P. (1981). The dimensionality of tests and items. British Journal of mathematical and statistical Psychology, 34(1), 100-117.

McDonald, R. P., & Ahlawat, K. S. (1974). Difficulty factors in binary data. British Journal of mathematical and statistical Psychology, 27(1), 82-99.

Messick, S. (1989). Validity. In R.L.Linn(Ed.), Educational measurement (3rd ed., pp. 13-103). New "York: Macmillan. M

Mislevy, R. J. (1986). Recent developments in the factor analysis of categorical variables. Journal of educational statistics, 11(1), 3-31.

Morris, C. N. (1982). On the foundations of test equating. Test equating, 169-191.

Park, S., Kim, K. Y., & Lee, W. C. (2022). Estimating Classification Accuracy and Consistency Indices for Multiple Measures with the Simple Structure MIRT Model. Journal of Educational Measurement.

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(2), 137-156.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), Educational measurement (pp. 221–262). Macmillan Publishing Co, Inc; American Council on Education.

Petersen, N. S., Marco, G. L., & Stewart, E. E. (1982). A test of the adequacy of linear score equating models. Test equating, 71-135.

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Reckase, M. D., Ackerman, T. A., & Carlson, J. E. (1988). Building a unidimensional test using multidimensional items. Journal of Educational Measurement, 25(3), 193-203.

Reckase, M. D. (2009). Multidimensional item response theory models. In Multidimensional item response theory. Springer, New York, NY.

Rijmen, F. (2009). Three multidimensional models for testlet-based tests: Formal relations and an empirical comparison. ETS Research Report Series, 2009(2), i-13.

Roussos, L., & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. Applied psychological measurement, 20(4), 355-371.

Rudner, L. M. (2000). Computing the expected proportions of misclassified examinees. Practical Assessment, Research, and Evaluation, 7(1), 14.

Rudner, L. M. (2005). Expected classification accuracy. Practical Assessment, Research, and Evaluation, 10(1), 13.

Rupp, A. A., & Zumbo, B. D. (2006). Understanding parameter invariance in unidimensional irt models. Educational and Psychological Measurement, 66(1), 63–84.

Sawatdirakpong, S. (1993). Native languages and differential dimensionality of English as a second language test proficiency: An exploratory study (Doctoral dissertation, University of Illinois at Urbana-Champaign).

Sinharay, S. (2018). On the choice of anchor tests in equating. Educational Measurement: Issues and Practice, 37(2), 64-69.

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

Skaggs, G., & Lissitz, R. W. (1986). IRT test equating: Relevant issues and a review of recent research. Review of Educational Research, 56, 495-529.

Spence, P. D. (1996). The effect of multidimensionality on unidimensional equating with item response theory. University of Florida.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. Psychometrika, 52(4), 589-617.

Strachan, T., Cho, U. H., Ackerman, T., Chen, S. H., de la Torre, J., & Ip, E. H. (2022). Evaluation of the Linear Composite Conjecture for Unidimensional IRT Scale for Multidimensional Responses. Applied Psychological Measurement, 46(5), 347-360.

Stucky, B. D. (2011). Logistic approximations of marginal trace lines for bifactor item response theory models (Doctoral dissertation, The University of North Carolina at Chapel Hill).

Svetina, D., & Levy, R. (2014). A framework for dimensionality assessment for multidimensional item response models. Educational Assessment, 19(1), 35-57.

Sykes, R. C., Hou, L., Hanson, B., & Wang, Z. (2002). Multidimensionality and the Equating of a Mixed-Format Math Examination.

Tate, R. (2003). A comparison of selected empirical methods for assessing the structure of responses to test items. Applied Psychological Measurement, 27(3), 159-203.

Tong, Y., & Kolen, M. J. (2005). Assessing equating results on different equating criteria. Applied Psychological Measurement, 29(6), 418-432.

Thomasson, G. (1993). The asymptotic equating methodology and other test equating evaluation procedures. Unpublished doctoral dissertation, University of Illinois, Urbana–Champaign

Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 law school admissions test as an example. Applied Measurement in Education, 8 (2), 157–186.

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

Wang, M.M. (1987). Fitting a unidimensional model to multidimensional item response data (ONR Rep. 042286). Iowa City, IA: University of Iowa.

Wang, W., Song, L., Ding, S., & Meng, Y. (2016). Estimating classification accuracy and consistency indices for multidimensional latent ability. In Quantitative Psychology Research (pp. 89-103). Springer, Cham.

Weeks, J. P. (2022). plink: IRT separate calibration linking methods. R package version 1.5-1

Yao, L. (2011). Multidimensional linking for domain scores and overall scores for nonequivalent groups. Applied Psychological Measurement, 35(1), 48-66.

Yao, L. (2013). Classification accuracy and consistency indices for summed scores enhanced using MIRT for test of mixed item types. Retrieved March, 1, 2015.

Yao, L., & Boughton, K. A. (2007). A multidimensional item response modeling approach for improving subscale proficiency estimation and classification. Applied Psychological Measurement, 31(2), 83-105.

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

Zhang, J. (1996). Some fundamental issues in item response theory with applications. University of Illinois at Urbana-Champaign.

Zhang, J., & Wang, M. M. (1998). Relating reported scores to latent traits in a multidimensional test. Paper presented at the annual meeting of the American Educational Research Association, April, San Diego, CA.

Zhang, J., & Stout, W. (1999a). Conditional covariance structure of generalized compensatory multidimensional items. Psychometrika, 64(2), 129-152.

Zhang, J., & Stout, W. F. (1999b). The theoretical DETECT index of dimensionality and its application to approximate simple structure. Psychometrika, 64, 213–249.

APPENDIX A: GENERATING ITEM PARAMETERS

MC-I
<X BASE>

Item ID	a1	a2	d	g	alpha	MDISC	MID	RC_Angle
1	0.6	0.6	-0.1	0.15	45	0.86	0.12	45
2	1.01	1.01	-0.13	0.11	45	1.43	0.09	45
3	0.57	0.57	0.22	0.18	45	0.81	-0.27	45
4	0.95	0.95	1.24	0.15	45	1.35	-0.92	45
5	0.83	0.83	1.17	0.15	45	1.17	-1	45
6	1.06	1.06	0.01	0.12	45	1.5	0	45
7	0.72	0.72	0.95	0.12	45	1.02	-0.94	45
8	0.77	0.77	1.4	0.16	45	1.09	-1.29	45
9	0.94	0.94	0.11	0.16	45	1.33	-0.09	45
10	0.8	0.8	0.8	0.09	45	1.14	-0.7	45
11	1.04	1.04	0.03	0.07	45	1.46	-0.02	45
12	0.67	0.67	-0.15	0.17	45	0.95	0.16	45
13	0.64	0.64	0.9	0.19	45	0.9	-1	45
14	0.66	0.66	-0.45	0.16	45	0.94	0.48	45
15	0.73	0.73	0.79	0.16	45	1.04	-0.76	45
16	1.02	1.02	-0.65	0.16	45	1.45	0.45	45
17	0.85	0.85	-0.35	0.2	45	1.2	0.29	45
18	1.11	1.11	-0.6	0.15	45	1.56	0.38	45
19	0.7	0.7	-0.36	0.17	45	0.99	0.37	45
20	0.98	0.98	-0.81	0.16	45	1.38	0.58	45
21	1.12	1.12	0.62	0.12	45	1.58	-0.39	45
22	0.79	0.79	1.2	0.12	45	1.12	-1.07	45
23	0.59	0.59	0.96	0.18	45	0.84	-1.14	45
24	0.63	0.63	1	0.15	45	0.89	-1.13	45
25	1.13	1.13	0.17	0.11	45	1.6	-0.1	45
26	0.86	0.86	0.05	0.18	45	1.22	-0.04	45
27	0.87	0.87	1.04	0.1	45	1.23	-0.84	45
28	0.92	0.92	0.53	0.19	45	1.3	-0.41	45
29	0.9	0.9	0.65	0.16	45	1.27	-0.52	45
30	0.81	0.81	-0.08	0.16	45	1.15	0.07	45
31	0.93	0.93	-0.29	0.13	45	1.32	0.22	45
32	0.84	0.84	1.48	0.15	45	1.18	-1.25	45
33	0.65	0.65	0.6	0.15	45	0.92	-0.65	45
34	0.97	0.97	-0.43	0.11	45	1.37	0.31	45

35	0.99	0.99	-0.63	0.16	45	1.4	0.45	45
36	1	1	0.22	0.09	45	1.41	-0.16	45
37	1.08	1.08	1.31	0.13	45	1.53	-0.86	45
38	0.78	0.78	0.52	0.12	45	1.1	-0.47	45
39	1.05	1.05	1.44	0.13	45	1.48	-0.97	45
40	1.09	1.09	0.84	0.19	45	1.55	-0.55	45
41	0.88	0.88	-0.08	0.15	45	1.25	0.06	45
42	0.71	0.71	0.11	0.18	45	1	-0.11	45
43	1.07	1.07	-0.59	0.19	45	1.51	0.39	45
44	0.74	0.74	-0.92	0.14	45	1.05	0.87	45
45	0.58	0.58	0.8	0.08	45	0.82	-0.98	45
46	0.62	0.62	-0.7	0.13	45	0.87	0.8	45
47	0.69	0.69	0.14	0.17	45	0.97	-0.15	45
48	1.14	1.14	0.62	0.13	45	1.61	-0.38	45
49	0.76	0.76	1.48	0.17	45	1.07	-1.38	45
50	0.91	0.91	0.26	0.19	45	1.28	-0.21	45

Alpha= item measurement direction, MDISC=item discrimination in MIRT, MID=item difficulty in MIRT, CVI = validity sector index, RC_Angle = test measurement direction, Common Items in bold

<Y BASE>

Item ID	a1	a2	d	g	alpha	MDISC	MID	RC_Angle
1	0.88	0.88	-0.15	0.08	45	1.25	0.12	45
2	0.67	0.67	-0.68	0.2	45	0.95	0.71	45
3	0.57	0.57	0.22	0.18	45	0.81	-0.27	45
4	0.95	0.95	1.24	0.15	45	1.35	-0.92	45
5	0.64	0.64	0.41	0.15	45	0.9	-0.45	45
6	0.71	0.71	-0.06	0.16	45	1	0.06	45
7	1.06	1.06	0.05	0.19	45	1.5	-0.03	45
8	0.85	0.85	-0.44	0.09	45	1.2	0.37	45
9	0.79	0.79	-0.85	0.18	45	1.12	0.76	45
10	0.76	0.76	0.39	0.16	45	1.07	-0.36	45
11	0.66	0.66	1.37	0.11	45	0.94	-1.46	45
12	1.07	1.07	0.18	0.06	45	1.51	-0.12	45
13	0.8	0.8	0.1	0.17	45	1.14	-0.09	45
14	0.72	0.72	0.43	0.13	45	1.02	-0.42	45
15	0.93	0.93	0.25	0.14	45	1.32	-0.19	45
16	0.87	0.87	-0.76	0.07	45	1.23	0.62	45
17	0.85	0.85	-0.35	0.2	45	1.2	0.29	45
18	0.58	0.58	0.35	0.19	45	0.82	-0.43	45

19	0.91	0.91	-0.01	0.15	45	1.28	0.01	45
20	0.9	0.9	1.46	0.1	45	1.27	-1.16	45
21	1.12	1.12	0.62	0.12	45	1.58	-0.39	45
22	1.13	1.13	-0.6	0.16	45	1.6	0.38	45
23	0.59	0.59	0.96	0.18	45	0.84	-1.14	45
24	0.74	0.74	0.06	0.1	45	1.05	-0.06	45
25	1.02	1.02	1.35	0.13	45	1.45	-0.93	45
26	0.86	0.86	0.05	0.18	45	1.22	-0.04	45
27	0.73	0.73	0.47	0.15	45	1.04	-0.45	45
28	1.08	1.08	0.56	0.19	45	1.53	-0.37	45
29	0.97	0.97	1.1	0.11	45	1.37	-0.8	45
30	0.62	0.62	0.69	0.15	45	0.87	-0.79	45
31	1	1	1.35	0.14	45	1.41	-0.96	45
32	0.98	0.98	1.11	0.17	45	1.38	-0.8	45
33	0.65	0.65	0.6	0.15	45	0.92	-0.65	45
34	1.11	1.11	-0.18	0.17	45	1.56	0.12	45
35	0.99	0.99	-0.63	0.16	45	1.4	0.45	45
36	0.59	0.59	-0.88	0.14	45	0.84	1.05	45
37	1.05	1.05	0.22	0.1	45	1.48	-0.15	45
38	0.78	0.78	0.52	0.12	45	1.1	-0.47	45
39	1.14	1.14	1.26	0.14	45	1.61	-0.78	45
40	0.92	0.92	0.91	0.2	45	1.3	-0.7	45
41	0.88	0.88	-0.08	0.15	45	1.25	0.06	45
42	0.83	0.83	-0.34	0.16	45	1.17	0.29	45
43	0.57	0.57	-0.82	0.18	45	0.81	1.02	45
44	1.09	1.09	1.41	0.18	45	1.55	-0.91	45
45	0.77	0.77	0.1	0.18	45	1.09	-0.09	45
46	0.7	0.7	0.72	0.19	45	0.99	-0.73	45
47	0.6	0.6	1.47	0.09	45	0.86	-1.72	45
48	0.95	0.95	-0.78	0.13	45	1.35	0.58	45
49	1.04	1.04	1.21	0.15	45	1.46	-0.83	45
50	0.78	0.78	0.15	0.14	45	1.1	-0.14	45

Alpha= item measurement direction, MDISC=item discrimination in MIRT, MID=item difficulty in MIRT, RC_Angle = test measurement direction, Common Items in bold

<Y SAME RC>

Item ID	a1	a2	d	g	alpha	MDISC	MID	RC_Angle
1	0.57	0.99	0.84	0.17	60	1.14	-0.74	45
2	0.59	1.03	1.44	0.18	60	1.19	-1.21	45
3	0.57	0.57	0.22	0.18	45	0.81	-0.27	45

4	0.95	0.95	1.24	0.15	45	1.35	-0.92	45
5	0.67	1.15	0.15	0.18	60	1.33	-0.12	45
6	0.69	1.19	1.13	0.1	60	1.38	-0.82	45
7	0.71	1.23	0.45	0.12	60	1.43	-0.31	45
8	0.74	1.28	1.41	0.19	60	1.47	-0.95	45
9	0.76	1.32	1.39	0.15	60	1.52	-0.91	45
10	0.78	1.36	-0.25	0.19	60	1.57	0.16	45
11	0.81	1.4	0.86	0.19	60	1.62	-0.53	45
12	0.83	1.44	-0.78	0.18	60	1.66	0.47	45
13	0.86	1.48	1.26	0.16	60	1.71	-0.73	45
14	0.88	1.52	-0.62	0.16	60	1.76	0.35	45
15	0.9	1.56	0.41	0.18	60	1.81	-0.23	45
16	0.93	1.6	0.52	0.15	60	1.85	-0.28	45
17	0.85	0.85	-0.35	0.2	45	1.2	0.29	45
18	0.97	1.69	0.3	0.16	60	1.95	-0.15	45
19	1	1.73	-0.05	0.11	60	2	0.03	45
20	1.02	1.77	-0.94	0.18	60	2.04	0.46	45
21	1.12	1.12	0.62	0.12	45	1.58	-0.39	45
22	1.07	1.85	1.21	0.1	60	2.14	-0.57	45
23	0.59	0.59	0.96	0.18	45	0.84	-1.14	45
24	1.12	1.93	0.06	0.1	60	2.23	-0.03	45
25	1.14	1.97	1.31	0.04	60	2.28	-0.57	45
26	0.86	0.86	0.05	0.18	45	1.22	-0.04	45
27	1.03	0.59	0.21	0.19	30	1.19	-0.18	45
28	1.07	0.62	0.66	0.09	30	1.24	-0.54	45
29	1.11	0.64	-0.55	0.1	30	1.28	0.43	45
30	1.15	0.67	-0.72	0.08	30	1.33	0.54	45
31	1.19	0.69	1.16	0.07	30	1.38	-0.84	45
32	1.23	0.71	-0.36	0.19	30	1.43	0.25	45
33	0.65	0.65	0.6	0.15	45	0.92	-0.65	45
34	1.32	0.76	1.13	0.11	30	1.52	-0.74	45
35	0.99	0.99	-0.63	0.16	45	1.4	0.45	45
36	1.4	0.81	1.44	0.13	30	1.62	-0.89	45
37	1.44	0.83	1.18	0.13	30	1.66	-0.71	45
38	0.78	0.78	0.52	0.12	45	1.1	-0.47	45
39	1.52	0.88	-0.4	0.14	30	1.76	0.23	45
40	1.56	0.9	0.94	0.17	30	1.81	-0.52	45
41	0.88	0.88	-0.08	0.15	45	1.25	0.06	45
42	1.65	0.95	-0.84	0.17	30	1.9	0.44	45
43	1.69	0.97	-0.46	0.19	30	1.95	0.24	45
44	1.73	1	1.14	0.19	30	2	-0.57	45

45	1.77	1.02	0.65	0.18	30	2.04	-0.32	45
46	1.81	1.05	1.18	0.13	30	2.09	-0.56	45
47	1.85	1.07	1.04	0.07	30	2.14	-0.49	45
48	1.89	1.09	-0.06	0.19	30	2.19	0.03	45
49	1.93	1.12	-0.73	0.17	30	2.23	0.33	45
50	1.97	1.14	0.46	0.16	30	2.28	-0.2	45

Alpha= item measurement direction, MDISC=item discrimination in MIRT, MID=item difficulty in MIRT, CVI = validity sector index, RC_Angle = test measurement direction, Common Items in bold

