COMPARISON OF CONCURRENT AND SEPARATE

MULTIDIMENSIONAL IRT LINKING OF ITEM PARAMETERS

A THESIS

SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL

OF THE UNIVERSITY OF MINNESOTA

BY

Mayuko Kanada Simon

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

Mark L. Davison, Adviser

November 2008

Acknowlegdment

I wish to express sincere gratitude to Dr. Mark L. Davison. He has been a great advisor, and I am fortunate to have the privilege to be his student. I want to thank him for his patience and countless hours he spent sharing his insights and guiding this dissertation. I could not have finished without him.

I would like to express my appreciation to Dr. Ernest C. Davenport, Jr. for mentoring me throughout my graduate experience. I enjoyed stopping by his office to discuss anything. I would like to thank Dr. David J. Weiss for giving me the scholarship to purchase the TESTFACT software and valuable inputs for my understanding of IRT and this dissertation. I am indebted to Dr. Michael R. Harwell for his advice and financial support when I changed my major to Educational Psychology. Without him, I would not have started this journey. I would like to thank Dr. Michael C. Rodriguez and Dr. Charles J. Geyer for being a wonderful teacher and being on my committee.

Finally, my deepest appreciation and love goes to my husband, Gyorgy, for his love, support, and patience and to my parents and brothers for their constant love and support of all my endeavors.

Abstract

With the No Child Left Behind Act of 2001 (NCLB) and the concept of adequate yearly progress, measuring growth across years is becoming more important. In vertical scaling where two tests have different difficulty levels and are given to different grade level students and there may be construct shift between grades, the IRT assumption of unidimensionality would appear implausible.

There are a few studies comparing separate Multidimensional Item Response Theory (MIRT) linking methods, however, none of them have compared concurrent calibration and separate MIRT linking. The purpose of this simulation research is to compare the performance of concurrent calibration and four separate linking methods. Based on the results from the studies of Unidimensional IRT (UIRT) concurrent and separate estimation methods, it was predicted that, in MIRT linking, concurrent linking would perform better than separate linking methods when groups are equivalent. As in the unidimensional IRT situation, separate estimation was expected to perform better than concurrent calibration with the nonequivalent groups design.

Independent variables were; sample size, test length, group equivalence, correlation between the two ability dimensions, and five estimation methods of MIRT linking (concurrent calibration, the test characteristic function (TCF), the item characteristic function (ICF), the direct method, and Mins methods). RMSE and bias were applied as the indices of linking quality.

The results of this study suggest that concurrent calibration generally performs

better than separate linking methods even when groups were non-equivalent with 0.5 standard deviation difference between group means and the correlation among ability dimensions was high. Concurrent calibration benefited more from a larger sample size than did separate linking methods with respect to all item parameters, especially with a shorter test form. Among separate linking methods, the ICF method tended to perform better than other separate linking methods when groups were non-equivalent, while Min's method did not perform as well as other methods. With equivalent groups, all separate linking methods performed similarly. A discussion of the limitations of the study and possibilities for future research is included.

Contents

List of Tables

List of Figures

Chapter **1**

Introduction

1.1   General Background

Scores on tests, especially standardized tests, are often used in making important decisions. At the individual level, a student decides which college to apply to and/or attend. At the institutional level, test scores are used to assess abilities and/or skills for college admissions, professional certification, employment, etc. At the public policy level, for example, scores are used to determine how to allocate funding, and how to improve students' achievement through school curricula. Since the decisions made based on scores are very important to individuals and the public, the scores should reflect the most accurate estimates of abilities and/or skills. The standardized test scores should provide a fair and equitable evaluation (Cook & Eignor, 1991).

Because standardized tests are administered on multiple occasions, the security of the tests is crucial to obtaining fair and equitable scores. Multiple standardized tests administered on different dates should not have the same item sets, otherwise

later examinees may have the advantage of knowing the items presented in the test before its subsequent administration. To avoid this situation in which the same items appear on a test booklet as appeared on another booklet which was administered previously, different test forms are developed. A test form is a set of test questions that is built according to content and statistical test specifications, the guidelines in making the test.

The use of different test forms helps secure the test items, however, it creates issues of fairness and equity across test forms. Different forms with the same test specification do not guarantee that the difficulties of each form are the same. It is almost impossible to create multiple forms of a test with exactly the same difficulty. The test scores from different forms should be on the same scale so that test scores on different forms can be used interchangeably. The statistical process of adjusting item parameter estimates from different forms so as to be on the same scale is called linking, and it is the focus of this paper.

Linking procedures exist in both classical test theory (CTT) and item response theory (IRT). An examinee's ability level is now usually estimated by IRT models in most of the standardized tests. Therefore this paper will focus on IRT linking procedures. The term linking is used in this paper to express explicitly the transformation of item parameters in IRT.

Most IRT models are used for tests that measure only one ability, which is unidimensional IRT (UIRT) with the assumption of unidimensionality of responses. How-

ever, often items and/or item sets may measure more than one ability of examinees no matter how carefully items are constructed (Ackerman, 1992; DeMars, 2006; Reckase, 1985). Multidimensionality might exist especially when multiple-choice and constructed-response items are calibrated together and when more grade levels are calibrated simultaneously (Patz & Yao, 2007). Consequently, multidimensional IRT (MIRT) has been developed (McKinley & Reckase, 1983; Sympson, 1978).

1.2   Application of MIRT

According to Reckase (1997), there are at least three useful applications of MIRT: 1) the investigation of the skill structure needed to respond to test items; 2) the description of differential item functioning (DIF); and 3) the selection of items that fit the unidimensional IRT model. One recent example of the first application is an investigation of testlet effects using MIRT(DeMars, 2006). A testlet is a group of items that center around a common stimulus (DeMars, 2006). The items in a testlet often violate the assumption of local independence due to the background knowledge, skills, interests, or motivation level specific to the testlet. (Sireci et al., 1991; Wainer & Kiely, 1987; Wang & Wilson, 2005; Yen, 1993). DeMars (2006) employed two-dimensional MIRT to examine the effect of testlets.

Reckase (1997) used MIRT to investigate the skill underlying test items. Walker and Beretvas (2003) used real data from a large-scale mathematics test and fit a two dimensional MIRT model. Dimensions considered were general mathematical ability

and the ability to communicate in mathematics. The study compared proficiency classifications under MIRT and UIRT and found that students with less mathematics communication ability were more likely to be placed in a lower general mathematics proficiency classification under UIRT than MIRT.

Yao and Boughton (2007) conducted a simulation study of dichotomous and polytomous MIRT for subscale score proficiency estimation using real data-derived parameters from a large-scale statewide assessment. Four dimensions with simple structure were considered in their study: 1) number sense and computational techniques, 2) algebra, patterns, and functions, 3) statistics and probability, and 4) geometry and measurement. The study found that to report accurate diagnostic information at the subscale level, the subscales need to be highly correlated, borrow information from other subscales, implement a multidimensional approach. With the No Child Left Behind Act of 2001 (NCLB) and the concept of adequate yearly progress, measuring growth across years is becoming more important. When tests given to different grades are linked, it is called vertical scaling. In vertical scaling, two tests have different difficulty levels and are given to different grade level students. There may be construct shifts between grades. There could be a situation where one test contains trigonometry, while another test to be linked (or scaled, in vertical scaling) does not contain trigonometry. In this situation, the IRT assumption of unidimensionality would appear implausible. Due to multidimensionality in vertical scaling, concurrent calibration of unidimensional IRT models may fail to perform well in practice (Patz

& Yao, 2007). Therefore, MIRT is considered useful in vertical scaling to account for any construct shift (Harris, 2007; Patz & Yao, 2007; Yon, 2006).

1.3   Linking: Placing parameter estimates on a common scale

Item parameters are generally estimated in such a way that the mean of the ability levels of the examinees taking a form equals zero with a standard deviation of one. Item parameters can be estimated using data from a common-item linking design either separately for each form or concurrently across forms. When two groups of examinees differ in ability levels, and when item parameters are estimated separately for each form, the units of the item parameters are not on the same scale because the examinees' mean ability levels and the standard deviations are not equal. When the distributions of ability levels are the same for two groups (equivalent groups) taking equivalent forms, no transformation of item parameters is necessary.

1.4   Concurrent Estimation

The transformation of item parameters is automatically performed when software can estimate item parameters simultaneously, which is called concurrent item calibration (Wingersky & Lord, 1984). In concurrent calibration, all estimated parameters are on the same scale since they are estimated simultaneously. To perform concurrent calibration, proper software such as MULTILOG (Thissen, 1991) or BILOG-MG (Zimowski et al., 1996) has to be used for UIRT. TESTFACT (Wood et al., 1987) can

be used for MIRT concurrent estimation.

The procedures used in concurrent calibration are joint maximum likelihood estimation (JMLE) and marginal maximum likelihood estimation (MMLE). Bayes estimation can also be used in either JMLE or MMLE. A concurrent procedure estimates item and ability level parameters combining data from more than one group and treating items not taken by a particular group as missing (Lord, 1980). There are variations of this procedure in which parameter estimates of the common items from the base groups are fixed and the uncommon item parameters are estimated using target group data (Kim & Cohen, 1998).

1.5   MIRT and Linking

Multidimensional item response theory (MIRT) is not yet popular in testing due, in part, to the lack of research on MIRT linking. MIRT linking is a relatively new area of interest, and research in this area has not been conducted as extensively as in unidimensional IRT linking. However, there are some papers that describe MIRT linking methods in detail (Davey et al., 1996; Hirsch, 1989; Li & Lissitz, 2000; Min, 2003; Oshima et al., 2000; Thompson et al., 1997; Yon, 2006).

In unidimensional IRT, the scale of Form Y is linearly transformed onto the scale of Form X. MIRT linking is also a linear transformation, but it is a linear transformation of matrices. MIRT linking involves, permutation, rotation, and reversal of reference systems, re-scaling the unit length (similar to slope in unidimensional IRT

linking), and shifting the point of origin (similar to the intercept in unidimensional IRT linking).

So far, the studies of MIRT linking are distinguished from each other by having different models and estimation procedures. The MIRT linking models differ in whether there is an orthogonal or non-orthogonal rotation matrix and in whether there is no dilating parameter, a single dilation parameter, or several dilation parameters to rescale the unit along each dimension. The estimation of transformation matrices and parameters can be done simultaneously or separately, and there is usually more than one way of obtaining the estimates.

The present study considered five different MIRT linking methods including concurrent calibration: (a) separate linking using Min's method (Min, 2003), (b) separate linking using the Multidimensional Test Characteristic Function (TCF) (Oshima et al., 2000), (c) separate linking using the Multidimensional Item Characteristic Function (ICF) (Oshima et al., 2000), (d) the Direct method (Oshima et al., 2000),, and (e) concurrent calibration.

In Li and Lissitz (2000), the transformation of one matrix onto another scale involves an orthogonal Procrustes rotation, a translation vector, and a single dilation parameter. Min's method is an extension of the model by Li and Lissitz (2000) obtained by replacing a single dilation parameter with a diagonal dilation matrix to allow different unit changes in different dimensions. The TCF method is an extension of the Stocking and Lord (1983) unidimensional IRT linking procedure (Oshima et al.,

2000). The Stocking and Lord procedure re-scales the parameters by minimizing the cumulative squared difference between the test characteristic curves over items for examinees of a particular ability. The ICF is an extension of the unidimensional IRT linking of Haebara(Haebara, 1980; Oshima et al., 2000). As in the unidimensional case, it minimizes the cumulative squared difference between the item characteristic curves for each item for examinees of a particular ability. The direct method is an extension of the unidimensional IRT equation that minimize the sum of squared differences between the two sets of common item parameter estimates (Divgi, 1985; Oshima et al., 2000). Min's method uses an orthogonal rotation of item discrimination parameters, while the TCF and the ICF allow a non-orthogonal rotation matrix. As in unidimensional IRT linking, concurrent calibration is possible for MIRT linking. Under concurrent calibration, since all items in different forms are calibrated simultaneously, parameter estimates from the two tests are on a common scale after one run.

The findings of unidimensional studies comparing separate linking and concurrent calibration are rather mixed. Kolen and Brennan (2004) concluded that studies comparing concurrent and separate estimation show that concurrent calibrations are more accurate than separate estimation when the data fit the IRT model. Concurrent calibration is, however, less robust to violations of the IRT assumptions than separate estimation using test characteristic curve methods such as the Stocking and Lord method (Kolen & Brennan, 2004).

However, there are some findings that concurrent calibration performs better than separate linking methods with assumption violations. Hanson and Béguin (2002) found that the bias in the item characteristic curve was larger with concurrent estimation than separate linking methods when the number of common items was small using MULTILOG, but not when BILOG-MG was used. Concurrent estimation generally had lower error than separate estimation (Hanson & Béguin, 1999; Béguin & Hanson, 2001; Hanson & Béguin, 2002). Recent studies found that concurrent estimation had smaller error than separate estimation with all conditions examined for the graded response model (Kim & Cohen, 2002).

1.6   Purpose of the Study

MIRT equating is not yet popular, due, in part, to the lack of research in MIRT linking. There are a few studies comparing separate linking methods, however, none of them have compared concurrent calibration and separate linking at the time this study was conducted. The purpose of this simulation research is assessing the performance of concurrent calibration and separate linking methods. Based on the results from the unidimensional studies on concurrent and separate estimations, it is expected that, in MIRT linking, concurrent estimation will perform better than separate estimation when the assumptions are met or when there is a slight violation of dimensionality. As in the unidimensional IRT situation, separate estimation is expected to perform better than concurrent estimation in the nonequivalent groups design. In this study,

the research questions are as follows:

1. How do separate linking methods and concurrent calibration compare in performance under various conditions: e.g., equivalent groups and zero correlation among abilities; or non-equivalent groups and non-zero correlation among abilities?

2. How is the performance of separate and concurrent calibration affected by varying sample sizes and test length?

3. Which separate linking method performs better than other separate linking methods?

Literature Review

2.1   Unidimensional Item Response Theory (UIRT) Models

There are many IRT models that differ in the form of the item characteristic curve.

Among unidimensional models, the three-parameter logistic model (3PL) is the most

popular (Kolen & Brennan, 2004). The model is as follows:

$$P\{X_{ij} = 1|\theta_i, a_j, b_j, c_j\} = c_j + (1 - c_j)\frac{\exp[a_j(\theta_i - b_j)]}{1 + \exp[a_j(\theta_i - b_j)]}, \qquad (2.1)$$

where $P\{X_{ij} = 1|\theta_i, a_j, b_j, c_j\}$ is the probability of the ith examinee answering the

jth item correctly given the ith examinee's ability level (θ_i), the jth item's discrim-

ination parameter (a_j), item difficulty (b_j), and the item guessing parameter (c_j).

The constant, typically set to 1.7, can be multiplied to a_j so that the logistic item

response curve and the normal ogive differ by no more than 0.01 for all values of θ in

the range -3 to +3 (Kolen & Brennan, 2004). The ability level, θ, is defined over the

range of negative infinity to positive infinity and often is assumed to be distributed

$N(0, 1)$. Therefore, almost all ability levels are within the range -3 to +3. The item

discrimination parameter is the slope of the item characteristic curve at the inflexion point. The item difficulty, or location, parameter corresponds to an inflexion point at $\theta = b$. When $c = 0$, the difficulty parameter is the level of ability where the probability of a correct answer is 0.5. When the guessing parameter is not zero, the difficulty parameter is halfway between the guessing parameter and 1.0. The guessing parameter (c_j) is also called the lower asymptote or pseudo-chance level parameter. The 3PL model with a unique guessing parameter for each item can lead to estimation problems, thus a common guessing parameter is often estimated for all items or for groups of similar items (Embretson & Reise, 2000).

2.1.1   UIRT Model Assumptions

There are two critical assumptions with IRT models: local independence and unidimensionality. The critical assumptions will be discussed next.

Local Independence

The assumption of local independence has strong and weak versions (Embretson & Reise, 2000). When the strong local independence assumption is met, the responses to all items of a test are independent of one another, except for abilities they have in common. Conditional on the latent ability level, responses to item pairs are independent. Local independence is obtained when the relationships among items (or persons) are fully characterized by the IRT model. In other words, local independence is obtained when the probability of solving any item is independent of the outcome of

any other items, controlling for latent ability levels and item parameters. This strong

local independence is expressed as below,

$$P\{X = x|\theta\} = \prod_{j=1}^{J} P_j\{X_j = x|\theta\}, \qquad (2.2)$$

where P stands for probability, J is the total number of items and θ is the latent

ability level. This equation above shows that the probability of an item pattern is

simply the product of the probability for each item when the ability level is accounted

for. That is, test items are completely independent once the vector of latent abilities

has been accounted for.

Weak local independence is met when, conditional on the dominant latent abili-

ties, the item covariances among all item pairs decreases toward zero as test length

approaches infinity. This weak local independence is expressed as below,

$$Cov\{X_j, X_{j*}|\theta\} = 0 \qquad j \neq j^*, \qquad (2.3)$$

where Cov stands for covariance and X_j and X_{j*} are variables for the jth and j^*th

item. This equation shows that pair-wise items have zero covariances once the latent

ability has been accounted for. In other words, this weak assumption concerns the

linear dependencies of two items, higher order dependencies among items are not

allowed.

Dimensionality

The assumption of dimensionality is that an IRT model has appropriate dimensional-

ity, with the majority of IRT models assuming unidimensionality (Embretson & Reise,

2000). There is strict and essential unidimensionality. With the strict unidimensionality assumption, one and only one latent ability underlies test performance. With the essential unidimensionality assumption, only one latent ability, called the dominant latent ability, among all latent abilities contained within a data set is needed to account for test performance. Thus, only one dominant latent ability is needed to model item responses, even though minor dimensions are present in the data.

Unidimensional IRT models define a single ability level for each person, and are appropriate for items that involve a single underlying ability or combination of abilities that are constant across items (Embretson & Reise, 2000). Therefore, even when we find multiple abilities (factors) that are contributing to the scores, as long as the set of abilities are constant across items or if examinees only vary significantly on the dimension being measured, we can still employ unidimensional IRT (Ackerman, 1992). Although it is not realistic, when the test is only one item long, the test is unidimensional (Ackerman, 1992).

2.1.2 Violation of Local Independence

The assumption of local independence is not met, for instance, when the content of an early item in a test provides clues as to the correct/keyed response for a later item. Then, the two items will correlate more highly than they should, based on only the ability that they have in common. Violation of local independence can lead to overestimates of reliability or information and underestimates of the standard error of the ability estimate(Sireci et al., 1991; Wainer, 1995; Wainer & Wang, 2000; Yen,

1993).

2.1.3   Violation of Unidimensionality

When each ability contributes to a person's item response differently, then we should use a multidimensional IRT model instead. In fact, multidimensionality almost always happens as stated by Ackerman (1992):

> This article takes the view that empirically two or more items will always produce multidimensionality, and, as such, their parameters will often need to be estimated using multidimensional models. (p. 90)

Although most items are designed to measure one ability, some items clearly require at least two abilities for a correct response (Reckase, 1985). An example of such items is a set of mathematical problems with a story, which requires the examinees to have both mathematical and verbal skills. When multidimensional data is modeled with unidimensional IRT, the resulting parameter estimates may be biased. For example, if two groups of examinees with different underlying multidimensional ability distributions take test items that can differentiate the groups based on underlying ability levels, any unidimensional models have the potential to produce item biases (Ackerman, 1992). Generally, ability level and item parameters are estimated using a large sample size assuming all examinees are homogeneous in the skills underlying the items. When two groups are not homogeneous and differ in a secondary dimension, a unidimensional scoring scheme may create item biases (Ackerman, 1992). Ackerman

(1992) explains that item bias and construct validity are interrelated because the number of skills measured and the degree to which comparisons between groups are appropriate are construct validity issues. A test with low construct validity contains items that are measuring skills other than those intended to be measured, thus, the potential for item bias also exists.

Item bias can be observed by linear transformation of item parameters on two groups of interest after estimating item parameters separately. When there is no item bias, that is, when the interaction between the examinee latent space and the item is unidimensional, the item parameter estimates should be invariant. When the only difference is the difference in ability levels, the estimates are invariant. Any difference in item characteristic curves (ICC) after a transformation could be due to bias (Ackerman, 1992).

2.2 Multidimensional IRT (MIRT)

When persons differ systematically in relation to items being hard or easy, multidimensional IRT (MIRT) can fit data better than unidimensional IRT (Embretson & Reise, 2000). MIRT helps to understand extraneous examinee characteristics, such as ethnicity, that affect the probability of response to items (Reckase, 1997). Research shows that the unidimensional assumption is often difficult to meet in real world contexts (Akerman, 1994; Nandakumar, 1994; Reckase, 1985) and the number of dimensions is often underestimated (Reckase & Hirsh, 1991). Therefore, an application

of MIRT models, which contains two or more parameters to represent each person, would often be appropriate.

There are exploratory and confirmatory MIRT models (Embretson & Reise, 2000). An exploratory MIRT estimates item and person parameters on more than one dimension to improve model fit. The number of dimensions is not determined by theory. With confirmatory MIRT, item and person parameters are estimated with specific dimensions. As in factor analysis, a confirmatory analysis involves specifying the relationship of the items to the dimensions.

There are two main forms of MIRT models: noncompensatory and compensatory models (Reckase, 1997). The partially compensatory or noncompensatory model with three parameters is as follows:

$$P\{x_{ij} = 1 | \boldsymbol{\theta}_i, \mathbf{a}_j, \mathbf{b}_j, c_j\} = c_j + (1 - c_j) \prod_{k=1}^{m} \frac{\exp[a_{jk}(\theta_{ik} - b_{jk})]}{1 + \exp[a_{jk}(\theta_{ik} - b_{jk})]}, \qquad (2.4)$$

where $\boldsymbol{\theta}_i = (\theta_{j1}, \theta_{j2}, ..., \theta_{jm})$ is a vector of the ith person's latent ability level, $\mathbf{a}_j = (a_{j1}, a_{j2}, ..., a_{jm})$ is a vector of discrimination parameters and $\mathbf{b}_j = (b_{j1}, b_{j2}, ..., b_{jm})$ is a vector of difficulty parameters for the jth item in m-dimensional space, and c_j is the lower asymptote (Sympson, 1978; Whitely, 1991). This model has the property that the probability of a correct response decreases with an increase in the number of dimensions for fixed values of the exponents. This is called noncompensatory because a high ability level on one dimension cannot fully compensate for a low ability level on another dimension.

The ability levels of different dimensions are additive in compensatory models, so

that a low ability level in one dimension can be compensated by a high ability level on another dimension (Lord & Novick, 1968; Reckase, 1985, 1995). The compensatory model with three parameters is as follows (Embretson & Reise, 2000; Reckase, 1997):

$$P\{X_{ij} = 1|\boldsymbol{\theta}_i, \mathbf{a}_j, d_j, c_j\} = c_j + (1 - c_j)\frac{\exp(\sum_{k=1}^m a_{jk}\theta_{ik} + d_j)}{1 + \exp(\sum_{k=1}^m a_{jk}\theta_{ik} + d_j)}, \qquad (2.5)$$

where d_j is the difficulty for the jth item. The easier the item, the higher the value of d_j. There is only one difficulty parameter in compensatory models as a composite parameter representing the difficulty of the item.

To have a better understanding of equation 2.5, the unidimensional IRT model (equation 3.1) can be expressed using the term d_j instead of b_j as follows:

$$
\begin{aligned}
P\{X_{ij} = 1|\theta_i, a_j, b_j, c_j\} &= c_j + (1 - c_j)\frac{\exp[a_j(\theta_i - b_j)]}{1 + \exp[a_j(\theta_i - b_j)]} & (2.6) \\
&= c_j + (1 - c_j)\frac{\exp[a_j\theta_i - a_jb_j)]}{1 + \exp[a_j\theta_i - a_jb_j)]} & (2.7) \\
&= c_j + (1 - c_j)\frac{\exp[a_j\theta_i + d_j]}{1 + \exp[a_j\theta_i + d_j]}, & (2.8)
\end{aligned}
$$

where $d_j = -a_jb_j$.

In this paper, the compensatory model with three parameters was used because it is well-studied, practical, and more frequently applied. Many other studies used this compensatory model (Ackerman, 1992; DeMars, 2006; Reckase, 1985) and estimation procedures are available for the parameters (McKinley & Reckase, 1983).

The assumption of local independence in MIRT is that within the group of examinees who have the same ability levels along every dimension measured, θ_1, θ_2,..., θ_m, the conditional distributions of the item scores are all independent of each other,

where m is the number of dimensions being measured (Lord & Novick, 1968).

2.3 Multidimensional Item Factor Analysis

MIRT models can be considered either as a special case of nonlinear factor analysis, or as an expansion of unidimensional models (Reckase, 1997). In this section, the relationship between nonlinear factor analysis and MIRT is discussed. When a single latent ability factor has a normal distribution, fitting the linear factor model to the tetrachoric correlations of the items is tantamount to fitting a unidimensional normal-ogive model (Lord, 1952). The extension of Lord's result to multiple factors also has been studied (Christoffersson, 1975; Muthen, 1978). To show the nonlinear item factor analysis model (the item response equivalent of the factor model), let X_{ij}^* be the ith person's latent quantitative response tendency for the jth item that behaves as follows:

$$X_{ij} = \begin{cases} 1 & \text{if} \qquad X_{ij}^* > \tau_j, \\ 0 & \text{if} \qquad X_{ij}^* \leq \tau_j. \end{cases} \tag{2.9}$$

where τ_j is a threshold value for the jth item. Suppose also that the underlying response tendencies fit the multiple factor model as follows:

$$X_{ij}^* = \sum_{k=1}^{m} \lambda_{jk} F_{ik} + E_{ij}^*, \tag{2.10}$$

where λ_{jk} is the jth item's kth factor loading, F_{ik} is the ith person's kth common factor score which is the same as latent ability score θ_{ik}, and E_{ij}^* is an error for the jth item and ith person. Since X_{ij}^* is standardized, the following decomposition of

variance appears:

$$\sum_{k=1}^{m} \lambda_{jk}^2 + \psi_j^2 = 1, \tag{2.11}$$

where ψ_j^2 is the error variance. This shows that the variance of the response tendency for one item is the sum of the squared factor loadings across dimensions and error variance. One difference between item factor analysis and item response theory is that in factor analysis the response variable X_{ij}^* is standardized to have mean zero and variance 1.0, unlike item response theory in which the error variance is standardized. The assumption here is that F_{ik} and E_{ij}^* have a normal distribution. Therefore, the dependent variable, X_{ij}^*, also has a normal distribution, and the following relationship exists with inclusion of the guessing parameter, c_j:

$$P\{X_{ij} = 1 | \mathbf{F} = \mathbf{f}\} = P\{X_{ij}^* > \tau_j | \mathbf{F} = \mathbf{f}\} = c_j + (1 - c_j)N(z) \tag{2.12}$$

for an examinee with a given vector $\mathbf{f}$ of the common factors $\mathbf{F}$, where

$$z = \sum_{k=1}^{m} [(\lambda_{jk}/\psi_j)f_{ik}] - \tau_j/\psi_j. \tag{2.13}$$

$N(z)$ is the cumulative probability of the normal distribution given the value of z. The IRT item parameters are derived from factor parameters as follows:

$$a_{jk} = \lambda_{jk}/\psi_j, \qquad d_j = -\tau_j/\psi_j, \tag{2.14}$$

Using the model of equation 2.14, the equivalent of z with item response function parameters is as follows:

$$P\{X_{ij} = 1 | \mathbf{F} = \mathbf{f}\} = c_j + (1 - c_j)N(\sum_{k=1}^{m} a_{jk}f_{ik} + d_j), \tag{2.15}$$

where a_{jk} and d_j are the item discriminations on each dimension and the item difficulty in the IRT model based on factor analysis. Using the logistic distribution instead of the normal distribution, the equation with three parameters can be expressed as follows:

$$P\{X_{ij} = 1 | \mathbf{F} = \mathbf{f}\} = c_j + (1 - c_j) \frac{\exp(\sum_{k=1}^{m} a_{jk} f_{ik} + d_j)}{1 + \exp(\sum_{k=1}^{m} a_{jk} f_{ik} + d_j)}. \qquad (2.16)$$

These item parameters from factor item analysis have the following relationships with the item parameters from IRT:

$$f_{ik} = \theta_{ik} \qquad (2.17)$$

$$\lambda_{jk} = \frac{a_{jk}}{\sqrt{1 + \sum a_{jk}^2}}, \qquad (2.18)$$

$$\tau_j = \frac{d_j}{\sqrt{1 + \sum a_{jk}^2}} \qquad (2.19)$$

TESTFACT (Wood et al., 1987) provides unrotated item parameters and rotated factor loadings when requested. To obtain rotated (varimax or promax) item parameters, they have to be computed using the rotated factor loadings. Then, τ is obtained as in equation 2.19, using the unrotated item parameters, a_{jk} and d_j. The value of τ is the same when the solution is unrotated and when rotated since it is the threshold value of getting the item correct. Using the obtained value of τ and rotated factor loadings, rotated item parameters are obtained as follows when factors are not correlated.

$$a_{jk} = \frac{\lambda_{jk}}{\sqrt{1 - \sum \lambda_{jk}^2}}, \qquad (2.20)$$

$$d_j = -\frac{\tau_j}{\sqrt{1 - \sum \lambda_{jk}^2}}. \qquad (2.21)$$

When factors are correlated, the following equations are used instead.

$$a_{jk} = \frac{\lambda_{jk}}{\sqrt{1 - \sum \lambda_{jk}^2 - 2 \sum \sum r_{kk'} \lambda_{jk} \lambda_{jk'}}}, \tag{2.22}$$

$$d_j = -\frac{\tau_j}{\sqrt{1 - \sum \lambda_{jk}^2 - 2 \sum \sum r_{kk'} \lambda_{jk} \lambda_{jk'}}}, \tag{2.23}$$

where $k \neq k'$ and $r_{kk'}$ is the correlation between factor scores.