<Y DIFF RC>

Item ID	a1	a2	d	g	alpha	MDISC	MID	RC_Angle
1	0.29	0.78	1.32	0.17	70	0.83	-1.59	51.68
2	0.3	0.82	-0.39	0.1	70	0.87	0.44	51.68
3	0.57	0.57	0.22	0.18	45	0.81	-0.27	45
4	0.95	0.95	1.24	0.15	45	1.35	-0.92	45
5	0.33	0.91	-0.13	0.09	70	0.97	0.14	51.68
6	0.34	0.95	1.17	0.19	70	1.01	-1.16	51.68
7	0.36	0.98	-0.52	0.19	70	1.04	0.5	51.68
8	0.37	1.01	0.26	0.19	70	1.08	-0.24	51.68
9	0.38	1.04	0.1	0.19	70	1.11	-0.09	51.68
10	0.39	1.08	0.43	0.09	70	1.15	-0.38	51.68
11	0.4	1.11	0.66	0.14	70	1.18	-0.56	51.68
12	0.42	1.14	1.45	0.16	70	1.22	-1.19	51.68
13	0.43	1.17	-0.38	0.1	70	1.25	0.3	51.68
14	0.44	1.21	-0.36	0.18	70	1.28	0.28	51.68
15	0.45	1.24	1	0.18	70	1.32	-0.76	51.68
16	0.46	1.27	1.09	0.08	70	1.35	-0.8	51.68
17	0.85	0.85	-0.35	0.2	45	1.2	0.29	45
18	0.49	1.34	1.41	0.19	70	1.42	-0.99	51.68
19	0.5	1.37	-0.59	0.07	70	1.46	0.4	51.68
20	0.51	1.4	-0.6	0.13	70	1.49	0.4	51.68
21	1.12	1.12	0.62	0.12	45	1.58	-0.39	45
22	0.53	1.47	0.29	0.16	70	1.56	-0.19	51.68
23	0.59	0.59	0.96	0.18	45	0.84	-1.14	45
24	0.56	1.53	-0.1	0.14	70	1.63	0.06	51.68
25	0.57	1.57	0.67	0.14	70	1.67	-0.4	51.68
26	0.86	0.86	0.05	0.18	45	1.22	-0.04	45
27	0.82	0.68	-0.09	0.1	40	1.06	0.08	51.68

28	0.85	0.71	-0.59	0.09	40	1.11	0.53	51.68
29	0.88	0.74	0.66	0.2	40	1.15	-0.58	51.68
30	0.91	0.77	-0.5	0.15	40	1.19	0.42	51.68
31	0.95	0.79	1.39	0.14	40	1.24	-1.12	51.68
32	0.98	0.82	1.25	0.11	40	1.28	-0.98	51.68
33	0.65	0.65	0.6	0.15	45	0.92	-0.65	45
34	1.04	0.88	0.82	0.17	40	1.36	-0.6	51.68
35	0.99	0.99	-0.63	0.16	45	1.4	0.45	45
36	1.11	0.93	0.96	0.2	40	1.45	-0.67	51.68
37	1.14	0.96	-0.92	0.16	40	1.49	0.62	51.68
38	0.78	0.78	0.52	0.12	45	1.1	-0.47	45
39	1.21	1.01	1.48	0.11	40	1.58	-0.94	51.68
40	1.24	1.04	-0.07	0.15	40	1.62	0.05	51.68
41	0.88	0.88	-0.08	0.15	45	1.25	0.06	45
42	1.31	1.1	0.99	0.15	40	1.7	-0.58	51.68
43	1.34	1.12	0.78	0.12	40	1.75	-0.45	51.68
44	1.37	1.15	0.22	0.19	40	1.79	-0.12	51.68
45	1.4	1.18	0.26	0.17	40	1.83	-0.14	51.68
46	1.44	1.2	-0.2	0.11	40	1.87	0.1	51.68
47	1.47	1.23	0.75	0.15	40	1.92	-0.39	51.68
48	1.5	1.26	1.07	0.17	40	1.96	-0.54	51.68
49	1.53	1.29	0.12	0.15	40	2	-0.06	51.68
50	1.57	1.31	0.31	0.2	40	2.04	-0.15	51.68

Alpha= item measurement direction, MDISC=item discrimination in MIRT, MID=item difficulty in MIRT, CVI = validity sector index, RC_Angle = test measurement direction, Common Items in bold

MC-II
<X BASE>

Item	ID	a1	a2	d	g	alpha	MDISC	MID	CVI	RC_Angle
	1	1.05	0.38	-0.03	0.15	19.95	1.12	0.02	0.95	6.95
	2	1.15	0.41	1	0.15	19.52	1.22	-0.82	0.95	6.95
	3	1.43	0.5	0.78	0.17	19.09	1.51	-0.51	0.96	6.95
	4	1.19	0.4	1.21	0.09	18.65	1.26	-0.96	0.96	6.95
	5	0.75	0.25	-0.66	0.17	18.19	0.79	0.84	0.96	6.95
	6	0.91	0.29	-0.16	0.18	17.73	0.96	0.17	0.97	6.95
	7	0.62	0.19	-0.68	0.12	17.25	0.65	1.05	0.97	6.95
	8	0.63	0.19	-0.51	0.07	16.76	0.66	0.77	0.97	6.95
	9	1.14	0.33	1.09	0.14	16.26	1.19	-0.92	0.98	6.95
	10	0.85	0.24	-1.17	0.13	15.74	0.88	1.33	0.98	6.95

11	0.92	0.25	-0.29	0.15	15.2	0.96	0.31	0.98	6.95
12	1.44	0.38	-0.28	0.14	14.65	1.49	0.19	0.98	6.95
13	0.88	0.22	-0.19	0.19	14.07	0.91	0.21	0.99	6.95
14	0.73	0.17	-0.97	0.11	13.47	0.75	1.29	0.99	6.95
15	1.19	0.27	-0.7	0.1	12.84	1.22	0.58	0.99	6.95
16	0.51	0.11	0.48	0.16	12.18	0.52	-0.92	0.99	6.95
17	1.19	0.24	0.95	0.06	11.48	1.21	-0.78	0.99	6.95
18	1.29	0.24	-0.24	0.08	10.73	1.31	0.18	1	6.95
19	1.01	0.18	-1.02	0.18	9.94	1.02	1	1	6.95
20	0.81	0.13	-0.63	0.19	9.07	0.82	0.77	1	6.95
21	1.24	0.18	0.91	0.12	8.11	1.25	-0.73	1	6.95
22	0.8	0.1	-1.01	0.17	7.02	0.81	1.24	1	6.95
23	1.27	0.13	0.87	0.13	5.73	1.28	-0.68	1	6.95
24	0.61	0.04	0.07	0.2	4.05	0.61	-0.12	1	6.95
25	1.3	0	-0.54	0.18	0	1.3	0.42	0.98	6.95
26	1	0	-0.67	0.18	0	1	0.67	0.98	6.95
27	1	0	0.91	0.12	0	1	-0.91	0.98	6.95
28	1	0	-1.19	0.08	0	1	1.19	0.98	6.95
29	1	0	1.01	0.14	0	1	-1.01	0.98	6.95
30	1	0	-1.26	0.18	0	1	1.26	0.98	6.95
31	1	0	0.65	0.19	0	1	-0.65	0.98	6.95
32	1	0	1.09	0.17	0	1	-1.09	0.98	6.95
33	1	0	1.44	0.17	0	1	-1.44	0.98	6.95
34	1	0	-0.75	0.19	0	1	0.75	0.98	6.95
35	1	0	-1.43	0.17	0	1	1.43	0.98	6.95
36	1	0	-0.83	0.19	0	1	0.83	0.98	6.95
37	1	0	-1.13	0.15	0	1	1.13	0.98	6.95
38	1	0	-0.34	0.17	0	1	0.34	0.98	6.95
39	1	0	-1.25	0.1	0	1	1.25	0.98	6.95
40	1	0	-0.65	0.18	0	1	0.65	0.98	6.95
41	1	0	0.2	0.08	0	1	-0.2	0.98	6.95
42	1	0	-0.51	0.13	0	1	0.51	0.98	6.95
43	1	0	1.33	0.14	0	1	-1.33	0.98	6.95
44	1	0	-0.37	0.19	0	1	0.37	0.98	6.95
45	1	0	1.36	0.2	0	1	-1.36	0.98	6.95
46	1	0	-0.68	0.07	0	1	0.68	0.98	6.95
47	1	0	0.56	0.16	0	1	-0.56	0.98	6.95
48	1	0	-1.09	0.07	0	1	1.09	0.98	6.95
49	1	0	0.93	0.16	0	1	-0.93	0.98	6.95
50	1	0	0.79	0.09	0	1	-0.79	0.98	6.95

Alpha= item measurement direction, MDISC=item discrimination in MIRT, MID=item difficulty in MIRT, CVI = validity sector index, CVI = validity index, RC_Angle = test measurement direction, Common Items in bold

<Y BASE>

Item ID	a1	a2	d	g	alpha	MDISC	MID	CVI	RC_Angle
1	0.82	0.3	0.2	0.13	19.95	0.87	-0.22	0.95	7.69
2	1.15	0.41	1	0.15	19.52	1.22	-0.82	0.95	6.95
3	1.39	0.48	-0.04	0.08	19.09	1.47	0.03	0.96	7.69
4	1.33	0.45	1.41	0.12	18.65	1.4	-1	0.96	7.69
5	0.75	0.25	-0.66	0.17	18.19	0.79	0.84	0.96	6.95
6	1.2	0.38	-0.46	0.17	17.73	1.25	0.36	0.97	7.69
7	0.62	0.19	-0.68	0.12	17.25	0.65	1.05	0.97	6.95
8	1.48	0.45	-1.15	0.1	16.76	1.55	0.74	0.97	7.69
9	1.1	0.32	-1.43	0.16	16.26	1.14	1.25	0.98	7.69
10	0.61	0.17	-0.86	0.14	15.74	0.63	1.37	0.98	7.69
11	0.92	0.25	-0.29	0.15	15.2	0.96	0.31	0.98	6.95
12	0.68	0.18	0.33	0.05	14.65	0.7	-0.47	0.98	7.69
13	1.15	0.29	-0.78	0.15	14.07	1.18	0.66	0.99	7.69
14	0.59	0.14	-0.53	0.11	13.47	0.61	0.87	0.99	7.69
15	1.46	0.33	-1.17	0.14	12.84	1.5	0.78	0.99	7.69
16	0.77	0.17	-0.91	0.14	12.18	0.79	1.16	0.99	7.69
17	1.43	0.29	0.54	0.11	11.48	1.46	-0.37	0.99	7.69
18	1.48	0.28	0.32	0.1	10.73	1.51	-0.22	1	7.69
19	1.16	0.2	0.57	0.15	9.94	1.17	-0.48	1	7.69
20	0.75	0.12	1.4	0.13	9.07	0.76	-1.84	1	7.69
21	0.95	0.14	0.92	0.12	8.11	0.96	-0.95	1	7.69
22	0.75	0.09	1.31	0.08	7.02	0.76	-1.73	1	7.69
23	1.27	0.13	0.87	0.13	5.73	1.28	-0.68	1	6.95
24	0.62	0.04	1.11	0.14	4.05	0.62	-1.79	1	7.69
25	0.97	0	0.71	0.09	0	0.97	-0.73	0.98	7.69
26	1	0	-0.67	0.18	0	1	0.67	0.98	6.95
27	1	0	-0.06	0.14	0	1	0.06	0.98	7.69
28	1	0	0.05	0.2	0	1	-0.05	0.98	7.69
29	1	0	1.01	0.14	0	1	-1.01	0.98	6.95
30	1	0	0.41	0.2	0	1	-0.41	0.98	7.69
31	1	0	-0.43	0.14	0	1	0.43	0.98	7.69
32	1	0	-1.16	0.18	0	1	1.16	0.98	7.69
33	1	0	-0.34	0.18	0	1	0.34	0.98	7.69
34	1	0	-0.18	0.11	0	1	0.18	0.98	7.69

35	1	0	0.8	0.18	0	1	-0.8	0.98	7.69
36	1	0	-0.27	0.17	0	1	0.27	0.98	7.69
37	1	0	-1.13	0.15	0	1	1.13	0.98	6.95
38	1	0	0.19	0.12	0	1	-0.19	0.98	7.69
39	1	0	-1.43	0.19	0	1	1.43	0.98	7.69
40	1	0	-0.65	0.18	0	1	0.65	0.98	6.95
41	1	0	0.2	0.08	0	1	-0.2	0.98	6.95
42	1	0	0.59	0.2	0	1	-0.59	0.98	7.69
43	1	0	-0.72	0.16	0	1	0.72	0.98	7.69
44	1	0	-1.09	0.19	0	1	1.09	0.98	7.69
45	1	0	-1.37	0.15	0	1	1.37	0.98	7.69
46	1	0	0.1	0.19	0	1	-0.1	0.98	7.69
47	1	0	0.64	0.19	0	1	-0.64	0.98	7.69
48	1	0	0.22	0.17	0	1	-0.22	0.98	7.69
49	1	0	0.22	0.16	0	1	-0.22	0.98	7.69
50	1	0	-0.1	0.19	0	1	0.1	0.98	7.69

Alpha= item measurement direction, MDISC=item discrimination in MIRT, MID=item difficulty in MIRT, CVI = validity sector index, CVI = validity index, RC_Angle = test measurement direction, Common Items in bold

<Y 60>

Item	ID	a1	a2	d	g	alpha	MDISC	MID	CVI	RC_Angle
	1	1.13	1.96	0.44	0.15	60	2.26	-0.2	0.37	29.4
	2	1.15	0.41	1	0.15	19.52	1.22	-0.82	0.95	6.95
	3	0.85	1.32	-1.34	0.06	57.2	1.57	0.85	0.41	29.4
	4	0.63	0.93	-1.17	0.16	55.77	1.13	1.04	0.44	29.4
	5	0.75	0.25	-0.66	0.17	18.19	0.79	0.84	0.96	6.95
	6	0.85	1.12	0.37	0.11	52.83	1.41	-0.26	0.49	29.4
	7	0.62	0.19	-0.68	0.12	17.25	0.65	1.05	0.97	6.95
	8	1.03	1.22	-0.51	0.2	49.77	1.6	0.32	0.54	29.4
	9	0.84	0.93	-1.05	0.18	48.19	1.25	0.84	0.57	29.4
	10	0.81	0.86	-0.58	0.09	46.57	1.18	0.49	0.6	29.4
	11	0.92	0.25	-0.29	0.15	15.2	0.96	0.31	0.98	6.95
	12	0.52	0.49	0.74	0.16	43.18	0.71	-1.04	0.66	29.4
	13	0.54	0.47	-1.31	0.15	41.41	0.72	1.83	0.68	29.4
	14	1.38	1.14	-1.23	0.04	39.57	1.79	0.69	0.71	29.4
	15	0.9	0.7	0.35	0.15	37.66	1.14	-0.31	0.74	29.4
	16	0.97	0.7	-0.75	0.15	35.66	1.19	0.63	0.77	29.4
	17	0.9	0.6	0.27	0.17	33.56	1.08	-0.25	0.8	29.4
	18	1.2	0.73	-0.93	0.17	31.33	1.41	0.66	0.83	29.4

19	0.9	0.5	-0.54	0.18	28.96	1.03	0.52	0.86	29.4
20	1	0.5	-0.26	0.1	26.38	1.12	0.23	0.89	29.4
21	0.54	0.23	0.8	0.11	23.56	0.59	-1.37	0.92	29.4
22	0.8	0.3	-1.36	0.18	20.36	0.85	1.6	0.95	29.4
23	1.27	0.13	0.87	0.13	5.73	1.28	-0.68	1	6.95
24	1.44	0.3	-0.91	0.17	11.72	1.47	0.62	0.99	29.4
25	1.46	0	-0.56	0.1	0	1.46	0.38	0.98	29.4
26	1	0	-0.67	0.18	0	1	0.67	0.98	6.95
27	1	0	0.62	0.13	0	1	-0.62	0.98	29.4
28	1	0	0.16	0.11	0	1	-0.16	0.98	29.4
29	1	0	1.01	0.14	0	1	-1.01	0.98	6.95
30	1	0	-1.1	0.11	0	1	1.1	0.98	29.4
31	1	0	0.47	0.09	0	1	-0.47	0.98	29.4
32	1	0	0.24	0.07	0	1	-0.24	0.98	29.4
33	1	0	0.78	0.19	0	1	-0.78	0.98	29.4
34	1	0	-1.48	0.14	0	1	1.48	0.98	29.4
35	1	0	-0.04	0.19	0	1	0.04	0.98	29.4
36	1	0	-0.6	0.09	0	1	0.6	0.98	29.4
37	1	0	-1.13	0.15	0	1	1.13	0.98	6.95
38	1	0	0.37	0.18	0	1	-0.37	0.98	29.4
39	1	0	-0.86	0.16	0	1	0.86	0.98	29.4
40	1	0	-0.65	0.18	0	1	0.65	0.98	6.95
41	1	0	0.2	0.08	0	1	-0.2	0.98	6.95
42	1	0	-0.31	0.2	0	1	0.31	0.98	29.4
43	1	0	-0.57	0.13	0	1	0.57	0.98	29.4
44	1	0	-1.36	0.1	0	1	1.36	0.98	29.4
45	1	0	0.7	0.11	0	1	-0.7	0.98	29.4
46	1	0	0.46	0.19	0	1	-0.46	0.98	29.4
47	1	0	-0.32	0.1	0	1	0.32	0.98	29.4
48	1	0	-1.42	0.13	0	1	1.42	0.98	29.4
49	1	0	-0.25	0.11	0	1	0.25	0.98	29.4
50	1	0	-0.59	0.09	0	1	0.59	0.98	29.4

Alpha= item measurement direction, MDISC=item discrimination in MIRT, MID=item difficulty in MIRT, CVI = validity sector index, CVI = validity index, RC_Angle = test measurement direction, Common Items in bold

APPENDIX B: EVALUATION OF EQUATING

MAB AND RMSE OF TSE
MC-I
<Y BASE>

Focal_Mean	Focal_Sigma	Linking Method	MAB	RMSE	DTM
(0, 0)	(1, 0.3, 1)	cc	0.6	0.65	1
		fc	0.41	0.47	0
		HB	0.3	0.37	0
		SL	0.29	0.35	0
	(1, 0.32, 1.14)	cc	0.59	0.63	1
		fc	0.4	0.46	0
		HB	0.27	0.33	0
		SL	0.28	0.33	0
	(1, 0.7, 1)	cc	0.58	0.62	1
		fc	0.41	0.47	0
		HB	0.3	0.36	0
		SL	0.3	0.36	0
	(1, 0.9, 1)	cc	0.54	0.59	1
		fc	0.42	0.49	0
		HB	0.34	0.4	0
		SL	0.34	0.4	0
(0, 1)	(1, 0.3, 1)	cc	0.67	0.72	1
		fc	0.46	0.53	0
		HB	0.29	0.35	0
		SL	0.27	0.34	0
	(1, 0.32, 1.14)	cc	0.65	0.7	1
		fc	0.45	0.5	0
		HB	0.26	0.32	0
		SL	0.28	0.34	0
	(1, 0.7, 1)	cc	0.61	0.66	1
		fc	0.45	0.51	0
		HB	0.28	0.34	0
		SL	0.29	0.36	0
	(1, 0.9, 1)	cc	0.54	0.6	1
		fc	0.45	0.51	0
		HB	0.28	0.35	0
		SL	0.28	0.34	0

		cc	0.68	0.73	1
	(1, 0.3, 1)	fc	0.49	0.55	0
		HB	0.27	0.32	0
		SL	0.28	0.34	0
		cc	0.67	0.72	1
	(1, 0.32, 1.14)	fc	0.45	0.51	0
		HB	0.28	0.34	0
(1, 0)		SL	0.25	0.32	0
		cc	0.61	0.66	1
	(1, 0.7, 1)	fc	0.45	0.52	0
		HB	0.27	0.34	0
		SL	0.26	0.31	0
		cc	0.52	0.57	1
	(1, 0.9, 1)	fc	0.41	0.47	0
		HB	0.3	0.37	0
		SL	0.27	0.34	0
		cc	0.93	0.98	1
	(1, 0.3, 1)	fc	0.81	0.86	1
		HB	0.48	0.56	0
		SL	0.49	0.56	0
		cc	0.89	0.94	1
	(1, 0.32, 1.14)	fc	0.78	0.83	1
		HB	0.44	0.51	0
(1, 1)		SL	0.46	0.54	0
		cc	0.72	0.77	1
	(1, 0.7, 1)	fc	0.64	0.7	1
		HB	0.4	0.47	0
		SL	0.39	0.47	0
		cc	0.56	0.62	1
	(1, 0.9, 1)	fc	0.51	0.59	1
		HB	0.29	0.35	0
		SL	0.32	0.39	0

<Y SAME RC>

Focal_Mean	Focal_Sigma	Linking Method	MAB	RMSE	DTM
		cc	0.65	0.7	1
(0, 0)	(1, 0.3, 1)	fc	0.82	0.87	1
		HB	1.14	1.19	1

		SL	1.17	1.21	1
		cc	0.63	0.69	1
	(1, 0.32, 1.14)	fc	0.81	0.86	1
		HB	1.16	1.2	1
		SL	1.16	1.21	1
		cc	0.74	0.79	1
	(1, 0.7, 1)	fc	0.96	1.01	1
		HB	1.28	1.32	1
		SL	1.29	1.33	1
		cc	0.78	0.82	1
	(1, 0.9, 1)	fc	0.95	0.99	1
		HB	1.3	1.34	1
		SL	1.26	1.31	1
		cc	0.25	0.31	0
	(1, 0.3, 1)	fc	0.38	0.46	0
		HB	0.95	1.01	1
		SL	0.94	1	1
		cc	0.25	0.31	0
	(1, 0.32, 1.14)	fc	0.43	0.5	0
		HB	0.96	1.01	1
(0, 1)		SL	1.02	1.08	1
		cc	0.38	0.45	0
	(1, 0.7, 1)	fc	0.57	0.64	1
		HB	1.05	1.1	1
		SL	1.06	1.11	1
		cc	0.54	0.6	1
	(1, 0.9, 1)	fc	0.72	0.78	1
		HB	1.15	1.2	1
		SL	1.12	1.17	1
		cc	0.26	0.32	0
	(1, 0.3, 1)	fc	0.37	0.43	0
		HB	0.9	0.96	1
		SL	0.91	0.97	1
(1, 0)		cc	0.25	0.31	0
	(1, 0.32, 1.14)	fc	0.37	0.43	0
		HB	0.9	0.95	1
		SL	0.93	0.99	1
	(1, 0.7, 1)	cc	0.37	0.43	0

		fc	0.58	0.65	1
		HB	1.06	1.11	1
		SL	1.08	1.13	1
		cc	0.43	0.49	0
	(1, 0.9, 1)	fc	0.56	0.61	1
		HB	0.99	1.04	1
		SL	1.02	1.08	1
		cc	0.83	0.9	1
	(1, 0.3, 1)	fc	0.82	0.9	1
		HB	1.19	1.25	1
		SL	1.22	1.28	1
		cc	0.71	0.77	1
	(1, 0.32, 1.14)	fc	0.75	0.82	1
		HB	1.18	1.24	1
(1, 1)		SL	1.16	1.23	1
		cc	0.65	0.71	1
	(1, 0.7, 1)	fc	0.74	0.8	1
		HB	1.3	1.36	1
		SL	1.28	1.35	1
		cc	0.64	0.7	1
	(1, 0.9, 1)	fc	0.75	0.82	1
		HB	1.22	1.27	1
		SL	1.23	1.29	1

<Y DIFF RC>

Focal_Mean	Focal_Sigma	Linking Method	MAB	RMSE	DTM
		cc	0.38	0.44	0
	(1, 0.3, 1)	fc	0.29	0.35	0
		HB	0.39	0.45	0
		SL	0.41	0.47	0
		cc	0.35	0.41	0
	(1, 0.32, 1.14)	fc	0.32	0.38	0
(0, 0)		HB	0.4	0.47	0
		SL	0.41	0.47	0
		cc	0.34	0.4	0
	(1, 0.7, 1)	fc	0.29	0.35	0
		HB	0.42	0.48	0
		SL	0.42	0.48	0

	(1, 0.9, 1)	cc	0.28	0.34	0
		fc	0.32	0.38	0
		HB	0.52	0.58	1
		SL	0.51	0.57	1
(0, 1)	(1, 0.3, 1)	cc	1.37	1.41	1
		fc	1.34	1.37	1
		HB	1.02	1.07	1
		SL	0.94	1	1
	(1, 0.32, 1.14)	cc	1.29	1.32	1
		fc	1.21	1.24	1
		HB	0.85	0.9	1
		SL	0.89	0.95	1
	(1, 0.7, 1)	cc	1.36	1.39	1
		fc	1.29	1.32	1
		HB	1.02	1.07	1
		SL	1.01	1.07	1
	(1, 0.9, 1)	cc	1.06	1.1	1
		fc	0.99	1.03	1
		HB	0.78	0.84	1
		SL	0.77	0.84	1
(1, 0)	(1, 0.3, 1)	cc	1.03	1.08	1
		fc	1.03	1.07	1
		HB	1.31	1.35	1
		SL	1.28	1.32	1
	(1, 0.32, 1.14)	cc	1.1	1.14	1
		fc	1.13	1.17	1
		HB	1.39	1.42	1
		SL	1.38	1.42	1
	(1, 0.7, 1)	cc	1.02	1.06	1
		fc	1.07	1.1	1
		HB	1.34	1.38	1
		SL	1.29	1.33	1
	(1, 0.9, 1)	cc	1.24	1.27	1
		fc	1.28	1.31	1
		HB	1.5	1.54	1
		SL	1.51	1.55	1
(1, 1)	(1, 0.3, 1)	cc	0.75	0.81	1
		fc	0.63	0.7	1

		HB	0.64	0.71	1
		SL	0.62	0.7	1
		cc	0.62	0.68	1
	(1, 0.32, 1.14)	fc	0.5	0.57	0
		HB	0.62	0.7	1
		SL	0.62	0.7	1
		cc	0.56	0.63	1
	(1, 0.7, 1)	fc	0.49	0.56	0
		HB	0.53	0.61	1
		SL	0.57	0.66	1
		cc	0.29	0.36	0
	(1, 0.9, 1)	fc	0.32	0.39	0
		HB	0.64	0.71	1
		SL	0.6	0.68	1

MAB AND RMSE OF OSE
<Y BASE>

Focal_Mean	Focal_Sigma	Linking Method	MAB	RMSE	DTM
		cc	0.6	0.65	1
	(1, 0.3, 1)	fc	0.37	0.43	0
		HB	0.28	0.35	0
		SL	0.27	0.32	0
		cc	0.58	0.63	1
	(1, 0.32, 1.14)	fc	0.37	0.43	0
		HB	0.29	0.35	0
		SL	0.29	0.35	0
(0, 0)		cc	0.54	0.6	1
	(1, 0.7, 1)	fc	0.38	0.44	0
		HB	0.29	0.35	0
		SL	0.29	0.36	0
		cc	0.51	0.56	1
	(1, 0.9, 1)	fc	0.36	0.41	0
		HB	0.32	0.38	0
		SL	0.3	0.36	0
		cc	0.7	0.74	1
(0, 1)	(1, 0.3, 1)	fc	0.49	0.55	0
		HB	0.27	0.35	0
		SL	0.27	0.34	0

	(1, 0.32, 1.14)	cc	0.7	0.75	1
		fc	0.52	0.57	1
		HB	0.27	0.33	0
		SL	0.27	0.33	0
	(1, 0.7, 1)	cc	0.6	0.66	1
		fc	0.46	0.51	0
		HB	0.27	0.33	0
		SL	0.27	0.34	0
	(1, 0.9, 1)	cc	0.55	0.6	1
		fc	0.44	0.5	0
		HB	0.26	0.32	0
		SL	0.28	0.34	0
(1, 0)	(1, 0.3, 1)	cc	0.72	0.76	1
		fc	0.54	0.59	1
		HB	0.28	0.35	0
		SL	0.29	0.36	0
	(1, 0.32, 1.14)	cc	0.71	0.75	1
		fc	0.52	0.58	1
		HB	0.26	0.33	0
		SL	0.28	0.35	0
	(1, 0.7, 1)	cc	0.65	0.7	1
		fc	0.48	0.54	0
		HB	0.26	0.32	0
		SL	0.24	0.3	0
	(1, 0.9, 1)	cc	0.54	0.59	1
		fc	0.41	0.48	0
		HB	0.29	0.36	0
		SL	0.29	0.36	0
(1, 1)	(1, 0.3, 1)	cc	0.98	1.03	1
		fc	0.89	0.95	1
		HB	0.48	0.56	0
		SL	0.45	0.53	0
	(1, 0.32, 1.14)	cc	0.96	1.01	1
		fc	0.84	0.89	1
		HB	0.42	0.5	0
		SL	0.44	0.51	0
	(1, 0.7, 1)	cc	0.83	0.88	1
		fc	0.67	0.74	1

		HB	0.39	0.47	0
		SL	0.37	0.44	0
	(1, 0.9, 1)	cc	0.6	0.66	1
		fc	0.53	0.6	1
		HB	0.33	0.41	0
		SL	0.3	0.38	0

<Y SAME RC>

Focal_Mean	Focal_Sigma	Linking Method	MAB	RMSE	DTM
(0, 0)	(1, 0.3, 1)	cc	0.66	0.72	1
		fc	0.85	0.9	1
		HB	1.17	1.22	1
		SL	1.21	1.26	1
	(1, 0.32, 1.14)	cc	0.69	0.73	1
		fc	0.87	0.92	1
		HB	1.16	1.2	1
		SL	1.19	1.23	1
	(1, 0.7, 1)	cc	0.73	0.78	1
		fc	0.99	1.03	1
		HB	1.34	1.38	1
		SL	1.36	1.4	1
	(1, 0.9, 1)	cc	0.76	0.81	1
		fc	1	1.04	1
		HB	1.32	1.37	1
		SL	1.33	1.36	1
(0, 1)	(1, 0.3, 1)	cc	0.26	0.31	0
		fc	0.39	0.46	0
		HB	0.93	0.99	1
		SL	0.91	0.96	1
	(1, 0.32, 1.14)	cc	0.26	0.32	0
		fc	0.41	0.48	0
		HB	0.93	0.98	1
		SL	0.96	1.02	1
	(1, 0.7, 1)	cc	0.38	0.45	0
		fc	0.6	0.66	1
		HB	1.07	1.12	1
		SL	1.08	1.12	1
	(1, 0.9, 1)	cc	0.54	0.59	1

		fc	0.75	0.8	1
		HB	1.17	1.22	1
		SL	1.17	1.22	1
	(1, 0.3, 1)	cc	0.24	0.29	0
		fc	0.38	0.45	0
		HB	0.92	0.98	1
		SL	0.91	0.97	1
	(1, 0.32, 1.14)	cc	0.24	0.3	0
		fc	0.35	0.42	0
		HB	0.92	0.99	1
(1, 0)		SL	0.92	0.97	1
	(1, 0.7, 1)	cc	0.36	0.43	0
		fc	0.58	0.64	1
		HB	1.09	1.14	1
		SL	1.07	1.12	1
	(1, 0.9, 1)	cc	0.41	0.47	0
		fc	0.6	0.66	1
		HB	1.04	1.09	1
		SL	1.06	1.1	1
	(1, 0.3, 1)	cc	0.89	0.95	1
		fc	0.83	0.9	1
		HB	1.05	1.12	1
		SL	1.07	1.13	1
	(1, 0.32, 1.14)	cc	0.78	0.84	1
		fc	0.78	0.84	1
		HB	1.03	1.09	1
(1, 1)		SL	1.07	1.14	1
	(1, 0.7, 1)	cc	0.62	0.69	1
		fc	0.73	0.8	1
		HB	1.22	1.28	1
		SL	1.2	1.27	1
	(1, 0.9, 1)	cc	0.63	0.7	1
		fc	0.76	0.82	1
		HB	1.19	1.25	1
		SL	1.21	1.26	1

<Y DIFF RC>

Focal_Mean	Focal_Sigma	Linking Method	MAB	RMSE	DTM

(0, 0)	(1, 0.3, 1)	cc	0.38	0.44	0
		fc	0.27	0.33	0
		HB	0.37	0.43	0
		SL	0.4	0.46	0
	(1, 0.32, 1.14)	cc	0.36	0.41	0
		fc	0.28	0.33	0
		HB	0.42	0.48	0
		SL	0.44	0.5	0
	(1, 0.7, 1)	cc	0.31	0.37	0
		fc	0.28	0.34	0
		HB	0.41	0.48	0
		SL	0.45	0.5	0
	(1, 0.9, 1)	cc	0.29	0.35	0
		fc	0.34	0.39	0
		HB	0.51	0.57	1
		SL	0.54	0.61	1
(0, 1)	(1, 0.3, 1)	cc	1.47	1.5	1
		fc	1.39	1.42	1
		HB	1.05	1.09	1
		SL	1	1.07	1
	(1, 0.32, 1.14)	cc	1.42	1.45	1
		fc	1.25	1.28	1
		HB	0.9	0.95	1
		SL	0.92	0.98	1
	(1, 0.7, 1)	cc	1.47	1.5	1
		fc	1.39	1.41	1
		HB	1	1.06	1
		SL	1.03	1.09	1
	(1, 0.9, 1)	cc	1.02	1.06	1
		fc	1	1.05	1
		HB	0.77	0.83	1
		SL	0.76	0.81	1
(1, 0)	(1, 0.3, 1)	cc	1.04	1.09	1
		fc	1.04	1.08	1
		HB	1.29	1.33	1
		SL	1.31	1.34	1
	(1, 0.32, 1.14)	cc	1.11	1.15	1
		fc	1.13	1.17	1

		HB	1.38	1.42	1
		SL	1.37	1.41	1
		cc	1.03	1.07	1
	(1, 0.7, 1)	fc	1.09	1.13	1
		HB	1.35	1.39	1
		SL	1.36	1.4	1
		cc	1.28	1.31	1
	(1, 0.9, 1)	fc	1.32	1.34	1
		HB	1.56	1.6	1
		SL	1.54	1.57	1
		cc	0.8	0.86	1
	(1, 0.3, 1)	fc	0.72	0.8	1
		HB	0.56	0.65	1
		SL	0.56	0.64	1
		cc	0.67	0.74	1
	(1, 0.32, 1.14)	fc	0.56	0.63	1
		HB	0.59	0.67	1
(1, 1)		SL	0.58	0.67	1
		cc	0.62	0.68	1
	(1, 0.7, 1)	fc	0.54	0.62	1
		HB	0.5	0.59	1
		SL	0.51	0.59	1
		cc	0.29	0.36	0
	(1, 0.9, 1)	fc	0.3	0.37	0
		HB	0.64	0.71	1
		SL	0.66	0.74	1

MC-I

MAB AND RMSE OF TSE
<Y BASE>

Focal_Mean	Focal_Sigma	Linking Method	MAB	RMSE	DTM
		cc	0.89	0.93	1
	(1, 0.3, 1)	fc	0.66	0.7	1
		HB	0.39	0.45	0
(0, 0)		SL	0.39	0.45	0
		cc	0.91	0.94	1
	(1, 0.32, 1.14)	fc	0.67	0.72	1

		HB	0.41	0.47	0
		SL	0.4	0.45	0
		cc	0.93	0.97	1
	(1, 0.7, 1)	fc	0.67	0.72	1
		HB	0.4	0.47	0
		SL	0.4	0.46	0
		cc	0.93	0.97	1
	(1, 0.9, 1)	fc	0.73	0.77	1
		HB	0.39	0.44	0
		SL	0.4	0.45	0
		cc	0.91	0.94	1
	(1, 0.3, 1)	fc	0.7	0.74	1
		HB	0.38	0.44	0
		SL	0.42	0.47	0
		cc	0.91	0.95	1
	(1, 0.32, 1.14)	fc	0.68	0.72	1
		HB	0.37	0.42	0
(0, 1)		SL	0.4	0.46	0
		cc	0.91	0.94	1
	(1, 0.7, 1)	fc	0.71	0.75	1
		HB	0.36	0.42	0
		SL	0.42	0.48	0
		cc	0.91	0.95	1
	(1, 0.9, 1)	fc	0.7	0.75	1
		HB	0.39	0.45	0
		SL	0.42	0.48	0
		cc	1.1	1.13	1
	(1, 0.3, 1)	fc	0.89	0.93	1
		HB	0.48	0.53	0
		SL	0.52	0.58	1
		cc	1.08	1.11	1
	(1, 0.32, 1.14)	fc	0.9	0.93	1
(1, 0)		HB	0.5	0.57	1
		SL	0.51	0.57	1
		cc	1.07	1.11	1
	(1, 0.7, 1)	fc	0.89	0.93	1
		HB	0.5	0.55	1
		SL	0.49	0.56	0