When there are two dimensions, equations 2.22 and 2.23 can be written as follows.

$$a_{jk} = \frac{\lambda_{jk}}{\sqrt{1 - \lambda_{j1}^2 - \lambda_{j2}^2 - 2r_{12}\lambda_{j1}\lambda_{j2}}}, \tag{2.24}$$

$$d_j = -\frac{\tau_j}{\sqrt{1 - \lambda_{j1}^2 - \lambda_{j2}^2 - 2r_{12}\lambda_{j1}\lambda_{j2}}}, \tag{2.25}$$

The above transformation of the item parameters in item factor analysis to/from the item parameters in IRT is possible because the difference between them is the variable to be standardized: the latent response variable, X_i^* is standardized in factor analysis, while the error is standardized in IRT.

2.4   UIRT Linking Techniques

2.4.1   Linking: Placing Parameter Estimates on a Common Scale

Item parameters are estimated by setting the mean of the the examinees' ability levels to zero and the standard deviation to one. Item parameters can be estimated using data from a common-item linking design either separately for each form or concurrently across forms. When two groups of examinees differ in ability levels, and when item parameters are estimated separately for each form, the units of the item

parameters are not on the same scale because the examinees' mean ability levels and standard deviations are not equal. When the distributions of ability levels are the same for two groups (equivalent groups) taking different forms that are equivalent, no transformation of item parameters is necessary. However, there may be a benefit of performing item parameter scaling even with equivalent groups since it may reduce estimation errors by adjusting for small differences that arise from sampling error or test difficulty (Hanson & Béguin, 2002).

2.4.2   Separate Estimation

When item parameter estimates differ because of the differences in ability level mean and standard deviation, item parameter estimates need to be transformed onto the same scale. If the relationship of item parameters in two forms is linear, the relationship can be expressed using the ability level from two different forms as follows:

$$\theta_X = A\theta_Y + B, \tag{2.26}$$

where A and B are the slope and the intercept of the linear relationship and θ_X and θ_Y are the ability levels expressed on the scale of Form X and Form Y for individuals who have the same ability level. The linear relationship of the item parameters for 2PL and 3PL IRT can also be expressed through the item difficulty estimates. Therefore, the transformation of the item difficulty is as follows:

$$b_X = Ab_Y + B, \tag{2.27}$$

where b_X and b_Y are the item difficulties for the items. The discrimination parameter is transformed by multiplying by the reciprocal of the A parameter of the linear transformation:

$$a_X = \frac{a_Y}{A}. \tag{2.28}$$

The guessing parameter is independent of the scale transformation:

$$c_X = c_Y. \tag{2.29}$$

Consider the situation where there are multiple common items. The linear relationship of the item difficulty parameter could be expressed as follows Cook & Eignor (1991):

$$\frac{\hat{b}_{Xj} - \bar{b}_{X.}}{SD_{\hat{b}_{X.}}} = \frac{\hat{b}_{Yj} - \bar{b}_{Y.}}{SD_{\hat{b}_{Y.}}}, \tag{2.30}$$

or equivalently,

$$\hat{b}_{Xj} = A\hat{b}_{Yj} + B, \tag{2.31}$$

where, $\hat{b}_{Xj}$ and $\hat{b}_{Yj}$ are the item difficulty estimates for the jth common item, $\bar{b}_{X.}$ and $\bar{b}_{Y.}$ are the mean difficulty estimates of all common items, $SD_{\hat{b}_{X.}}$ and $SD_{\hat{b}_{Y.}}$ are the standard deviations of the item difficulty estimates based on all common items on Form X and Form Y, respectively. The parameters A and B are the slope and the intercept obtained by plugging in the means and the standard deviations of the item difficulty estimates in equations 2.34 and 2.35 below. The same transformation is applied to item discrimination estimates and ability estimates as follows:

$$\hat{a}_{Xj} = \frac{\hat{a}_{Yj}}{A}, \tag{2.32}$$

$$\hat{\theta}_{Xi} = A\hat{\theta}_{Yi} + B, \tag{2.33}$$

where $\hat{a}_{Xj}$ and $\hat{a}_{Yj}$ are the item discrimination estimates for the jth item, $\hat{\theta}_{Xi}$ and $\hat{\theta}_{Yi}$ are the ability level estimates for the ith examinee on the scale of Form X and Form Y, respectively. The transformation of the guessing parameter is not necessary because it is the intercept of the item characteristic curve.

The equations used to find A and B parameters using the above notation are as follows Kolen & Brennan (2004):

$$A = \frac{SD_{\hat{b}_{X.}}}{SD_{\hat{b}_{Y.}}} = \frac{\bar{a}_{Y.}}{\bar{a}_{X.}}, \tag{2.34}$$

$$B = \bar{b}_{X.} - A\bar{b}_{Y.}, \tag{2.35}$$

where $\bar{a}_{X.}$ and $\bar{a}_{Y.}$ are the mean item discrimination parameter estimates of all common items, $\bar{b}_{X.}$ and $\bar{b}_{Y.}$ are the mean item difficulty estimates of all common items, and $SD_{\hat{b}_{X.}}$ and $SD_{\hat{b}_{Y.}}$ are the standard deviations of the item difficulty estimates based on all common items on Form X and Form Y, respectively. There is only one way to estimate the parameter B using the item difficulty or discrimination estimates. However, there are two ways to estimate the A parameter. When the A parameter is estimated using the standard deviations of the difficulty levels, the method is called the mean/sigma method in this paper following Kolen and Brennan (2004). When the A parameter is estimated using the mean of the item discrimination parameter, the method is called the mean/mean method.

After transformation, the item parameter estimates are said to be calibrated. Theoretically, the linear relationship of the item parameters can be determined through

item discrimination estimates or ability estimates. However, item difficulty estimates are frequently used to find the linear relationship because of its superior stability over other estimates (Cook & Eignor, 1991).

The mean/sigma and mean/mean methods described above have a shortcoming in that they do not consider all of the item parameters simultaneously (Kolen & Brennan, 2004). To compensate for this problem, there are two approaches that consider all of the item parameters simultaneously: the Haebara approach and the Stocking and Lord approach (Haebara, 1980; Stocking & Lord, 1983).

The Haebara approach minimizes the cumulative squared difference between the item characteristic curves for each item for examinees of a particular ability. This is also called the item characteristic function (ICF) method.

$$Hdiff_{Y_X} = \frac{1}{JN_X} \sum_{j=1}^{J} \sum_{i_X}^{N_X} [E_{j,X}(\hat{\theta}_{i_X,X}) - E_{j,Y}(\hat{\theta}_{i_X,Y_X})]^2 \tag{2.36}$$

where J is the number of common items, and $E_{j,X}(\hat{\theta}_{i_X,X})$ is the expected value for item j for the Form X calibration given the i_Xth individual's latent ability estimate calculated in that calibration. $E_{j,Y}(\hat{\theta}_{i_X,Y_X})$ is the expected value for item j from the Form Y calibration after being transformed to the metric for the Form X calibration given the i_Xth individual's latent ability estimate from the Form X calibration. N_X is the number of examinees who were administered Form X. More precisely, the elements of the above equation can be expressed as follows:

$$E_{j,X}(\hat{\theta}_{i_X,X}) = P_{ij}(X = 1|\theta_{i_X}, \hat{a}_{Xj}, \hat{b}_{Xj}, \hat{c}_{Xj}) \tag{2.37}$$

$$E_{j,Y}(\hat{\theta}_{i_X,Y_X}) = P_{ij}(X = 1|\theta_{i_X}, \frac{\hat{a}_{Yj}}{A}, A\hat{b}_{Yj} + B, \hat{c}_{Yj}). \tag{2.38}$$

The expectation of the common item responses should be the same for Form X and Form Y for an examinee. The Haebara approach solves for the linking parameters A and B by minimizing the difference in the expectations of the item responses between Form X and Form Y (in the metric of Form X) for those who took Form X.

In contrast to the Haebara approach, the Stocking and Lord approach, which is called the test characteristic function (TCF) method, minimizes the cumulative squared difference between the test characteristic curves over items for examinees of a particular ability. Accordingly, the loss function to be minimized is

$$SLdiff_{Y_X} = \frac{1}{N_X} \sum_{i_X}^{N_X} [T_X(\hat{\theta}_{i_X,X}) - T_Y(\hat{\theta}_{i_X,Y_X})]^2 \tag{2.39}$$

where the subscript Y_X indicates that parameter estimates from Form Y are placed on the metric of Form X. $T_X(\hat{\theta}_{i_X,X})$ is the true score on the set of common items on the Form X calibration, given the ith individual taking Form X has the latent ability estimate derived in that calibration. $T_Y(\hat{\theta}_{i_X,Y_X})$ is the true score on the common items based on the Form Y calibration transformed onto the metric of Form X, given the ith individual's latent ability estimate from the Form X calibration. More precisely, the elements of the above equation can be expressed as follows:

$$T_X(\hat{\theta}_{i_X,X}) = \sum_j P_{ij}(X = 1|\theta_{i_X}, \hat{a}_{Xj}, \hat{b}_{Xj}, \hat{c}_{Xj}) \tag{2.40}$$

$$T_Y(\hat{\theta}_{i_X,Y_X}) = \sum_j P_{ij}(X = 1|\theta_{i_X}, \frac{\hat{a}_{Yj}}{A}, A\hat{b}_{Yj} + B, \hat{c}_{Yj}). \tag{2.41}$$

Note that the test characteristic curve in IRT refers to the following function, the

sum of the probability of getting an item correct given an examinee's ability level.

$$\tau(\theta_i) = \sum_j P_{ij}(\theta_i) \tag{2.42}$$

Since it should not make any difference whether examinees to take Form X or Form Y, the sums of the probabilities of getting an item correct, T_X and T_Y, across common items should be the same for Form X and Form Y for an examinee. The Stocking and Lord approach solves for the linking parameters A and B by minimizing the difference in the sum of the probability of getting an item correct across common items between Form X and Form Y (in metric of Form X) for those who took Form X.

The function for the Stocking and Lord approach, equation 2.39, is minimized when the function's partial derivatives with respect to each of the linking parameters is equal to zero as below.

$$\frac{\partial SLdiff_{Y_X}}{\partial A} = -\frac{2}{N_X} \sum_{i_X}^{N_X} [T_X(\hat{\theta}_{i_X,X}) - T_Y(\hat{\theta}_{i_X,Y_X})]\frac{\partial T_Y(\hat{\theta}_{i_X,Y_X})}{\partial A} = 0 \tag{2.43}$$

$$\frac{\partial SLdiff_{Y_X}}{\partial B} = -\frac{2}{N_X} \sum_{i_X}^{N_X} [T_X(\hat{\theta}_{i_X,X}) - T_Y(\hat{\theta}_{i_X,Y_X})]\frac{\partial T_Y(\hat{\theta}_{i_X,Y_X})}{\partial B} = 0, \tag{2.44}$$

where the partial derivative of the true score (the sum of the probability of getting each item correct) with respect to each linking parameter is expressed as the sum of partial derivative for each item as follows:

$$\begin{aligned} \frac{\partial T_Y(\hat{\theta}_{i_X,Y_X})}{\partial A} &= \sum_{i_X}^{N_X} \left[\frac{\partial P_Y(\hat{\theta}_{i_X,Y_X})}{\partial b_{Y_X}}\frac{\partial b_{Y_X}}{\partial A} + \frac{\partial P_Y(\hat{\theta}_{i_X,Y_X})}{\partial a_{Y_X}}\frac{\partial a_{Y_X}}{\partial A} \right] \\ &= \sum_{i_X}^{N_X} \left[b_Y\frac{\partial P_Y(\hat{\theta}_{i_X,Y_X})}{\partial b_{Y_X}} - \frac{a_Y}{A^2}\frac{\partial P_Y(\hat{\theta}_{i_X,Y_X})}{\partial a_{Y_X}} \right] \end{aligned} \tag{2.45}$$

$$
\begin{aligned}
\frac{\partial T_Y(\hat{\theta}_{i_X,Y_X})}{\partial B} \;\; &= \;\; \sum_{i_X}^{N_X} \frac{\partial P_Y(\hat{\theta}_{i_X,Y_X})}{\partial b_{Y_X}} \frac{\partial b_{Y_X}}{\partial B} \\
&= \;\; \sum_{i_X}^{N_X} \frac{\partial P_Y(\hat{\theta}_{i_X,Y_X})}{\partial b_{Y_X}}
\end{aligned}
\tag{2.46}
$$

The Haebara approach and the Stocking and Lord approach utilize the indicators of errors in the transformation of item parameters described in this section. The iterative minimization algorithm would be used to generate better estimates, A and B, using the slopes and the intercepts described in the mean/mean and mean/sigma methods as the starting values of the iteration.

2.4.3 Concurrent Estimation

The transformation of item parameters is automatically performed when software can estimate item parameters simultaneously, which is referred to as concurrent item calibration (Wingersky & Lord, 1984). In concurrent estimation, all estimated parameters for Form X and Form Y are on the same scale since they are estimated simultaneously, however they may not be on the same scale as the generating parameters in a simulation. To perform concurrent estimation, proper software such as MULTILOG (Thissen, 1991) or BILOG-MG (Zimowski et al., 1996) has to be used.

The procedures used in concurrent calibration are joint maximum likelihood estimation (JMLE) and marginal maximum likelihood estimation (MMLE). Bayes estimation can also be used in either JMLE or MMLE. Concurrent calibration estimates item and ability level parameters combining data from more than one group and treat-

ing items not taken by a particular group as planned missingness (Lord, 1980). There are variations of this procedure in which parameter estimates of the common items from the base groups are fixed and the uncommon item parameters are estimated using target group data (Kim & Cohen, 1998).

2.5   Comparison of Separate and Concurrent Estimation: Unidimensional Case

Some researchers have compared the item parameter estimates using separate and concurrent estimation processes (Béguin et al., 2000a; Béguin & Hanson, 2001; Hanson & Béguin, 2002; Kim & Cohen, 1998, 2002; Tsai et al., 2001). Tsai et al. (2001) investigated IRT linking with the Stocking and Lord method and concurrent calibrations for true score and observed score linking. Their results showed smaller standard errors with concurrent calibration.

Kim and Cohen (1998) examined separate and concurrent estimation processes with simulated unidimensional data using the computer program BILOG (Mislevy & Bock, 1990) for separate item parameter estimation and MULTILOG (Thissen, 1991) for concurrent estimation based on marginal maximum a posteriori estimation and MMLE. The Stocking and Lord (1983) method was used for separate estimation. The study had four different numbers of common items (5, 10, 25, and 50) in a test that contained 50 items in total. The evaluation criteria was root mean square difference (RMSD) between the item parameter estimates and the generating parameters and

the mean Euclidean distance (MED) that considers all parameters simultaneously. The authors concluded that the three methods produced similar estimates when the number of common items was large (more than 5 common items in a test of 50 items). When the number of common items was small, Kim and Cohen (1998) found that separate estimation provided more accurate results. However, their results employed different programs for the separate and concurrent calibration, which introduces a confounding factor. Other studies found that concurrent estimation provides more stable linking results with smaller numbers of common items (Hanson & Béguin, 2002; Peterson et al., 1983).

Hanson and Béguin (2002) pointed out that the study by Kim and Cohen (1998) had a confounding effect of difference between computer programs. Therefore, Hanson and Béguin (2002) used MULTILOG (Thissen, 1991) and BILOG-MG (Zimowski et al., 1996) for both concurrent and separate estimation of unidimensional data in their study. The study had five factors: (a) equivalent versus nonequivalent groups, (b) concurrent versus separate estimation using four item parameter scaling methods, (c) estimation program (BIOLG-MG and MULTILOG), (d) sample size (3,000 and 1,000), and (e) the number of common items (20 and 10). In the separate estimation conditions, four methods were used: (i) Mean/Mean, (ii) Mean/Sigma, (iii) Stocking-Lord, and (iv) Haebara method. Each condition was replicated 50 times.

Given nonequivalent groups with different standard deviations, the standard deviations of the latent distributions for the two groups were allowed to differ in BILOG-

MG, but not in MULTILOG. For both program the means of the latent variable for the two groups were allowed to differ. The criteria used to assess the linking were; 1) a criterion based on the IRT true score linking function (MSE based on the true score), and 2) a criterion based on how close the estimated item characteristic curves were to the true item characteristic curves (MSE based on the weighted and unweighted ICC). Note that the criteria used in Hanson and Béguin (2002) are based on true score, not item parameters as in Kim and Cohen (1998). The item parameter estimates of the base form were used as the parameter estimates of the common items for the purpose of computing the evaluation criteria.

Hanson and Béguin (2002) found that concurrent estimation resulted in lower error than separate estimation except when groups were non-equivalent. The MSE (based on the true score and ICC) of nonequivalent groups was substantially larger than equivalent groups with both concurrent and separate estimations. The MSE was less when the number of common items was larger. However, given the same number of common items, the MSE was smaller when the sample size was larger. Hanson and Béguin (2002) examined the bias and the variance component of each criterion. The differences in MSE among estimation procedures were largely due to the difference in variance rather than the difference in bias. They found that the bias was larger with concurrent estimation than other methods when the number of common items was small and groups were non-equivalent using MULTILOG, but not when BILOG-MG was used. This result has an implication for Kim and Cohen's

(1998) conclusion that separate estimation provided more accurate results than concurrent when the number of common items was small. Hanson and Béguin (2002) reported that the performance of the program affects the comparison of concurrent and separate estimations. Using BILOG-MG, concurrent estimates performed better than the separate estimates, except with the MSE based on the unweighted ICC when the number of common items was small. Using MULTILOG, the concurrent estimates had lower MSE than the separate estimate, except when the groups were nonequivalent.

Kim and Cohen (2002) conducted a study of the comparison of separate and concurrent estimation in graded items using unidimensional simulated data. In Kim and Cohen (2002), they used only MULTILOG (Thissen, 1991) to perform separate and concurrent estimations. The conditions of the study were: (a) sample size (small, large, and unbalanced between two forms), (b) equivalent versus nonequivalent groups, (c) different numbers of common items, and (d) concurrent and separate estimation. For the separate estimations, the authors estimated the item parameters separately using MULTILOG. Default MULTILOG options were used for the calibrations. Then, the Stocking and Lord method, which is the TCF method, for linking under the graded response model (Baker, 1992) with the computer program EQUATE (Baker, 1993) was used. They added another evaluation criteria: RMSE for ability level estimates. Kim and Cohen (2002) performed second stage linking to place linked item parameter estimates onto the metric of generated item parameters.

As expected, the error was smaller when the number of common items was larger. The finding showed concurrent estimation had smaller error than separate estimation in all conditions. However, the difference between the two estimation methods was very small.

The literatures reviewed above on the comparison of concurrent and separate calibration used unidimensional data. Béguin, Hanson, and Glas (2000) and Béguin and Hanson (2001) simulated multidimensional data but fit a unidimensional model. Béguin, Hanson, and Glas (2000) simulated multidimensional compensatory data and examined the performance of separate and concurrent estimation of unidimensional IRT score distributions using BILOG-MG. The Stocking and Lord (1983) method was used as the separate method. The study had four factors: (a) equivalent versus nonequivalent groups administered the two forms, (b) concurrent versus separate estimation, (c) three different covariance levels between two forms, and the corresponding three levels of variance on the second dimension, and (d) multidimensional and unidimensional estimation. Two evaluation criteria were used. One was the difference between estimated score distributions and the population score distributions. Another was the difference between estimated score points and the population score points. The first criterion used the difference of frequency of score point to obtain MSE, and the second used the difference of score points to obtain WMSE where the weight was the proportion of the score points. Both criteria were decomposed into the variance and the bias.

In the nonequivalent groups conditions, Béguin, Hanson, and Glas (2000) found that the error increased as the covariance and variance of the second dimension increased. Consequently, the MSE and bias were larger for the concurrent estimation when the covariance between the two dimensions was larger. The general finding from the study by Béguin, Hanson, and Glas (2000) is that multidimensional data with equivalent groups is not much of a problem as compared to nonequivalent groups. The larger covariance and second dimension variance increases the error especially with concurrent estimation.

Béguin and Hanson (2001) simulated multidimensional noncompensatory data and examined the performance of separate and concurrent estimation of unidimensional IRT using BILOG-MG and EPDIRM. The Stocking and Lord (1983) method was used as the separate estimation. The study had five factors: (a) equivalent versus nonequivalent groups administered the two forms, (b) concurrent versus separate estimation method, (c) three different covariance levels between the two forms, (d) EPDIRM versus BILOG-MG concurrent estimation, and (e) multidimensional and unidimensional estimation.The design was very similar to Béguin, Hanson, and Glas (2000), however, the mean difference between the nonequivalent groups was smaller, and the covariance between the two dimensions was smaller in Béguin and Hanson (2001). Also, in Béguin and Hanson (2001), the variances of the two dimensions were equal. The two evaluation criteria were the same as Béguin, Hanson, and Glas (2000).

The results of the study by Béguin and Hanson (2001) were consistent with their

earlier studies (Béguin et al., 2000a; Hanson & Béguin, 1999). That is, the concurrent estimation method had smaller MSE than the separate estimation method in the equivalent groups conditions. The study by Béguin and Hanson (2001) had smaller covariances of the two dimensions than Béguin, Hanson, and Glas (2000), thus the MSE of the concurrent estimation method was smaller than the separate estimation method in the nonequivalent groups condition. Béguin and Hanson (2001) state that the bias in the unidimensional results was less with the noncompensatory model than with the compensatory model examined in Béguin, Hanson, and Glas (2000). However, since the degree of covariance was different in the two studies, the comparison of noncompensatory and compensatory models cannot be made with confidence.

While almost all comparisons of concurrent and separate estimations used simulated data, one study compared concurrent and separate estimation for vertical scaling (non-equivalent groups) using a real data set with more complex assumption violations than the simulated data (Kerkee et al., 2003). The Stocking and Lord (1983) method was used for separate estimation. Kerkee et al. (2003) reported that the concurrent estimation method had more non-converging items and items flagged for potential poor fit than separate linking. Consequently, Karkee et al. (2003) found that the separate estimation method had better average fit for the items in every grade.

Kolen and Brennan (2004) conclude that concurrent calibrations are more accurate than separate estimation when the data fit the IRT model. Concurrent calibration is,

however, less robust to violations of the IRT assumptions than separate estimation using test characteristic curve methods such as the Stocking and Lord method, as evidenced by larger errors in non-equivalent group designs and a study using real data (Hanson & Béguin, 2002; Kerkee et al., 2003). The situation when separate estimation performs better than concurrent calibration depends on the degree of assumption violations and software used to estimate the item parameters. The conditions examined in the reviewed studies are small number of common items, nonequivalent groups, and multidimensional latent ability levels. Based on the studies reviewed here, concurrent item calibrations are more accurate than separate estimation when the data fit the IRT model and even when the data is multidimensional. However, Béguin and Hanson (2001) noted that the effect of multidimensionality depends on the kind of multidimensionality. As the results of Béguin, Hanson, and Glas (2000) show, when the covariance between the two dimensions are very large, separate estimation is better than concurrent estimation.

2.6 Multidimensional IRT (MIRT) Linking

Research on MIRT linking is a relatively new area of interest and has not been conducted as intensively as unidimensional IRT linking. However, there are some papers that describe MIRT linking methods in detail (Davey et al., 1996; Hirsch, 1989; Li & Lissitz, 2000; Min, 2003; Oshima et al., 2000; Thompson et al., 1997; Yon, 2006).

When more than two forms of multidimensional tests are calibrated separately, there is more than one set of item parameters for each form. For example, consider the situation when there are Forms X and Y and the scale of Form Y needs to be re-scaled to be the same as the scale of Form X. In unidimensional IRT, the scale of Form Y is linearly transformed onto the scale of Form X as described in the section on IRT linking techniques. MIRT linking is also a linear transformation. However, since there are multiple sets of parameter estimates, it is a linear transformation of matrices.

MIRT linking is possible when the data set has one or both of the following conditions (Angoff, 1982; Davey et al., 1996): 1) the tests contain common items; 2) there are some examinees who take both forms. In unidimenstional IRT linking, it is possible to equate test forms that do not contain common items and the groups of examinees of each form are randomly equivalent. Thompson et al. (1997) attempted MIRT linking without common items or common groups of examinees (randomly equivalent groups). The study illustrated the Proscrustes orthogonal rotation which is used in many of the later MIRT linking studies such as Li and Lissitz (2000) and Min (2003). The study by Thompson et al. (1997) made an assumption that each test form measured exactly the same unidimensional reference composite, meaning that, each test form measured the same constructs and the composite constructs formed were the same across test forms. This is a strong assumption, and it is questionable whether the assumption is actually met. Therefore, in MIRT linking, such a design

(no common items nor common examinees) is not recommended (Davey et al., 1996).

2.6.1   Thompson, Nering, and Davey

The study conducted by Thompson et al. (1997) had eight test forms needing to be calibrated to form an item pool for a computer adaptive test simulation. The forms were constructed to be equivalent, and contained English, math, reading, and science, with a total of over 200 multiple choice items on each form. Item parameters were estimated separately for each form that contained four subjects using a modified version of the NOHARM program (Fraser, 1988). The number of dimensions in each form was 50 since each subject had more than one dimension. The authors assumed the groups were equivalent, thus no rescaling was carried out. Since each form was calibrated separately, the direction of the dimensions was arbitrary between forms. Then, a Procrustes rotation procedure was used to rotate reference composites between the base and linked forms to match the dimensions.

The solution of the orthogonal procrustes rotation is explained in Schönemann (1966), and the use in MIRT linking is illustrated in Thompson et al. (1997). In orthogonal Procrustes rotation, the matrices of interest are the matrix to be rotated ($\mathbf{R}$), the transformation matrix ($\mathbf{T}$), and the target matrix ($\mathbf{B}$) (Schönemann, 1966; Thompson et al., 1997). When item parameters are rotated, the matrix to be rotated is the matrix with item discrimination parameters on equated form, and the target matrix is the item discrimination parameters on base form. The transformation matrix ($\mathbf{T}$) is found when $\text{tr}(\mathbf{E}'\mathbf{E})$ is minimized in $\mathbf{E} = \mathbf{B} - \mathbf{R}\mathbf{T}$. The solution $\mathbf{T}$ is an

orthogonal matrix ($\mathbf{T'T = I}$). The orthogonal rotation matrix is obtained as below:

$$\mathbf{S = R}^t\mathbf{B}, \tag{2.47}$$

thus $\mathbf{S}$ is the product of the matrix to be rotated ($\mathbf{R}$) and the target matrix ($\mathbf{B}$). With singular value decomposition, the matrix $\mathbf{S}$ is decomposed as following:

$$\mathbf{S = UQV}^t, \tag{2.48}$$

where $\mathbf{Q}$ is a diagonal matrix of the square root of the eigenvalues of $\mathbf{S}$, and $\mathbf{U}$ and $\mathbf{V}$ are matrices of eigenvectors of $\mathbf{S}$. The orthogonal rotation matrix is

$$\mathbf{T = UV}^t. \tag{2.49}$$

Since Thompson et al. (1997) had no common items, the authors made composite of item set, which was 26-35 depending on the form based on their content analysis. The rotation was conducted to match up reference of composite items. Thus the matrix to be rotated was 26-35 rows by 50 columns (50 dimensions). Consequently, the evaluation is based on the reference composite, not the item parameter estimates. The angles between the reference composite of the base and linked forms was one way to measure the adequacy of the rotation. The results showed that the rotation substantially improved the correspondence between the ability spaces. The authors contended that the Procrustes rotation was reasonably successful in aligning the ability spaces but with room for improvement.

2.6.2   Hirsch

Hirsch (1989) conducted a MIRT linking study using the common examinee design, with the number of common examinees being 100, 200, and 300, out of 2000 total examinees. The study used both simulated and real data sets with two dimensional compensatory MIRT model. The evaluation was based on the true scores and ability estimates found on both tests for the common examinees used in the linking. The item parameters used in the simulation were based on the *College Level Academic Skills Test* (CLAST) reading test which had 36 multiple choice items. By duplicating 4 items out of the 36 items, the total number of items on a form in the simulation study was 40. The ability levels of base and linked groups were equvialent and non-equivalent with 0.1 standard deviation difference.

The linking procedure in Hirsch (1989) was: (1) estimate item parameters and abilities on both dimensions for both tests; (2) identify common basis vectors; (3) align basis vectors through orthogonal procrustes rotation; (4) for each dimension, equate means and standard deviations of the ability estimates for the base and linked forms. The software used for the item parameter estimation was MIRTE (Carlson, 1987). The second step was conducted so that the discrimination parameter vectors of the linked form had a set of basis vectors with the same angle as the basis vector of the base test. The third step was orthogonal procrustes rotation of the ability estimates. In the fourth step, the means and standard deviations of the ability estimates for the common examinees on base and linked forms were determined. A linear linking

procedure was used for each ability dimension so that the resulting mean and standard deviation were the same between base and linked forms. The evaluation of the study was based on the true scores for common examinees. The author claimed that the linking results were satisfactory compared to past studies, with small error and a high correlation between the base and linked true scores with as few as 100 common examinees. However, the study also showed that the ability estimates were unstable probably due to the small number of items.

2.6.3   Davey, Oshima and Lee

Davey et al. (1996) explored MIRT linking methods that were extensions of UIRT linking methods. The methods the authors introduced were as follows:

1. 'matching scaling function', later called the 'equated function method', in Oshima et al. (2000), which is the extension of the mean/sigma method in UIRT

2. 'matching test response functions or surfaces', later called 'the test characteristic function (TCF) method' in Oshima et al. (2000), which is the extension of the Stocking-Lord method

3. 'minimizied differences between common-item parameter estimates', later called the 'Direct methods' in Oshima et al. (2000)

These methods will be described in more detail in a later section, but all of these methods use an orthogonal rotation matrix.

2.6.4   Oshima, Davey and Lee's (ODL) Method

MIRT linking models

The study by Oshima et al. (2000) extended an earlier study by Davey et al. (1996) and, although it is not the best model for MIRT equating, makes clear connections between unidimensional IRT equating and MIRT equating. Oshima et al. (2000) state that, in MIRT, the ability level in multiple dimensions complicates the unidimensional IRT equating only slightly. As in unidimensional IRT equating, the relationship between two sets of item parameters is linear as follows (Oshima et al., 2000):

$$\mathbf{a}_{jX} = \mathbf{A}^{-T}\mathbf{a}_{jY}, \tag{2.50}$$

$$\theta_{iX} = \mathbf{A}\theta_{iY} + \boldsymbol{\beta}, \tag{2.51}$$

and

$$d_{jX} = d_{jY} - \mathbf{a}_{jY}^{T}\mathbf{A}^{-1}\boldsymbol{\beta}, \tag{2.52}$$

where $m \times m$ rotation matrix $\mathbf{A}$ adjusts the variances, covariances and orientation of the θ dimensions, and the $m \times 1$ translation vector $\boldsymbol{\beta}$ shifts the means. Therefore, in the model by Oshima et al. (2000), a non-orthogonal rotation matrix and translation vector are included, but not a dilation parameter.

MIRT linking estimation procedures

In Oshima et al. (2000), there are four extensions of the unidimensional IRT equating estimation methods used to estimate the rotation matrices, $\mathbf{A}$, and the translation

vector, β, simultaneously. The four procedures are: 1) the equated function method; 2) the test characteristic function method; 3) the direct method; and 4) the item characteristic function method. Each of the four procedures will be described below.

Equated Function Method

The equated function method is the multidimensional extension of the unidimensional IRT equating using the mean/mean method (Oshima et al., 2000). The system of scale equating is as in Equations 2.50, 2.51, 2.52 (Oshima et al., 2000).

When it is two dimensional, six functions are needed to estimate the four values in the two by two rotation matrix and the two values in the translation matrix. The functions specified are:

$$h_1 = \frac{1}{N/2} \sum_{j=1}^{N/2} a_{j1}, \tag{2.53}$$

$$h_2 = \frac{1}{N/2} \sum_{j=1}^{N/2} a_{j2}, \tag{2.54}$$

$$h_3 = \frac{1}{N/2} \sum_{j=1}^{N/2} d_j, \tag{2.55}$$

$$h_4 = \frac{1}{N/2} \sum_{j=N/2+1}^{N} a_{j1}, \tag{2.56}$$

$$h_5 = \frac{1}{N/2} \sum_{j=N/2+1}^{N} a_{j2}, \tag{2.57}$$

$$h_6 = \frac{1}{N/2} \sum_{j=N/2+1}^{N} d_j, \tag{2.58}$$

where N is the total number of common items, and j is the jth item number. The total item set is split into two blocks of the same size, and the means of item discrimination

parameters and item difficulties are obtained within each block. These six functions are defined for both groups. $\mathbf{A}$ and $\boldsymbol{\beta}$ are then found so that the functions computed on the transformed estimates form linked groups that are equal to the same function for the base group (Davey et al., 1996). The function to be minimized is as follows:

$$\frac{1}{p}\sum_{1}^{p}(h_p - h_p^*)^2, \tag{2.59}$$

where p is the number of elements to be estimated: six for the two dimensional case, and h_p is the means of separate sets of item parameters.

The factors that affect the quality of the estimates are; 1) the choice of scaling functions, 2) the character of the common item sets, and 3) the values of the scaling constants being estimated (Davey et al., 1996). When the covariance matrices are the same in groups of examinees, no rotation is necessary.

The Test Characteristic (TCF) Function Method

The test characteristic function method is the extension of the Stocking and Lord (1983) unidimensional IRT equating procedure (Oshima et al., 2000). The Stocking and Lord procedure re-scales the parameters by minimizing the cumulative squared difference between the test characteristic curves over items for examinees of a particular ability. The $T(\theta) = \sum_{j=1}^{n} P_j(\theta)$. In unidimensional IRT equating, the Stocking and Lord procedure finds scaling constants minimizing the following:

$$\sum_{q=1}^{Q} w_q[T_X(\theta_q) - T_Y^*(\theta_q)]^2, \tag{2.60}$$

where Q is the number of ability levels. The w_q are allowed to differentially weight each ability level. T_Y^* indicates the transformed test characteristic curve of Form Y onto the Form X scale. The extension of the Stocking and Lord procedure to the MIRT two dimensional case is as follows:

$$\sum_{q_1=1}^{Q_1}\sum_{q_2=1}^{Q_2} w_{q_1 q_2}[T_X(\theta_{q_1},\theta_{q_2}) - T_Y^*(\theta_{q_1},\theta_{q_2})]^2, \tag{2.61}$$

where Q_1 and Q_2 are the number of ability levels on the first and the second dimension, respectively. In the unidimensional case, the Stocking and Lord procedure was shown to yield better estimates than the mean/mean or mean/sigma procedure since it considers all parameters at the same time (Kolen & Brennan, 2004).

Direct Method

The direct method is an extension of the unidimensional IRT equation that minimizes the sum of squared differences between the two sets of common item parameter estimates (Divgi, 1985; Oshima et al., 2000). In the unidimensional case, the item parameters are found such that following function is minimized:

$$\sum_{j=1}^{n}(a_{Xj} - a_{Yj}^*) + \sum_{j=1}^{n}(b_{Xj} - b_{Yj}^*), \tag{2.62}$$

where a_Y^* and b_Y^* are the transformed common item parameter estimates of Form Y. The extension of the function to be minimized to the multidimensional case is as follows:

$$\sum_{j=1}^{n}(\mathbf{a}_{Xj} - \mathbf{a}_{Yj}^*)^T \mathbf{W}_j(\mathbf{a}_{Xj} - \mathbf{a}_{Yj}^*) + \sum_{j=1}^{n} w_j(d_{Xj} - d_{Yj}^*), \tag{2.63}$$

where the matrices W_j and the scalars w_j are weights across items and θ dimensions. A reasonable choice of the weights is that they are inversely proportional to the asymptotic sampling variance of the parameter estimates (Davey et al., 1996).

The Item Characteristic Function Method

This is the extension of the unidimensional IRT equating of Haebara (Haebara, 1980; Oshima et al., 2000). As in the unidimensional case, it minimizes the cumulative squared difference between the item characteristic curves for each item for examinees of a particular ability. The function to be minimized is

$$\sum_{j=1}^{J}\sum_{q_1=1}^{Q_1}\sum_{q_2=1}^{Q_2} w_{q_1 q_2}[E_{Xj}(\theta_{q_1},\theta_{q_2}) - E^*_{Yj}(\theta_{q_1},\theta_{q_2})]^2, \tag{2.64}$$

where Q_1 and Q_2 are the number of ability levels on the first and the second dimension, respectively, E_{Xj} is the expected value for item j on form X given the ability levels, E^*_{Yj} is the transformed expected value for item j on form Y given the ability levels; θ_{q_1} and θ_{q_2}.

The Research Design and Results

Oshima et al. (2000) compared four estimation procedures: 1) the equated function method; 2) the test characteristic function method; 3) the direct method; and 4) the item characteristic function method. Oshima et al. (2000) showed that the test characteristic function method performed better than the other three estimation procedures.

2.6.5 Li and Lissitz's (LL) Method

MIRT Linking Models

The MIRT equating model (equations 2.50, 2.51, and 2.52) presented in Oshima et al. (2000) is straightforward. However, Li and Lissitz (2000) argue that no dilation parameter was found or defined, and the rotation matrix could have multiple forms (rotational indeterminacy) in Oshima et al. (2000) because the rotation matrix is non-orthogonal; however, Oshima et al. (2000) noted that their transformation matrix is orthogonal when the correlation of latent abilities are the same between two forms to be equated, and non-orthogonal (oblique) when the correlation of latent abilities are not the same between the two forms to be equated. In Li and Lissitz (2000), the transformation of one matrix onto another scale involves an orthogonal Procrustes rotation, a translation vector, and a single dilation parameter. Thus, Li and Lissitz (2000) express the item parameter transformation in a slightly different way from Oshima et al. (2000) Their transformation of item parameters is as follows Li & Lissitz (2000):

$$\mathbf{a}_{jX} = k\mathbf{a}_{jY}^{T}\mathbf{T}, \tag{2.65}$$

$$\theta_{iX} = \frac{1}{k}(\mathbf{T}^{-1}\theta_{iY} - \boldsymbol{\beta}), \tag{2.66}$$

and

$$d_{jX} = d_{jY} + (\mathbf{a}_{jY}^{T}\mathbf{T})\boldsymbol{\beta}, \tag{2.67}$$

where k is a dilation parameter, $\mathbf{T}$ is an orthogonal rotation matrix, and $\boldsymbol{\beta}$ is a translation vector. Comparing the transformation expressions by Oshima et al. (2000),

T is similar to **A**, where **T** is an orthogonal matrix while **A** is not, and $\boldsymbol{\beta}$ in equations 2.66 and 2.67 is the same as $\boldsymbol{\beta}$ in Oshima et al. (2000). It should be noted here that the rotation matrix in Oshima et al. (2000) rotates and dilates the matrix at the same time. The orthogonal rotation matrix in Li and Lissitz (2000) only rotates the matrix. Thus, the dilation parameter was added in Li and Lissitz (2000): the dilation parameter was included in the rotation matrix in the study by Oshima et al. (2000). The dilation parameter in Li and Lissitz (2000) is a scalar, not a vector, that is common across different dimensions.

The solution of the orthogonal procrustes rotation is explained in Schönemann (1966), and the use in MIRT equating is illustrated in Thompson et al. (1997), which avoids the rotational indeterminacy problem that happens in Oshima et al. (2000). The dilation parameter is the ratio of the unit for the equated form to the unit in the base form. Li and Lissitz (2000) contend that there should be a single dilation parameter for two reasons: 1) it provides a more tractable mathematical problem, and 2) the variance across dimensions will be similar enough that a single dilation parameter will provide reasonable accuracy. Thus, having a single dilation parameter means that the ratio of the length of the two vectors is the same for both test forms. Having a single dilation parameter makes the assumption that "the variance across dimensions will be *similar enough*," and this could be questionable in some situations. Min (2003) argued for multiple dilation parameters which allows re-scaling multiple dimensions individually.

MIRT Linking Estimation Procedures

The rotation matrix and scaling coefficient (translation vector and dilation parameter) are estimated separately. There is only one way of estimating the rotation matrix, however Li and Lissitz (2000) used three sets of methods to estimate the scaling coefficients: 1) matching the test response surface for the simultaneous estimation of scaling coefficients; 2) least squares estimation for the translation parameters combined with the ratio of the eigenvalues for estimating the dilation parameter; and 3) least squares estimation for the translation parameters combined with the ratio of the trace for estimating the dilation parameter.

Estimation of the Rotation Matrix

Ordinary orthogonal Procrustes rotation (Schönemann, 1966) is used to obtain the transformation matrix, $\mathbf{T}$, to minimize the sum of the squared differences between each item's pair of item discrimination estimates between forms (Li & Lissitz, 2000).

$$\mathbf{E}_1 = \mathbf{A}_Y \mathbf{T} - \mathbf{A}_X, \tag{2.68}$$

where $\mathbf{E}_1$ is the residual matrix, $\mathbf{A}_X$ and $\mathbf{A}_Y$ are the discrimination parameter matrices for form X and Y, respectively.

Matching Test Response Surface for Estimating Translation and Dilation Parameters

The matching test response surface is an extension of Stocking and Lord's (1983) procedure for MIRT models (Oshima et al., 2000). The scaling coefficients are obtained

so as to minimize the difference in the expected number correct on base test and linked tests. Therefore, the function minimized is similar to the test characteristic function method (equation 2.60) in Oshima et al. (2000).

Least Squares Estimates of Translation Parameters

Least squares are used to estimate the elements in $\boldsymbol{\beta}$ (translation parameters), β_1 and β_2 (Li & Lissitz, 2000). The criterion Q is:

$$Q = \sum_{j=1}^{J}(d_{jX} - d_{jX}^*)^2, \tag{2.69}$$

which is written using equation 2.76 as below.

$$Q = \sum_{j=1}^{J}(d_{jX} - d_{jY} - (\mathbf{a}_{jY}^T\mathbf{T})\boldsymbol{\beta})^2 \tag{2.70}$$

where J is number of common items.

Ratio of eigenvalues for estimating the dilation parameter

This is for estimation of the dilation parameter, k, in Li and Lissitz (2000). This is the ratio of the square root of the maximum eigenvalue of $\mathbf{X^TX}$ and the square root of the maximum eigenvalue of $\mathbf{Y^TY}$. The singular value of X is the non-negative square root of the eigenvalue of $\mathbf{X^TX}$. The dilation parameter k is:

$$k = \frac{Max[Sig(X)]}{Max[Sig(Y)]}. \tag{2.71}$$

Least squares estimation for the dilation parameter

This is another estimation procedure for the dilation parameter, k (Li & Lissitz, 2000). In this method, the sum of squared error of the residual matrix $\mathbf{E}_2$ is minimized,

$$\mathbf{E}_2 = (k\mathbf{A}_Y\mathbf{T}) - \mathbf{A}_X. \tag{2.72}$$

Here, $\mathbf{T}$ and k are derived simultaneously through the following steps: (1) center all elements of $\mathbf{A}_Y$ to create the centered matrix, $\mathbf{A}_{CY}$; (2) center all elements of $\mathbf{A}_X$ to create the centered matrix, $\mathbf{A}_{CX}$; (3) perform a standard orthogonal Procrustes rotation to obtain $\mathbf{T}$, and (4) compute the dilation parameter k through following equation (Li & Lissitz, 2000):

$$k = \frac{trace(\mathbf{T}^{\mathbf{T}}\mathbf{A}^{\mathbf{T}}{}_{CY})\mathbf{A}_{CX}}{trace(\mathbf{A}^{\mathbf{T}}{}_{CY})\mathbf{A}_{CY}}. \tag{2.73}$$

The research design and results

Li and Lissitz (2000) conducted two studies to evaluate MIRT equating on the three transformation parameters: an orthogonal Procrustes rotation, a translation transformation, and a single dilation. Three sets of methods to estimate the scaling coefficients (the translation coefficients and the dilation coefficient) were used in the study: matching the test response surface, and least squares for the translation parameters combined with the ratio of the eigenvalues or the ratio of the trace for the dilation parameter.

In the first study of Li and Lissitz (2000) the effect of error in parameter estimates on the precision of transformation parameter estimation was examined. Based on the

results of the first study, the best MIRT equating method was selected for the second study to equate MIRT parameters onto a known scale.

In the first study, the variables considered were sample size, MIRT equating method, equating procedure, the number of anchor-test items, the number of common items, and the use of horizontal and vertical equating. The study used known parameters from an existing test.The results were analyzed using the average differences between true parameters and the corresponding estimates (BIAS), and the root-mean-square errors (RMSE)(Skaggs & Lissitz, 1988).

In their second study, two combinations of ability level composites in the equated group were generated, normal and skewed. The sample size was 2,000. Parameter estimates were obtained using TESTFACT (Wood et al., 1987). The accuracy of the equating method on equating ability level and item parameters onto a common scale was assessed using bias and RMSE.

The Li and Lissitz (2000) study showed that biases from each of the methods were all close to zero, however RMSE differed across the estimation methods examined. For the estimation of the dilation parameter, the ratio of trace method performed better than others. For the estimation of the translation matrix, the least square method consistently performed better. Therefore, the most appropriate MIRT linking method in Li and Lissitz (2000) is the combination of least squares estimation for estimating the rotation matrix and dilation parameter estimates [Procrustes rotation and the ratio of trace (equations 2.72 and 2.73)], and translation matrix (equation 2.69).

2.6.6 Min's (M) Method

MIRT linking models

Min (2003) extended the model by Li and Lissitz (2000) by replacing a single dilation parameter with a diagonal dilation matrix to allow different unit changes in different dimensions. For example, with 2 dimensions, the dilation matrix $\mathbf{K}$ has diagonal elements, k_1 and k_2 and off diagonal of zero. The dilation component for the first and second dimensions are k_1 and k_2, respectively. The model by Min (2003) is here re-written slightly to parallel the Li and Lissitz (2000) model (equations 2.65, 2.66, and 2.67):

$$\mathbf{a}_{jX} = k\mathbf{a}_{jY}^T\mathbf{T}, \tag{2.74}$$

$$\theta_{iX} = \mathbf{K}^{-1}(\mathbf{T}^{-1}\theta_{iY} - \boldsymbol{\beta}), \tag{2.75}$$

and

$$d_{jX} = d_{jY} + (\mathbf{a}_{jY}^T\mathbf{T})\boldsymbol{\beta}, \tag{2.76}$$

where $\mathbf{K}$ is the dilation matrix.

MIRT linking estimation procedures

Min's dilation

The dilation constant k in Li and Lissitz (2000) is replaced with a diagonal dilation matrix to model different unit changes along with an orthogonal rotation in MIRT linking. The dilation matrix $\mathbf{K}$ has diagonal elements, k_1 and k_2, and off diagonal

elements of zero. The dilation component for the first and second dimensions are k_1 and k_2, respectively. The rotation matrix is the same as in Li and Lissitz method. The dilation vector is obtained using the following formula (Min, 2003):

$$\mathbf{K} = diag[\mathbf{A'_X A_Y T}] \times (diag[\mathbf{T' A'_Y A_Y T}])^{-1}, \tag{2.77}$$

where the notation is the same as described in other equations in this section.

The research design and results

Three equating methods were evaluated based on how closely item estimates for common items were transformed into item parameters in Min's study. The three methods were: 1) the test characteristic function procedure (Oshima et al., 2000); 2) the composite procedure with Procrustes orthogonal rotation (Li & Lissitz, 2000); and 3) the method with a dilation matrix following the criteria of Li and Lissitz (2000).

Li and Lissitz (2000) used item parameters from real data and underlying abilities were orthogonal in their study. Min (2003) simulated item parameters under two types of dimensional structures: approximate simple structure (APSS) and mixed structure (MS) (Roussos et al., 1998; Kim, 1994, 2001). Under APSS, a set of items has high discrimination on one dimension and low discriminations on the other dimension. In MS, a set of items has relatively high discrimination on the composite of two dimensions. Another factor considered was sample size (1,000 and 2,000). The distributions of ability were all normal. The accuracy of the equating method on

equating ability level and item parameters onto a common scale was assessed using
BIAS and RMSE. The results of MIRT using real data were also examined based
on mean difference and difference variation of item parameter estimates as well as
exploration of the largest/smallest discrepancy between base and linked test response
surfaces.

Min's study showed that, by having two dilation parameters for the two dimen-
sional test, the orthogonal Procrustes transformation is more accurate than the Li
and Lissitz (2000) method which had only one dilation parameter common to all di-
mensions. By comparing three MIRT linking methods, Min (2003) commented that
the oblique rotation of the Oshima, Davey, Lee (2000) method may provide closer
agreement of dimensional orientation.

2.6.7 Non-Orthogonal Procrustes (NOP) Method

MIRT linking models

Yon (2006) conducted a MIRT linking study specifically for the vertical equating
situation. The NOP method suggested by Reckase & Martineau (2004) is similar to
Min's method, but the rotation matrix is non-orthogonal. By using the NOP rotation
matrix, the dilation matrix is not needed as in the study by Oshima et al. (2000).
The difference between the NOP method and that of Oshima et al. (2000) is the way
they estimate the rotation matrix and translation vector. Oshima et al. (2000) es-
timates them simultaneously, and thus has an inderminancy problem that relates to
translation, rotation, and scaling. NOP by Yon (2006), Li & Lissitz (2000), and Min's

methods (Min, 2003) resolves the indeteminancy problem by estimating the rotation and translation matrces separately and by the use of Procrustean linear transformation. In Li & Lissitz (2000) and Min (2003), the rotation matrix is orthogonal; that is, all items are rotated by a fixed number of degrees. In the NOP, each item is not rotated by a fixed number of degrees, thus the relationships among item dimension vectors change before and after the rotation. The model for the NOP is here re-written slightly to parallel the Min (2003) model(equations 2.74, 2.75, and 2.76):

$$\mathbf{a}_{jX} = \mathbf{a}_{jY}^{T}\mathbf{T}, \tag{2.78}$$

$$\theta_{iX} = \mathbf{T}^{-1}\theta_{iY} - \boldsymbol{\beta}, \tag{2.79}$$

and

$$d_{jX} = d_{jY} + (\mathbf{a}_{jY}^{T}\mathbf{T})\boldsymbol{\beta}. \tag{2.80}$$

The difference from the Min method is that $\mathbf{K}$ in Min's method is eliminated and and the rotation matrix is non-orthogonal. Also, the process of obtaining the translation vector in NOP is different from Li & Lissitz (2000) and Min (2003).

MIRT linking estimation procedures

NOP rotation matrix

Yon (2006) used the oblique (non-orthogonal) procrustes rotation matrix which is obtained as follows (Mulaik,1972):

$$\mathbf{T} = (\mathbf{A}_{\mathbf{Y}}'\mathbf{A}_{\mathbf{Y}})^{-1}\mathbf{A}_{\mathbf{Y}}'\mathbf{A}_{\mathbf{X}} \tag{2.81}$$

Translation vector in NOP

Yon (2006) claims that the procedure for obtaining the translation vector in Li & Lissitz (2000) and Min (2003) seems to work well only in a low-dimensional space. Li & Lissitz (2000) and Min (2003) obtain the translation vector elements by taking the derivative of equation 2.69 with respect to each element of $\boldsymbol{\beta}$. In NOP, the translation vector elements are obtained by taking the derivative of the same equation 2.69 with respect to the entire vector. A problem of this procedure is that the decision as to which dimension of Form Y is to be matched with each dimension of the base Form X is left entirely up to the mathematical procedure.

The research design and results

Yon (2006) compared the TCF method in Oshima, et al. (2000) and the NOP method in vertical scaling. Grade levels considered were 6, 7, and 8; the 6th grade and 8th grade scales were linked to the 7th grade scale. The item dimension structure was approximate simple structure (APSS) and mixed structure (MS) as in Min (2003). The evaluation of vertical scaling was done by bias, correlation between estimates from the base and the equated forms, and RMSE. The simulation results were analyzed in a repeated measures design. Yon's study showed that NOP was a better linking method than TCF for the item discrimination matrix. The correlation between estimates from the base and equated forms and RMSE were also superior with NOP. Although the bias of NOP for item difficulty was larger than TCF, NOP was slightly better than TCF overall.

2.6.8   Summary of MIRT Simulation Studies

MIRT linking involves re-rotation of reference systems, re-scaling the unit length (similar to slope in unidimensional IRT linking), and shifting the point of origin (similar to the intercept in unidimensional IRT linking). So far, the studies of MIRT linking are distinguished from each other by having different models and estimation procedures. The MIRT linking models differ in whether the model has an orthogonal or non-orthogonal rotation matrix, having or not having a dilation parameter to re-scale unit length, and/or having single or multiple dilation parameters in the model. The estimation of transformation matrices and parameters can be done simultaneously or separately, and there is usually more than one way of obtaining the estimates.

As in unidimensional IRT linking, concurrent calibration should be possible for MIRT linking. Under concurrent calibration, since all items in different forms are calibrated simultaneously, parameter estimates are on a common scale after one run.

2.7   Summary

In the comparison of concurrent and separate linking in unidimensional IRT, the evaluation criteria were mostly in the form of standard errors. The standard error can be calculated for true score as well as for each item parameter estimate. The studies on the comparison of concurrent and separate UIRT linking show that concurrent calibrations are more accurate than separate linking when the data fit the UIRT model. Concurrent calibration is, however, less robust to violations of the UIRT assumptions

than separate linking using the test characteristic curve method of Stocking and Lord. The conditions examined in the reviewed studies are a small number of common items, nonequivalent groups, and multidimensional latent ability levels. When the number of common items is small, or when groups are nonequivalent, separate linking tended to show better performance than concurrent estimation. With a small difference between non-equivalent groups, concurrent calibration generally performed better than separate linking methods. When there is multidimensionality, concurrent estimation seems to perform better than separate linking when the violation is small.

Based on the results from the studies on concurrent and separate linking methods, it is expected that, in MIRT linking, concurrent calibration will perform better than separate linking when conditions are ideal; e.g., equivalent groups and zero correlation among ability dimensions. As in the unidimensional IRT situation, separate linking is expected to perform better than concurrent calibration in the nonequivalent groups design.

The MIRT studies by Oshima et al. (2000), Li and Lissitz (2000) and Min (2003) were reviewed in this chapter. Oshima et al. (2000) showed that the test characteristic function method performed better than the other three separate linking methods. Li and Lissitz (2000) had a dilation scalar that was not considered in Oshima et al. (2000) and concluded that the combination of least squares estimation for estimating the rotation matrix and dilation parameter estimates [Procrustes rotation and the ratio of trace (equations 2.72 and 2.73)], and translation matrix (equation 2.69) per-

formed better than other estimation procedures considered in Li and Lissitz (2000). Min (2003) improved the best procedure in Li and Lissitz (2000) by having unique dilation parameters for each dimension.

The research on MIRT linking is still in a developing phase where there are many uncertainties. In one simulation study, the researcher can examine the accuracy of only a limited number of MIRT equating procedures. The MIRT equating studies reviewed here looked closely at different linking methods; however, none of them compared separate linking methods against concurrent calibration. It is shown in unidimensional IRT linking that separate linking method (test characteristic method) tend to yield better linking results than concurrent calibration when there is severe multidimensionality or markedly non-equivalent groups. It is hypothesized that similar findings will be obtained with MIRT linking, however, the magnitude of the effect of group non-equvalence, for example, is unknown at this point.

Li and Lissitz (2000) examined skewed ability distributions; however, Min (2003) did not. A simulation study using Min's method with skewed ability distributions would give more information on Min's method. Thompson et al. (1997) conducted MIRT linking with multiple subjects in a test; however, most simulation studies used only one subject (e.g., mathematics) when they used real data. Having multiple subjects in a test to be equated (e.g., combination of mathematics and reading) may also be beneficial to gain more understanding of MIRT linking with real data since multiple subjects clearly give multiple dimensions. When a simulation study has

multiple subjects in a test, it is also possible to compare the linking results using MIRT linking and unidimensional IRT linking for each subject separately, which may provide a better understanding of MIRT linking.

2.8 Hypotheses in This Study

The purpose of this paper is to compare concurrent and separate MIRT linking methods. It is hypothesized that results will be similar to those in concurrent and separate UIRT linking methods. That is, concurrent calibration would have smaller error than separate linking methods when groups are equivalent and correlation between ability dimensions is zero. Concurrent calibration might still perform better than separate linking under minor departure from group equivalence and zero correlation between ability dimensions. Also, since concurrent calibration can benefit from having larger sample size than separate linking, concurrent calibration is expected to perform better than separate linking methods when sample size is small.

Chapter **3**

Methodology

3.1 Repeated Measures Design

In this study, four separate linking methods and concurrent calibration were applied to the same test response patterns. A repeated measures multivariate analysis of variance model (MANOVA) was used to evaluate effects of simulation conditions (between-factors) and linking method (within-factor) on each dependent variable. There were four between-factors: sample size, test length, equivalent/nonequivalent groups, and correlation between ability dimensions. The interaction terms for between- by within-factors were examined.

Besides repeated measure ANOVA, descriptive statistics on dependent variables and line plots using the mean of the dependent variables over replications were examined to explore more detailed patterns of linking errors and to compare the four separate linking methods with concurrent calibration.

3.2 Data Generation

Data for the simulation study were based on state standardized achievement tests for reading and mathematics. A form included 68 items and 40 items for mathematics and reading, respectively. Using ten forms for each subject, item parameters in the 3 parameter logistic (3PL) model were obtained using the computer software BILOG-MG (Zimowski et al., 1996) for each form separately. Although the forms were different, the original raw score and scaled score across forms were essentially the same (data not shown). Therefore, the item parameters obtained from each form were not equated with the assumption that they are already on the same scale.

From the BILOG output, only selected ranges of item parameters were kept for the simulation study. Items with item discrimination parameter values between 0.5 and 1.5, with item difficulty parameter values between -2 and 2 were kept in the item pool. Items with item dicrimination parameters and item difficulty parameters beyond these ranges were considered less representive of realistic achievement test items.