	(1, 0.9, 1)	cc	1.09	1.12	1
		fc	0.89	0.93	1
		HB	0.46	0.52	0
		SL	0.49	0.54	0
(1, 1)	(1, 0.3, 1)	cc	1.15	1.18	1
		fc	0.96	1	1
		HB	0.53	0.59	1
		SL	0.57	0.63	1
	(1, 0.32, 1.14)	cc	1.15	1.19	1
		fc	0.93	0.97	1
		HB	0.55	0.61	1
		SL	0.54	0.59	1
	(1, 0.7, 1)	cc	1.12	1.15	1
		fc	0.93	0.97	1
		HB	0.5	0.56	1
		SL	0.51	0.56	1
	(1, 0.9, 1)	cc	1.12	1.15	1
		fc	0.94	0.98	1
		HB	0.53	0.59	1
		SL	0.53	0.59	1

<Y 60>

Focal_Mean	Focal_Sigma	Linking Method	MAB	RMSE	DTM
(0, 0)	(1, 0.3, 1)	cc	0.89	0.92	1
		fc	0.71	0.74	1
		HB	0.49	0.54	0
		SL	0.49	0.54	0
	(1, 0.32, 1.14)	cc	0.9	0.94	1
		fc	0.72	0.75	1
		HB	0.48	0.55	0
		SL	0.49	0.55	0
	(1, 0.7, 1)	cc	0.76	0.81	1
		fc	0.55	0.6	1
		HB	0.39	0.45	0
		SL	0.43	0.49	0
	(1, 0.9, 1)	cc	0.73	0.77	1
		fc	0.55	0.6	1
		HB	0.48	0.53	0

		SL	0.48	0.54	0
(0, 1)	(1, 0.3, 1)	cc	1.91	1.94	1
		fc	1.75	1.78	1
		HB	1.62	1.65	1
		SL	1.61	1.64	1
	(1, 0.32, 1.14)	cc	1.87	1.89	1
		fc	1.73	1.75	1
		HB	1.6	1.63	1
		SL	1.61	1.64	1
	(1, 0.7, 1)	cc	1.84	1.87	1
		fc	1.71	1.74	1
		HB	1.58	1.6	1
		SL	1.61	1.64	1
	(1, 0.9, 1)	cc	1.77	1.79	1
		fc	1.6	1.62	1
		HB	1.5	1.53	1
		SL	1.52	1.55	1
(1, 0)	(1, 0.3, 1)	cc	0.96	0.99	1
		fc	0.73	0.78	1
		HB	0.57	0.63	1
		SL	0.58	0.63	1
	(1, 0.32, 1.14)	cc	0.96	0.99	1
		fc	0.73	0.78	1
		HB	0.59	0.64	1
		SL	0.59	0.65	1
	(1, 0.7, 1)	cc	0.82	0.86	1
		fc	0.64	0.69	1
		HB	0.66	0.72	1
		SL	0.62	0.67	1
	(1, 0.9, 1)	cc	0.86	0.9	1
		fc	0.68	0.73	1
		HB	0.75	0.8	1
		SL	0.75	0.8	1
(1, 1)	(1, 0.3, 1)	cc	1.64	1.67	1
		fc	1.51	1.54	1
		HB	1.21	1.24	1
		SL	1.21	1.24	1
	(1, 0.32, 1.14)	cc	1.57	1.59	1

		fc	1.44	1.47	1
		HB	1.17	1.21	1
		SL	1.15	1.19	1
		cc	1.31	1.34	1
	(1, 0.7, 1)	fc	1.2	1.24	1
		HB	0.92	0.97	1
		SL	0.98	1.04	1
		cc	1.03	1.06	1
	(1, 0.9, 1)	fc	0.94	0.97	1
		HB	0.71	0.77	1
		SL	0.69	0.74	1

MAB AND RMSE OF OSE
<Y BASE>

Focal_Mean	Focal_Sigma	Linking Method	MAB	RMSE	DTM
		cc	0.89	0.93	1
	(1, 0.3, 1)	fc	0.66	0.7	1
		HB	0.39	0.45	0
		SL	0.39	0.45	0
		cc	0.91	0.94	1
	(1, 0.32, 1.14)	fc	0.67	0.72	1
		HB	0.41	0.47	0
(0, 0)		SL	0.4	0.45	0
		cc	0.93	0.97	1
	(1, 0.7, 1)	fc	0.67	0.72	1
		HB	0.4	0.47	0
		SL	0.4	0.46	0
		cc	0.93	0.97	1
	(1, 0.9, 1)	fc	0.73	0.77	1
		HB	0.39	0.44	0
		SL	0.4	0.45	0
		cc	0.91	0.94	1
	(1, 0.3, 1)	fc	0.7	0.74	1
		HB	0.38	0.44	0
(0, 1)		SL	0.42	0.47	0
		cc	0.91	0.95	1
	(1, 0.32, 1.14)	fc	0.68	0.72	1
		HB	0.37	0.42	0

		SL	0.4	0.46	0
		cc	0.91	0.94	1
	(1, 0.7, 1)	fc	0.71	0.75	1
		HB	0.36	0.42	0
		SL	0.42	0.48	0
		cc	0.91	0.95	1
	(1, 0.9, 1)	fc	0.7	0.75	1
		HB	0.39	0.45	0
		SL	0.42	0.48	0
		cc	1.1	1.13	1
	(1, 0.3, 1)	fc	0.89	0.93	1
		HB	0.48	0.53	0
		SL	0.52	0.58	1
		cc	1.08	1.11	1
	(1, 0.32, 1.14)	fc	0.9	0.93	1
		HB	0.5	0.57	1
(1, 0)		SL	0.51	0.57	1
		cc	1.07	1.11	1
	(1, 0.7, 1)	fc	0.89	0.93	1
		HB	0.5	0.55	1
		SL	0.49	0.56	0
		cc	1.09	1.12	1
	(1, 0.9, 1)	fc	0.89	0.93	1
		HB	0.46	0.52	0
		SL	0.49	0.54	0
		cc	1.15	1.18	1
	(1, 0.3, 1)	fc	0.96	1	1
		HB	0.53	0.59	1
		SL	0.57	0.63	1
		cc	1.15	1.19	1
	(1, 0.32, 1.14)	fc	0.93	0.97	1
		HB	0.55	0.61	1
(1, 1)		SL	0.54	0.59	1
		cc	1.12	1.15	1
	(1, 0.7, 1)	fc	0.93	0.97	1
		HB	0.5	0.56	1
		SL	0.51	0.56	1
		cc	1.12	1.15	1
	(1, 0.9, 1)	fc	0.94	0.98	1

		HB	0.53	0.59	1
		SL	0.53	0.59	1

<Y 60>

Focal_Mean	Focal_Sigma	Linking Method	MAB	RMSE	DTM
(0, 0)	(1, 0.3, 1)	cc	0.93	0.97	1
		fc	0.72	0.76	1
		HB	0.47	0.53	0
		SL	0.48	0.53	0
	(1, 0.32, 1.14)	cc	0.94	0.98	1
		fc	0.73	0.77	1
		HB	0.47	0.53	0
		SL	0.48	0.54	0
	(1, 0.7, 1)	cc	0.75	0.79	1
		fc	0.54	0.6	1
		HB	0.43	0.48	0
		SL	0.41	0.47	0
	(1, 0.9, 1)	cc	0.76	0.81	1
		fc	0.53	0.58	1
		HB	0.46	0.52	0
		SL	0.45	0.51	0
(0, 1)	(1, 0.3, 1)	cc	2	2.02	1
		fc	1.85	1.88	1
		HB	1.66	1.69	1
		SL	1.69	1.71	1
	(1, 0.32, 1.14)	cc	2.01	2.03	1
		fc	1.84	1.87	1
		HB	1.68	1.7	1
		SL	1.67	1.69	1
	(1, 0.7, 1)	cc	1.91	1.93	1
		fc	1.75	1.78	1
		HB	1.57	1.6	1
		SL	1.64	1.67	1
	(1, 0.9, 1)	cc	1.82	1.85	1
		fc	1.68	1.71	1
		HB	1.49	1.52	1
		SL	1.51	1.55	1
(1, 0)	(1, 0.3, 1)	cc	0.99	1.03	1

	fc	0.76	0.81	1
	HB	0.55	0.61	1
	SL	0.59	0.65	1
	cc	1	1.03	1
(1, 0.32, 1.14)	fc	0.75	0.8	1
	HB	0.54	0.6	1
	SL	0.58	0.63	1
	cc	0.85	0.9	1
(1, 0.7, 1)	fc	0.63	0.67	1
	HB	0.56	0.62	1
	SL	0.59	0.65	1
	cc	0.89	0.92	1
(1, 0.9, 1)	fc	0.72	0.77	1
	HB	0.72	0.77	1
	SL	0.7	0.75	1
	cc	1.72	1.75	1
(1, 0.3, 1)	fc	1.55	1.58	1
	HB	1.23	1.26	1
	SL	1.25	1.29	1
	cc	1.66	1.69	1
(1, 0.32, 1.14)	fc	1.5	1.53	1
	HB	1.17	1.21	1
(1, 1)	SL	1.26	1.3	1
	cc	1.35	1.38	1
(1, 0.7, 1)	fc	1.21	1.24	1
	HB	0.98	1.02	1
	SL	0.9	0.95	1
	cc	1.09	1.12	1
(1, 0.9, 1)	fc	0.97	1	1
	HB	0.67	0.73	1
	SL	0.73	0.78	1

RESULTS OF CLASSIFICATION AND EQUITY PROPERTIES

MC-I
<Y BASE>

MU	Sigma	Linking	A	B	CC	CA	FOE_D1	SOE_D2	D12
(0, 0)	(1, 0.3, 1)	cc	1.37	-0.04	0.94	0.96	0.12	0.02	14.6

		fc	1.25	-0.05	0.94	0.96	0.09	0.02	9.04
		sc_HB	1.08	-0.03	0.94	0.95	0.05	0.01	5.26
		sc_SL	1.09	-0.03	0.94	0.95	0.05	0.01	4.93
		cc	1.34	-0.05	0.94	0.96	0.12	0.02	14
	(1, 0.32, 1.14)	fc	1.21	-0.04	0.94	0.96	0.08	0.02	9.22
		sc_HB	1.06	-0.03	0.93	0.95	0.05	0.01	5
		sc_SL	1.06	-0.03	0.93	0.95	0.05	0.01	4.97
		cc	1.47	-0.04	0.95	0.96	0.12	0.02	13.8
	(1, 0.7, 1)	fc	1.35	-0.05	0.95	0.96	0.09	0.02	9.34
		sc_HB	1.16	-0.03	0.94	0.96	0.05	0.01	5.23
		sc_SL	1.17	-0.03	0.94	0.96	0.06	0.01	5.32
		cc	1.59	-0.04	0.95	0.97	0.11	0.02	12.7
	(1, 0.9, 1)	fc	1.49	-0.04	0.95	0.96	0.09	0.02	8.85
		sc_HB	1.28	-0.02	0.94	0.96	0.06	0.01	5.67
		sc_SL	1.28	-0.02	0.94	0.96	0.06	0.01	5.41
		cc	1.35	0.67	0.94	0.96	0.13	0.02	14.9
	(1, 0.3, 1)	fc	1.23	0.59	0.94	0.96	0.09	0.01	9.28
		sc_HB	1.05	0.55	0.94	0.95	0.04	0.01	5.68
		sc_SL	1.05	0.56	0.94	0.95	0.04	0.01	5.49
		cc	1.31	0.67	0.94	0.96	0.13	0.02	14.5
	(1, 0.32, 1.14)	fc	1.2	0.6	0.94	0.96	0.1	0.01	9.57
		sc_HB	1.02	0.55	0.93	0.95	0.04	0.01	5.62
		sc_SL	1.03	0.55	0.93	0.95	0.05	0.01	5.61
(0, 1)		cc	1.45	0.65	0.95	0.96	0.12	0.02	13.5
	(1, 0.7, 1)	fc	1.34	0.6	0.95	0.96	0.09	0.01	9.18
		sc_HB	1.14	0.55	0.94	0.96	0.05	0.01	5.31
		sc_SL	1.15	0.56	0.94	0.96	0.05	0.01	5.35
		cc	1.56	0.63	0.95	0.96	0.11	0.02	12
	(1, 0.9, 1)	fc	1.48	0.58	0.95	0.96	0.08	0.01	8.84
		sc_HB	1.26	0.56	0.94	0.96	0.05	0.01	4.79
		sc_SL	1.27	0.56	0.94	0.96	0.05	0.01	4.98
		cc	1.35	0.67	0.94	0.96	0.13	0.02	14.8
	(1, 0.3, 1)	fc	1.24	0.6	0.94	0.96	0.1	0.01	9.93
		sc_HB	1.05	0.55	0.94	0.95	0.04	0.01	5.81
(1, 0)		sc_SL	1.06	0.56	0.94	0.95	0.05	0.01	5.75
		cc	1.32	0.67	0.94	0.96	0.13	0.02	14.5
	(1, 0.32, 1.14)	fc	1.21	0.6	0.94	0.96	0.1	0.01	9.7
		sc_HB	1.03	0.56	0.93	0.95	0.05	0.01	5.36

		sc_SL	1.03	0.55	0.93	0.95	0.04	0.01	5.51
		cc	1.45	0.65	0.95	0.96	0.12	0.02	13.4
	(1, 0.7, 1)	fc	1.33	0.6	0.95	0.96	0.09	0.01	9.51
		sc_HB	1.14	0.56	0.94	0.96	0.05	0.01	5.17
		sc_SL	1.14	0.56	0.94	0.96	0.05	0.01	4.88
		cc	1.57	0.63	0.95	0.97	0.11	0.02	11.9
	(1, 0.9, 1)	fc	1.48	0.6	0.95	0.96	0.08	0.01	8.65
		sc_HB	1.26	0.55	0.94	0.96	0.05	0.01	5.31
		sc_SL	1.27	0.57	0.95	0.96	0.05	0.01	5.14
		cc	1.29	1.33	0.96	0.97	0.15	0.02	16.6
	(1, 0.3, 1)	fc	1.22	1.24	0.96	0.97	0.13	0.02	13.5
		sc_HB	0.97	1.08	0.95	0.96	0.06	0.01	9.84
		sc_SL	0.98	1.09	0.95	0.96	0.07	0.01	9.61
		cc	1.27	1.34	0.95	0.97	0.16	0.02	16.2
	(1, 0.32, 1.14)	fc	1.19	1.24	0.96	0.97	0.14	0.02	13
		sc_HB	0.93	1.06	0.95	0.96	0.07	0.01	8.72
(1, 1)		sc_SL	0.95	1.08	0.95	0.96	0.07	0.01	9.31
		cc	1.37	1.3	0.96	0.97	0.13	0.02	14.1
	(1, 0.7, 1)	fc	1.3	1.22	0.96	0.97	0.12	0.01	11.2
		sc_HB	1.07	1.09	0.95	0.96	0.06	0.01	7.97
		sc_SL	1.08	1.1	0.95	0.97	0.06	0.01	7.61
		cc	1.47	1.24	0.96	0.97	0.1	0.02	11
	(1, 0.9, 1)	fc	1.43	1.2	0.96	0.97	0.09	0.01	9.4
		sc_HB	1.21	1.11	0.96	0.97	0.05	0.01	6.2
		sc_SL	1.21	1.11	0.96	0.97	0.05	0.01	6.17

cc=Concurrent Calibration, fc=Fixed Parameter Calibration, sc_HB: separate calibration with Haebara, sc_SL=separate calibration with Stocking and Lord , A= linking constant A, B = linking constant B, CC=Classification Consistency, CA=Classification Accuracy, FOE D1= First Order Equity Marginal Index, SOE D2= Second Order Equity Marginal Index, D12= combined index