To describe the difference in item difficulty between unidimensional IRT (UIRT) and MIRT, the equations for UIRT and MIRT are described next. The unidimensional IRT model is as follows:

$$P\{X_{ij} = 1|\theta_i, a_j, b_j, c_j\} = c_j + (1 - c_j)\frac{\exp[a_j(\theta_i - b_j)]}{1 + \exp[a_j(\theta_i - b_j)]}, \qquad (3.1)$$

where $P\{X_{ij} = 1|\theta_i, a_j, b_j, c_j\}$ is the probability of the ith examinee answering the jth item correctly given the ith examinee's ability level (θ_i), the jth item's discrimination

parameter (a_j), item difficulty (b_j), and the item guessing parameter (c_j).

The MIRT compensatory model with three parameters is as follows (Embretson & Reise, 2000; Reckase, 1997):

$$P\{X_{ij} = 1|\theta_i, \mathbf{a}_j, d_j, c_j\} = c_j + (1 - c_j)\frac{\exp(\sum_{k=1}^m a_{jk}\theta_{ik} + d_j)}{1 + \exp(\sum_{k=1}^m a_{jk}\theta_{ik} + d_j)}, \qquad (3.2)$$

where d_j is the difficulty for the jth item. The easier the item, the higher the value of d_j.

As can be seen from equation 3.1 and equation 3.2, the item difficulty parameter in the UIRT model is b and is opposite in sign to d in the multidimensional IRT model. In the unidimensional case, the parameters d and b are related as follows:

$$d_{MIRT} = -a_{UIRT}b_{UIRT}, \qquad (3.3)$$

where d_{MIRT} is the item difficulty in the MIRT model, a_{UIRT} is the item discrimination parameter obtained from BILOG, and b_{UIRT} is the item difficulty obtained from BILOG.

With ten forms, 384 and 281 items were obtained in mathematics and reading from which to select items for the simulation. For each combination of a base group and an equated group, two different total item sets were used with fixed common items: 60 items in total with 20 items as common items, and 40 items in total with 20 items as common items. To create the common items, 10 items were randomly selected without replacement from the item pools of mathematics and reading respectively, making 20 common items per form in total. A study by Hanson et. al (2000) randomly

selected common items. In this study, the common items were also selected randomly to represent the total test forms in content and statistical characteristics (Kolen & Brennan, 2004). The 10 unique items were randomly selected without replacement from each subject, making a total of 20 unique items, to make a total of 40 items (20 common and 20 unique). To make the comparison of test length with the same distributions of item difficulty and discrimination parameters for unique items, the unique items used in a 40-item test were duplicated to make a 60-item test. That is, a 60-item test form was constructed with 20 common items and two sets of the 20 unique items used in a 40-item test form. The two sets of 20 items were identical in their distributions of item parameters. Thus, the final 60 item test contained 10 common math items, 10 common reading items, 20 unique math items, and 20 unique reading items.

Since only multidimensional simple structure was considered, the generating item discrimination parameter for dimension 1 was non-zero for mathematics items, and zero for reading items. Likewise, the generating item discrimination parameter for dimension 2 was non-zero for reading items, and zero for mathematics items. Table 3.1 shows the means of the generating item parameters.

3.3   Independent Variables

Independent variables/factors in this study were; (1) 3 sample sizes, (2) 2 test lengths, (3) 3 levels of equivalent and non-equivalent ability levels between form X and Y, (4)

Table 3.1. Average values of generating item parameter statistics.

Item Group	Discrimination 1[1]	Discrimination 2[1]	Difficulty	Guessing
Common 1[1]	1.102	0.000	1.032	0.229
Common 2[1]	0.000	0.883	1.294	0.253
Unique1: x60[1]	0.820	0.000	0.560	0.233
Unique1: x40	0.820	0.000	0.560	0.233
Unique1: y60	0.895	0.000	0.942	0.230
Unique1: y40	0.895	0.000	0.942	0.230
Unique2: x60	0.000	0.793	0.907	0.224
Unique2: x40	0.000	0.793	0.907	0.224
Unique2: y60	0.000	0.933	1.207	0.242
Unique2: y40	0.000	0.933	1.207	0.242

[1] 1 indicates the item set for dimension 1; 2 indicates the item set for dimension 2; 'Unique1: x60' is the unique item set of base items for dimension 1, x, in 60-item form.

3 correlation levels between 2 ability dimensions, and (5) 5 estimation methods for MIRT linking (Min, Direct, TCF, ICF, and MIRT concurrent calibration).

In a typical equating situation, there were two groups of examinees, the base group and target (or equated) group. In this study, the base group took Form X and the equated group took Form Y. Sample sizes used were 500, 1000, and 3000, to represent vary small, small, and adequate sample sizes. The sample sizes for both forms were set to be equal. Kim and Cohen (2002) used 300 as their smallest sample size for their UIRT linking study. However, the minimum sample size suggested for the graded response models in UIRT is 500 (Reise & Yu, 1990). The sample size of 500 was chosen to represent a very small sample size since it is suggested as a small sample size for UIRT. A simulation study on MIRT item parameter recovery used

a sample size of 2000, and reports that item parameters were adequately recovered using MIRTE (Carlson, 1987) when dimensions were correlated and abilities were normally distributed (Batley & Boss, 1993). Therefore, a sample size of 3000 would likely to be sufficient. A sample size of 1000 was chosen to examine a somewhat small sample size for MIRT.

Two different lengths of total item sets were used with fixed common items: 60 items in total with 20 items as common items, and 40 items in total with 20 items as common items. The common items should be at least 20 percent of the test (Cook & Eignor, 1991). Having 20 items as common items in a 60-item test is more than 20 percent of the total items and similar to some other studies (Béguin et al., 2000b; Hanson & Béguin, 2002). Kim and Cohen (1998) used 5, 10 , and 25 items as common items in 50-item tests.

The ability parameters of the base and equated groups were drawn from a normal distribution, with nine conditions (3 different levels of correlations times 3 different levels of equivalence of average ability): (1) both groups had mean ability zero, and a standard deviation of 1, N(0,1) (call this condition equivalent groups) with zero correlation between dimensions, (2) equivalent groups with 0.5 correlation between dimensions, (3) equivalent groups with 0.8 correlation between dimensions, (4) base group had ability distribution N(0,1), equated group had N(0.5,1) (call this non-equivalent groups (.5SD)), with zero correlation between dimensions, (5) non-equivalent groups (.5SD) with 0.5 correlation between dimensions, (6) non-equivalent

groups (.5SD) with 0.8 correlation between dimensions, (7) base group had ability distribution N(0,1), equated group had N(1,1) (call this non-equivalent groups (1SD)), with zero correlation between dimensions, (8) non-equivalent groups (1SD) with 0.5 correlation between dimensions, (9) non-equivalent groups (1SD) with 0.8 correlation between dimensions. The ability distributions were all multivariate normal. The distribution of abilities was set to normal because the distributions of scaled scores in the real mathematics and reading data were very close to normal; the skewness and kurtosis values were less than absolute value of 0.3.

The size of correlations between ability dimensions, $r = 0, 0.5, 0.8$, were chosen to represent zero (very low), moderate, and high correlations. A simulation study with 4 dimensional MIRT by Yao and Boughton (2007) used correlations between dimensions as high as 0.9. A MIRT linking study by Min (2003) used 0.5 as the highest correlation condition. A study by Hanson et. al (2000), in which multidimensional data was simulated and unidimensional linking methods were applied, used a correlation of 0.71 as the highest correlation condition. With these studies in mind, the correlation for this study, $r = 0, 0.5, 0.8$, were considered appropriate.

In the equivalent groups condition, the average abilities of the two groups were set to zero along both dimensions. In the non-equivalent groups conditions, a 1 standard deviation difference and a 0.5 standard deviation difference between the two groups were used. In all conditions, the average abilities along dimension 1 and 2 were set equal. It is common to have a 1 standard deviation difference in simulation studies to

represent non-equivalent groups (Béguin et al., 2000b; Hanson & Béguin, 2002; Kim & Cohen, 1998, 2002). A MIRT linking study by Min (2003) used a half standard deviation difference to represent non-equivalent groups. Therefore, the difference in means for non-equivalent groups for this study was considered appropriate.

3.4 Dependent Variables

One of the ways to evaluate results from the different methods is a comparison between the estimated item parameters and the generating item parameters. RMSE and bias were applied as the indices of linking quality. RMSE shows the root mean square difference between the transformed item parameter and the generating parameter. Bias shows the average difference between the transformed item parameter and the generating parameter. Since RMSE and bias are comparisons of transformed item parameters to the generating parameters, they are indices of systematic error.

RMSE for the item discrimination parameter for a form was obtained as follows:

$$RMSE_{a_1} = \sqrt{\frac{\sum_{j=1}^{J}(\hat{a}_{1j} - a_{1j})^2}{J}},\qquad(3.4)$$

where a_{1j} is the generating item discrimination parameter for the jth item in the first dimension, $\hat{a}_{1j}$ is the transformed value of a for the jth item along the first dimension, and J is the total number of items. The item discrimination parameter bias for a form was calculated as follows:

$$BIAS_{a_1} = \frac{\sum_{j=1}^{J}(\hat{a}_{1j} - a_{1j})}{J}.\qquad(3.5)$$

Since this equation for bias is not squared, it preserves the sign of the discrepancy between the observed value and the generating parameter.

The correlation between transformed item parameter estimates and generating item parameters was also calculated as another dependent variable. Also, when the sample size is small and there are correlated ability dimensions and non-equivalent groups, TESTFACT (Wood et al., 1987) may not estimate the rotated factor loading well and separate linking methods may not converge. Therefore, the number of replications with successful convergence was also computed as a dependent variable.

3.5 Linking

This study considered five different MIRT linking methods including the concurrent method for MIRT: (a) MIRT separate linking using Min's method (Min, 2003), (b) separate linking using the Direct method (Oshima et al., 2000), (c) MIRT separate linking using the Multidimensional Test Characteristic Function (TCF)(Oshima et al., 2000), (d) MIRT separate linking using the Multidimensional Item Characteristic Function (ICF)(Oshima et al., 2000), and (e) concurrent calibration.

Using the simulated ability distribution, item responses were simulated using R (R Development Core Team, 2007). The computer program TESTFACT (Wood et al., 1987) was used to estimate the item parameters for separate and concurrent calibration. The rotated factor loadings were used to obtain the rotated item parameter estimate for each item.

Final estimates of item parameters from separate linking and concurrent calibrations were expressed on the metric of the base group (form X scale). Kim and Cohen (1998, 2002) state that final item parameter estimates from separate linking and concurrent calibrations may not be directly comparable. To compare the estimates, Kim and Cohen (1998, 2002) performed another separate linking to place both separate linking and concurrent calibration item parameter estimates onto the metric of generating item parameters.

Following the same idea as Kim and Cohen (1998, 2002), this study employed two steps of linking. The first step was the transformation of the scale of Y onto the scale of X, and the second step was the transformation of the transformed Y (Y star) onto the scale of the generating parameters. In the first step, the item estimations were equated onto the same scale using four separate linking methods using common items. The obtained linking parameters were used to transform all items on form Y. For concurrent calibration, unique items on form X taken by the base group were indicated as not presented for the group who took form Y and vice versa. In the second step, the linked item parameters from separate linking and item parameters from concurrent calibrations were equated onto the generating item parameters using Min's method. Among the four separate linking methods considered in this study, Min's method was chosen as the common linking method to transform onto the generating scale after the first step of linking to conserve the angle between the two dimensions in the first step since Min's method was the only method that

employs an orthogonal transformation matrix. In the second step of linking, all items on form Y were used to obtain the equating parameters as opposed to common items only for the first linking.

The code for Min's method was originally written in MATLAB, with two functions missing for which a Newton-Raphson procedure was implemented. The MATLAB code was converted into R and the missing functions were completed for this study. The converted code was shown to be accurate by having the same input and output as the example in Min's thesis. All separate linking procedures were written in R.

3.6   Correcting the Dimension and Direction

Since TESTFACT did not always identify dimension 1 as dimension 1 and dimension 2 as dimension 2 in the estimated results, the dimensions were permuted when necessary before any linking. The first half of the items loaded on dimension 1 in the generating parameters. Therefore, by checking the first half items' discrimination parameters, it is possible to tell if dimensions need to be permuted or not. TESTFACT is based on a nonlinear factor analysis method for binary items. Therefore, the estimation of the signs might also be reversed. That is, even though all true discrimination parameters were positive, TESTFACT may yield estimates in which they are all negative. When averages of item discrimination parameters for each dimensions were negative, the sign was reversed. Although the software used was different, a MIRT simulation study also corrected dimensions (Batley & Boss, 1993).

Chapter **4**

Results

4.1   Successful Replications

Tables 4.1 and 4.2 show the number of successful replications out of 50 replications. The number of successful replications was 50 for all equivalent conditions. There were a few unsuccessful replications when groups were non-equivalent with 1 standard deviation difference and small sample size. Based on inspection of outputs from TESTFACT, the unsuccessful replications were mostly due to non-convergencee within TESTFACT, not due to non-convergence of linking methods.

Table 4.1. The number of successful replications with 40-item forms

Condition	Sample size	Number of items	Min	Direct	TCF[1]	ICF[1]	Con[1]
Eqv, r = 0.0	500	40	50	50	50	50	50
Eqv, r = 0.0	1000	40	50	50	50	50	50
Eqv, r = 0.0	3000	40	50	50	50	50	50
Eqv, r = 0.5	500	40	50	50	50	50	50
Eqv, r = 0.5	1000	40	50	50	50	50	50
Eqv, r = 0.5	3000	40	50	50	50	50	50
Eqv, r = 0.8	500	40	50	50	50	50	50
Eqv, r = 0.8	1000	40	50	50	50	50	50
Eqv, r = 0.8	3000	40	50	50	50	50	50
Non-Eqv (.5SD), r = 0.0	500	40	50	50	50	50	50
Non-Eqv (.5SD), r = 0.0	1000	40	50	50	50	50	50
Non-Eqv (.5SD), r = 0.0	3000	40	50	50	50	50	50
Non-Eqv (.5SD), r = 0.5	500	40	49	49	49	49	50
Non-Eqv (.5SD), r = 0.5	1000	40	49	49	49	49	50
Non-Eqv (.5SD), r = 0.5	3000	40	49	49	49	49	50
Non-Eqv (.5SD), r = 0.8	500	40	49	49	49	49	50
Non-Eqv (.5SD), r = 0.8	1000	40	49	49	49	49	50
Non-Eqv (.5SD), r = 0.8	3000	40	49	49	49	49	50
Non-Eqv (1SD), r = 0.0	500	40	44	44	44	44	45
Non-Eqv (1SD), r = 0.0	1000	40	50	50	50	50	50
Non-Eqv (1SD), r = 0.0	3000	40	50	50	50	50	50
Non-Eqv (1SD), r = 0.5	500	40	45	45	45	45	48
Non-Eqv (1SD), r = 0.5	1000	40	49	49	49	49	50
Non-Eqv (1SD), r = 0.5	3000	40	49	49	49	49	50
Non-Eqv (1SD), r = 0.8	500	40	45	45	45	45	48
Non-Eqv (1SD), r = 0.8	1000	40	49	49	49	49	50
Non-Eqv (1SD), r = 0.8	3000	40	49	49	49	49	50

[1]TCF: Test Characteristic Function method; ICF: Item Characteristic Function method; Con:concurrent calibration.

Table 4.2. The number of successful replications with 60-item forms

Condition	Sample size	Number of items	Min	Direct	TCF[1]	ICF[1]	Con[1]
Eqv, r = 0.0	500	60	50	49	50	50	50
Eqv, r = 0.0	1000	60	50	50	50	50	50
Eqv, r = 0.0	3000	60	50	50	50	50	50
Eqv, r = 0.5	500	60	50	50	50	50	50
Eqv, r = 0.5	1000	60	50	50	50	50	50
Eqv, r = 0.5	3000	60	50	50	50	50	50
Eqv, r = 0.8	500	60	50	50	50	50	50
Eqv, r = 0.8	1000	60	50	50	50	50	50
Eqv, r = 0.8	3000	60	50	50	50	50	50
Non-Eqv (.5SD), r = 0.0	500	60	50	50	50	50	50
Non-Eqv (.5SD), r = 0.0	1000	60	50	50	50	50	50
Non-Eqv (.5SD), r = 0.0	3000	60	50	50	50	50	50
Non-Eqv (.5SD), r = 0.5	500	60	49	49	49	49	50
Non-Eqv (.5SD), r = 0.5	1000	60	49	49	49	49	50
Non-Eqv (.5SD), r = 0.5	3000	60	49	49	49	49	50
Non-Eqv (.5SD), r = 0.8	500	60	49	49	49	49	50
Non-Eqv (.5SD), r = 0.8	1000	60	49	49	49	49	50
Non-Eqv (.5SD), r = 0.8	3000	60	49	49	49	49	50
Non-Eqv (1SD), r = 0.0	500	60	37	37	37	37	38
Non-Eqv (1SD), r = 0.0	1000	60	50	50	50	50	50
Non-Eqv (1SD), r = 0.0	3000	60	50	50	50	50	50
Non-Eqv (1SD), r = 0.5	500	60	45	45	45	45	47
Non-Eqv (1SD), r = 0.5	1000	60	49	49	49	49	50
Non-Eqv (1SD), r = 0.5	3000	60	49	49	49	49	50
Non-Eqv (1SD), r = 0.8	500	60	46	46	46	46	50
Non-Eqv (1SD), r = 0.8	1000	60	49	49	49	49	50
Non-Eqv (1SD), r = 0.8	3000	60	49	49	49	49	50

[1]TCF: Test Characteristic Function method; ICF: Item Characteristic Function method; Con:concurrent calibration.

4.2 RMSE and Bias

Tables 4.3 and 4.4 show repeated measures analysis results. The within factors considered in the models were main effects of the linking methods and all of the two-way interactions. The multivariate repeated measures analysis does not require the sphericity assumption; equal variances and covariance among dependent measures, while univariate repeated measures does. There was evidence (sphericity test results such as Mauchly's) that the sphericity assumption was not met in all tests, therefore, multivariate results were shown for within factors. The between factors considered were equivalence of groups, correlation between ability dimensions, sample size, and test length.

The repeated measure analyses were run for RMSE and bias for each of $a1$, $a2$, and d. Tables 4.3 and 4.4 show F values, degrees of freedom for the denominator and numerator, p values and partial eta squared as an effect size. The partial eta squared, η_p^2, for within factors was obtained as follows (Tabachnick & Fidell, 2007):

$$\eta_p^2 = 1 - \sqrt{\Lambda}, \tag{4.1}$$

where Λ is Wilk's lambda. Following Yon (2006), the partial eta squared for between factors was obtained as follows:

$$\eta_p^2 = \frac{SS_{effect}}{\sum SS_{effects} + SS_{error}}, \tag{4.2}$$

where SS_{effect} is the sum of squares for the effect and SS_{error} is the sum of squared error. The partial eta squared is the proportion of the total variance explained by

the effect uniquely.

Table 4.3. Repeated measure analysis results for RMSE.

Statistic	factors		Value	F Value	DFn[2]	DFd[2]	P value	η_p^{2}[2]
RMSE a1	Within[1]	linking	228.14	4	2618	<.0001	0.14
		linking*equivalence	13.78	8	5236	<.0001	0.02
		linking*correlation	8.54	8	5236	<.0001	0.01
		linking*sample size	23.73	8	5236	<.0001	0.03
		linking* test length	14.81	4	2618	<.0001	0.01
	Between	equivalence	708.93	2	2621	<.0001	0.14
		correlation	858.21	2	2621	<.0001	0.17
		sample size	2073.84	2	2621	<.0001	0.42
		test length	66.68	1	2621	<.0001	<.01
RMSE a2	Within[1]	linking	248.18	4	2618	<.0001	0.15
		linking*equivalence	16.23	8	5236	<.0001	0.02
		linking*correlation	8.60	8	5236	<.0001	0.01
		linking*sample size	21.39	8	5236	<.0001	0.03
		linking* test length	20.32	4	2618	<.0001	0.02
	Between	equivalence	902.83	2	2621	<.0001	0.17
		correlation	871.29	2	2621	<.0001	0.16
		sample size	2198.16	2	2621	<.0001	0.42
		test length	25.59	1	2621	<.0001	<.01
RMSE d	Within[1]	linking	5267.24	4	2618	<.0001	0.67
		linking*equivalence	996.30	8	5236	<.0001	0.60
		linking*correlation	8.12	8	5236	<.0001	0.09
		linking*sample size	34.53	8	5236	<.0001	0.05
		linking* test length	3.89	4	2618	0.0038	<.01
	Between	equivalence	1116.97	2	2621	<.0001	0.42
		correlation	0.80	2	2621	0.4490	<.01
		sample size	225.99	2	2621	<.0001	0.09
		test length	14.17	1	2621	0.0002	<.01

[1]Within factor analysis shows multivariate results which do not require the sphericity assumption. [2] DFn: degrees of freedom for numerator; DFd: degrees of freedom for denominator; η_p^{2}: partial eta squared

Table 4.4. Repeated measure analysis results for BIAS.

Statistic	Factors	Value	F Value	DFn[2]	DFd[2]	P value	η_p^2[2]
BIAS a1	Within[1]	linking	144.18	4	2618	<.0001	0.09
		linking*equivalence	36.77	8	5236	<.0001	0.05
		linking*correlation	42.46	8	5236	<.0001	0.06
		linking*sample size	27.89	8	5236	<.0001	0.04
		linking* test length	6.08	4	2618	<.0001	<.01
	Between	equivalence	240.25	2	2621	<.0001	0.12
		correlation	30.13	2	2621	<.0001	0.01
		sample size	493.18	2	2621	<.0001	0.24
		test length	15.54	1	2621	<.0001	<.01
BIAS a2	Within[1]	linking	151.62	4	2618	<.0001	0.10
		linking*equivalence	37.18	8	5236	<.0001	0.05
		linking*correlation	45.31	8	5236	<.0001	0.06
		linking*sample size	27.28	8	5236	<.0001	0.04
		linking* test length	9.17	4	2618	<.0001	<.01
	Between	equivalence	329.19	2	2621	<.0001	0.14
		correlation	78.85	2	2621	<.0001	0.03
		sample size	621.17	2	2621	<.0001	0.26
		test length	19.84	1	2621	<.0001	<.01
BIAS d	Within[1]	linking	6119.74	4	2618	<.0001	0.69
		linking*equivalence	1053.04	8	5236	<.0001	0.62
		linking*correlation	16.77	8	5236	<.0001	0.02
		linking*sample size	45.21	8	5236	<.0001	0.06
		linking* test length	13.47	4	2618	<.0001	0.01
	Between	equivalence	12847.3	2	2621	<.0001	0.88
		correlation	31.69	2	2621	0.0453	<.01
		sample size	317.10	2	2621	<.0001	0.02
		test length	77.68	1	2621	<.0001	<.01

[1]Within factor analysis shows multivariate results which do not require the sphericity assumption. [2] DFn: degrees of freedom for numerator; DFd: degrees of freedom for denominator; η_p^2: partial eta squared

4.2.1   **Within Factors**

Statistical significances were found with all of the within and the between factors except the between factor of correlation for the RMSE of d. The within factor of linking method was significant with RMSE and bias for all item parameter statistics.

The effect sizes for within factors of linking with RMSE were 0.14, 0.15, and 0.67, for $a1$, $a2$, and d, respectively. The differences among linking methods were larger with RMSE of d than RMSE of $a1$ and $a2$.

The effect sizes for the interaction of linking methods and between factors were very small for RMSE of $a1$ and $a2$, however, the effect size of linking and group equivalence for RMSE of d was large, 0.60. This means that the relationship between RMSE d and linking method was different depending on groups equivalence.

Figures 4.1 and 4.2 show box plots of RMSE $a1$ and $a2$. Both figures are under equivalent groups, sample size of 3000, and test length of 60 items. Figures 4.1 (i) and (ii) are with the correlation between ability dimensions equal to zero, while Figure 4.2 is with a correlation of 0.8. Figures 4.1 and 4.2 show that concurrent calibration performed better (lower RMSE) than all separate linking methods. This was true even when groups are not equivalent with high correlation (refer to tables in Appendix). These figures also suggest that separate linking methods were not very different from each other in terms of RMSE for $a1$ and $a2$ under equivalent groups.

Figures 4.3 (i) and (ii) show the RMSE d for non-equivalent groups. RMSE d values were very similar in the equivalent groups condition, however, in the non-

equivalent groups condition, the difference among linking methods was very clear. Concurrent, the Direct, and TCF were very similar to each other, however, Min's method had very large RMSE d in the non-equivalent groups conditions. The ICF generally had lower RMSE d than other linking methods.

The effect sizes of the within factor linking method for bias were 0.09 and 0.10, for $a1$ and $a2$, respectively. Figures 4.4, 4.5, and 4.6 show the relationship between bias of $a1$ and $a2$ and correlation for each linking method with different levels of group equivalence. Figure 4.4 (i), which is for an equivalent groups condition, shows that the bias of $a1$ with concurrent calibration was slightly better than with other linking methods for all correlation levels. Figure 4.4 (ii) shows that the average values of $a2$ bias were very similar among concurrent calibration, Min, the TCF, and the ICF linking methods. However, the differences seen in Figure 4.4 (i) and (ii) were very small.

Figures 4.5 (i) and (ii), which are for a non-equivalent groups condition with 0.5 standard deviation difference, show that the mean values of bias for $a1$ with concurrent calibration was smaller than separate linking methods when the correlation was zero and 0.5, but the bias of $a2$ was larger with concurrent calibrations compared to separate linking methods. Thus, differences in bias of $a1$ and $a2$ among concurrent and separate linking methods were not very clear under non-equivalent conditions with 0.5 standard deviation difference. However, with larger non-equivalence (1SD) with high correlation among ability dimensions, concurrent calibration had smaller

bias for both of $a1$ and $a2$ (Figures 4.6). Therefore, Figures 4.4 to 4.6 show that concurrent and separate linking methods performed similarly under equivalent and non-equivalent conditions with 0.5 standard deviation difference among groups. However, concurrent performed better than separate linking methods when groups were very non-equivalent (1SD) and ability dimensions had high correlations.

Among within factors, the interaction between linking and test length had the smallest effect size for both RMSE and bias for $a1$, $a2$, and d. Therefore, the difference among linking methods differed little depending on test length examined in this study.

4.2.2 Between Factors

The largest effect size for between factors with RMSE of $a1$ and $a2$ was the sample size followed by the correlation between ability dimensions and group equivalence (Table 4.3). Figures 4.7 and 4.8 show the relationship between sample sizes and RMSE for $a1$ and $a2$ for each linking method. Since the within factor of interaction of sample size and linking methods for RMSE of $a1$ and $a2$ had a small effect size, Figures 4.7 and 4.8 also show that the relationship among linking methods did not change across sample sizes. As the between factor of sample size was large (0.42 and 0.42 for RMSE $a1$ and $a2$, respectively), Figures 4.7 and 4.8 show that RMSE decreased as sample size increased. The decline of RMSE from sample size of 500 to 1000 was larger than from 1000 to 3000. Figures 4.7 and 4.8 also show that the RMSE of $a1$ and $a2$ for concurrent calibration was consistently lower than for separate linking methods across sample sizes. Also, by comparing Figure 4.7 (i) and 4.7 (ii),

it can be seen that non-equivalent groups and high correlation increased the RMSE of $a1$ with all levels of sample size, and the difference between concurrent calibration and separate linkings became larger.

Figures 4.9 and 4.10 show RMSE for $a1$ and $a2$ across correlation levels when groups were equivalent and non-equivalent (0.5 standard deviation difference), respectively. Both figures show RMSE of $a1$ and $a2$ increased as correlation increased. The increase was larger when the correlation increased from 0.5 to 0.8 than when the correlation increased from zero to 0.5. The mean RMSE of $a1$ and $a2$ for concurrent calibration was consistently lower than for other linking methods (Figure 4.9 and 4.10) across correlation levels.

With between factors for RMSE d, group equivalence had a larger effect size than other between factors with the effect size equal to 0.42. Correlation and test length had very small effect sizes for RMSE d. Figure 4.11 (i) and (ii) show RMSE d across different levels of group equivalence when the correlations between ability dimensions were zero and 0.8. Although the mean value of RMSE d did not distinguish the linking methods when groups were equivalent, they were very different depending on the linking methods when groups were non-equivalent. Min's method had the largest average RMSE d values with non-equivalent groups, while the ICF performed better than other linking methods when there was a 1 standard deviation difference between groups.

With between factors for bias, sample size had a larger effect size than other

between factors with effect sizes 0.24 and 0.26 for $a1$ and $a2$, respectively, while correlation and test length had very small effect sizes. Figures 4.12 and 4.13 show the relationship between bias and sample size for $a1$ and $a2$. As sample size increased, the biases became closer to zero. The change in bias was larger when sample size increased from 500 to 1000 than from 1000 to 3000. Given equivalent groups and zero correlation between ability dimensions (Figure 4.12), concurrent calibration had a smaller bias of $a1$ and $a2$ with sample sizes of 500 and 1000 than separate linking methods. With sample size 3000, concurrent calibration and separate linking methods had similar average values of bias for $a1$ and $a2$. With non-equivalent groups and high correlation between ability dimensions (Figure 4.13), the Direct method performed better than other linking methods for bias of $a1$ and $a2$ with sample size 500 and 1000. However, with non-equivalent groups and correlated dimensions, concurrent calibration still performed better than the TCF and Min methods with sample sizes of 500 and 1000. When sample size was 3000, concurrent and all separate linking methods had bias of $a1$ and $a2$ very close to zero.

The largest effect size with between factors for bias of d was equivalence of group, with the effect size equal to 0.88. Correlation between ability dimensions and test length had very small effect sizes. Figure 4.14 (i) and (ii) show bias of d across level of equivalence. The mean values of d bias were very similar among linking methods when groups were equivalent. Bias increased as groups became more non-equivalent. The average value of bias for d under non-equialent conditions was largest with Min's

method. When groups were not equivalent, the bias of d for concurrent calibration was larger than for the TCF, the ICF and direct methods. The smallest bias of d under non-equivalent conditions was observed with the ICF condition. The increased bias of d from equivalent groups to non-equivalent groups with 0.5 standard deviation was about the same as that of from 0.5 standard deviation to 1.0 standard deviation difference between groups.

Figure 4.1.   RMSE $a1$ and $a2$ for equivalent groups and zero correlation condition for the 60-item form when sample size is 3000.

i) Dimension 1

ii) Dimension 2
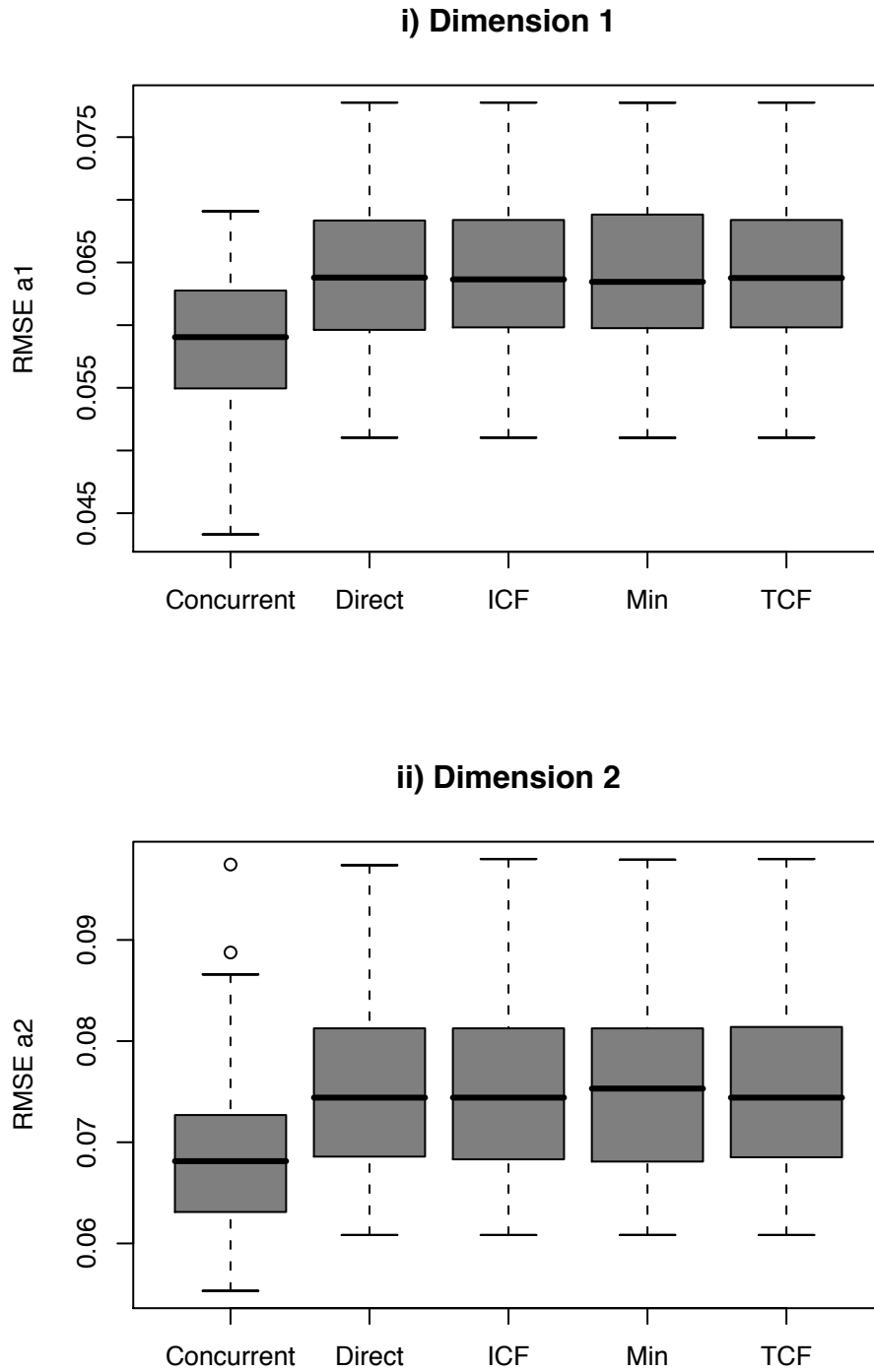
Figure 4.2. RMSE $a1$ and $a2$ for equivalent groups and 0.8 correlation condition for the 60-item form when sample size is 3000.

i) Dimension 1

ii) Dimension 2
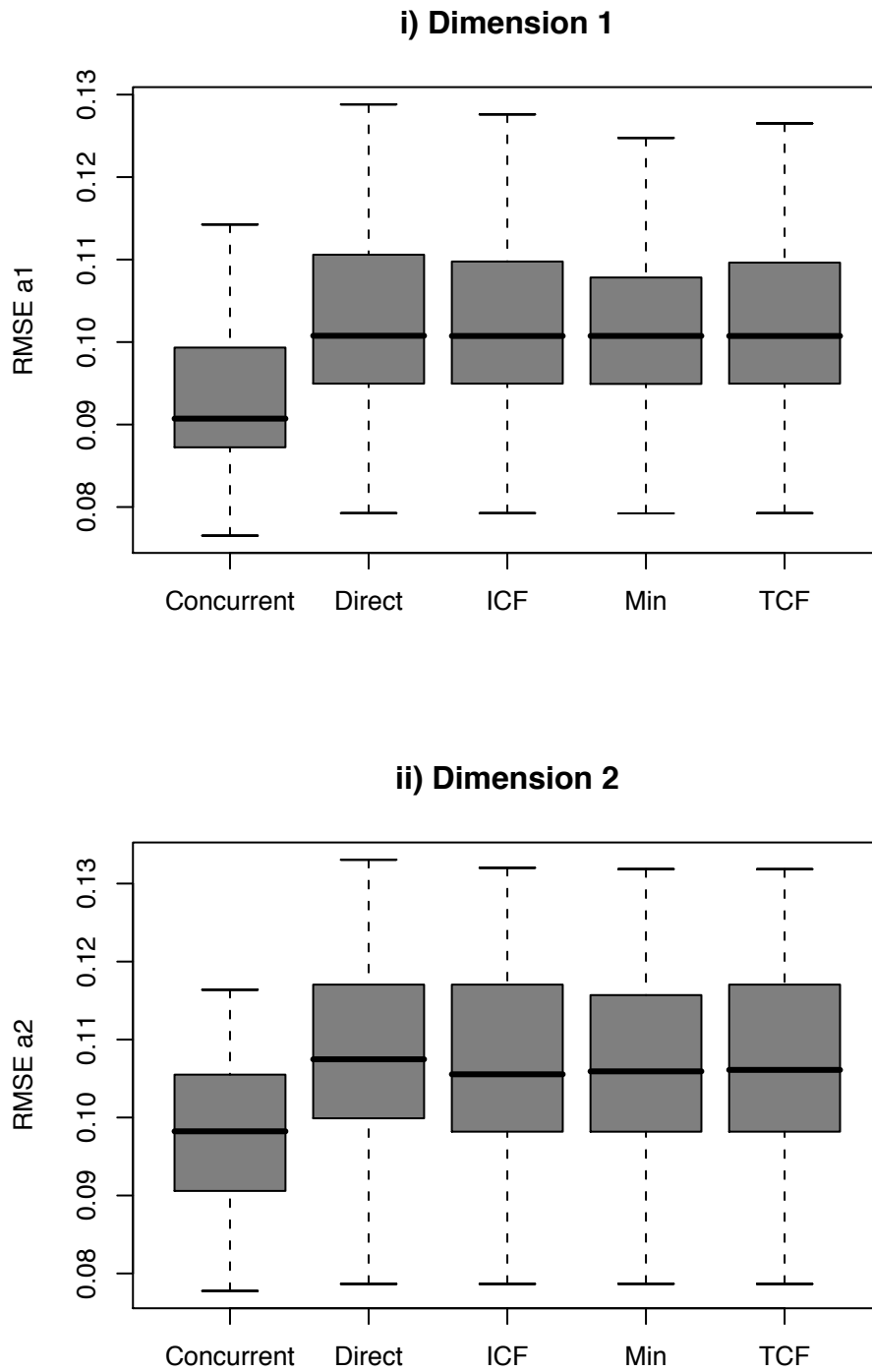
Figure 4.3. RMSE d for non-equivalent groups (.5SD) with 0 and 0.8 correlation condition for the 60-item form when sample size is 3000.

i) r = 0

ii) r = 0.8
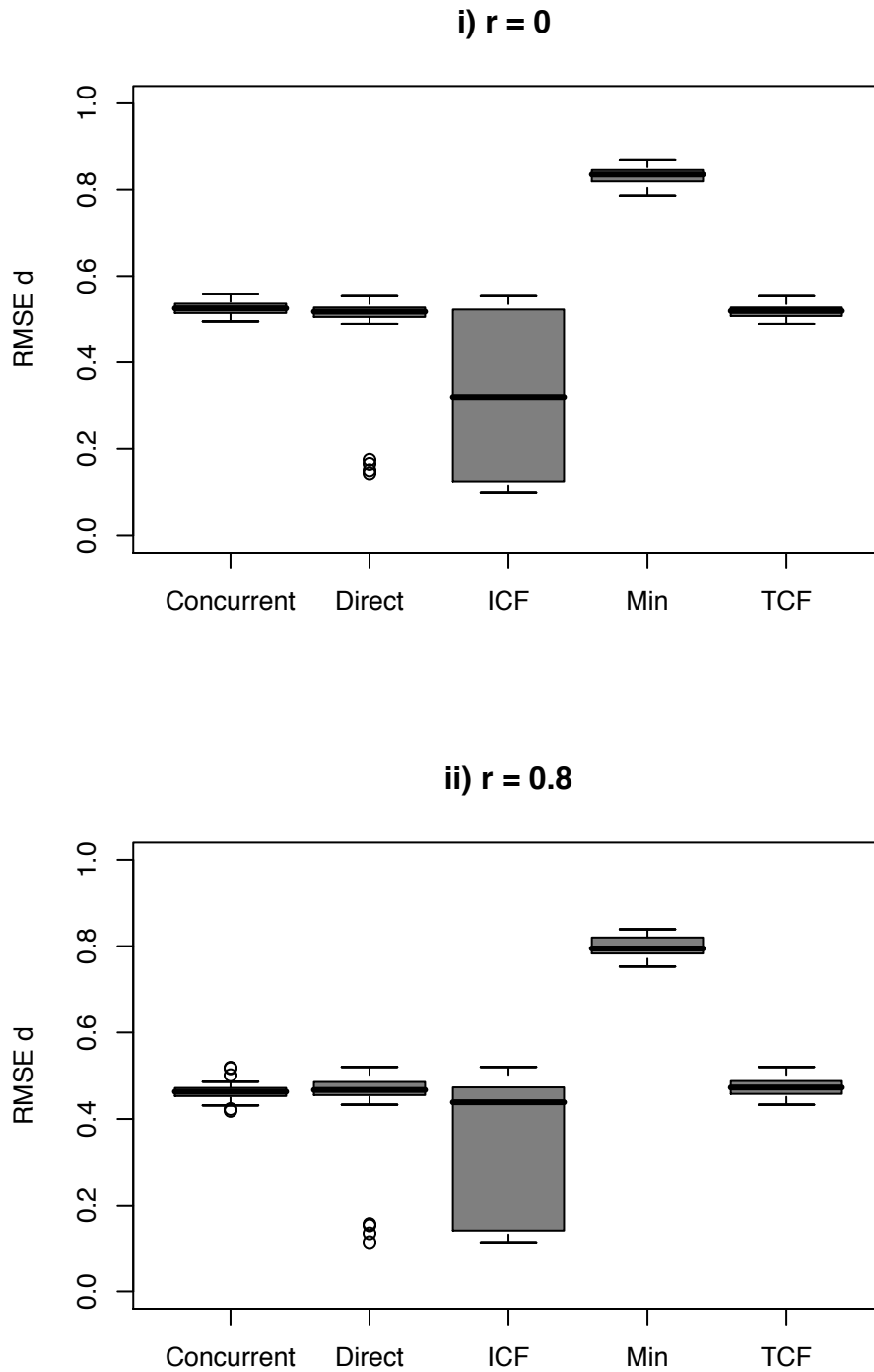
Figure 4.4. BIAS of $a1$ and $a2$ across correlation levels for equivalent groups for the 60-item form when sample size is 3000.

i) Dimension 1

ii) Dimension 2
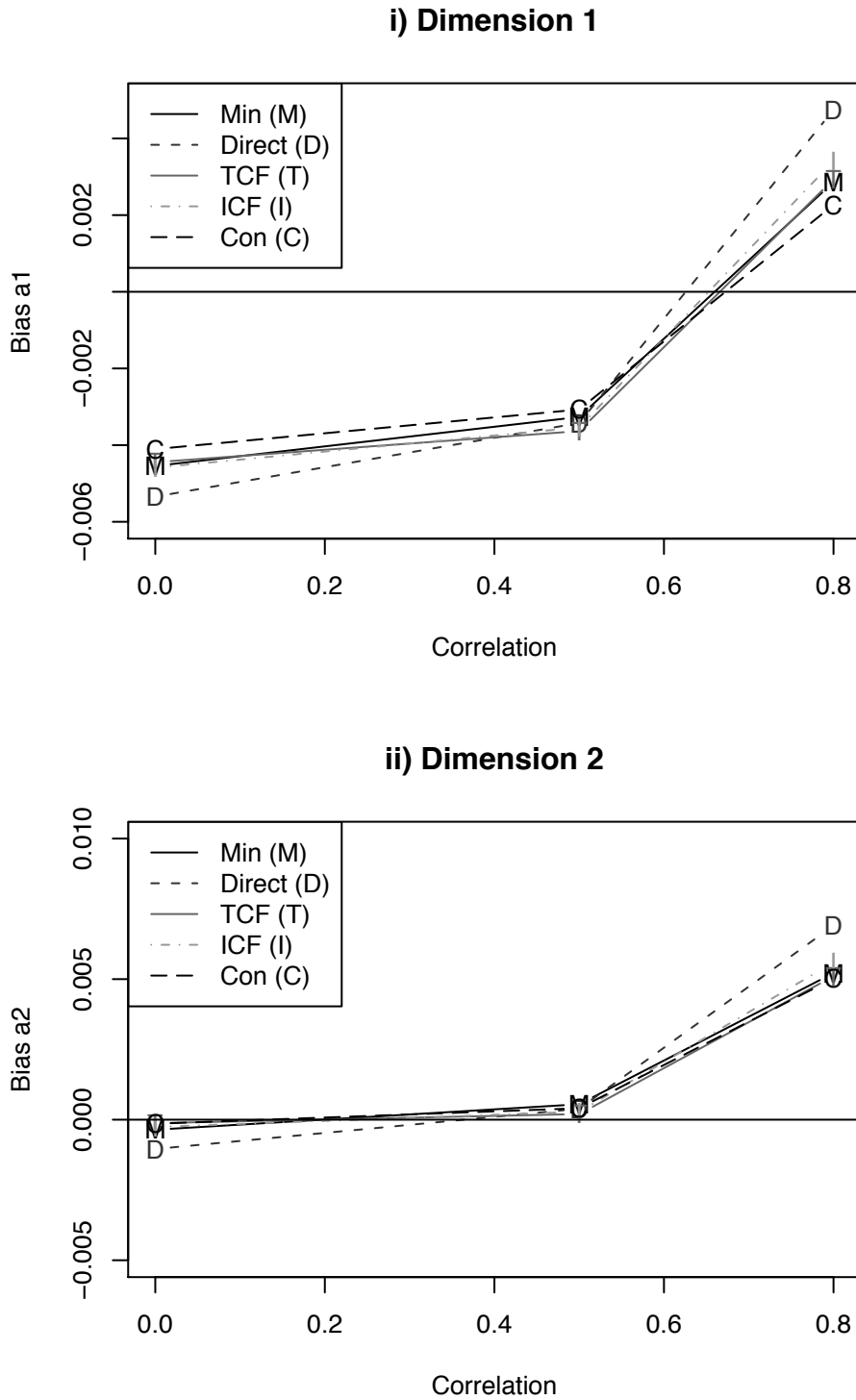
Figure 4.5.   BIAS of $a1$ and $a2$ across correlation levels for non-equivalent groups (.5SD) for the 60-item form when sample size is 3000.

i) Dimension 1

ii) Dimension 2
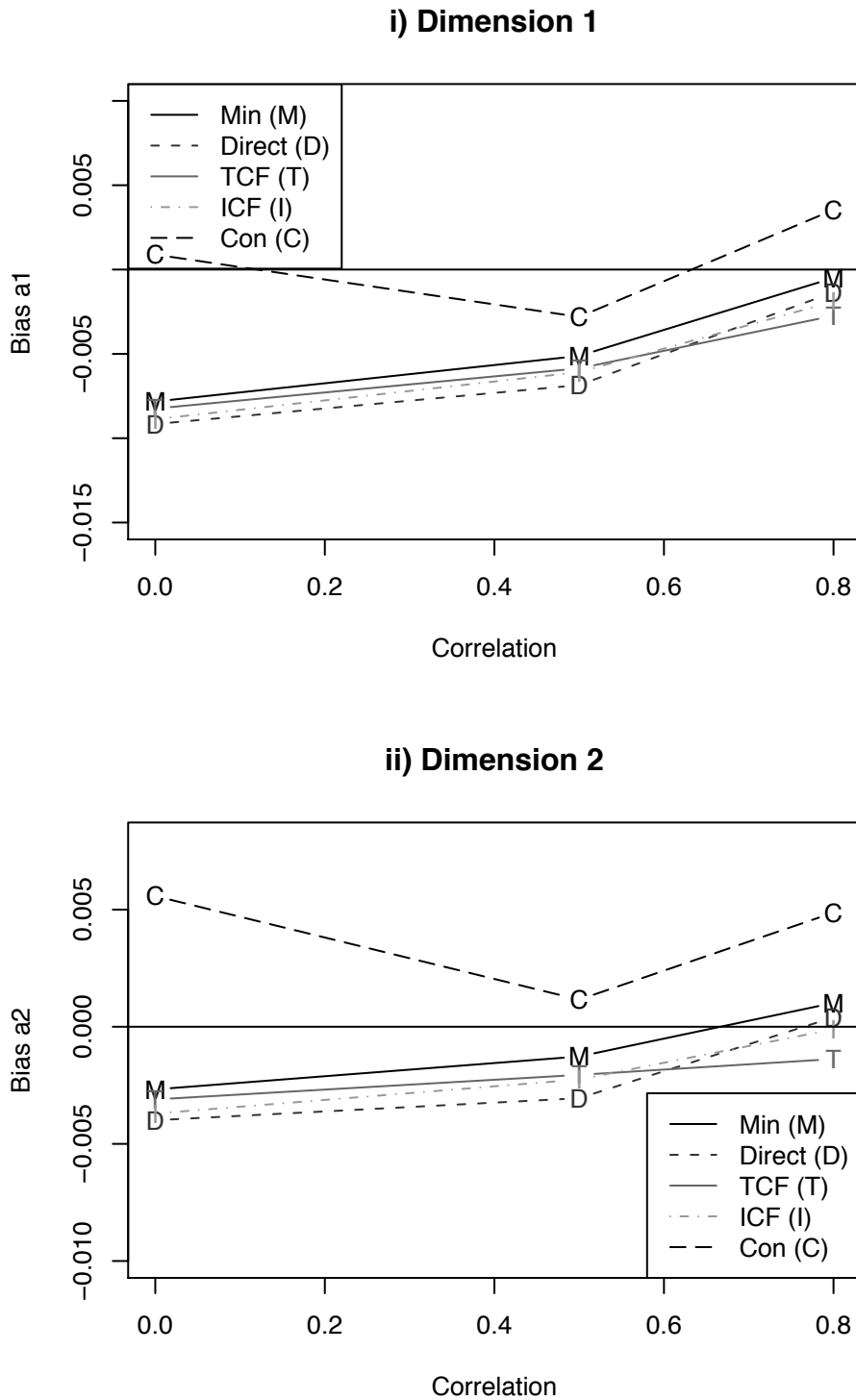
Figure 4.6.   BIAS of $a1$ and $a2$ across correlation levels for non-equivalent groups (1SD) for the 60-item form when sample size is 3000.

i) Dimension 1

ii) Dimension 2
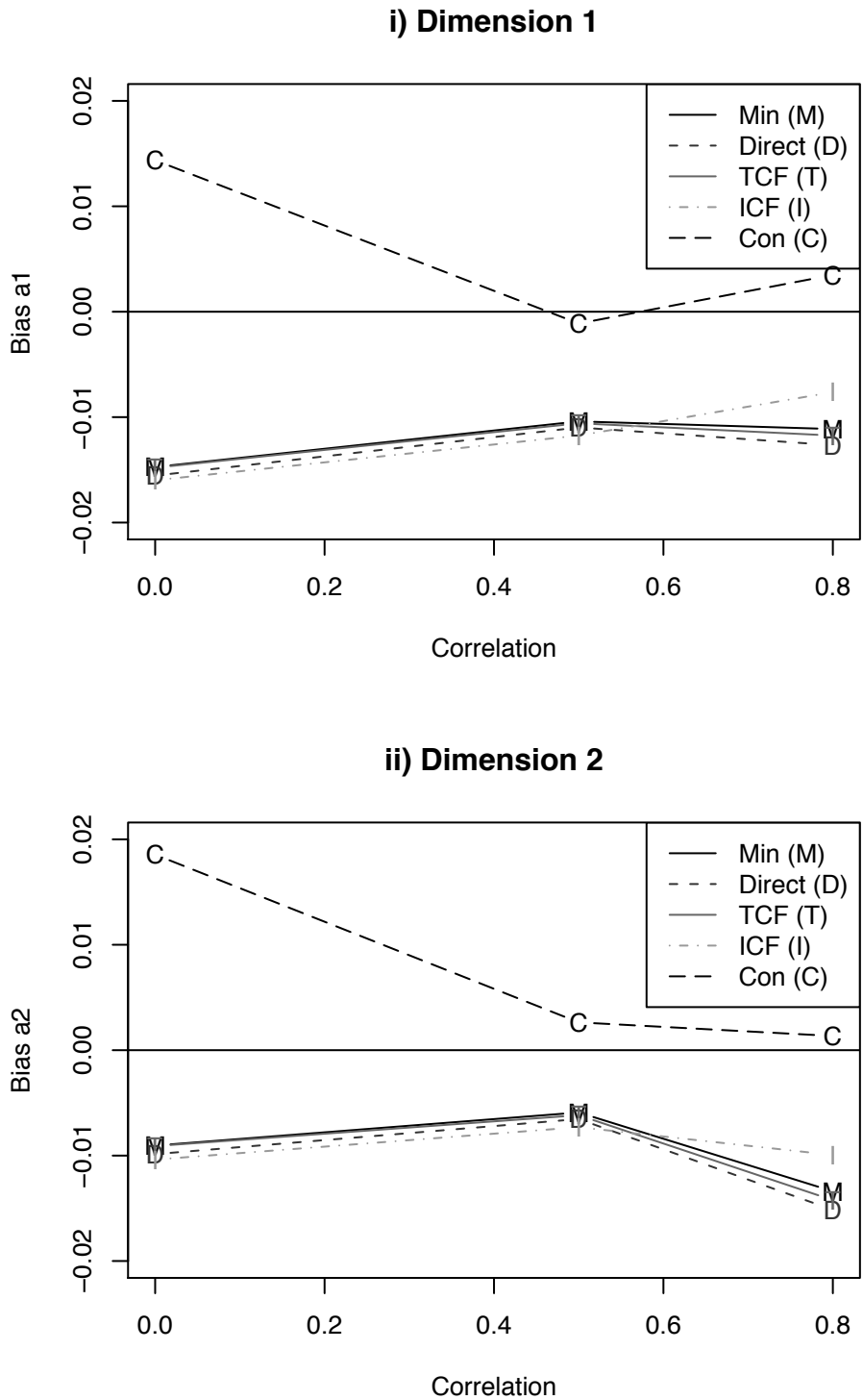
Figure 4.7. RMSE of $a1$ across sample sizes for equivalent groups with zero correlation condition and for non-equivalent groups (.5SD) with 0.8 correlation condition for the 60-item form when sample size is 3000.

i) Equivalent groups with r = 0
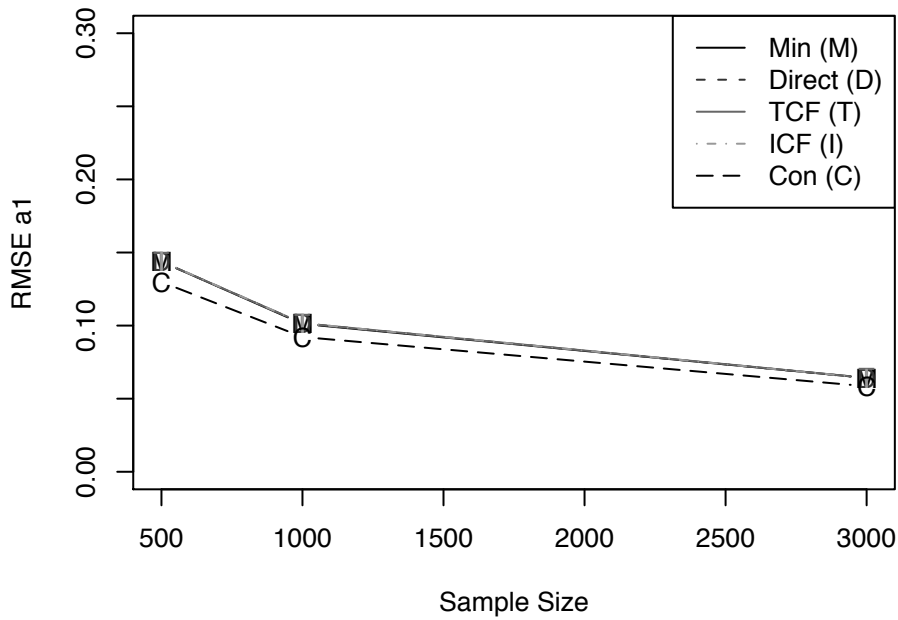
ii) Non–equivalent groups (.5SD) with r = 0.8
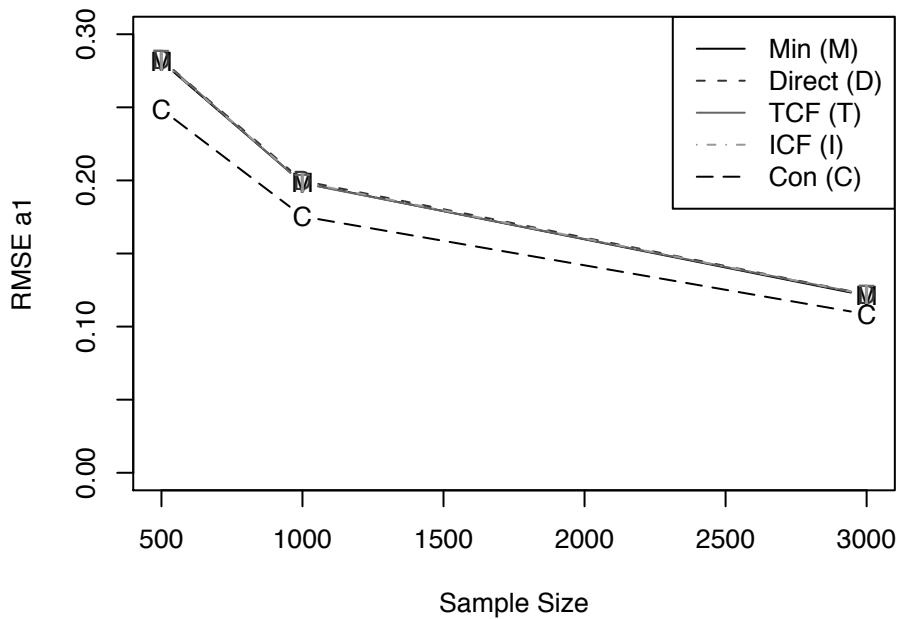
Figure 4.8. RMSE of $a2$ across sample sizes for equivalent groups with zero correlation condition and non-equivalent groups (.5SD) with 0.8 correlation condition for the 60-item form when sample size is 3000.