<Y SAME RC>

MU	Sigma	Linking	A	B	CC	CA	FOE_D1	SOE_D2	D12
		cc	1.58	0.05	0.95	0.97	0.09	0.02	9.8
	(1, 0.3, 1)	fc	1.46	0.04	0.95	0.96	0.05	0.01	13.5
(0, 0)		sc_HB	1.22	0.05	0.94	0.96	0.04	0.01	21.3
		sc_SL	1.25	0.05	0.94	0.96	0.04	0.01	21.1
	(1, 0.32, 1.14)	cc	1.54	0.05	0.95	0.97	0.09	0.02	9.74

		fc	1.42	0.04	0.95	0.96	0.06	0.01	13.4
		sc_HB	1.2	0.05	0.94	0.96	0.04	0.01	20.8
		sc_SL	1.22	0.05	0.94	0.96	0.04	0.01	20.7
		cc	1.7	0.06	0.96	0.97	0.07	0.01	11.3
	(1, 0.7, 1)	fc	1.62	0.05	0.96	0.97	0.05	0.01	16.8
		sc_HB	1.34	0.06	0.95	0.96	0.04	0.01	25
		sc_SL	1.37	0.07	0.95	0.96	0.04	0.01	24.7
		cc	1.81	0.06	0.96	0.97	0.06	0.01	12
	(1, 0.9, 1)	fc	1.74	0.04	0.96	0.97	0.04	0.01	16.9
		sc_HB	1.46	0.06	0.95	0.96	0.04	0.01	24.5
		sc_SL	1.49	0.07	0.95	0.96	0.04	0.01	24
		cc	1.5	0.85	0.95	0.97	0.09	0.02	5.79
	(1, 0.3, 1)	fc	1.41	0.78	0.95	0.97	0.08	0.01	8.93
		sc_HB	1.08	0.63	0.95	0.96	0.08	0.02	20.5
		sc_SL	1.16	0.71	0.95	0.96	0.04	0.01	20.4
		cc	1.47	0.86	0.95	0.97	0.1	0.02	5.65
	(1, 0.32, 1.14)	fc	1.38	0.79	0.95	0.97	0.08	0.01	9.49
		sc_HB	1.03	0.62	0.95	0.96	0.08	0.02	20
(0, 1)		sc_SL	1.13	0.71	0.95	0.96	0.05	0.01	21
		cc	1.6	0.84	0.96	0.97	0.07	0.01	7.71
	(1, 0.7, 1)	fc	1.53	0.79	0.96	0.97	0.06	0.01	12.7
		sc_HB	1.18	0.65	0.95	0.96	0.08	0.02	22.6
		sc_SL	1.29	0.74	0.95	0.97	0.04	0.01	22.8
		cc	1.72	0.8	0.96	0.97	0.05	0.01	9.86
	(1, 0.9, 1)	fc	1.67	0.77	0.96	0.97	0.05	0.01	14.6
		sc_HB	1.34	0.68	0.96	0.96	0.07	0.01	23.1
		sc_SL	1.45	0.75	0.96	0.97	0.04	0.01	22.9
		cc	1.5	0.85	0.95	0.97	0.1	0.02	5.71
	(1, 0.3, 1)	fc	1.42	0.78	0.95	0.97	0.09	0.01	8.65
		sc_HB	1.02	0.58	0.95	0.96	0.11	0.03	20.1
		sc_SL	1.14	0.68	0.95	0.96	0.05	0.01	20.4
		cc	1.47	0.85	0.95	0.97	0.1	0.02	5.69
(1, 0)	(1, 0.32, 1.14)	fc	1.39	0.79	0.95	0.97	0.09	0.01	8.43
		sc_HB	1.01	0.6	0.95	0.96	0.1	0.02	19.6
		sc_SL	1.12	0.69	0.95	0.96	0.05	0.01	20.7
		cc	1.61	0.84	0.96	0.97	0.08	0.01	7.29
	(1, 0.7, 1)	fc	1.55	0.8	0.96	0.97	0.07	0.01	12.7
		sc_HB	1.15	0.63	0.95	0.96	0.1	0.02	23.5

		sc_SL	1.28	0.73	0.95	0.96	0.04	0.01	23.1
		cc	1.71	0.79	0.96	0.97	0.07	0.01	8.11
	(1, 0.9, 1)	fc	1.67	0.76	0.96	0.97	0.07	0.01	11.9
		sc_HB	1.21	0.58	0.96	0.97	0.13	0.03	21.6
		sc_SL	1.39	0.7	0.96	0.97	0.04	0.01	21.7
		cc	1.4	1.58	0.97	0.97	0.1	0.01	20.7
	(1, 0.3, 1)	fc	1.37	1.54	0.97	0.97	0.1	0.01	22.3
		sc_HB	1.11	1.36	0.97	0.97	0.07	0.01	29.4
		sc_SL	1.12	1.38	0.97	0.97	0.06	0.01	29.8
		cc	1.37	1.6	0.97	0.98	0.1	0.01	18.8
	(1, 0.32, 1.14)	fc	1.35	1.56	0.97	0.97	0.11	0.01	21
		sc_HB	1.09	1.37	0.97	0.97	0.07	0.01	27.9
		sc_SL	1.09	1.38	0.97	0.97	0.07	0.01	28.9
(1, 1)		cc	1.48	1.54	0.97	0.97	0.06	0.01	18
	(1, 0.7, 1)	fc	1.46	1.53	0.97	0.97	0.07	0.01	21.2
		sc_HB	1.24	1.4	0.97	0.97	0.06	0.01	32.6
		sc_SL	1.26	1.43	0.97	0.97	0.05	0.01	31.9
		cc	1.57	1.46	0.97	0.97	0.05	0.01	17
	(1, 0.9, 1)	fc	1.57	1.47	0.97	0.97	0.05	0.01	20.3
		sc_HB	1.37	1.38	0.97	0.97	0.06	0.01	29.3
		sc_SL	1.41	1.41	0.97	0.97	0.05	0.01	29.8

cc=Concurrent Calibration, fc=Fixed Parameter Calibration, sc_HB: separate calibration with Haebara, sc_SL=separate calibration with Stocking and Lord , A= linking constant A, B = linking constant B, CC=Classification Consistency, CA=Classification Accuracy, FOE D1= First Order Equity Marginal Index, SOE D2= Second Order Equity Marginal Index, D12= combined index

<Y DIFF RC>

MU	Sigma	Linking	A	B	CC	CA	FOE_D1	SOE_D2	D12
		cc	1.4	0.02	0.94	0.96	0.09	0.02	11.3
	(1, 0.3, 1)	fc	1.27	0.01	0.94	0.96	0.06	0.01	6.59
		sc_HB	1.09	0.02	0.94	0.95	0.04	0.01	6.03
		sc_SL	1.11	0.02	0.94	0.95	0.04	0.01	6.29
(0, 0)		cc	1.36	0.02	0.94	0.96	0.1	0.02	10.3
	(1, 0.32, 1.14)	fc	1.23	0.01	0.94	0.96	0.06	0.01	6.6
		sc_HB	1.06	0.02	0.94	0.95	0.04	0.01	6.33
		sc_SL	1.08	0.02	0.94	0.95	0.04	0.01	6.55
	(1, 0.7, 1)	cc	1.51	0.03	0.95	0.96	0.09	0.02	8.69

		fc	1.39	0.01	0.95	0.96	0.06	0.01	5.5
		sc_HB	1.18	0.03	0.94	0.96	0.04	0.01	6.49
		sc_SL	1.2	0.04	0.94	0.96	0.04	0.01	6.63
		cc	1.65	0.03	0.95	0.97	0.08	0.02	6.48
	(1, 0.9, 1)	fc	1.56	0.02	0.95	0.97	0.06	0.01	5.26
		sc_HB	1.32	0.03	0.95	0.96	0.05	0.01	7.94
		sc_SL	1.34	0.04	0.95	0.96	0.05	0.01	8.22
		cc	1.35	0.84	0.95	0.96	0.11	0.02	22.4
	(1, 0.3, 1)	fc	1.25	0.76	0.95	0.96	0.09	0.01	24.6
		sc_HB	0.97	0.62	0.94	0.95	0.07	0.02	24.8
		sc_SL	1.03	0.68	0.94	0.96	0.05	0.01	24.4
		cc	1.32	0.85	0.95	0.96	0.11	0.02	21.5
	(1, 0.32, 1.14)	fc	1.22	0.77	0.95	0.96	0.09	0.01	23
		sc_HB	0.93	0.62	0.94	0.95	0.08	0.02	22.2
(0, 1)		sc_SL	0.99	0.68	0.94	0.96	0.05	0.01	23
		cc	1.45	0.83	0.95	0.97	0.1	0.02	23.9
	(1, 0.7, 1)	fc	1.35	0.77	0.95	0.96	0.08	0.01	25.4
		sc_HB	1.04	0.63	0.95	0.96	0.08	0.02	24.6
		sc_SL	1.12	0.71	0.95	0.96	0.05	0.01	24.8
		cc	1.59	0.8	0.96	0.97	0.08	0.01	19.5
	(1, 0.9, 1)	fc	1.52	0.76	0.96	0.97	0.06	0.01	20.9
		sc_HB	1.23	0.67	0.95	0.96	0.05	0.01	19.6
		sc_SL	1.29	0.7	0.95	0.96	0.04	0.01	19.5
		cc	1.34	0.6	0.94	0.96	0.1	0.02	29.4
	(1, 0.3, 1)	fc	1.22	0.52	0.94	0.96	0.08	0.01	25.4
		sc_HB	0.96	0.43	0.94	0.95	0.08	0.02	24.5
		sc_SL	1.04	0.48	0.94	0.95	0.04	0.01	24.5
		cc	1.31	0.6	0.94	0.96	0.1	0.02	29.8
	(1, 0.32, 1.14)	fc	1.19	0.53	0.94	0.96	0.08	0.01	26.8
		sc_HB	0.94	0.43	0.93	0.95	0.07	0.02	26.1
(1, 0)		sc_SL	1.01	0.48	0.93	0.95	0.04	0.01	25.7
		cc	1.45	0.61	0.95	0.96	0.09	0.02	27.1
	(1, 0.7, 1)	fc	1.32	0.53	0.95	0.96	0.07	0.01	25
		sc_HB	1.06	0.45	0.94	0.95	0.07	0.01	25.2
		sc_SL	1.15	0.51	0.94	0.95	0.04	0.01	25.2
		cc	1.59	0.58	0.95	0.97	0.09	0.02	29.3
	(1, 0.9, 1)	fc	1.5	0.53	0.95	0.96	0.07	0.01	27.7
		sc_HB	1.18	0.45	0.95	0.96	0.08	0.01	28.7

		sc_SL	1.27	0.51	0.95	0.96	0.05	0.01	28.5
(1, 1)	(1, 0.3, 1)	cc	1.31	1.4	0.96	0.97	0.13	0.02	13.1
		fc	1.24	1.29	0.96	0.97	0.12	0.01	10.9
		sc_HB	0.97	1.11	0.96	0.97	0.07	0.02	11.7
		sc_SL	0.99	1.14	0.96	0.97	0.06	0.01	11.8
	(1, 0.32, 1.14)	cc	1.28	1.41	0.96	0.97	0.14	0.02	12.2
		fc	1.2	1.3	0.96	0.97	0.13	0.01	9.11
		sc_HB	0.94	1.11	0.96	0.97	0.07	0.02	10.8
		sc_SL	0.96	1.13	0.96	0.97	0.07	0.01	10.9
	(1, 0.7, 1)	cc	1.39	1.38	0.96	0.97	0.1	0.01	9.89
		fc	1.32	1.3	0.96	0.97	0.1	0.01	9.1
		sc_HB	1.07	1.16	0.96	0.97	0.06	0.01	11.1
		sc_SL	1.1	1.18	0.96	0.97	0.05	0.01	11.6
	(1, 0.9, 1)	cc	1.5	1.3	0.96	0.97	0.07	0.01	6.04
		fc	1.46	1.26	0.96	0.97	0.07	0.01	5.5
		sc_HB	1.24	1.16	0.96	0.97	0.05	0.01	10.3
		sc_SL	1.25	1.18	0.96	0.97	0.04	0.01	10.2

cc=Concurrent Calibration, fc=Fixed Parameter Calibration, sc_HB: separate calibration with Haebara, sc_SL=separate calibration with Stocking and Lord , A= linking constant A, B = linking constant B, CC=Classification Consistency, CA=Classification Accuracy, FOE D1= First Order Equity Marginal Index, SOE D2= Second Order Equity Marginal Index, D12= combined index

MC-II
<Y BASE>

MU	Sigma	Linking	A	B	CC	CA	FOE_D1	SOE_D2	D12
(0, 0)	(1, 0.3, 1)	cc	1.17	-0.01	0.93	0.93	0.93	0.95	0.9
		fc	1.07	0	0.93	0.93	0.94	0.95	0.9
		sc_HB	0.93	0.01	0.93	0.93	0.94	0.95	0.9
		sc_SL	0.93	0.01	0.93	0.93	0.94	0.95	0.9
	(1, 0.32, 1.14)	cc	1.18	-0.01	0.93	0.93	0.93	0.95	0.9
		fc	1.07	0	0.93	0.93	0.94	0.95	0.9
		sc_HB	0.93	0.01	0.93	0.93	0.94	0.95	0.9
		sc_SL	0.93	0.02	0.93	0.93	0.94	0.95	0.9
	(1, 0.7, 1)	cc	1.21	-0.02	0.93	0.93	0.93	0.95	0.9
		fc	1.1	0	0.93	0.93	0.94	0.95	0.9
		sc_HB	0.96	0.01	0.93	0.93	0.94	0.95	0.9
		sc_SL	0.96	0.02	0.93	0.93	0.94	0.95	0.9
	(1, 0.9, 1)	cc	1.22	-0.01	0.93	0.93	0.93	0.95	0.9

		fc	1.12	0	0.93	0.93	0.94	0.95	0.9
		sc_HB	0.96	0.02	0.93	0.93	0.94	0.95	0.9
		sc_SL	0.97	0.02	0.93	0.93	0.94	0.95	0.9
(0, 1)	(1, 0.3, 1)	cc	1.16	0.1	0.93	0.93	0.93	0.95	0.9
		fc	1.06	0.1	0.93	0.93	0.94	0.95	0.9
		sc_HB	0.9	0.1	0.93	0.93	0.94	0.95	0.9
		sc_SL	0.92	0.11	0.93	0.93	0.94	0.95	0.9
	(1, 0.32, 1.14)	cc	1.16	0.1	0.93	0.93	0.93	0.95	0.9
		fc	1.06	0.09	0.93	0.93	0.94	0.95	0.9
		sc_HB	0.91	0.1	0.93	0.93	0.94	0.95	0.9
		sc_SL	0.92	0.11	0.93	0.93	0.94	0.95	0.9
	(1, 0.7, 1)	cc	1.19	0.1	0.93	0.93	0.93	0.95	0.9
		fc	1.09	0.1	0.93	0.93	0.94	0.95	0.9
		sc_HB	0.93	0.1	0.93	0.93	0.94	0.95	0.9
		sc_SL	0.95	0.11	0.93	0.93	0.94	0.95	0.9
	(1, 0.9, 1)	cc	1.21	0.09	0.93	0.93	0.93	0.95	0.9
		fc	1.1	0.1	0.93	0.93	0.94	0.95	0.9
		sc_HB	0.95	0.1	0.93	0.93	0.94	0.95	0.9
		sc_SL	0.96	0.11	0.93	0.93	0.94	0.95	0.9
(1, 0)	(1, 0.3, 1)	cc	1.17	0.84	0.93	0.93	0.93	0.95	0.9
		fc	1.07	0.79	0.93	0.93	0.94	0.95	0.9
		sc_HB	0.89	0.72	0.93	0.93	0.94	0.95	0.9
		sc_SL	0.9	0.72	0.93	0.93	0.94	0.95	0.9
	(1, 0.32, 1.14)	cc	1.17	0.85	0.93	0.93	0.93	0.95	0.9
		fc	1.06	0.78	0.93	0.93	0.94	0.95	0.9
		sc_HB	0.89	0.72	0.93	0.93	0.94	0.95	0.9
		sc_SL	0.89	0.72	0.93	0.93	0.94	0.95	0.9
	(1, 0.7, 1)	cc	1.2	0.85	0.93	0.93	0.93	0.95	0.9
		fc	1.1	0.79	0.93	0.93	0.94	0.95	0.9
		sc_HB	0.91	0.72	0.93	0.93	0.94	0.95	0.9
		sc_SL	0.92	0.72	0.93	0.93	0.94	0.95	0.9
	(1, 0.9, 1)	cc	1.21	0.85	0.93	0.93	0.93	0.95	0.9
		fc	1.11	0.78	0.93	0.93	0.94	0.95	0.9
		sc_HB	0.92	0.71	0.93	0.93	0.94	0.95	0.9
		sc_SL	0.93	0.72	0.93	0.93	0.94	0.95	0.9
(1, 1)	(1, 0.3, 1)	cc	1.17	0.95	0.93	0.93	0.93	0.95	0.9
		fc	1.07	0.88	0.93	0.93	0.94	0.95	0.9
		sc_HB	0.88	0.79	0.93	0.93	0.94	0.95	0.9

		sc_SL	0.88	0.79	0.93	0.93	0.94	0.95	0.9
		cc	1.17	0.95	0.93	0.93	0.93	0.95	0.9
	(1, 0.32, 1.14)	fc	1.07	0.88	0.93	0.93	0.94	0.95	0.9
		sc_HB	0.87	0.8	0.93	0.93	0.94	0.95	0.9
		sc_SL	0.88	0.8	0.93	0.93	0.94	0.95	0.9
		cc	1.2	0.96	0.93	0.93	0.93	0.95	0.9
	(1, 0.7, 1)	fc	1.1	0.87	0.93	0.93	0.94	0.95	0.9
		sc_HB	0.9	0.8	0.93	0.93	0.94	0.95	0.9
		sc_SL	0.91	0.8	0.93	0.93	0.94	0.95	0.9
		cc	1.21	0.95	0.93	0.93	0.93	0.95	0.9
	(1, 0.9, 1)	fc	1.11	0.88	0.93	0.93	0.94	0.95	0.9
		sc_HB	0.92	0.8	0.93	0.93	0.94	0.95	0.9
		sc_SL	0.92	0.8	0.93	0.93	0.94	0.95	0.9

cc=Concurrent Calibration, fc=Fixed Parameter Calibration, sc_HB: separate calibration with Haebara, sc_SL=separate calibration with Stocking and Lord , A= linking constant A, B = linking constant B, CC=Classification Consistency, CA=Classification Accuracy, FOE D1= First Order Equity Marginal Index, SOE D2= Second Order Equity Marginal Index, D12= combined index

<Y 60>

MU	Sigma	Linking			CC	CA	FOE_D1	SOE_D2	D12
		cc	1.17	-0.26	0.93	0.93	0.93	0.95	0.94
	(1, 0.3, 1)	fc	1.07	-0.23	0.93	0.93	0.94	0.95	0.94
		sc_HB	0.94	-0.19	0.93	0.93	0.94	0.95	0.94
		sc_SL	0.95	-0.20	0.93	0.93	0.94	0.95	0.94
		cc	1.17	-0.25	0.93	0.93	0.93	0.95	0.94
	(1, 0.32, 1.14)	fc	1.07	-0.23	0.93	0.93	0.94	0.95	0.94
		sc_HB	0.94	-0.20	0.93	0.93	0.94	0.95	0.94
(0, 0)		sc_SL	0.94	-0.20	0.93	0.93	0.94	0.95	0.94
		cc	1.21	-0.26	0.93	0.93	0.93	0.95	0.94
	(1, 0.7, 1)	fc	1.10	-0.23	0.93	0.93	0.94	0.95	0.94
		sc_HB	0.97	-0.19	0.93	0.93	0.94	0.95	0.94
		sc_SL	0.97	-0.19	0.93	0.93	0.94	0.95	0.94
		cc	1.22	-0.26	0.93	0.93	0.93	0.95	0.94
	(1, 0.9, 1)	fc	1.11	-0.22	0.93	0.93	0.94	0.95	0.94
		sc_HB	0.97	-0.19	0.93	0.93	0.94	0.95	0.94
		sc_SL	0.98	-0.19	0.93	0.93	0.94	0.95	0.94
		cc	1.19	-0.17	0.93	0.93	0.93	0.95	0.94
(0, 1)	(1, 0.3, 1)	fc	1.09	-0.14	0.93	0.93	0.94	0.95	0.94
		sc_HB	0.92	-0.11	0.93	0.93	0.94	0.95	0.93
		sc_SL	0.94	-0.12	0.93	0.93	0.94	0.95	0.93

	(1, 0.32, 1.14)	cc	1.19	-0.17	0.93	0.93	0.93	0.95	0.94
		fc	1.08	-0.14	0.93	0.93	0.94	0.95	0.94
		sc_HB	0.92	-0.11	0.93	0.93	0.94	0.95	0.93
		sc_SL	0.94	-0.12	0.93	0.93	0.94	0.95	0.93
	(1, 0.7, 1)	cc	1.21	-0.16	0.93	0.93	0.93	0.95	0.94
		f					0	0	
		c	.10	0.14	.93	.93	.94	.95	.94
		sc_HB	0.94	-0.11	0.93	0.93	0.94	0.95	0.93
		sc_SL	0.96	-0.11	0.93	0.93	0.94	0.95	0.93
	(1, 0.9, 1)	cc	1.22	-0.16	0.93	0.93	0.93	0.95	0.94
		fc	1.11	-0.13	0.93	0.93	0.94	0.95	0.94
		sc_HB	0.93	-0.11	0.93	0.93	0.94	0.95	0.93
		sc_SL	0.95	-0.11	0.93	0.93	0.94	0.95	0.93
(1, 0)	(1, 0.3, 1)	cc	1.15	0.61	0.93	0.93	0.93	0.95	0.93
		fc	1.05	0.56	0.93	0.93	0.94	0.95	0.92
		sc_HB	0.90	0.53	0.93	0.93	0.94	0.95	0.92
		sc_SL	0.91	0.54	0.93	0.93	0.94	0.95	0.92
	(1, 0.32, 1.14)	cc	1.15	0.61	0.93	0.93	0.93	0.95	0.92
		fc	1.05	0.57	0.93	0.93	0.94	0.95	0.93
		sc_HB	0.90	0.53	0.93	0.93	0.94	0.95	0.92
		sc_SL	0.91	0.54	0.93	0.93	0.94	0.95	0.92
	(1, 0.7, 1)	cc	1.20	0.62	0.93	0.93	0.93	0.95	0.93
		fc	1.08	0.57	0.93	0.93	0.94	0.95	0.93
		sc_HB	0.93	0.53	0.93	0.93	0.94	0.95	0.93
		sc_SL	0.94	0.54	0.93	0.93	0.94	0.95	0.93
	(1, 0.9, 1)	cc	1.20	0.62	0.93	0.93	0.93	0.95	0.93
		fc	1.10	0.56	0.93	0.93	0.94	0.95	0.93
		sc_HB	0.92	0.53	0.93	0.93	0.94	0.95	0.93
		sc_SL	0.95	0.54	0.93	0.93	0.94	0.95	0.93
(1, 1)	(1, 0.3, 1)	cc	1.15	0.71	0.93	0.93	0.93	0.95	0.92
		fc	1.06	0.65	0.93	0.93	0.94	0.95	0.92
		sc_HB	0.88	0.61	0.93	0.93	0.94	0.95	0.92
		sc_SL	0.89	0.61	0.93	0.93	0.94	0.95	0.92
	(1, 0.32, 1.14)	cc	1.16	0.72	0.93	0.93	0.93	0.95	0.92
		fc	1.06	0.66	0.93	0.93	0.94	0.95	0.92
		sc_HB	0.88	0.60	0.93	0.93	0.94	0.95	0.92
		sc_SL	0.89	0.61	0.93	0.93	0.94	0.95	0.92
	(1, 0.7, 1)	cc	1.19	0.72	0.93	0.93	0.93	0.95	0.93
		fc	1.09	0.67	0.93	0.93	0.94	0.95	0.93
		sc_HB	0.92	0.61	0.93	0.93	0.94	0.95	0.92
		sc_SL	0.92	0.62	0.93	0.93	0.94	0.95	0.92
	(1, 0.9, 1)	cc	1.21	0.72	0.93	0.93	0.93	0.95	0.93
		fc	1.10	0.66	0.93	0.93	0.94	0.95	0.93

sc_HB	0.93	0.61	0.93	0.93	0.94	0.95	0.92
sc_SL	0.93	0.62	0.93	0.93	0.94	0.95	0.92

cc=Concurrent Calibration, fc=Fixed Parameter Calibration, sc_HB: separate calibration with Haebara, sc_SL=separate calibration with Stocking and Lord , A= linking constant A, B = linking constant B, CC=Classification Consistency, CA=Classification Accuracy, FOE D1= First Order Equity Marginal Index, SOE D2= Second Order Equity Marginal Index, D12= combined index

APPENDIX C: APPROXIMATE MIRT TSE

ITEM PARAMETERS OF MC-I
<X BASE> REFERS TO <X BASE> DISPLAYED EARLIER.

<Y BASE>

Item ID	a1	a2	d	g	alpha	MDISC	MID	RC_Angle
1	0.88	0.88	-0.15	0.08	45	1.25	0.12	45
2	0.67	0.67	-0.68	0.2	45	0.95	0.71	45
3	0.84	0.84	0.8	0.18	45	1.18	-0.68	45
4	0.81	0.81	-0.35	0.1	45	1.15	0.3	45
5	0.64	0.64	0.41	0.15	45	0.9	-0.45	45
6	0.71	0.71	-0.06	0.16	45	1	0.06	45
7	1.06	1.06	0.05	0.19	45	1.5	-0.03	45
8	0.85	0.85	-0.44	0.09	45	1.2	0.37	45
9	0.79	0.79	-0.85	0.18	45	1.12	0.76	45
10	0.76	0.76	0.39	0.16	45	1.07	-0.36	45
11	0.66	0.66	1.37	0.11	45	0.94	-1.46	45
12	1.07	1.07	0.18	0.06	45	1.51	-0.12	45
13	0.8	0.8	0.1	0.17	45	1.14	-0.09	45
14	0.72	0.72	0.43	0.13	45	1.02	-0.42	45
15	0.93	0.93	0.25	0.14	45	1.32	-0.19	45
16	0.87	0.87	-0.76	0.07	45	1.23	0.62	45
17	1.01	1.01	0.08	0.15	45	1.43	-0.05	45
18	0.58	0.58	0.35	0.19	45	0.82	-0.43	45
19	0.91	0.91	-0.01	0.15	45	1.28	0.01	45
20	0.9	0.9	1.46	0.1	45	1.27	-1.16	45
21	0.63	0.63	-0.17	0.11	45	0.89	0.19	45
22	1.13	1.13	-0.6	0.16	45	1.6	0.38	45
23	0.69	0.69	0.65	0.12	45	0.97	-0.67	45
24	0.74	0.74	0.06	0.1	45	1.05	-0.06	45
25	1.02	1.02	1.35	0.13	45	1.45	-0.93	45
26	1.12	1.12	0.81	0.17	45	1.58	-0.51	45
27	0.73	0.73	0.47	0.15	45	1.04	-0.45	45
28	1.08	1.08	0.56	0.19	45	1.53	-0.37	45
29	0.97	0.97	1.1	0.11	45	1.37	-0.8	45
30	0.62	0.62	0.69	0.15	45	0.87	-0.79	45
31	1	1	1.35	0.14	45	1.41	-0.96	45
32	0.98	0.98	1.11	0.17	45	1.38	-0.8	45

33	0.65	0.65	-0.46	0.13	45	0.92	0.5	45
34	1.11	1.11	-0.18	0.17	45	1.56	0.12	45
35	0.94	0.94	0.08	0.12	45	1.33	-0.06	45
36	0.59	0.59	-0.88	0.14	45	0.84	1.05	45
37	1.05	1.05	0.22	0.1	45	1.48	-0.15	45
38	0.99	0.99	0.7	0.18	45	1.4	-0.5	45
39	1.14	1.14	1.26	0.14	45	1.61	-0.78	45
40	0.92	0.92	0.91	0.2	45	1.3	-0.7	45
41	0.86	0.86	-0.7	0.18	45	1.22	0.58	45
42	0.83	0.83	-0.34	0.16	45	1.17	0.29	45
43	0.57	0.57	-0.82	0.18	45	0.81	1.02	45
44	1.09	1.09	1.41	0.18	45	1.55	-0.91	45
45	0.77	0.77	0.1	0.18	45	1.09	-0.09	45
46	0.7	0.7	0.72	0.19	45	0.99	-0.73	45
47	0.6	0.6	1.47	0.09	45	0.86	-1.72	45
48	0.95	0.95	-0.78	0.13	45	1.35	0.58	45
49	1.04	1.04	1.21	0.15	45	1.46	-0.83	45
50	0.78	0.78	0.15	0.14	45	1.1	-0.14	45

Alpha= item measurement direction, MDISC=item discrimination in MIRT, MID=item difficulty in MIRT, RC_Angle = test measurement direction

<Y SAME RC>

Item ID	a1	a2	d	g	alpha	MDISC	MID	RC_Angle
1	0.57	0.99	0.84	0.17	60	1.14	-0.74	45
2	0.59	1.03	1.44	0.18	60	1.19	-1.21	45
3	0.62	1.07	-0.7	0.15	60	1.24	0.57	45
4	0.64	1.11	1.09	0.2	60	1.28	-0.85	45
5	0.67	1.15	0.15	0.18	60	1.33	-0.12	45
6	0.69	1.19	1.13	0.1	60	1.38	-0.82	45
7	0.71	1.23	0.45	0.12	60	1.43	-0.31	45
8	0.74	1.28	1.41	0.19	60	1.47	-0.95	45
9	0.76	1.32	1.39	0.15	60	1.52	-0.91	45
10	0.78	1.36	-0.25	0.19	60	1.57	0.16	45
11	0.81	1.4	0.86	0.19	60	1.62	-0.53	45
12	0.83	1.44	-0.78	0.18	60	1.66	0.47	45
13	0.86	1.48	1.26	0.16	60	1.71	-0.73	45
14	0.88	1.52	-0.62	0.16	60	1.76	0.35	45
15	0.9	1.56	0.41	0.18	60	1.81	-0.23	45
16	0.93	1.6	0.52	0.15	60	1.85	-0.28	45
17	0.95	1.65	-0.62	0.14	60	1.9	0.33	45
18	0.97	1.69	0.3	0.16	60	1.95	-0.15	45

19	1	1.73	-0.05	0.11	60	2	0.03	45
20	1.02	1.77	-0.94	0.18	60	2.04	0.46	45
21	1.05	1.81	-0.83	0.15	60	2.09	0.4	45
22	1.07	1.85	1.21	0.1	60	2.14	-0.57	45
23	1.09	1.89	1.07	0.1	60	2.19	-0.49	45
24	1.12	1.93	0.06	0.1	60	2.23	-0.03	45
25	1.14	1.97	1.31	0.04	60	2.28	-0.57	45
26	0.99	0.57	0.22	0.12	30	1.14	-0.19	45
27	1.03	0.59	0.21	0.19	30	1.19	-0.18	45
28	1.07	0.62	0.66	0.09	30	1.24	-0.54	45
29	1.11	0.64	-0.55	0.1	30	1.28	0.43	45
30	1.15	0.67	-0.72	0.08	30	1.33	0.54	45
31	1.19	0.69	1.16	0.07	30	1.38	-0.84	45
32	1.23	0.71	-0.36	0.19	30	1.43	0.25	45
33	1.28	0.74	0.41	0.19	30	1.47	-0.28	45
34	1.32	0.76	1.13	0.11	30	1.52	-0.74	45
35	1.36	0.78	-0.09	0.15	30	1.57	0.06	45
36	1.4	0.81	1.44	0.13	30	1.62	-0.89	45
37	1.44	0.83	1.18	0.13	30	1.66	-0.71	45
38	1.48	0.86	0.9	0.1	30	1.71	-0.52	45
39	1.52	0.88	-0.4	0.14	30	1.76	0.23	45
40	1.56	0.9	0.94	0.17	30	1.81	-0.52	45
41	1.6	0.93	1.19	0.13	30	1.85	-0.64	45
42	1.65	0.95	-0.84	0.17	30	1.9	0.44	45
43	1.69	0.97	-0.46	0.19	30	1.95	0.24	45
44	1.73	1	1.14	0.19	30	2	-0.57	45
45	1.77	1.02	0.65	0.18	30	2.04	-0.32	45
46	1.81	1.05	1.18	0.13	30	2.09	-0.56	45
47	1.85	1.07	1.04	0.07	30	2.14	-0.49	45
48	1.89	1.09	-0.06	0.19	30	2.19	0.03	45
49	1.93	1.12	-0.73	0.17	30	2.23	0.33	45
50	1.97	1.14	0.46	0.16	30	2.28	-0.2	45

Alpha= item measurement direction, MDISC=item discrimination in MIRT, MID=item difficulty in MIRT, RC_Angle = test measurement direction

<Y DIFF RC>

Item ID	a1	a2	d	g	alpha	MDISC	MID	RC_Angle
1	0.29	0.78	1.32	0.17	70	0.83	-1.59	51.68
2	0.3	0.82	-0.39	0.1	70	0.87	0.44	51.68
3	0.31	0.85	-0.41	0.13	70	0.9	0.45	51.68

4	0.32	0.88	0.08	0.14	70	0.94	-0.08	51.68
5	0.33	0.91	-0.13	0.09	70	0.97	0.14	51.68
6	0.34	0.95	1.17	0.19	70	1.01	-1.16	51.68
7	0.36	0.98	-0.52	0.19	70	1.04	0.5	51.68
8	0.37	1.01	0.26	0.19	70	1.08	-0.24	51.68
9	0.38	1.04	0.1	0.19	70	1.11	-0.09	51.68
10	0.39	1.08	0.43	0.09	70	1.15	-0.38	51.68
11	0.4	1.11	0.66	0.14	70	1.18	-0.56	51.68
12	0.42	1.14	1.45	0.16	70	1.22	-1.19	51.68
13	0.43	1.17	-0.38	0.1	70	1.25	0.3	51.68
14	0.44	1.21	-0.36	0.18	70	1.28	0.28	51.68
15	0.45	1.24	1	0.18	70	1.32	-0.76	51.68
16	0.46	1.27	1.09	0.08	70	1.35	-0.8	51.68
17	0.48	1.31	-0.67	0.15	70	1.39	0.48	51.68
18	0.49	1.34	1.41	0.19	70	1.42	-0.99	51.68
19	0.5	1.37	-0.59	0.07	70	1.46	0.4	51.68
20	0.51	1.4	-0.6	0.13	70	1.49	0.4	51.68
21	0.52	1.44	1.32	0.15	70	1.53	-0.86	51.68
22	0.53	1.47	0.29	0.16	70	1.56	-0.19	51.68
23	0.55	1.5	-0.57	0.12	70	1.6	0.36	51.68
24	0.56	1.53	-0.1	0.14	70	1.63	0.06	51.68
25	0.57	1.57	0.67	0.14	70	1.67	-0.4	51.68
26	0.78	0.66	-0.19	0.19	40	1.02	0.18	51.68
27	0.82	0.68	-0.09	0.1	40	1.06	0.08	51.68
28	0.85	0.71	-0.59	0.09	40	1.11	0.53	51.68
29	0.88	0.74	0.66	0.2	40	1.15	-0.58	51.68
30	0.91	0.77	-0.5	0.15	40	1.19	0.42	51.68
31	0.95	0.79	1.39	0.14	40	1.24	-1.12	51.68
32	0.98	0.82	1.25	0.11	40	1.28	-0.98	51.68
33	1.01	0.85	1.36	0.08	40	1.32	-1.03	51.68
34	1.04	0.88	0.82	0.17	40	1.36	-0.6	51.68
35	1.08	0.9	-0.04	0.16	40	1.41	0.03	51.68
36	1.11	0.93	0.96	0.2	40	1.45	-0.67	51.68
37	1.14	0.96	-0.92	0.16	40	1.49	0.62	51.68
38	1.17	0.99	1.35	0.08	40	1.53	-0.88	51.68
39	1.21	1.01	1.48	0.11	40	1.58	-0.94	51.68
40	1.24	1.04	-0.07	0.15	40	1.62	0.05	51.68
41	1.27	1.07	0.88	0.04	40	1.66	-0.53	51.68
42	1.31	1.1	0.99	0.15	40	1.7	-0.58	51.68
43	1.34	1.12	0.78	0.12	40	1.75	-0.45	51.68
44	1.37	1.15	0.22	0.19	40	1.79	-0.12	51.68

45	1.4	1.18	0.26	0.17	40	1.83	-0.14	51.68
46	1.44	1.2	-0.2	0.11	40	1.87	0.1	51.68
47	1.47	1.23	0.75	0.15	40	1.92	-0.39	51.68
48	1.5	1.26	1.07	0.17	40	1.96	-0.54	51.68
49	1.53	1.29	0.12	0.15	40	2	-0.06	51.68
50	1.57	1.31	0.31	0.2	40	2.04	-0.15	51.68

Alpha= item measurement direction, MDISC=item discrimination in MIRT, MID=item difficulty in MIRT, RC_Angle = test measurement direction