i) Equivalent groups with r = 0
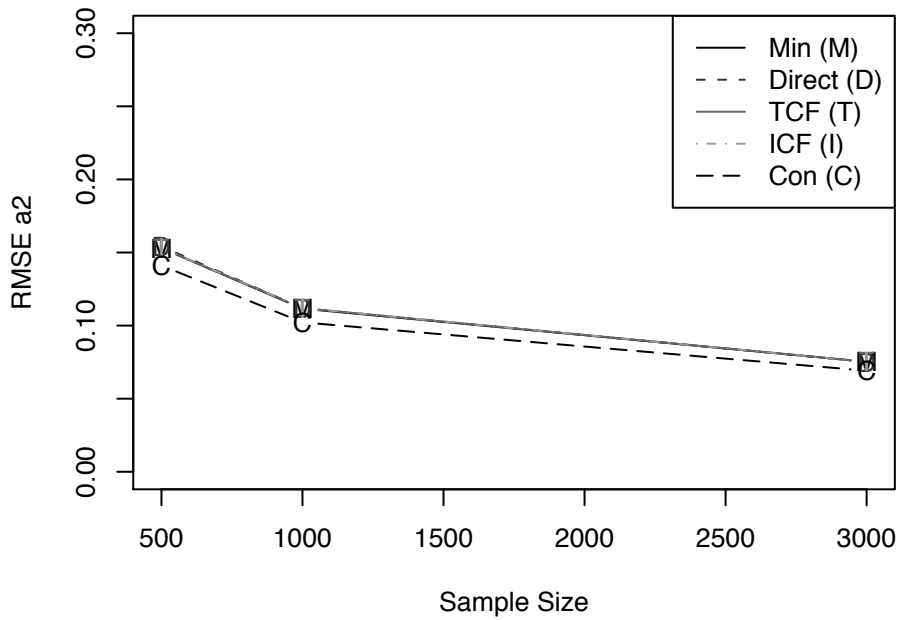
ii) Non–equivalent groups (.5SD) with r = 0.8
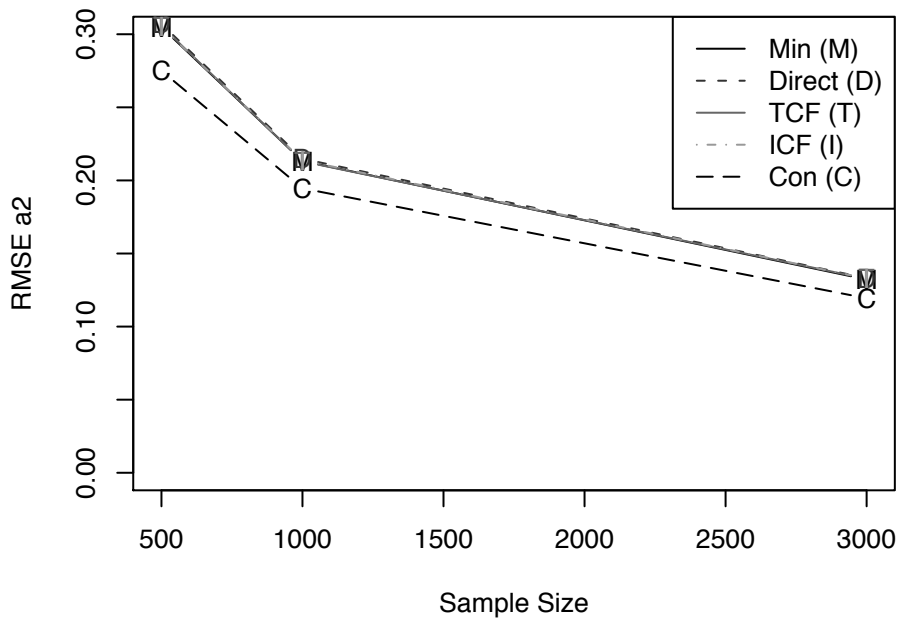
Figure 4.9. RMSE of $a1$ and $a2$ across correlation levels for equivalent groups for the 60-item form when sample size is 3000.

i) Dimension 1

ii) Dimension 2
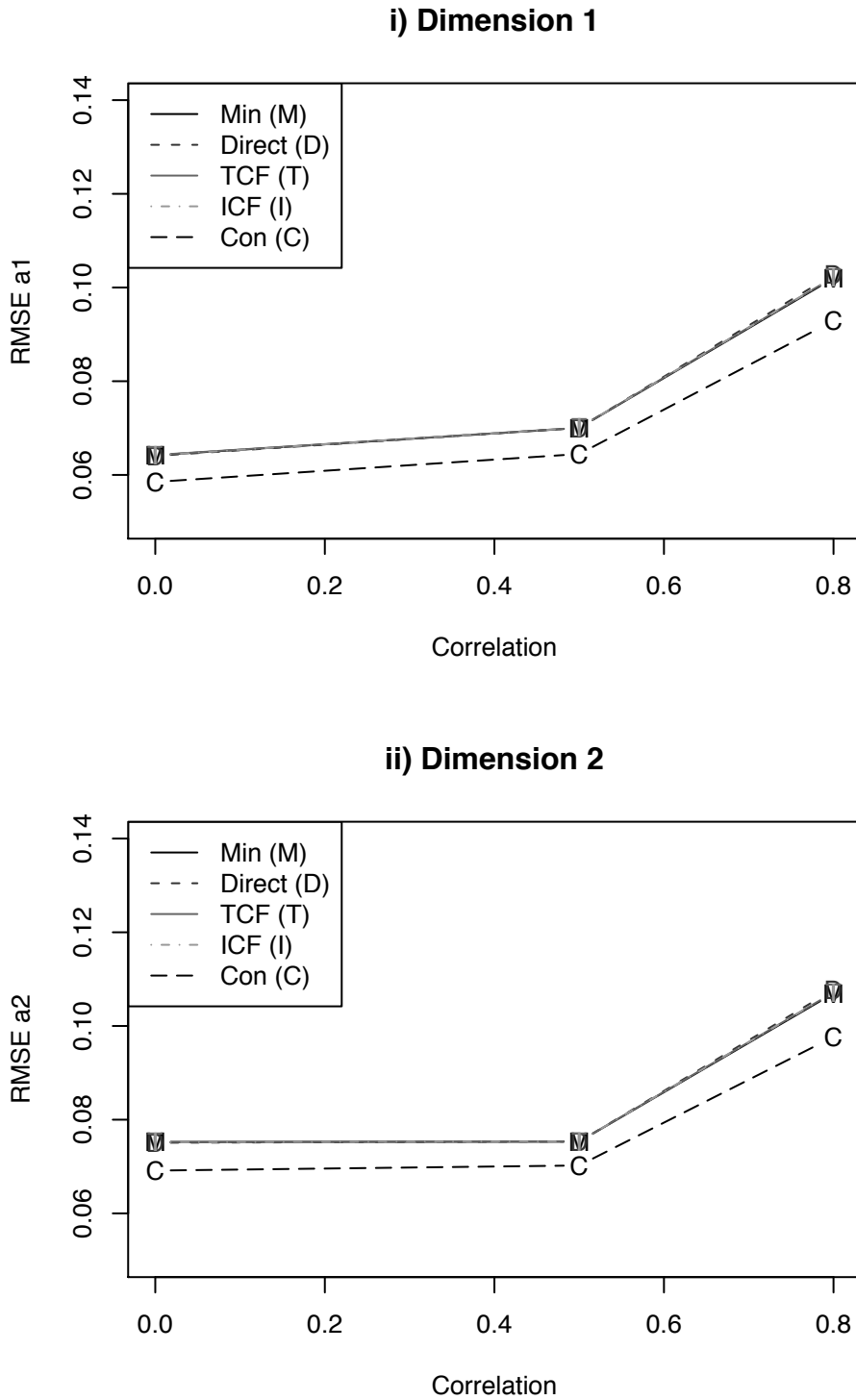
Figure 4.10. RMSE of $a1$ and $a2$ across correlation levels for non-equivalent groups (.5SD) for the 60-item form when sample size is 3000.

i) Dimension 1

ii) Dimension 2
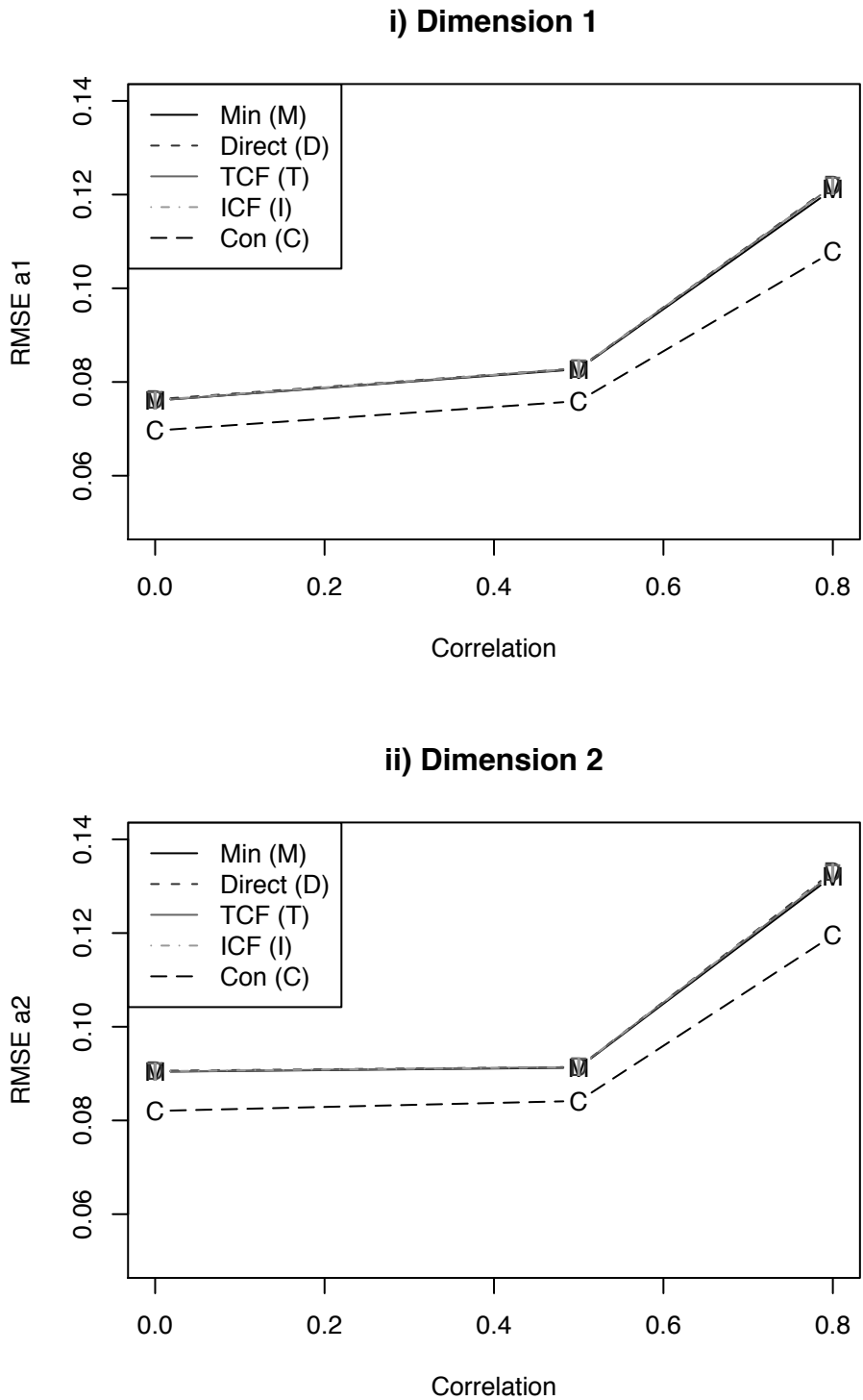
Figure 4.11. RMSE d across equivalence levels for zero and 0.8 correlation conditions for the 60-item form when sample size is 3000.
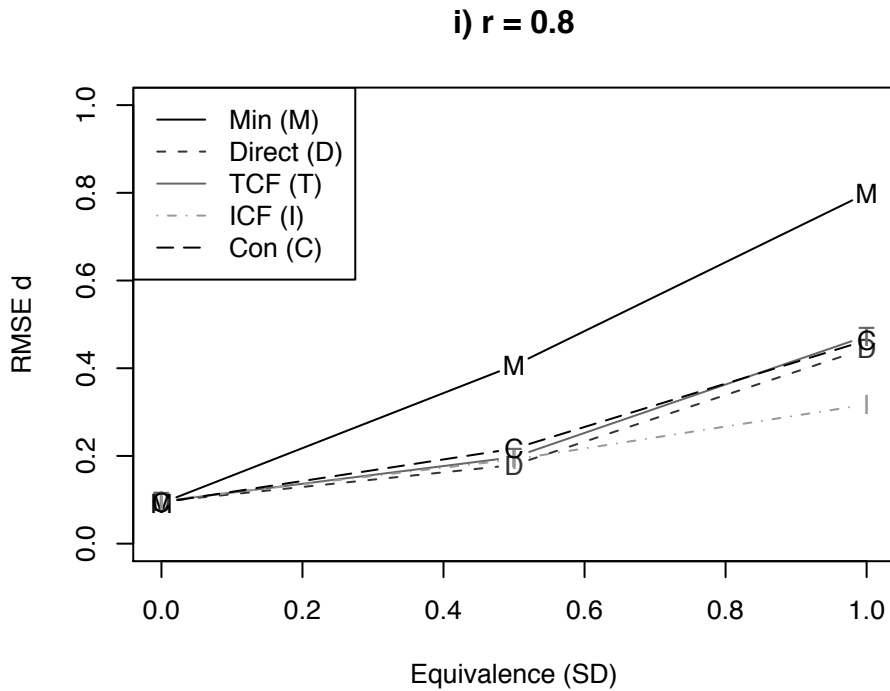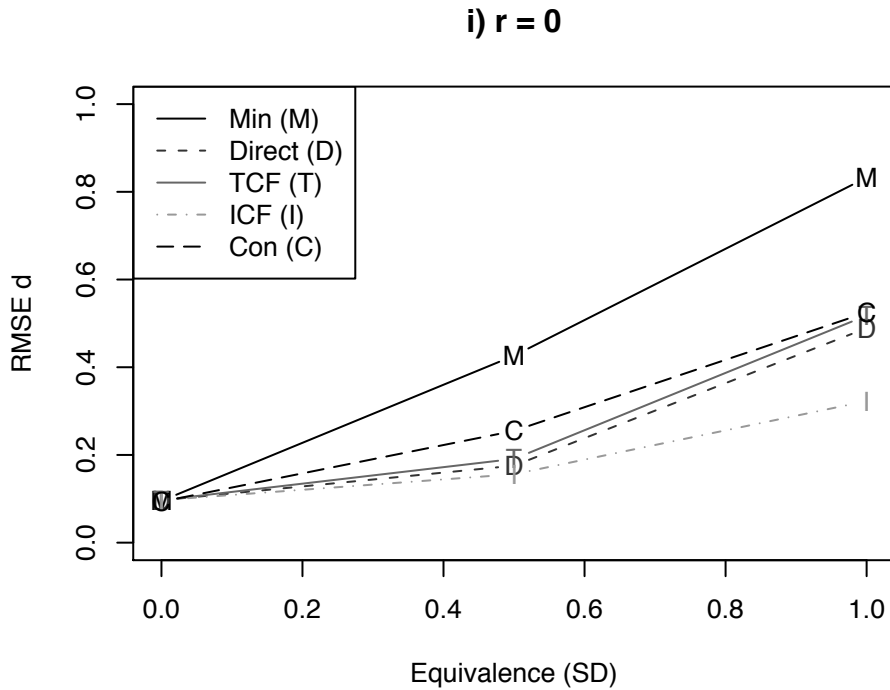
Figure 4.12. BIAS $a1$ and $a2$ across sample sizes for the equivalent groups with zero correlation condition for the 60-item form when sample size is 3000.

i) Dimension 1

ii) Dimension 2
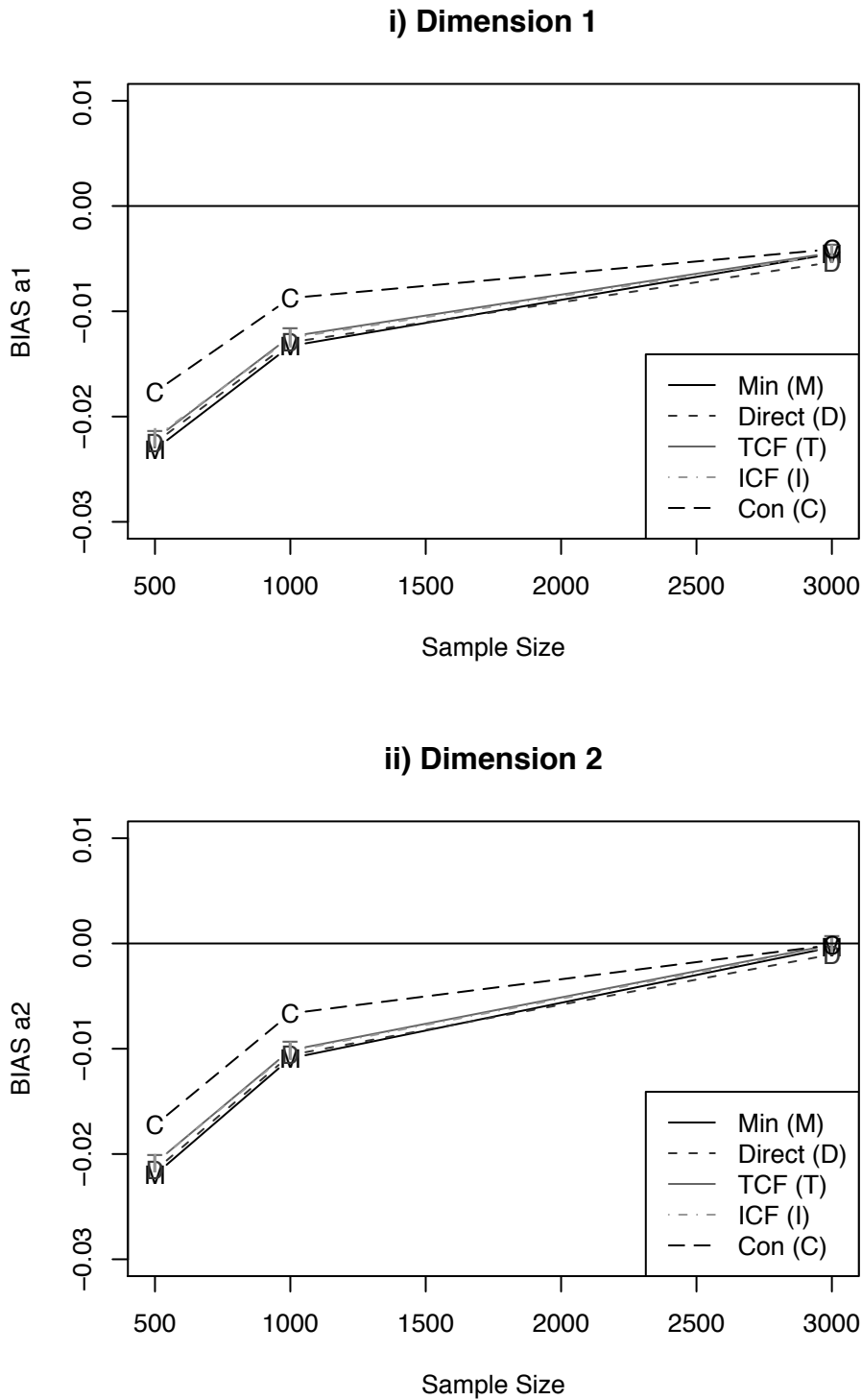
Figure 4.13. BIAS $a1$ and $a2$ across sample sizes for non-equivalent groups (.5SD) with 0.8 correlation condition for the 60-item form when sample size is 3000.

i) Dimension 1

ii) Dimension 2
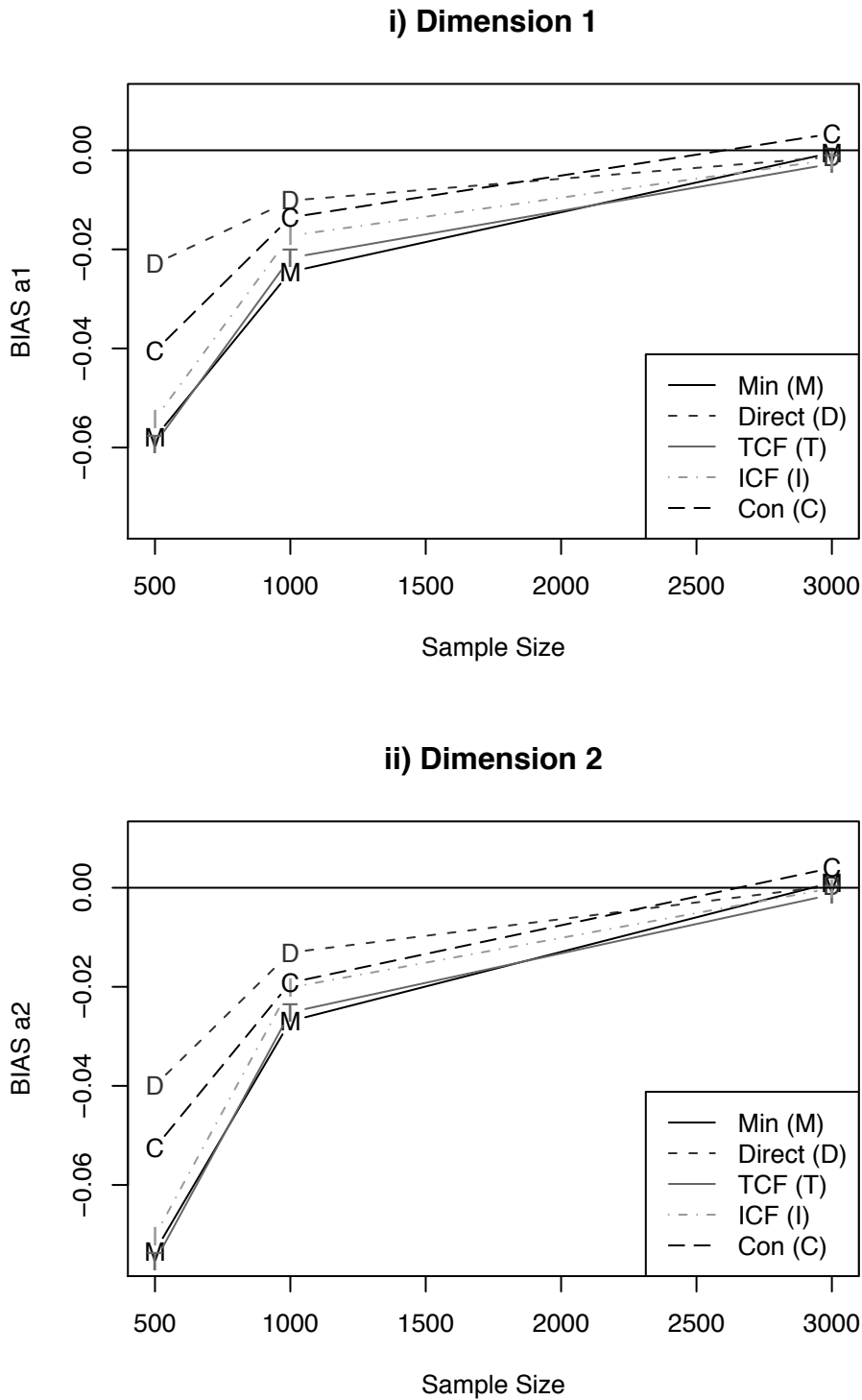
Figure 4.14. BIAS d across equivalence levels for zero and 0.8 correlation conditions with the 60-item form when sample size is 3000.

i) r = 0

i) r = 0.8
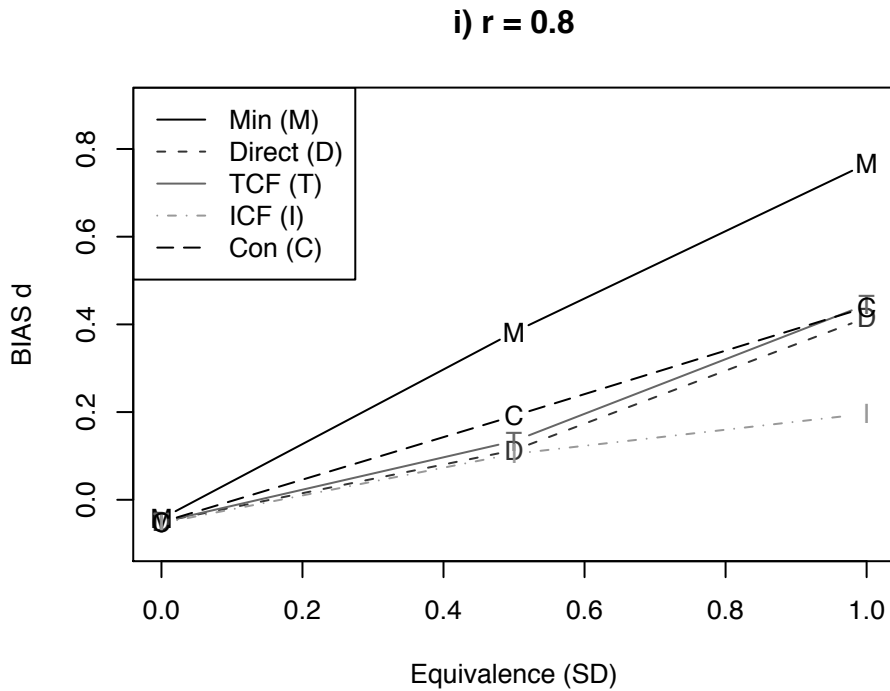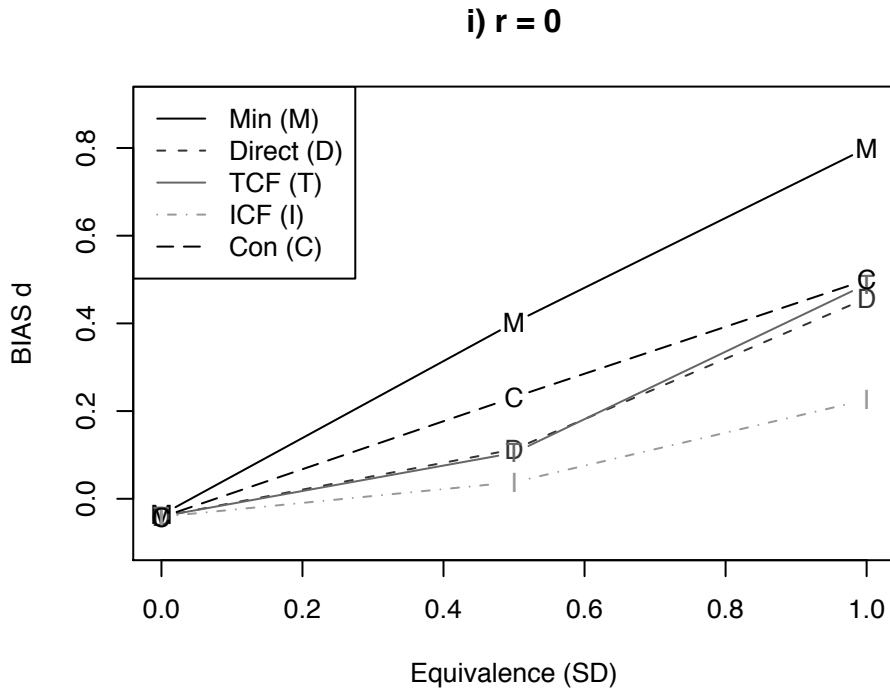
4.3   Correlation Between Final Estimate and Generating Parameters

The correlation between transformed item parameter estimates and generating item parameters was also calculated as another dependent variable. Transformed correlations (Fisher's r to z transformation) were used for repeated measures analysis. Table 4.5 shows the results of the repeated measures analysis. Tables D.1 to D.6 show the means of the untransformed correlation values.

The effect sizes for linking methods were 0.27, 0.29, and 0.64, for $a1$, $a2$, and d, respectively. Figures 4.15, 4.16, and 4.17 show box plots of the untransformed correlation between estimated and generating item parameters for $a1$, $a2$, and d, respectively, with equivalent groups and zero correlation between ability dimensions, as well as with non-equivalent groups and 0.8 correlation under the test length of 60 items and sample size of 3000. Figures 4.15 and 4.16 show that concurrent calibration had higher correlations between estimated and generating item parameters for $a1$ and $a2$ even when groups were non-equivalent and ability dimensions were correlated. Figures 4.17 (i) and (ii) show concurrent calibration had higher correlations between estimated and generating item parameters of d than separate linking methods with equivalent groups and zero correlation, however, when groups were non-equivalent and dimensions were correlated, concurrent calibration, the ICF, the Direct, and the TCF performed very similarly. Min's method had a lower correlation for estimated and generating parameters of d than other methods.

The largest effect size among between factors for $a1$ and $a2$ was sample size followed by correlation between ability dimensions. Figures 4.18 (i) and (ii) show the correlations between estimated and generating parameters across sample sizes for $a1$ and $a2$ with equivalent groups and zero correlation between dimensions. Figures 4.19 (i) and (ii) show the correlations with non-equivalent groups (0.5 standard deviation difference) and 0.8 correlation between dimensions. Concurrent calibration had consistently higher average correlations across all sample sizes for $a1$ and $a2$ even with non-equivalent groups and correlated dimensions.