<LINEAR COMPOSITE IRT>

	X Base			Y Base			Y Same RC			Y Diff RC		
Item ID	a	b	g	a	b	g	a	b	g	a	b	g
1	0.85	0.12	0.15	1.24	0.12	0.08	1.06	-0.76	0.17	0.77	-1.67	0.17
2	0.43	0.09	0.11	0.95	0.72	0.2	1.09	-1.26	0.18	0.8	0.47	0.1
3	0.81	-0.27	0.18	1.19	-0.67	0.18	1.14	0.59	0.15	0.83	0.48	0.13
4	1.34	-0.92	0.15	1.15	0.31	0.1	1.17	-0.88	0.2	0.85	-0.09	0.14
5	1.17	-1	0.15	0.91	-0.45	0.15	1.22	-0.12	0.18	0.88	0.14	0.09
6	1.5	-0.01	0.12	1	0.06	0.16	1.25	-0.85	0.1	0.91	-1.22	0.19
7	1.02	-0.93	0.12	1.5	-0.03	0.19	1.29	-0.33	0.12	0.94	0.52	0.19
8	1.09	-1.29	0.16	1.2	0.37	0.09	1.33	-0.99	0.19	0.97	-0.25	0.19
9	1.33	-0.08	0.16	1.12	0.76	0.18	1.37	-0.95	0.15	0.99	-0.1	0.19
10	1.13	-0.71	0.09	1.07	-0.36	0.16	1.4	0.17	0.19	1.02	-0.39	0.09
11	1.47	-0.02	0.07	0.93	-1.47	0.11	1.44	-0.55	0.19	1.05	-0.59	0.14
12	0.95	0.16	0.17	1.51	-0.12	0.06	1.47	0.49	0.18	1.08	-1.26	0.16
13	0.91	-0.99	0.19	1.13	-0.09	0.17	1.52	-0.76	0.16	1.1	0.32	0.1
14	0.93	0.48	0.16	1.02	-0.42	0.13	1.55	0.37	0.16	1.13	0.29	0.18
15	1.03	-0.77	0.16	1.32	-0.19	0.14	1.58	-0.24	0.18	1.16	-0.8	0.18
16	1.44	0.45	0.16	1.23	0.62	0.07	1.62	-0.29	0.15	1.18	-0.85	0.08
17	1.2	0.29	0.2	1.43	-0.06	0.15	1.65	0.34	0.14	1.22	0.51	0.15
18	1.57	0.38	0.15	0.82	-0.43	0.19	1.68	-0.16	0.16	1.24	-1.04	0.19
19	0.99	0.36	0.17	1.29	0.01	0.15	1.72	0.03	0.11	1.26	0.43	0.07
20	1.39	0.58	0.16	1.27	-1.15	0.1	1.74	0.48	0.18	1.28	0.42	0.13
21	1.58	-0.39	0.12	0.89	0.19	0.11	1.78	0.41	0.15	1.31	-0.91	0.15
22	1.12	-1.07	0.12	1.6	0.38	0.16	1.81	-0.59	0.1	1.33	-0.2	0.16
23	0.83	-1.15	0.18	0.98	-0.67	0.12	1.83	-0.51	0.1	1.36	0.38	0.12
24	0.89	-1.12	0.15	1.05	-0.06	0.1	1.87	-0.03	0.1	1.38	0.06	0.14
25	1.6	-0.11	0.11	1.44	-0.94	0.13	1.9	-0.6	0.04	1.4	-0.42	0.14
26	1.22	-0.04	0.18	1.58	-0.51	0.17	1.06	-0.2	0.12	0.98	0.19	0.19
27	1.23	-0.85	0.1	1.03	-0.46	0.15	1.09	-0.18	0.19	1.02	0.09	0.1
28	1.3	-0.41	0.19	1.53	-0.37	0.19	1.14	-0.55	0.09	1.06	0.54	0.09

29	1.27	-0.51	0.16	1.37	-0.8	0.11	1.17	0.44	0.1	1.1	-0.59	0.2
30	1.15	0.07	0.16	0.88	-0.79	0.15	1.22	0.56	0.08	1.14	0.43	0.15
31	1.32	0.22	0.13	1.41	-0.95	0.14	1.25	-0.87	0.07	1.17	-1.15	0.14
32	1.19	-1.25	0.15	1.39	-0.8	0.17	1.29	0.26	0.19	1.21	-1	0.11
33	0.92	-0.65	0.15	0.92	0.5	0.13	1.33	-0.29	0.19	1.25	-1.05	0.08
34	1.37	0.31	0.11	1.57	0.11	0.17	1.37	-0.77	0.11	1.29	-0.61	0.17
35	1.4	0.45	0.16	1.33	-0.06	0.12	1.4	0.06	0.15	1.32	0.03	0.16
36	1.41	-0.16	0.09	0.83	1.05	0.14	1.44	-0.92	0.13	1.36	-0.68	0.2
37	1.53	-0.86	0.13	1.48	-0.15	0.1	1.47	-0.74	0.13	1.4	0.63	0.16
38	1.1	-0.47	0.12	1.4	-0.5	0.18	1.52	-0.54	0.1	1.44	-0.9	0.08
39	1.48	-0.97	0.13	1.61	-0.78	0.14	1.55	0.24	0.14	1.47	-0.96	0.11
40	1.54	-0.54	0.19	1.3	-0.7	0.2	1.58	-0.54	0.17	1.51	0.04	0.15
41	1.24	0.06	0.15	1.22	0.58	0.18	1.62	-0.67	0.13	1.54	-0.54	0.04
42	1	-0.11	0.18	1.17	0.29	0.16	1.65	0.46	0.17	1.58	-0.59	0.15
43	1.51	0.39	0.19	0.81	1.02	0.18	1.68	0.24	0.19	1.61	-0.46	0.12
44	1.05	0.88	0.14	1.54	-0.91	0.18	1.72	-0.59	0.19	1.65	-0.13	0.19
45	0.82	-0.98	0.08	1.09	-0.09	0.18	1.74	-0.33	0.18	1.68	-0.14	0.17
46	0.88	0.8	0.13	0.99	-0.73	0.19	1.78	-0.58	0.13	1.71	0.11	0.11
47	0.98	-0.14	0.17	0.85	-1.73	0.09	1.81	-0.5	0.07	1.75	-0.4	0.15
48	1.61	-0.38	0.13	1.34	0.58	0.13	1.83	0.03	0.19	1.78	-0.56	0.17
49	1.07	-1.38	0.17	1.47	-0.82	0.15	1.87	0.34	0.17	1.82	-0.06	0.15
50	1.29	-0.2	0.19	1.1	-0.14	0.14	1.9	-0.21	0.16	1.85	-0.15	0.2

ITEM PARAMETERS OF MC-II

<X BASE> REFERS TO <X BASE> DISPLAYED EARLIER.
<Y BASE>

Item ID	1	2			alpha	MDISC	MID	CVI	RC_Angle
1	0.82	0.30	0.20	0.13	19.95	0.87	-0.22	0.95	7.69
2	1.05	0.37	-1.43	0.14	19.52	1.12	1.28	0.95	7.69
3	1.39	0.48	-0.04	0.08	19.09	1.47	0.03	0.96	7.69
4	1.33	0.45	1.41	0.12	18.65	1.40	-1.00	0.96	7.69
5	1.28	0.42	-0.19	0.12	18.19	1.35	0.14	0.96	7.69
6	1.20	0.38	-0.46	0.17	17.73	1.25	0.36	0.97	7.69
7	1.12	0.35	-0.37	0.19	17.25	1.18	0.32	0.97	7.69
8	1.48	0.45	-1.15	0.10	16.76	1.55	0.74	0.97	7.69
9	1.10	0.32	-1.43	0.16	16.26	1.14	1.25	0.98	7.69
10	0.61	0.17	-0.86	0.14	15.74	0.63	1.37	0.98	7.69
11	0.72	0.20	1.10	0.14	15.20	0.74	-1.48	0.98	7.69
12	0.68	0.18	0.33	0.05	14.65	0.70	-0.47	0.98	7.69

13	1.15	0.29	-0.78	0.15	14.07	1.18	0.66	0.99	7.69
14	0.59	0.14	-0.53	0.11	13.47	0.61	0.87	0.99	7.69
15	1.46	0.33	-1.17	0.14	12.84	1.50	0.78	0.99	7.69
16	0.77	0.17	-0.91	0.14	12.18	0.79	1.16	0.99	7.69
17	1.43	0.29	0.54	0.11	11.48	1.46	-0.37	0.99	7.69
18	1.48	0.28	0.32	0.10	10.73	1.51	-0.22	1.00	7.69
19	1.16	0.20	0.57	0.15	9.94	1.17	-0.48	1.00	7.69
20	0.75	0.12	1.40	0.13	9.07	0.76	-1.84	1.00	7.69
21	0.95	0.14	0.92	0.12	8.11	0.96	-0.95	1.00	7.69
22	0.75	0.09	1.31	0.08	7.02	0.76	-1.73	1.00	7.69
23	1.43	0.14	0.16	0.11	5.73	1.44	-0.11	1.00	7.69
24	0.62	0.04	1.11	0.14	4.05	0.62	-1.79	1.00	7.69
25	0.97	0.00	0.71	0.09	0.00	0.97	-0.73	0.98	7.69
26	1.00	0.00	0.85	0.12	0.00	1.00	-0.85	0.98	7.69
27	1.00	0.00	-0.06	0.14	0.00	1.00	0.06	0.98	7.69
28	1.00	0.00	0.05	0.20	0.00	1.00	-0.05	0.98	7.69
29	1.00	0.00	0.00	0.19	0.00	1.00	0.00	0.98	7.69
30	1.00	0.00	0.41	0.20	0.00	1.00	-0.41	0.98	7.69
31	1.00	0.00	-0.43	0.14	0.00	1.00	0.43	0.98	7.69
32	1.00	0.00	-1.16	0.18	0.00	1.00	1.16	0.98	7.69
33	1.00	0.00	-0.34	0.18	0.00	1.00	0.34	0.98	7.69
34	1.00	0.00	-0.18	0.11	0.00	1.00	0.18	0.98	7.69
35	1.00	0.00	0.80	0.18	0.00	1.00	-0.80	0.98	7.69
36	1.00	0.00	-0.27	0.17	0.00	1.00	0.27	0.98	7.69
37	1.00	0.00	0.88	0.18	0.00	1.00	-0.88	0.98	7.69
38	1.00	0.00	0.19	0.12	0.00	1.00	-0.19	0.98	7.69
39	1.00	0.00	-1.43	0.19	0.00	1.00	1.43	0.98	7.69
40	1.00	0.00	-0.15	0.17	0.00	1.00	0.15	0.98	7.69
41	1.00	0.00	0.21	0.17	0.00	1.00	-0.21	0.98	7.69
42	1.00	0.00	0.59	0.20	0.00	1.00	-0.59	0.98	7.69
43	1.00	0.00	-0.72	0.16	0.00	1.00	0.72	0.98	7.69
44	1.00	0.00	-1.09	0.19	0.00	1.00	1.09	0.98	7.69
45	1.00	0.00	-1.37	0.15	0.00	1.00	1.37	0.98	7.69
46	1.00	0.00	0.10	0.19	0.00	1.00	-0.10	0.98	7.69
47	1.00	0.00	0.64	0.19	0.00	1.00	-0.64	0.98	7.69
48	1.00	0.00	0.22	0.17	0.00	1.00	-0.22	0.98	7.69
49	1.00	0.00	0.22	0.16	0.00	1.00	-0.22	0.98	7.69
50	1.00	0.00	-0.10	0.19	0.00	1.00	0.10	0.98	7.69

Alpha= item measurement direction, MDISC=item discrimination in MIRT, MID=item difficulty in MIRT, CVI = validity sector index, RC_Angle = test measurement direction

<Y 60>

Item ID	a1	a2	d	g	alpha	MDISC	MID	CVI	RC_Angle
1	1.13	1.96	0.44	0.15	60	2.26	-0.2	0.37	29.4
2	1.32	2.16	0.35	0.14	58.61	2.53	-0.14	0.39	29.4
3	0.85	1.32	-1.34	0.06	57.2	1.57	0.85	0.41	29.4
4	0.63	0.93	-1.17	0.16	55.77	1.13	1.04	0.44	29.4
5	0.58	0.81	-1.2	0.11	54.31	1	1.2	0.46	29.4
6	0.85	1.12	0.37	0.11	52.83	1.41	-0.26	0.49	29.4
7	1.5	1.87	0.71	0.2	51.32	2.39	-0.3	0.52	29.4
8	1.03	1.22	-0.51	0.2	49.77	1.6	0.32	0.54	29.4
9	0.84	0.93	-1.05	0.18	48.19	1.25	0.84	0.57	29.4
10	0.81	0.86	-0.58	0.09	46.57	1.18	0.49	0.6	29.4
11	0.88	0.88	-0.26	0.15	44.9	1.25	0.21	0.63	29.4
12	0.52	0.49	0.74	0.16	43.18	0.71	-1.04	0.66	29.4
13	0.54	0.47	-1.31	0.15	41.41	0.72	1.83	0.68	29.4
14	1.38	1.14	-1.23	0.04	39.57	1.79	0.69	0.71	29.4
15	0.9	0.7	0.35	0.15	37.66	1.14	-0.31	0.74	29.4
16	0.97	0.7	-0.75	0.15	35.66	1.19	0.63	0.77	29.4
17	0.9	0.6	0.27	0.17	33.56	1.08	-0.25	0.8	29.4
18	1.2	0.73	-0.93	0.17	31.33	1.41	0.66	0.83	29.4
19	0.9	0.5	-0.54	0.18	28.96	1.03	0.52	0.86	29.4
20	1	0.5	-0.26	0.1	26.38	1.12	0.23	0.89	29.4
21	0.54	0.23	0.8	0.11	23.56	0.59	-1.37	0.92	29.4
22	0.8	0.3	-1.36	0.18	20.36	0.85	1.6	0.95	29.4
23	1.46	0.44	-0.75	0.15	16.6	1.53	0.49	0.97	29.4
24	1.44	0.3	-0.91	0.17	11.72	1.47	0.62	0.99	29.4
25	1.46	0	-0.56	0.1	0	1.46	0.38	0.98	29.4
26	1	0	0.37	0.19	0	1	-0.37	0.98	29.4
27	1	0	0.62	0.13	0	1	-0.62	0.98	29.4
28	1	0	0.16	0.11	0	1	-0.16	0.98	29.4
29	1	0	0.16	0.17	0	1	-0.16	0.98	29.4
30	1	0	-1.1	0.11	0	1	1.1	0.98	29.4
31	1	0	0.47	0.09	0	1	-0.47	0.98	29.4
32	1	0	0.24	0.07	0	1	-0.24	0.98	29.4
33	1	0	0.78	0.19	0	1	-0.78	0.98	29.4
34	1	0	-1.48	0.14	0	1	1.48	0.98	29.4
35	1	0	-0.04	0.19	0	1	0.04	0.98	29.4
36	1	0	-0.6	0.09	0	1	0.6	0.98	29.4
37	1	0	0.32	0.15	0	1	-0.32	0.98	29.4
38	1	0	0.37	0.18	0	1	-0.37	0.98	29.4
39	1	0	-0.86	0.16	0	1	0.86	0.98	29.4

40	1	0	-0.88	0.16	0	1	0.88	0.98	29.4
41	1	0	0.16	0.19	0	1	-0.16	0.98	29.4
42	1	0	-0.31	0.2	0	1	0.31	0.98	29.4
43	1	0	-0.57	0.13	0	1	0.57	0.98	29.4
44	1	0	-1.36	0.1	0	1	1.36	0.98	29.4
45	1	0	0.7	0.11	0	1	-0.7	0.98	29.4
46	1	0	0.46	0.19	0	1	-0.46	0.98	29.4
47	1	0	-0.32	0.1	0	1	0.32	0.98	29.4
48	1	0	-1.42	0.13	0	1	1.42	0.98	29.4
49	1	0	-0.25	0.11	0	1	0.25	0.98	29.4
50	1	0	-0.59	0.09	0	1	0.59	0.98	29.4

Alpha= item measurement direction, MDISC=item discrimination in MIRT, MID=item difficulty in MIRT, CVI = validity sector index, RC_Angle = test measurement direction

<Parameters of Linear Composite IRT>

	X Base			Y Base			Y 60		
Item ID	a	b	c	a	b	c	a	b	c
1	1.06	-0.03	0.15	0.84	0.19	0.13	1.28	0.29	0.15
2	1.15	0.97	0.15	1.07	-1.40	0.14	1.39	0.22	0.14
3	1.41	0.74	0.17	1.38	-0.04	0.08	1.12	-1.08	0.06
4	1.19	1.17	0.09	1.33	1.36	0.12	0.90	-1.05	0.16
5	0.77	-0.65	0.17	1.29	-0.18	0.12	0.83	-1.11	0.11
6	0.92	-0.16	0.18	1.21	-0.45	0.17	1.13	0.32	0.11
7	0.63	-0.68	0.12	1.14	-0.37	0.19	1.66	0.53	0.20
8	0.64	-0.51	0.07	1.48	-1.12	0.10	1.31	-0.45	0.20
9	1.15	1.07	0.14	1.11	-1.41	0.16	1.10	-0.98	0.18
10	0.86	-1.16	0.13	0.62	-0.86	0.14	1.06	-0.55	0.09
11	0.93	-0.29	0.15	0.74	1.09	0.14	1.14	-0.25	0.15
12	1.45	-0.27	0.14	0.70	0.33	0.05	0.68	0.73	0.16
13	0.89	-0.19	0.19	1.17	-0.78	0.15	0.69	-1.30	0.15
14	0.74	-0.97	0.11	0.61	-0.53	0.11	1.68	-1.17	0.04
15	1.20	-0.69	0.10	1.48	-1.16	0.14	1.11	0.35	0.15
16	0.52	0.48	0.16	0.78	-0.91	0.14	1.18	-0.74	0.15
17	1.20	0.95	0.06	1.45	0.54	0.11	1.08	0.27	0.17
18	1.31	-0.24	0.08	1.50	0.32	0.10	1.40	-0.93	0.17
19	1.02	-1.02	0.18	1.17	0.57	0.15	1.03	-0.54	0.18
20	0.82	-0.63	0.19	0.76	1.40	0.13	1.11	-0.26	0.10
21	1.25	0.91	0.12	0.96	0.91	0.12	0.58	0.80	0.11
22	0.81	-1.01	0.17	0.76	1.31	0.08	0.84	-1.35	0.18
23	1.28	0.87	0.13	1.43	0.16	0.11	1.41	-0.71	0.15

24	0.61	0.07	0.20	0.62	1.11	0.14	1.28	-0.83	0.17
25	1.27	-0.53	0.18	0.95	0.70	0.09	1.03	-0.46	0.10
26	0.99	-0.67	0.18	0.98	0.84	0.12	0.78	0.33	0.19
27	0.99	0.90	0.12	0.98	-0.06	0.14	0.78	0.56	0.13
28	0.99	-1.18	0.08	0.98	0.05	0.20	0.78	0.14	0.11
29	0.99	1.00	0.14	0.98	0.00	0.19	0.78	0.14	0.17
30	0.99	-1.25	0.18	0.98	0.41	0.20	0.78	-0.99	0.11
31	0.99	0.65	0.19	0.98	-0.43	0.14	0.78	0.42	0.09
32	0.99	1.08	0.17	0.98	-1.15	0.18	0.78	0.22	0.07
33	0.99	1.43	0.17	0.98	-0.34	0.18	0.78	0.70	0.19
34	0.99	-0.74	0.19	0.98	-0.18	0.11	0.78	-1.33	0.14
35	0.99	-1.42	0.17	0.98	0.80	0.18	0.78	-0.04	0.19
36	0.99	-0.82	0.19	0.98	-0.27	0.17	0.78	-0.54	0.09
37	0.99	-1.12	0.15	0.98	0.87	0.18	0.78	0.29	0.15
38	0.99	-0.34	0.17	0.98	0.18	0.12	0.78	0.33	0.18
39	0.99	-1.24	0.10	0.98	-1.42	0.19	0.78	-0.77	0.16
40	0.99	-0.65	0.18	0.98	-0.15	0.17	0.78	-0.79	0.16
41	0.99	0.20	0.08	0.98	0.21	0.17	0.78	0.14	0.19
42	0.99	-0.51	0.13	0.98	0.58	0.20	0.78	-0.28	0.20
43	0.99	1.32	0.14	0.98	-0.71	0.16	0.78	-0.51	0.13
44	0.99	-0.37	0.19	0.98	-1.08	0.19	0.78	-1.22	0.10
45	0.99	1.35	0.20	0.98	-1.36	0.15	0.78	0.63	0.11
46	0.99	-0.68	0.07	0.98	0.10	0.19	0.78	0.41	0.19
47	0.99	0.56	0.16	0.98	0.64	0.19	0.78	-0.29	0.10
48	0.99	-1.08	0.07	0.98	0.22	0.17	0.78	-1.27	0.13
49	0.99	0.92	0.16	0.98	0.21	0.16	0.78	-0.22	0.11
50	0.99	0.78	0.09	0.98	-0.10	0.19	0.78	-0.53	0.09

APPENDIX D: EQUATING RESULTS

EQUATING RESULTS of MC-I
<MIRT OSE>

X Base	Identity	Y Base	Y Same RC	Y Diff RC
0	0	0	0	0
1	1	1.11	0.67	1.96
2	2	2.13	1.62	3.16
3	3	3.14	2.56	4.32
4	4	4.16	3.49	5.44
5	5	5.18	4.34	6.52
6	6	6.2	5.22	7.55
7	7	7.22	6.12	8.58
8	8	8.24	7.02	9.58
9	9	9.26	7.93	10.56
10	10	10.27	8.85	11.51
11	11	11.29	9.77	12.43
12	12	12.3	10.7	13.32
13	13	13.31	11.63	14.18
14	14	14.31	12.58	15.03
15	15	15.31	13.54	15.87
16	16	16.3	14.5	16.69
17	17	17.29	15.48	17.51
18	18	18.28	16.47	18.31
19	19	19.27	17.45	19.12
20	20	20.26	18.44	19.93
21	21	21.24	19.42	20.75
22	22	22.23	20.4	21.57
23	23	23.22	21.39	22.4
24	24	24.21	22.37	23.25
25	25	25.2	23.35	24.1
26	26	26.19	24.34	24.96
27	27	27.18	25.34	25.84
28	28	28.17	26.35	26.72
29	29	29.17	27.37	27.62
30	30	30.16	28.4	28.53
31	31	31.16	29.46	29.45
32	32	32.16	30.54	30.38
33	33	33.16	31.64	31.32
34	34	34.16	32.77	32.28

35	35	35.17	33.93	33.24
36	36	36.18	35.11	34.21
37	37	37.19	36.32	35.19
38	38	38.2	37.54	36.18
39	39	39.22	38.78	37.17
40	40	40.23	40.02	38.17
41	41	41.25	41.25	39.17
42	42	42.27	42.45	40.17
43	43	43.28	43.62	41.18
44	44	44.29	44.75	42.2
45	45	45.29	45.81	43.22
46	46	46.28	46.82	44.25
47	47	47.26	47.76	45.31
48	48	48.22	48.63	46.39
49	49	49.17	49.43	47.54
50	50	50.1	50.24	48.94

<AMT>

	No Weight				Integer Value Weight			
X Base	Identity	Y Base	Y Same RC	Y Diff RC	Identity	Y Base	Y Same RC	Y Diff RC
8	8	8.26	9.06	8.58	8	8	9.07	8.53
9	9	9.48	10.44	9.74	9	9.52	10.09	9.63
10	10	10.64	11.56	10.78	10	11	11.6	10.27
11	11.02	11.78	12.54	11.79	11.02	12	12.25	11.73
12	11.99	12.87	13.45	12.73	11.99	13	13.22	12.39
13	13.01	13.96	14.28	13.65	13.01	14	14.18	13.06
14	14.03	14.97	15.09	14.52	14.03	15	14.81	13.73
15	15	16	15.89	15.48	15	16	15.76	15.18
16	16	17.04	16.67	16.37	16	17	16.42	15.87
17	17	18.03	17.45	17.34	17	18	17.35	16.83
18	18.03	19.03	18.26	18.21	18.03	19	18.31	17.5
19	19.03	19.98	19.07	19.17	19.03	20	18.5	18.45
20	19.98	20.98	19.84	20.2	19.98	21	19.45	19.2
21	20.98	21.9	20.73	21.1	20.98	22	20.43	20.63
22	22.01	22.91	21.59	22.36	22.01	23	21.41	21.48
23	22.96	23.89	22.5	23.5	22.96	24	22.4	22.42
24	24	24.84	23.45	24.53	24	25	22.9	23.89
25	25.01	25.86	24.43	25.87	25.01	26	23.91	25.43
26	25.97	26.84	25.46	27.29	25.97	27	25.23	26.78

27	27.01	27.83	26.66	28.34	27.01	28	26.28	28.06
28	28.01	28.77	27.95	29.48	28.01	29	27.67	29.64
29	29	29.83	29.5	30.43	29	30	30.03	30.88
30	30.01	30.77	31.17	31.39	30.01	31	32.04	32.11
31	31.01	31.77	32.52	32.33	31.01	32	33.51	32.95
32	32.01	32.76	33.66	33.16	32.01	33	34.57	34.05
33	32.99	33.8	34.69	34.02	32.99	34	35.57	35.3
34	34.03	34.82	35.68	34.93	34.03	35	36.57	35.75
35	34.99	35.82	36.62	35.83	34.99	36	37.55	37.13
36	36.04	36.79	37.55	36.62	36.04	37	38.55	38.02
37	37	37.84	38.39	37.49	37	38	39.51	38.55
38	38.04	38.84	39.21	38.46	38.04	39	39.77	39.58
39	39.03	39.79	40.05	39.37	39.03	40	40.75	40.67
40	39.98	40.83	40.92	40.28	39.98	41	41.75	41.28
41	41	41.84	41.7	41.1	41	42	42.3	41.86
42	42	42.84	42.5	42.04	42	43	43.28	43.41
43	42.99	43.81	43.28	42.9	42.99	44	43.82	44.01
44	44.01	44.79	44.03	43.82	44.01	45	44.8	44.67
45	44.99	45.76	44.81	44.71	44.99	46	45.31	45.61
46	46	46.69	45.6	45.68	46	47	46.32	46.29
47	47	47.58	46.41	46.68	47	47.63	46.82	47.25
48	48	48.45	47.28	47.71	48	48.67	47.88	48.49
49	49	49.27	48.3	48.81	49	49	48.64	49.37
50	49.95	49.97	49.87	49.94	49.95	50	50	50

Equated scores were truncated at the observed score 8 which is the largest integer score larger than the sum of guessings

<OSE OF LINEAR COMPOSITE IRT>

X Base	Identity	Y Base	Y Same RC	Y Diff RC
0	0	0	0	0
1	1	1.1	0.72	0.87
2	2	2.12	1.68	1.85
3	3	3.13	2.63	2.84
4	4	4.16	3.57	3.82
5	5	5.18	4.51	4.8
6	6	6.21	5.41	5.78
7	7	7.24	6.32	6.75
8	8	8.27	7.24	7.72
9	9	9.3	8.17	8.69
10	10	10.33	9.1	9.65

11	11	11.36	10.04	10.6
12	12	12.38	10.99	11.54
13	13	13.39	11.94	12.48
14	14	14.4	12.91	13.42
15	15	15.41	13.88	14.36
16	16	16.41	14.86	15.3
17	17	17.4	15.83	16.24
18	18	18.4	16.8	17.18
19	19	19.39	17.77	18.14
20	20	20.39	18.74	19.1
21	21	21.38	19.7	20.07
22	22	22.38	20.66	21.05
23	23	23.37	21.63	22.03
24	24	24.37	22.59	23.02
25	25	25.37	23.56	24.03
26	26	26.37	24.54	25.03
27	27	27.36	25.52	26.05
28	28	28.36	26.52	27.08
29	29	29.36	27.53	28.11
30	30	30.35	28.56	29.15
31	31	31.35	29.61	30.19
32	32	32.34	30.69	31.24
33	33	33.33	31.78	32.3
34	34	34.33	32.9	33.37
35	35	35.32	34.05	34.44
36	36	36.32	35.22	35.52
37	37	37.31	36.41	36.61
38	38	38.32	37.63	37.7
39	39	39.32	38.87	38.79
40	40	40.32	40.11	39.88
41	41	41.33	41.34	40.97
42	42	42.34	42.56	42.05
43	43	43.35	43.75	43.12
44	44	44.35	44.9	44.17
45	45	45.35	45.99	45.21
46	46	46.34	47.02	46.24
47	47	47.31	47.97	47.24
48	48	48.26	48.84	48.21
49	49	49.18	49.6	49.16
50	50	50.09	50.25	50.08

<TSE OF LINEAR COMPOSITE IRT>

X Base	Identity	Y Base	Y Same RC	Y Diff RC
0	0	0	0	0
1	1	0.98	0.99	0.97
2	2	1.97	1.98	1.94
3	3	2.95	2.97	2.91
4	4	3.93	3.96	3.87
5	5	4.92	4.95	4.84
6	6	5.9	5.94	5.81
7	7	6.89	6.93	6.78
8	8	8.08	7.56	7.78
9	9	9.24	8.23	8.77
10	10	10.34	9.04	9.72
11	11	11.39	9.94	10.66
12	12	12.42	10.88	11.59
13	13	13.43	11.86	12.51
14	14	14.43	12.85	13.43
15	15	15.43	13.84	14.35
16	16	16.42	14.83	15.28
17	17	17.41	15.82	16.21
18	18	18.4	16.8	17.15
19	19	19.39	17.77	18.1
20	20	20.38	18.73	19.06
21	21	21.37	19.69	20.02
22	22	22.37	20.65	21
23	23	23.37	21.61	21.99
24	24	24.36	22.57	22.99
25	25	25.36	23.54	23.99
26	26	26.36	24.52	25.01
27	27	27.36	25.51	26.03
28	28	28.35	26.51	27.06
29	29	29.35	27.52	28.1
30	30	30.34	28.56	29.14
31	31	31.33	29.62	30.19
32	32	32.33	30.7	31.25
33	33	33.32	31.8	32.31
34	34	34.31	32.94	33.39
35	35	35.31	34.1	34.47
36	36	36.31	35.29	35.56
37	37	37.31	36.51	36.65

38	38	38.31	37.75	37.75
39	39	39.32	39	38.84
40	40	40.33	40.26	39.94
41	41	41.34	41.5	41.03
42	42	42.36	42.73	42.11
43	43	43.37	43.91	43.18
44	44	44.38	45.05	44.23
45	45	45.38	46.12	45.27
46	46	46.36	47.12	46.28
47	47	47.33	48.03	47.27
48	48	48.26	48.83	48.22
49	49	49.15	49.52	49.14
50	50	50	50	50

EQUATING RESULTS of MC-II
<MIRT OSE>

Item ID	X Base	Identity	Y Base	Y 60 RC
1	0	0	0	0
2	1	1	0.85	1.14
3	2	2	1.82	2.17
4	3	3	2.79	3.2
5	4	4	3.76	4.23
6	5	5	4.72	5.27
7	6	6	5.67	6.32
8	7	7	6.62	7.38
9	8	8	7.57	8.45
10	9	9	8.5	9.52
11	10	10	9.42	10.6
12	11	11	10.34	11.69
13	12	12	11.27	12.8
14	13	13	12.2	13.92
15	14	14	13.13	15.05
16	15	15	14.07	16.19
17	16	16	15.02	17.33
18	17	17	15.97	18.49
19	18	18	16.92	19.63
20	19	19	17.89	20.78
21	20	20	18.87	21.93
22	21	21	19.86	23.08
23	22	22	20.86	24.23

24	23	23	21.88	25.38
25	24	24	22.92	26.52
26	25	25	23.97	27.66
27	26	26	25.05	28.8
28	27	27	26.15	29.93
29	28	28	27.26	31.06
30	29	29	28.4	32.19
31	30	30	29.54	33.31
32	31	31	30.7	34.42
33	32	32	31.87	35.51
34	33	33	33.04	36.59
35	34	34	34.19	37.65
36	35	35	35.34	38.69
37	36	36	36.47	39.69
38	37	37	37.57	40.67
39	38	38	38.65	41.6
40	39	39	39.69	42.5
41	40	40	40.71	43.36
42	41	41	41.69	44.19
43	42	42	42.65	44.97
44	43	43	43.59	45.72
45	44	44	44.5	46.42
46	45	45	45.4	47.11
47	46	46	46.29	47.77
48	47	47	47.18	48.39
49	48	48	48.08	49.03
50	49	49	49	49.63
51	50	50	49.97	50.27

<AMT>

	No Weight			Integer Value Weight		
X Base	Identity	Y Base2	Y 60 RC2	Identity	Y Base	Y 60 RC
8	8	7.63	8.66	8	8	8.53
9	9	8.44	10	9	8.52	9.67
10	10	9.43	11.11	10	9.52	10.84
11	11	10.48	11.97	11	10.52	11.41
12	12	11.56	12.63	12	11.53	12.37
13	13	12.65	13.2	13	12.53	12.74
14	14	13.72	13.78	14	13.53	13.64
15	15	14.77	14.57	15	14.53	14.14

16	16	15.8	15.71	16	15.53	14.86
17	17	16.8	16.89	17	16.53	16.5
18	18	17.77	18.14	18	17.53	17.62
19	19	18.72	19.41	19	18.53	18.7
20	20	19.65	20.74	20	19.53	20.4
21	21	20.57	22.04	21	20.53	21.48
22	22	21.47	23.38	22	21.09	22.55
23	23	22.36	24.67	23	22.09	23.83
24	24	23.24	26.01	24	23.08	25.48
25	25	24.11	27.2	25	24.07	26.67
26	26	24.98	28.54	26	25.06	28.3
27	27	25.85	29.75	27	25.53	29.24
28	28	26.73	31.02	28	26.53	31.1
29	29	27.6	32.26	29	27.53	32.1
30	30	28.49	33.4	30	28.53	33.62
31	31	29.4	34.65	31	29.53	34.99
32	32	30.3	35.77	32	30	35.95
33	33	31.23	36.97	33	31	37
34	34	32.17	38.05	34	32	38.68
35	35	33.12	39.11	35	33	39.78
36	36	34.09	40.05	36	34.52	40.11
37	37	35.07	40.94	37	35.52	41.61
38	38	36.06	41.76	38	36.09	42.79
39	39	37.08	42.4	39	37.1	43.13
40	40	38.11	42.65	40	38.11	43.37
41	41	39.15	43.07	41	39.12	43.76
42	42	40.23	43.58	42	40.13	44.23
43	43	41.33	44.13	43	41.54	45.08
44	44	42.45	44.81	44	42.55	45.66
45	45	43.61	45.56	45	44.11	46.26
46	46	44.8	46.33	46	45.15	46.88
47	47	46.03	47.11	47	46.56	47.74
48	48	47.3	47.94	48	47.6	48.27
49	49	48.62	48.84	49	48.69	48.72

Equated scores were truncated at the observed score 8 which is the largest integer score larger than the sum of guessings.

<OSE OF LINEAR COMPOSITE IRT>

X Base	Identity	Y Base	Y 60
0	0	0	0

1	1	0.85	1.08
2	2	1.82	2.1
3	3	2.79	3.12
4	4	3.75	4.15
5	5	4.71	5.19
6	6	5.67	6.23
7	7	6.61	7.28
8	8	7.56	8.34
9	9	8.49	9.41
10	10	9.41	10.5
11	11	10.33	11.59
12	12	11.26	12.69
13	13	12.19	13.82
14	14	13.12	14.96
15	15	14.06	16.11
16	16	15.01	17.27
17	17	15.96	18.44
18	18	16.93	19.61
19	19	17.9	20.77
20	20	18.88	21.94
21	21	19.87	23.1
22	22	20.87	24.26
23	23	21.89	25.41
24	24	22.92	26.56
25	25	23.98	27.7
26	26	25.05	28.84
27	27	26.14	29.97
28	28	27.25	31.1
29	29	28.37	32.22
30	30	29.51	33.32
31	31	30.66	34.42
32	32	31.82	35.5
33	33	32.97	36.56
34	34	34.12	37.6
35	35	35.26	38.6
36	36	36.38	39.58
37	37	37.48	40.52
38	38	38.56	41.43
39	39	39.61	42.3
40	40	40.63	43.13
41	41	41.62	43.92

42	42	42.58	44.68
43	43	43.52	45.41
44	44	44.44	46.11
45	45	45.35	46.8
46	46	46.26	47.45
47	47	47.16	48.13
48	48	48.07	48.8
49	49	48.99	49.44
50	50	49.97	50.2

<TSE OF LINEAR COMPOSITE IRT>

X Base	Identity	Y Base	Y 60
0	0	0	0
1	1	1.02	0.97
2	2	2.04	1.95
3	3	3.05	2.92
4	4	4.07	3.89
5	5	5.09	4.87
6	6	6.11	5.84
7	7	7.13	6.82
8	8	7.92	8
9	9	8.71	9.14
10	10	9.55	10.26
11	11	10.42	11.38
12	12	11.3	12.51
13	13	12.21	13.65
14	14	13.13	14.81
15	15	14.07	15.98
16	16	15.01	17.17
17	17	15.96	18.35
18	18	16.92	19.54
19	19	17.88	20.72
20	20	18.86	21.9
21	21	19.84	23.08
22	22	20.84	24.24
23	23	21.86	25.41
24	24	22.89	26.56
25	25	23.94	27.71
26	26	25.01	28.85
27	27	26.11	29.99

28	28	27.23	31.12
29	29	28.36	32.24
30	30	29.52	33.35
31	31	30.68	34.45
32	32	31.85	35.54
33	33	33.02	36.6
34	34	34.18	37.64
35	35	35.33	38.64
36	36	36.46	39.61
37	37	37.55	40.54
38	38	38.62	41.43
39	39	39.66	42.28
40	40	40.66	43.08
41	41	41.64	43.85
42	42	42.58	44.57
43	43	43.51	45.27
44	44	44.41	45.93
45	45	45.3	46.58
46	46	46.18	47.21
47	47	47.07	47.83
48	48	47.96	48.47
49	49	48.9	49.14
50	50	50	50