Figures 4.20 (i) and (ii) show the correlations between estimated and generating item parameters for $a1$ and $a2$ across correlation levels between ability dimensions with equivalent groups. Figures 4.21 (i) and (ii) show the correlations with non-equivalent groups (0.5 standard deviation difference). The correlation between estimated and generating item parameters decreased as the correlation between ability dimensions increased. Concurrent calibration had higher correlations between estimated and generating item discrimination parameters than separate linking methods even when groups were non-equivalent and when the correlation between ability dimensions was high. The difference between concurrent calibration and separate linking methods increased with higher correlation and non-equivalent groups.

The parameter d, item difficulty, was greatly influenced by group equivalence. The linking methods effect was larger with d than with $a1$ and $a2$. Figures 4.22 (i) and (ii) show the correlations between estimated d and generating item parameter across

levels of group equivalence when the correlation between ability dimensions was zero or 0.8. With equivalent groups, all methods had essentially the same average value of the correlation. However, as groups departed from equivalence, the correlation for Min's method became lower than for other methods. This was true when correlation between ability dimensions was zero and when the correlation was 0.8. There were only small differences in the correlations among the concurrent calibration, TCF, ICF, and Direct methods.

Figures 4.23 (i) and (ii) show the correlations between estimated and generating item parameters for d across sample sizes with equivalent groups and zero correlation between ability dimensions (Figure 4.23(i)), and with non-equivalent groups and correlation between ability dimensions equal to 0.8 (Figure 4.23(ii)). Min's method performed as well as other separate linking methods with equivalent groups and zero correlation between ability dimensions (Figure 4.23(i)). However, with non-equivalent groups and non-zero correlation between ability dimensions, Min's method performed poorly compared to other methods. Results were similar with non-equivalent groups and zero correlation between ability dimensions (Tables D.5 and D.6). With equivalent groups and non-zero correlation between ability dimensions, Min's method performed as well as other methods.

Table 4.5. Repeated measure analysis results for correlation between estimates and generating item parameters.

Statistic	factors		Value	F Value	DFn[2]	DFd[2]	P value	η_p^{2}[2]
a1	Within[1]	linking	631.74	4	2618	<.0001	0.27
		linking*equivalence	6.59	8	5236	<.0001	<.01
		linking*correlation	8.00	8	5236	<.0001	0.01
		linking*sample size	7.72	8	5236	<.0001	0.01
		linking* test length	37.50	4	2618	<.0001	0.03
	Between	equivalence	1841.30	2	2621	<.0001	0.15
		correlation	2761.40	2	2621	<.0001	0.22
		sample size	6650.53	2	2621	<.0001	0.53
		test length	26.41	1	2621	<.0001	<.01
a2	Within[1]	linking	661.74	4	2618	<.0001	0.29
		linking*equivalence	8.78	8	5236	<.0001	0.01
		linking*correlation	14.30	8	5236	<.0001	0.02
		linking*sample size	11.18	8	5236	<.0001	0.02
		linking* test length	45.62	4	2618	<.0001	0.03
	Between	equivalence	1996.82	2	2621	<.0001	0.18
		correlation	2254.63	2	2621	<.0001	0.20
		sample size	5763.63	2	2621	<.0001	0.51
		test length	57.71	1	2621	<.0001	<.01
d	Within[1]	linking	4327.46	4	2618	<.0001	0.64
		linking*equivalence	674.88	8	5236	<.0001	0.51
		linking*correlation	9.67	8	5236	<.0001	0.01
		linking*sample size	147.39	8	5236	<.0001	0.18
		linking* test length	35.87	4	2618	0.0038	0.03
	Between	equivalence	4420.65	2	2621	<.0001	0.45
		correlation	2.11	2	2621	0.1208	<.01
		sample size	4044.21	2	2621	<.0001	0.41
		test length	64.96	1	2621	<.0001	<.01

[1]Within factor analysis shows multivariate results which do not require sphericity assumption. [2] DFn: degrees of freedom for numerator; DFd: degrees of freedom for denominator; η_p^{2}: partial eta squared

Figure 4.15. Correlation between estimated and generating parameter of $a1$ for equivalent groups with zero correlation and for non-equivalent groups with 0.8 correlation conditions for the 60-item form when sample size is 3000.

i) Equivalent groups with r= 0

ii) Non−equivalent groups (.5SD) with r = 0.8
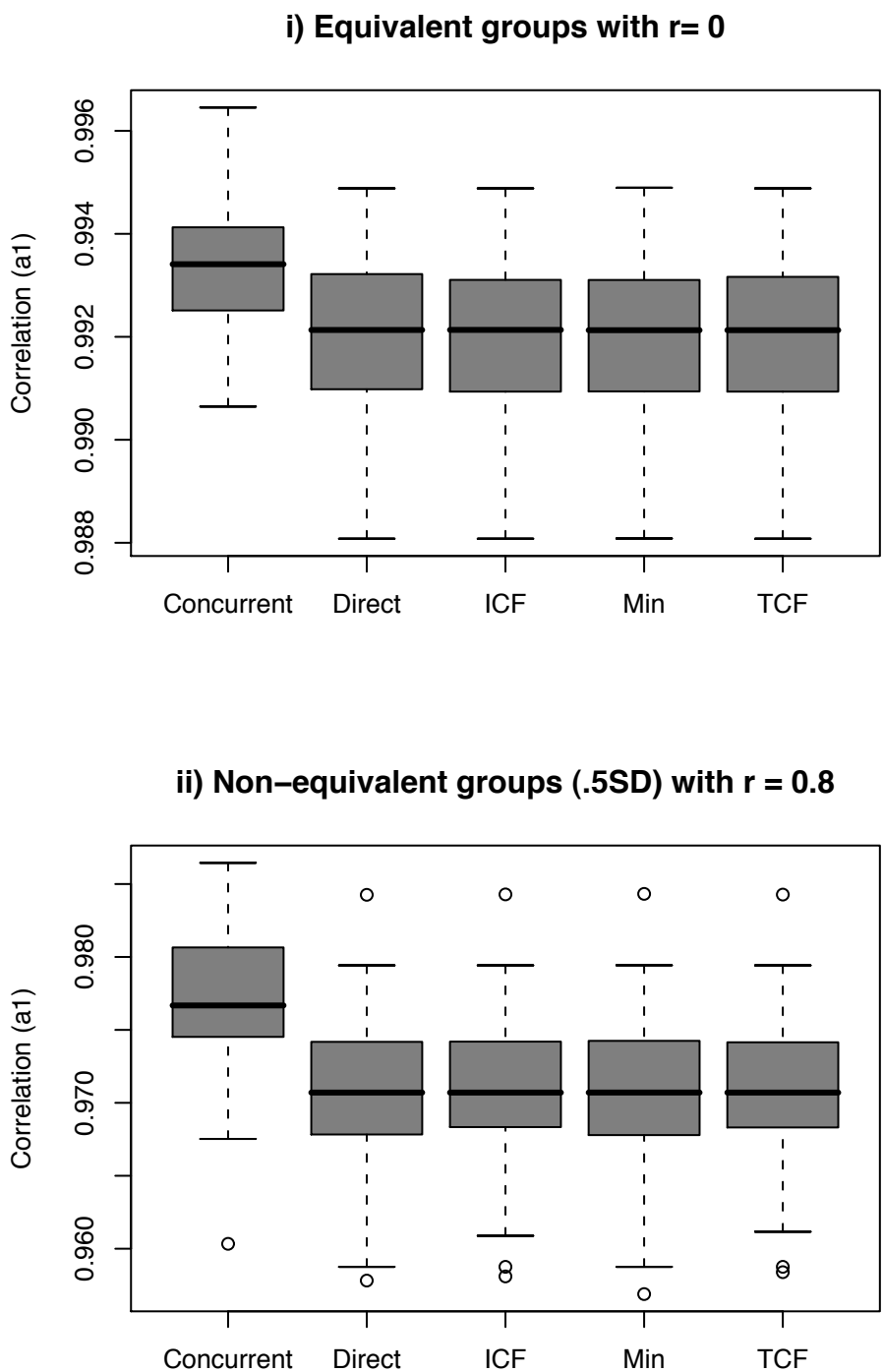
Figure 4.16. Correlation between estimated and generating parameter of $a2$ for equivalent groups with zero correlation and for non-equivalent groups with 0.8 correlation conditions for the 60-item form when sample size is 3000.

i) Equivalent groups with r = 0

ii) Non–equivalent groups (.5SD) with r = 0.8
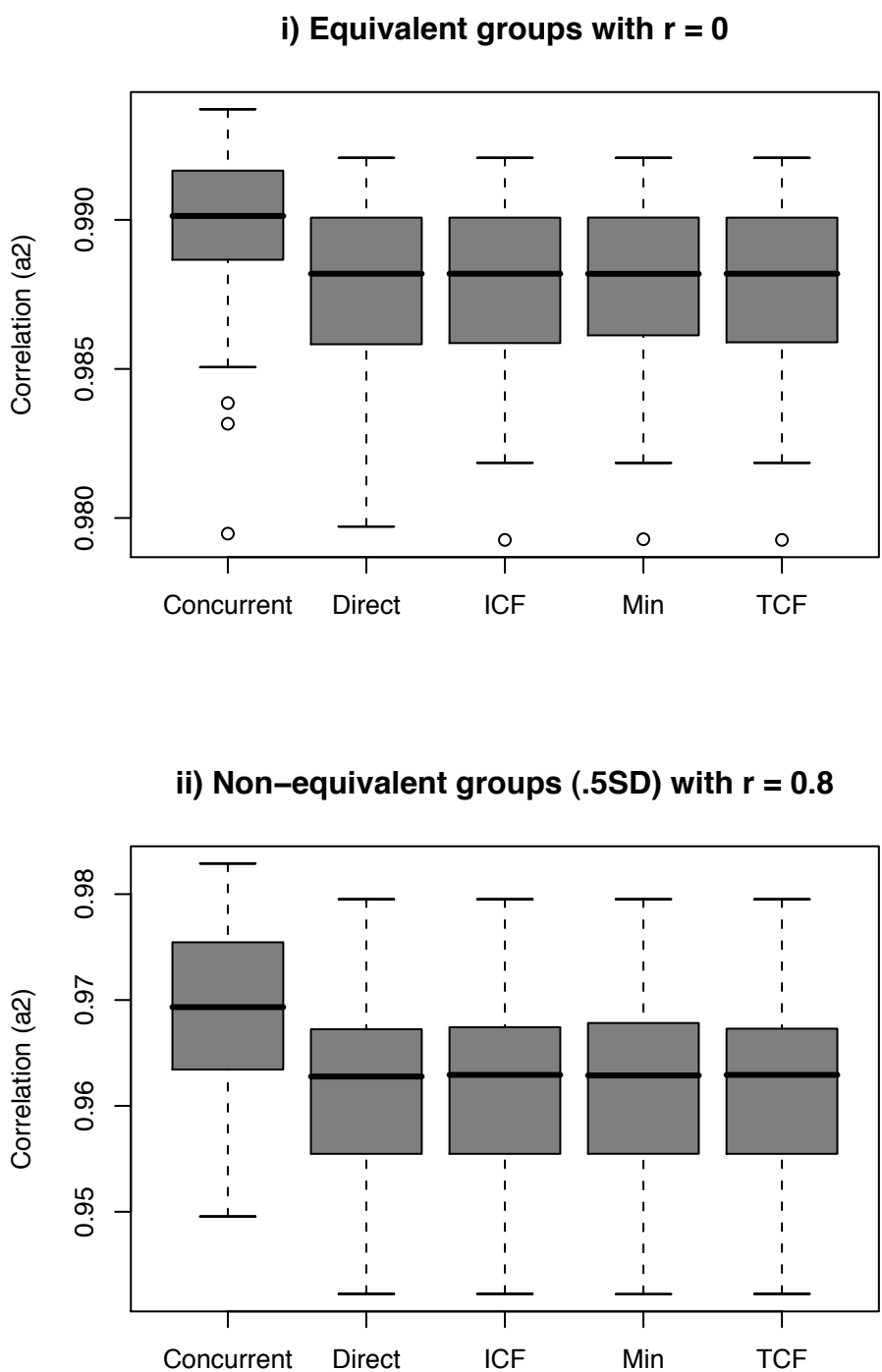
Figure 4.17.   Correlation between estimated and generating parameter of d for equivalent groups with zero correlation and for non-equivalent groups with 0.8 correlation conditions for the 60-item form when sample size is 3000.

i) Equivalent groups with r = 0

ii) Non–equivalent groups (.5SD) with r= 0.8
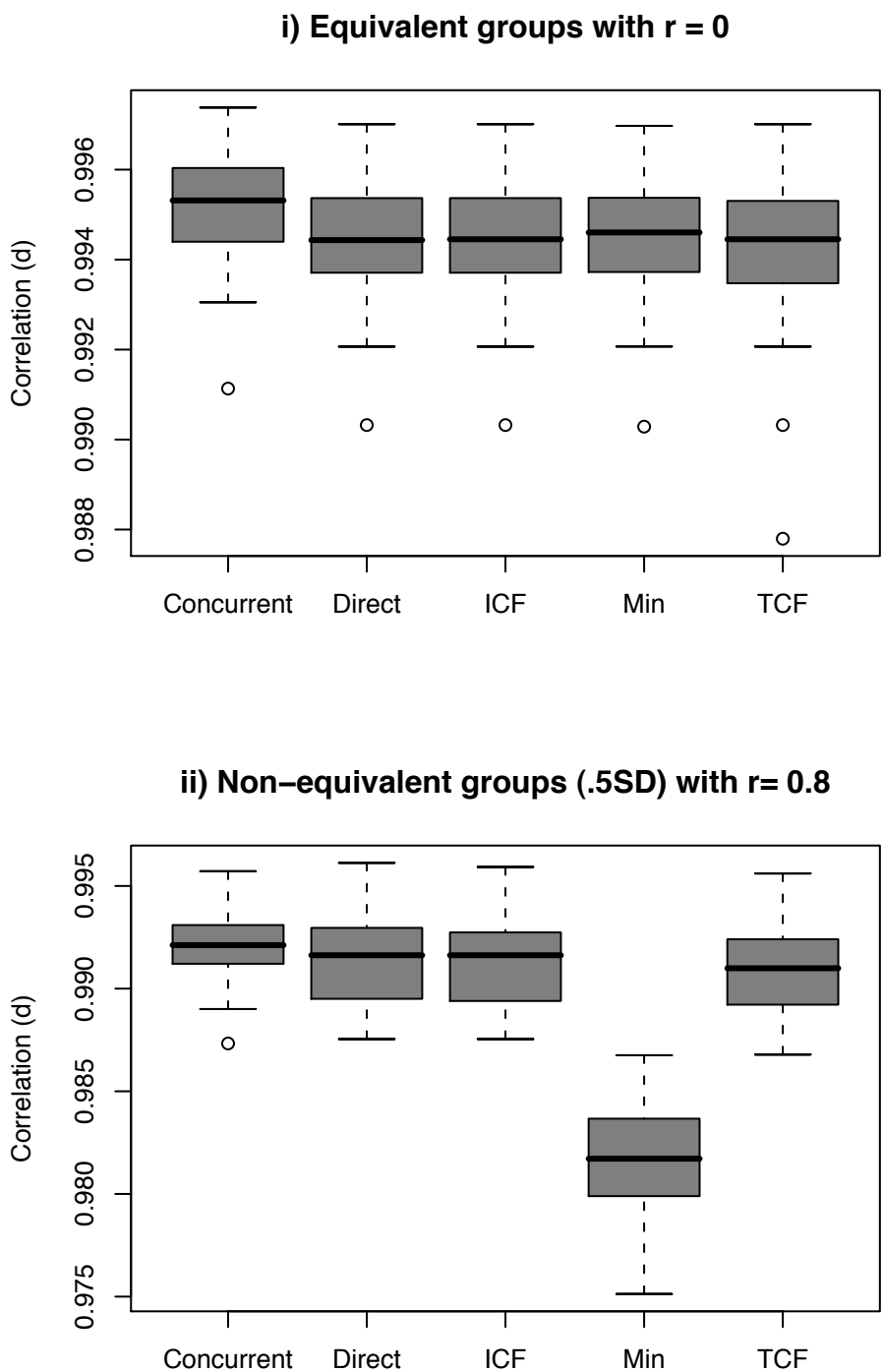
Figure 4.18. Correlation between estimated and generating parameter of $a1$ and $a2$ across sample sizes with equivalent groups condition for the 60-item form when sample size is 3000.
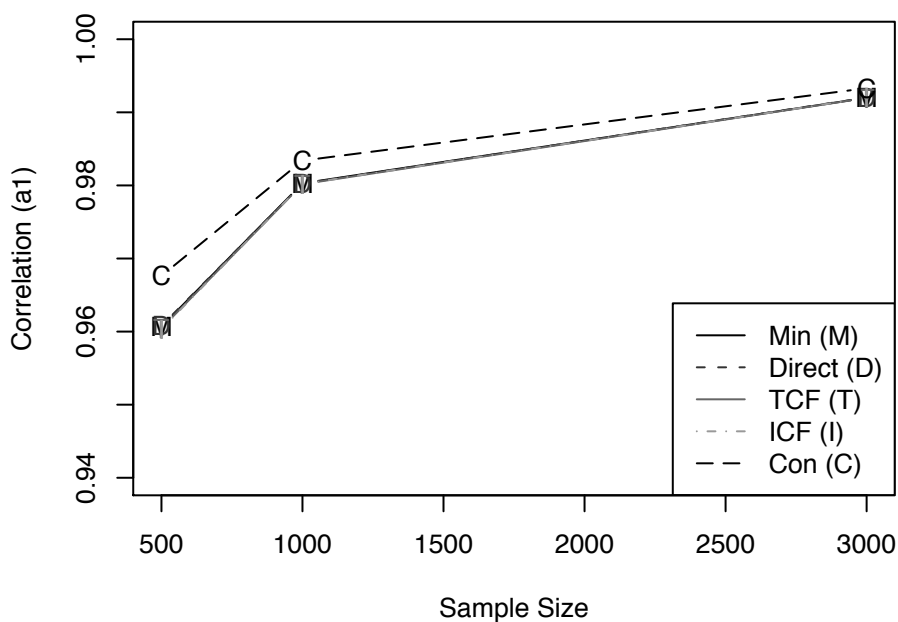
i) Dimension 1

ii) Dimension 2
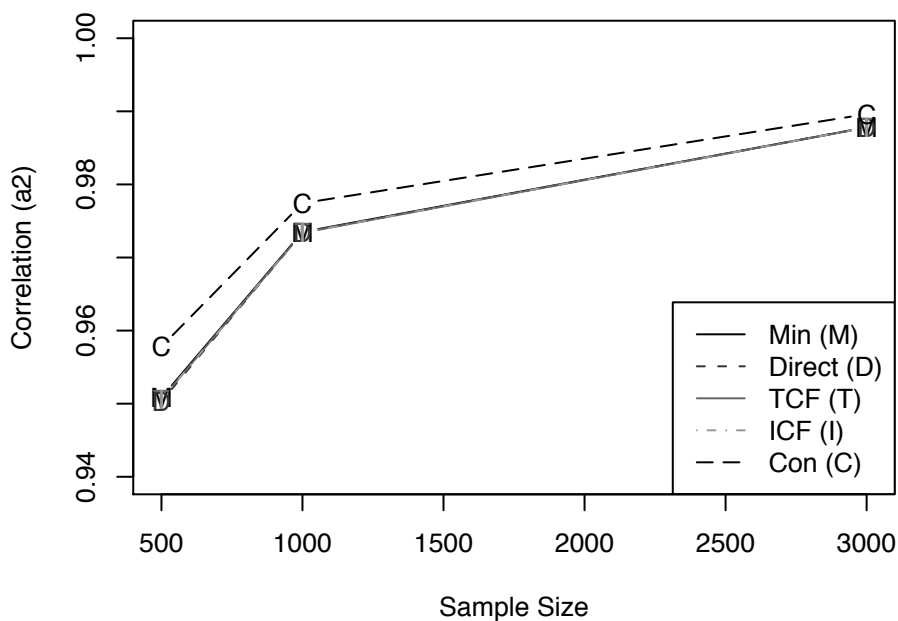
Figure 4.19.   Correlation between estimated and generating parameter of $a1$ and $a2$ across sample sizes with non-equivalent groups (.5SD) condition with 0.8 correlation condition for the 60-item form when sample size is 3000.

i) Dimension 1

ii) Dimension 2
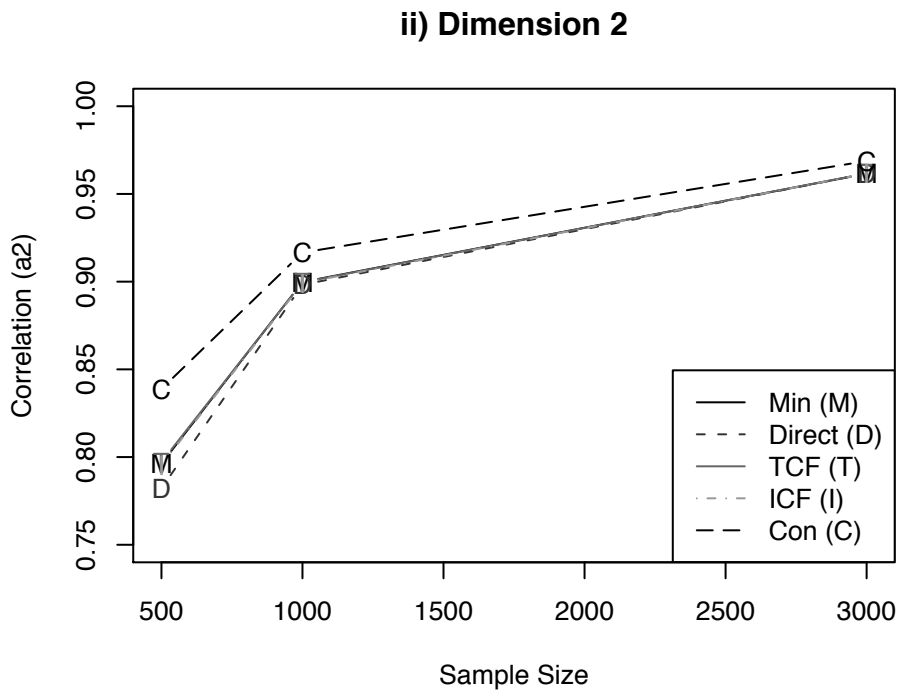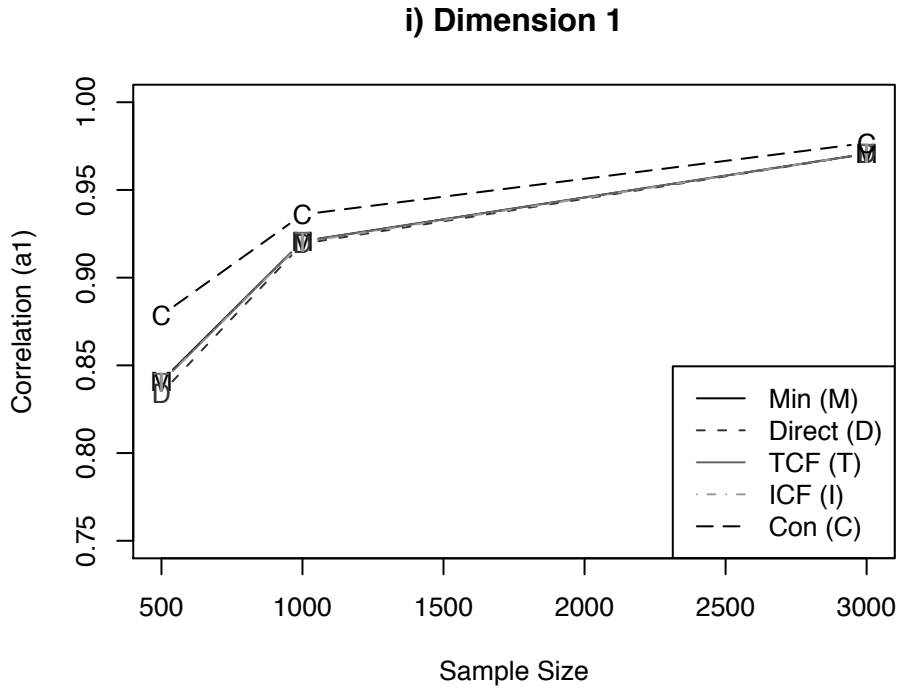
Figure 4.20.   Correlation between estimated and generating parameter of $a1$ and $a2$ across correlation levels with equivalent groups condition for the 60-item form when sample size is 3000.

i) Dimension 1

ii) Dimension 2
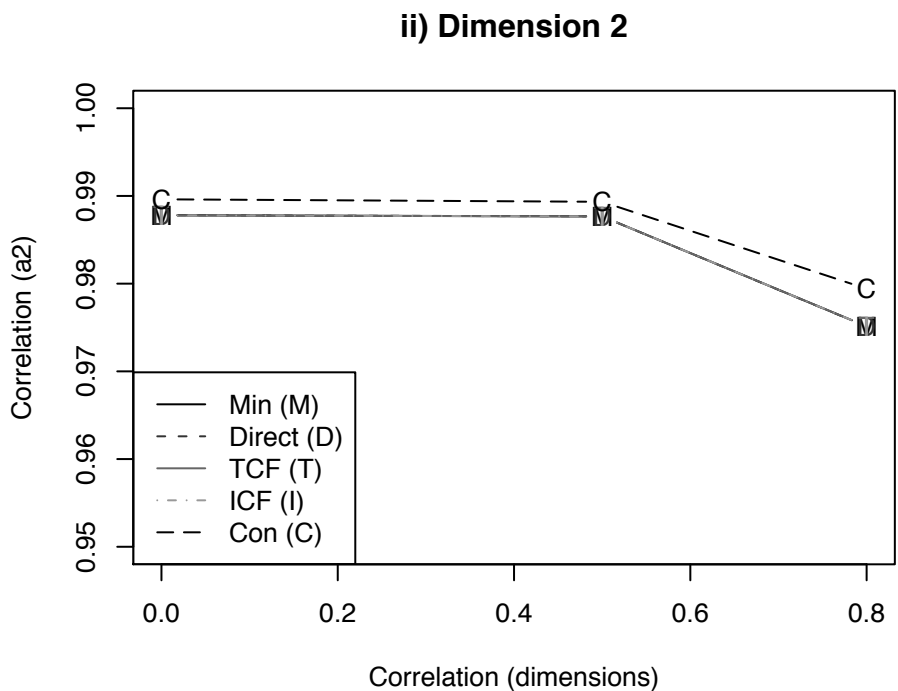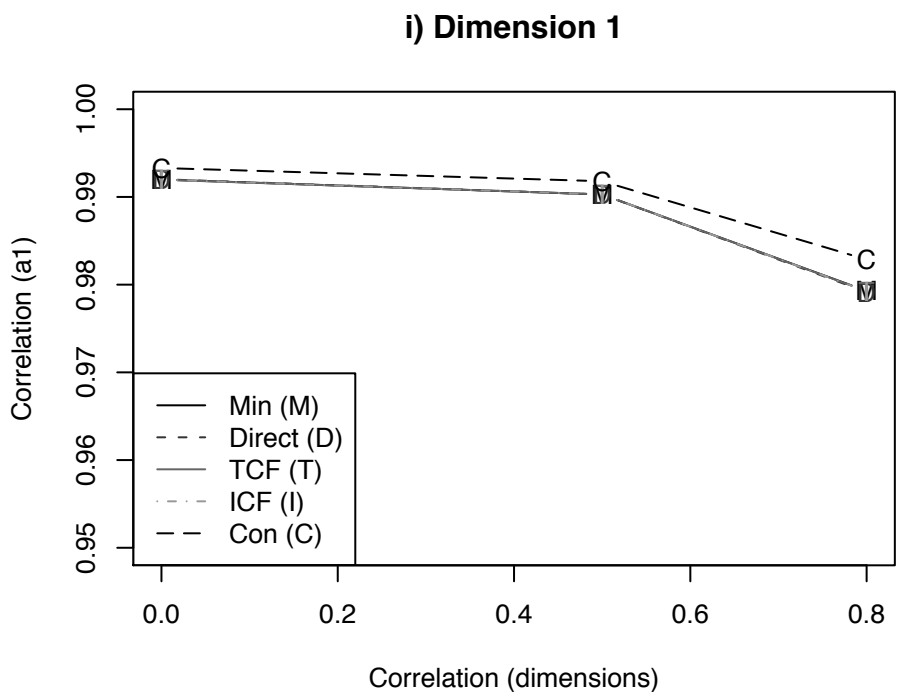
Figure 4.21. Correlation between estimated and generating parameter of $a1$ and $a2$ across correlation levels with the non-equivalent groups (.5SD) condition for the 60-item form when sample size is 3000.

i) Dimension 1

ii) Dimension 2
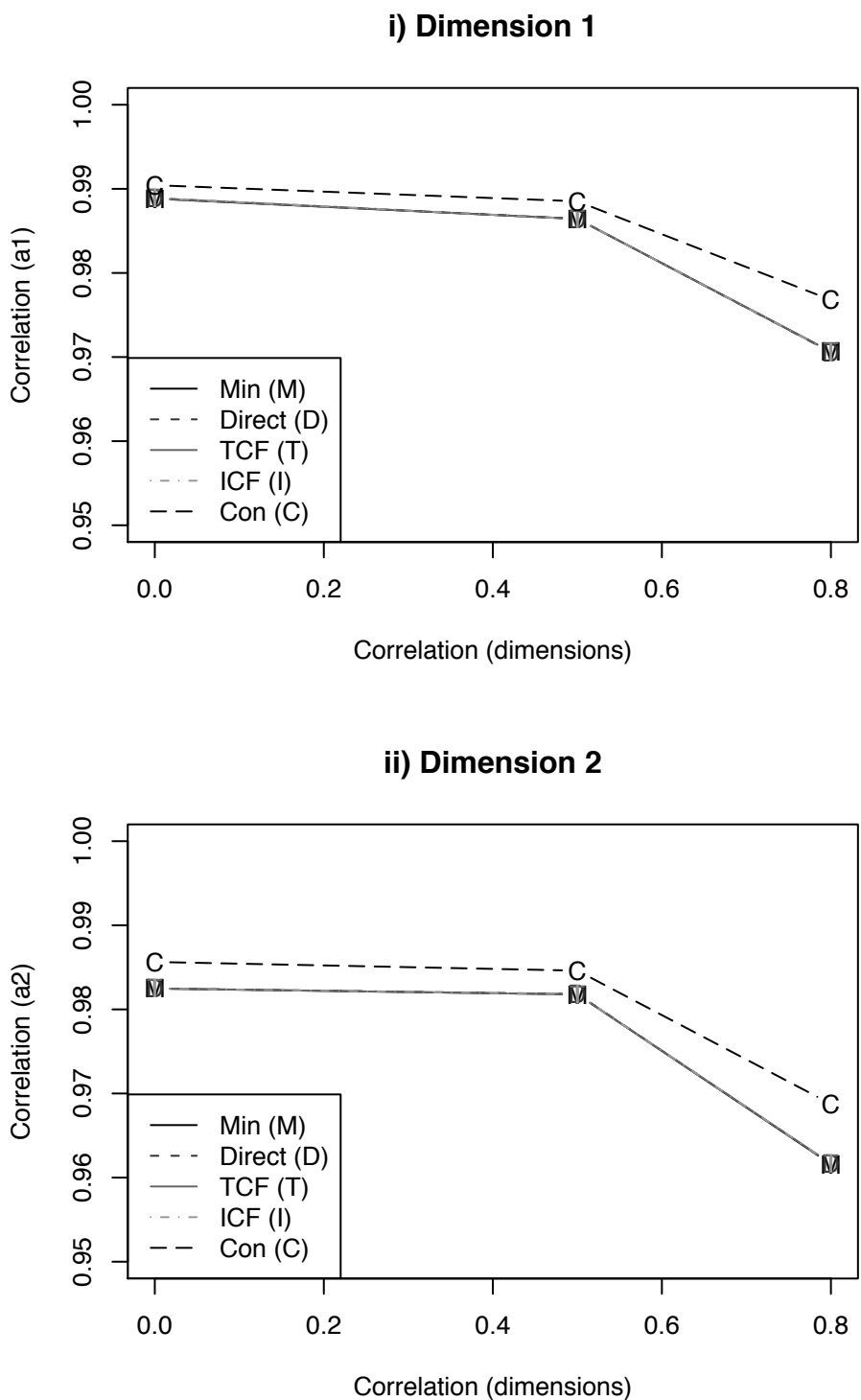
Figure 4.22. Correlation between estimated and generating parameter of d across group equivalence with zero and 0.8 correlation conditions for the 60-item form when sample size is 3000.
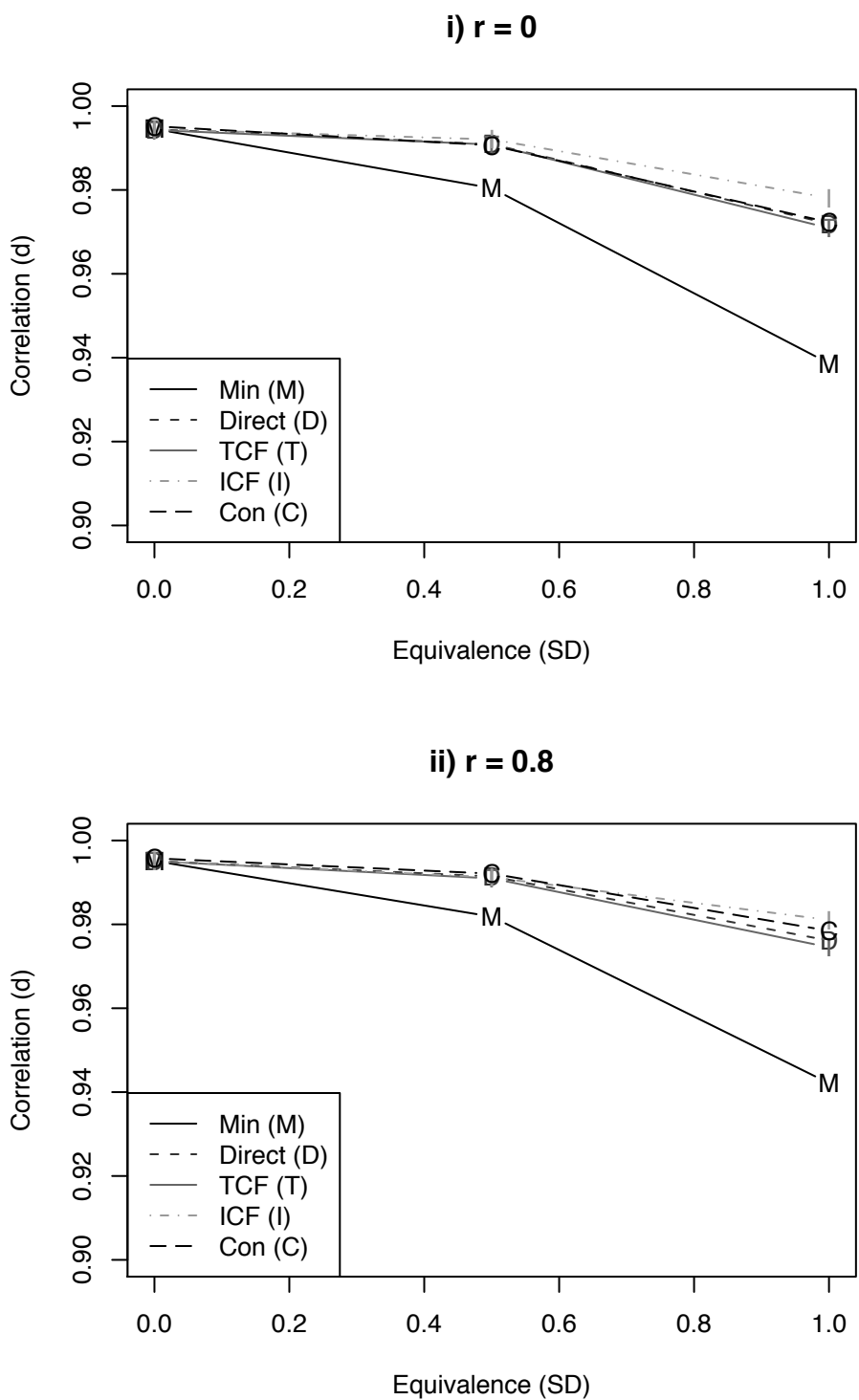
Figure 4.23.   Correlation between estimated and generating parameter of d across sample size for equivalent groups with zero correlation, and for non-equivalent groups (.5SD) with 0.8 correlation conditions for the 60-item form when sample size is 3000.

i) Equivalent groups with r = 0

i) Non−eqvuivalent groups (.5SD) with r = 0.8

Chapter **5**
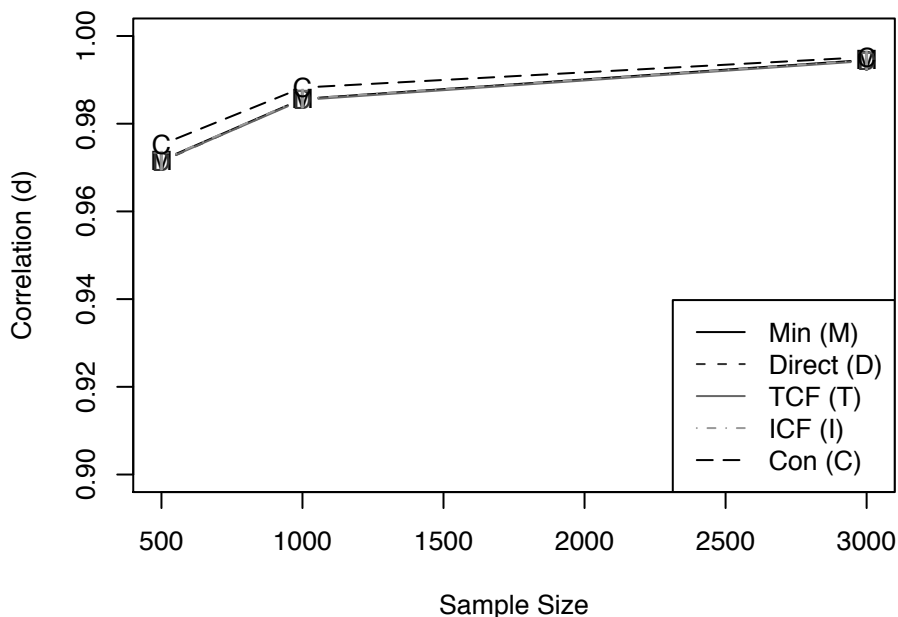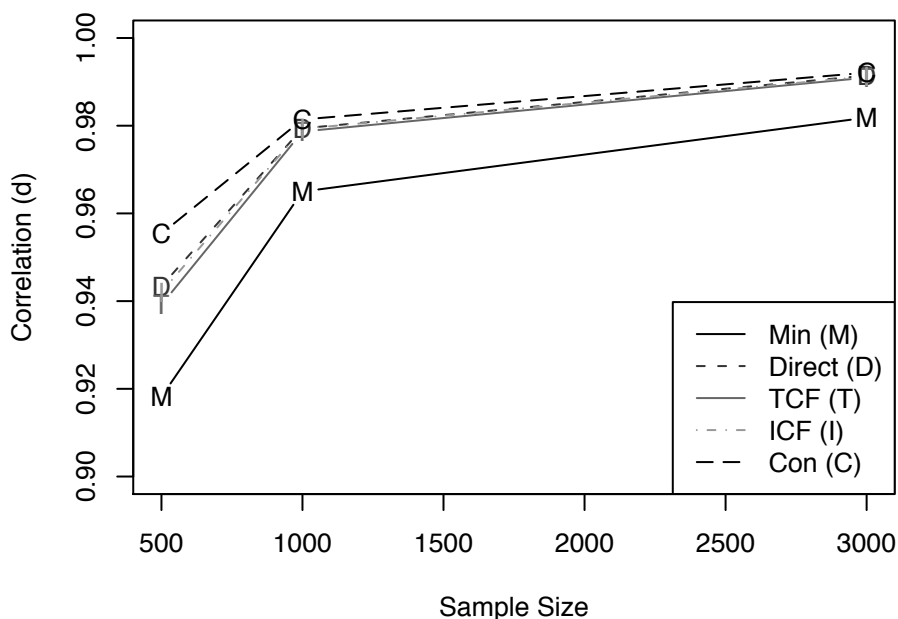
Discussion and

Conclusion

The main purpose of this simulation research was to investigate the performance of concurrent calibration and separate linking methods while varying group equivalence, correlation levels among ability dimensions, sample sizes, and test lengths. This study employed four separate MIRT linking methods and the comparison among them was also of interest in the study. The research questions are presented in the first chapter of this dissertation. Simulated data based on actual state standardized tests were analyzed to address the research questions.

In this chapter, each of the research questions is revisited, and results addressing these questions are described and discussed. Following the discussion for each research question, limitations of the study are discussed and suggestions for further research in MIRT linking are proposed. Lastly, the major conclusions are stated.

5.1 Research Question 1

The first research question involved the comparison of concurrent and separate linking methods varying group equivalence and the correlation level among ability dimensions. The research question was as follows:

- How do separate linking methods and concurrent calibration compare in performance with various conditions: e.g., equivalent groups and zero correlation among abilities; or non-equivalent groups and non-zero correlation among abilities?

Four separate linking methods were considered: the Direct, the TCF, the ICF, and Min's methods. The first three allow non-orthogonal rotation matrices, and have a dilation matrix separate from the rotation matrix. Min employs an orthogonal rotation matrix and a dilation matrix. The comparisons of concurrent and separate linking methods were made based on RMSE, bias, and the correlation between the estimated item parameters and the generating item parameters.

5.1.1 Item Discrimination Parameters

Concurrent calibration had generally lower item discrimination RMSE and bias than separate linking methods even when groups were non-equivalent and the correlations between ability dimensions were high. This dissertation is the first study to compare concurrent calibration and separate linking methods with MIRT. However, there are some studies comparing concurrent calibration and separate linking methods with

UIRT. For example, Kim & Cohen (2002) compared concurrent calibration and separate linking under the graded response UIRT model using MULTILOG. Their finding was that concurrent calibration performed slightly better than separate linking methods with item discrimination parameter even when groups were non-equivalent with 1 standard deviation difference. The study by Hanson & Béguin (2002) compared concurrent calibration and separate UIRT linking methods using MULTILOG and BILOG-MG. Although their evaluation was with true score, not item parameters, they found that concurrent calibration generally resulted in lower error than separate linking methods, which is consistent with this study.

Non-equivalent groups and the high correlation between ability dimensions increased the RMSE of item discrimination parameters with all levels of sample size, and the differences between concurrent calibration and separate linkings were larger. Therefore, concurrent calibration might be a better choice than separate linking with respect to item discrimination parameters, even when there are non-equivalent groups or high correlation between ability dimensions. Concurrent calibration generally had a higher correlation between the estimated and the generating item discrimination parameters even when groups were non-equivalent and ability dimensions were highly correlated.

5.1.2   Item Difficulty Parameter

The ICF generally performed better than concurrent calibration and other separate linking methods with respect to the RMSE and bias of item difficulty parameters

in the non-equivalent groups condition. When groups were non-equivalent, the item difficulty RMSE and bias for concurrent calibration was slightly larger than for the TCF, the ICF and the Direct methods. With the study by Kim & Cohen (2002), the authors found that concurrent calibration had slightly smaller error in UIRT than the Stocking and Lord method: TCF of UIRT, with the item difficulty parameter. Differences between this dissertation and the study by Kim & Cohen (2002) could be due to the difference between MIRT and UIRT, as well as the difference in the software used. The software used in this dissertation was TESTFACT, and Kim & Cohen (2002) used MULTILOG. Hanson & Béguin (1999) found that with non-equivalent groups, concurrent estimation tended to produce less error than separate estimation using BILOG-MG, but the opposite was true with MULTILOG. Thus software used to estimate item parameters influence the performance of concurrent and separate linking with respect to item difficulty.

When groups were equivalent, the item difficulty RMSE and bias were similar among concurrent and separate linking methods when sample size was large. The level of correlation did not affect the relationship among methods. The small effect sizes for the interaction of linking method and correlation for item difficulty RMSE and bias indicate the relationship among methods with zero correlation and with 0.8 correlation was very similar. The ICF had slightly higher correlation between estimated and the generating item difficulty parameter than concurrent calibration and other linking methods when groups were non-equivalent with 1 standard deviation difference. With

non-equivalent groups and 0.5 standard deviation difference, concurrent calibration, TCF, ICF, and the Direct methods performed equally well.

5.2   Research Question 2

The second research question involved the comparison of concurrent and separate linking methods with various sample sizes and test lengths. As in the first research question, the comparisons were made based on RMSE, bias, and the correlation between estimated and generating item parameters. The research question was as follows:

- How is the performance of separate and concurrent calibration affected by varying sample sizes and test lengths?

5.2.1   Item Discrimination Parameters

With smaller sample size, concurrent calibration had lower item discrimination RMSE and bias when groups were equivalent and the correlation between ability dimensions was zero. Thus, with smaller sample size, concurrent calibration would be a better choice than separate linking methods with equivalent groups and low correlations among ability dimensions. The study by Hanson & Béguin (2002) also showed that concurrent calibration performed better than separate linking methods especially when sample size was small. The same is true with Kim & Cohen (2002).

With small sample size, non-equivalent groups and high correlation between ability dimensions, the item discrimination RMSE of concurrent calibration was lower than

separate linking methods, but the bias of concurrent calibration was larger than for the Direct method. Since the RMSE is the sum of bias and standard error, concurrent calibration would be better overall than the Direct method even with small sample size, non-equivalent groups and a high correlation between ability dimensions.

With non-equivalent groups or high correlation between ability dimensions, concurrent calibration still performed better than the TCF and Min methods with sample sizes of 500 and 1000. When sample size was 3000, concurrent and all separate linking methods had item discrimination mean bias very close to zero. Therefore, with non-equivalent groups and high correlation among ability dimensions, concurrent is a more reasonable option than separate linking methods across all sample sizes.

Test length had a very small effect either as a main effect or as an interaction with linking method. Therefore, the differences between the 40-item form and the 60-item form were very small and the relationship among linking methods did not differ depending on test length. Regardless of the test length, concurrent calibration generally performed slightly better than separate linking methods with respect to item discrimination RMSE and bias (refer to tables in Appendix).

Concurrent calibration has twice larger sample in item parameter estimation than with separate linking methods. The benefit of having large sample in item parameter estimation with concurrent calibration was larger with the 40-item form than with 60-item form. The proportion of the common items with 40-item form is larger than with 60-item form. The study by Kim & Cohen (2002) also showed that concurrent

calibration benefited more when the proportion of the common items was large. How-ever, the study by Hanson & Béguin (2002) showed that the the difference between concurrent calibration and separate linking methods was large when the number of common items was small, that is when the proportion of common items is small. Since simulation studies have results specific to the generating item parameters used in each study, the difference seen with Kim & Cohen (2002) and Hanson & Béguin (2002) could be due to the difference between property of generating items used in each study. Therefore, although this dissertation found that concurrent calibration benefited more when the proportion of common items was large, that is with 40-item form, the results could be different if different generating item sets were used.

5.2.2   Item Difficulty Parameter

The effect sizes of the interaction between the linking methods and sample size, and the interaction between the linking methods and test length were very small with respect to item difficulty RMSE and bias. The relationship among linking methods with respect to the item difficulty RMSE and bias was largely influenced by group equivalence, but little by sample size and test length. The relationships among linking methods with the 40-item form and the 60-item form were very similar. With a 40-item form, concurrent calibration generally benefited more from a larger sample than did separate linking methods. With a 60-item form, concurrent calibration had smaller item difficulty RMSE than separate linking methods when sample size was small and the correlation between ability dimensions was zero. With non-zero

correlation among ability dimensions, concurrent calibration did not benefit more from a larger sample size than separate linking methods with 60-item forms.

The effect size of test length and the interaction between linking methods and test length were small with the repeated measure ANOVA for the correlation between estimates and generating parameters. Thus, the correlation of the estimated and generating item difficulty was similar and the relationship among the linking methods differed little between 40-item and 60-item forms. Regardless of the test length, concurrent calibration generally performed as well or better than TCF, ICF and the Direct methods.

The effect size of the interaction between linking methods and sample size for the correlation of the item difficulty parameter was 0.18. With large sample size, concurrent calibration performed as well as TCF, ICF and the direct methods, however, with smaller sample size, concurrent calibration performed better than separate linking methods.

5.3   Research Question 3

The third research question involved the comparison among the separate linking methods. As in the first two research questions, the comparisons were made based on RMSE, bias, and the correlation between the estimated item parameters and the generating item parameters. The research question was as follows:

- Which separate linking method performs better than other separate linking

methods?

5.3.1   Item Discrimination Parameters

With respect to the RMSE and the bias of item discrimination parameters, all separate MIRT linking methods performed very similarly. Min (2003) compared the TCF and Min methods with approximate simple and mixed structure. It should be remembered that this dissertation and Min (2003) are not directly comparable since the software, item structure, and generating items employed are different. Also, in Min (2003), the non-equivalent groups were non-equivalent not only in ability levels, but also in the correlation between ability dimensions. With approximate simple structure in Min (2003), the TCF method had a smaller item discrimination parameter RMSE than the Mins method with non-equivalent groups, while the Min method had smaller RMSE than TCF when groups were equivalent.

5.3.2   Item Difficulty Parameter

With item difficulty, as described in the discussion of the first research question, the ICF had smaller item difficulty RMSE and bias than concurrent calibration and other separate linking methods when groups were non-equivalent. The ICF may have benefited by minimizing the cumulative squared difference between the item characteristic curves for each item for examinees of a particular ability, unlike the TCF, which uses the test characteristic function, or the Direct method which minimizes the sum of squared differences between the two sets of common item parameter estimates. Min's

method performed poorly with non-equivalent groups. When groups were equivalent, Min's method performed as well as other methods.

As discussed in regard to the first research question, the ICF had higher correlations between estimated and generating item difficulty parameters than other separate linking methods when groups were non-equivalent. As groups departed from equivalence, the correlation for Min's method became lower than other methods. This was true when correlations between trait dimensions were zero and when the correlation was 0.8.

The linking of item difficulty involves a rotation matrix and translation vector. The larger item difficulty RMSE with non-equivalent groups for Min's method could be due to a lack of dilation vector in the linking of item difficulty, and/or poor estimation of the translation vectors.

Unlike this dissertation, Min (2003) showed that the Min method had less RMSE than TCF regardless of group equivalence with approximate simple item structure. Item difficulty RMSE and bias were larger in this study than in Min (2003). In Min (2003), the average item difficulty values for the common items was -0.32, while in this study it was 1.163. The non-equivalent groups in this study were high achieving groups. Because most students answered the common items correctly, it was probably difficult to estimate the item parameters in TESTFACT.

Also, Min (2003) used only a half standard deviation difference for non-equivalent groups, while this study used a 0.5 and 1 standard deviation difference. The different

correlation values between ability dimensions for base and linked groups may have added more discrepancy between this study and Min (2003). Thus the differences seen in item difficulty RMSE and bias between this study and Min (2003) may be due, in part, to differences in the software used, the average item difficulty for common items, or the level of non-equivalence between groups.

5.4 Future Research and Limitations

There are several limitations of this study. First, this study considered only simple structure in which each item measures only one dimension. In MIRT, one item can measure multiple dimensions. Thus, further study of this kind is needed with a mixed item structure where items measure multiple dimensions.

Second, this study had two linking stages to place linked parameters onto the same scale. The second step is very important especially when groups are non-equivalent since concurrent calibration shifts the scale. However, depending on the choice of the linking method used in the second stage of linking, the results might be different. In this study, Min's method was used in thesecond stage of linking. To compare among studies, another way of shifting the scale may be needed.

Third, the only MIRT model considered in this study was the compensatory model. There would be situations in which a non-compensatory MIRT model would be more appropriate. Therefore, a study with non-compensatory MIRT would be beneficial in the future.

Fourth, the software used in this study was TESTFACT. As seen in the results section, there were problems in estimating item parameters with MIRT when there were high correlations between ability dimensions and non-equivalent groups. In this study, the amount of error due to the software was not examined separately from linking errors. Therefore, the amount of error in item parameter estimates from the software needs to be examined, possibly using multiple programs for MIRT, e.g., TESTFACT, MIRTE, and MULTIFACT. Moreover, better estimation procedures with correlated ability dimensions and non-equivalent groups are needed.

Fifth, the evaluation of linking was conducted only using the item parameter statistics. The estimated ability level for each examinee was not examined. Further studies that examine the ability estimates of examinees are needed as well.

Lastly, the conditions examined in this study included three levels of sample size and three levels of correlation between ability dimensions. These levels were the same between groups to be linked. In the non-equivalent groups condition, the only difference was the ability level. Further study is needed when the two groups differ in sample sizes and correlation between ability dimensions.

5.5 Conclusions

Concurrent calibration performed better than separate linking methods when groups were equivalent and the ability dimension had zero correlation. Although it appeared that the ICF performed much better than concurrent calibration with respect to

the item difficulty RMSE and bias when groups were non-equivalent, the correlation between the estimated and the generating item parameters showed that concurrent calibration was performing almost as well as the ICF with non-equivalent groups and even with 1 standard deviation difference.

Test length generally had small effect sizes and little influence on the relationship among linking methods. Thus the relationships among linking methods were similar for both 40-item and 60-item forms. Sample size had larger effects than test length. Concurrent calibration benefited more from a larger sample size than did separate linking methods with respect to all item parameters, especially with a shorter test form.

All the separate linking methods examined in this study were very similar with respect to the item discrimination parameters. The differences among the separate linking methods were seen clearly with the item difficulty parameter. The ICF method tended to perform better than other separate linking methods when groups were non-equivalent, while Min did not perform as well as other methods. With equivalent groups, all separate linking methods performed similarly.

This study examined the performance of concurrent calibration and separate linking methods for MIRT. The results of this study suggest that concurrent calibration generally performs better than separate linking methods even when groups were non-equivalent with 0.5 standard deviation difference and the correlation among ability dimensions was high. The results are generally consistent with UIRT studies by Han-

son & Béguin (1999), Hanson & Béguin (2002), Kim & Cohen (2002). As seen with UIRT studies by Hanson & Béguin (1999), concurrent calibration benefited by having larger sample size more than did separate linking methods. Concurrent calibration should not be blindly chosen given the results of this dissertation. There is a benefit of separate linking methods as discussed in Hanson & Béguin (1999) and Kolen & Brennan (2004). The potential benefit of separate linking is that having two sets of item parameter estimates can help to identify potential problems. There is a chance that serious problem might be buried and remain undetected if a single set of item parameters existed for the common items. Using multiple linking methods could help identify potential problems.

Bibliography

Ackerman, T. (1992). A didactic explanation of item bias item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29*, 67–91.

Akerman, T. (1994). Using multidimensional item response theory to understand what items and test are measureing. *Applied Measurement in Education, 7*(4), 255–278.

Angoff, W. (1982). *Summary and derivation of equating methods used at ETS*, (pp. 55 – 70). New York: Academy Press.

Baker, F. (1992). Equating tests under the graded response model. *Applied Psychological Measurement, 16*, 87–96.

Baker, F. (1993). EQUATE 2.0: A computer program for the characteristic curve method of IRT equating. *Applied Psychological Measurement, 17*, 20.

Batley, R. & Boss, M. (1993). The effects on parameter estimation of correlated

dimensions and a distribution-restricted trait in a multidimensional item response model. *Applied Psychological Measurement, 17*, 131–141.

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

Béguin, A., Hanson, B., & Glas, C. (2000a). Effect of multidimensionality on separate and concurrent estimation in IRT equating. Paper presented at the annual meeting of the National Council of Measurement in Education, New Orleans, LA. (Available at http://www.b-a-h.com/papers/paper0002.html).

Béguin, A., Hanson, B., & Glas, C. (2000b). Effect of multidimensionality on separate and concurrent estimation in IRT equating. Paper presented at the American Educational Research Association, New Orleans, LA.

Carlson, J. E. (1987). *Multidimensional item response theory estimation: A computer program (ACT Research Rep. No. 87-19).* Iowa City IA: American College Testing Program.

Christoffersson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika, 40*, 5–32.

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

Davey, T., Oshima, T., & Lee, K. (1996). Linking multidimensional item calibrations. *Applied Psychological Measurement*, *20*, 405–416.

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

Divgi, D. (1985). A minimum chi-square method for developing a common metric in item response theory. *Applied Psychological Measurement*, *9*, 413–415.

Embretson, S. E. & Reise, S. P. (2000). *Item Response Theory for Psychologists* (First ed.). NJ: Lawrence Erlbaum Associates.

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

Hanson, B. & Béguin, A. (1999). Separate versus concurrent estimation of IRT item parameters in the common item equating design. Technical report, Iowa City, IA: ACT inc.

Hanson, B. & Béguin, 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. (2007). *Linking and aligning scores and scales*, chapter 13: Practical Issues in Vertical Scaling, (pp. 233 –251). New York: Springer.

Hirsch, T. (1989). Mutidimensional equating. *Journal of Educational Measurement*, *26*, 337–349.

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

Kim, H. (1994). *New techniques for the dimensionality assessment of standardized test data*. PhD thesis, University of Illinois at Urbana-Champain.

Kim, J. (2001). *Proximity measures and luster analyses in multidimensional item response theory*. PhD thesis, Michigan State University.

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

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

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

Li, Y. & Lissitz, R. (2000). An evaluatin of the accuracy of multidimensional IRT linking. *Applied Psychological Measurement*, *24*, 115–138.

Lord, F. (1952). A theory of test scores. *Psychometric Monograph*, *7*.

Lord, F. (1980). *Applications of item response theory to practical testing problems*. NJ: Erlbaum.

Lord, F. & Novick, M. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison Wesley.

McKinley, R. & Reckase, M. (1983). An extension of the two-parameter logistic model to the multidimensional latent space. Technical report, Iowa City IA: The American College Testing Program.

Min, K. (2003). *The impact of scale dilation on the quality of the linking of multidimensional item response theory calibrations*. PhD thesis, Michigan State University.

Mislevy, R. & Bock, R. (1990). *BILOG 3: Item analysis and test scoring with binary logistic models*. Computer program.

Muthen, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika, 43*, 551–560.

Nandakumar, R. (1994). Assessing dimensionality of a set of item responses: Comparison of different approaches. *Journal of Educational Measurement, 31*(1), 17–35.

Oshima, T., Davey, T., & Lee, K. (2000). Mutidimensional linking: four practical approaches. *Journal of Educational Measurement, 37*, 357–373.

Patz, R. & Yao, L. (2007). *Linking and aligning scores and scales*, chapter 14: Methods and models for verstical scaling, (pp. 253–272). New York: Springer.

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

R Development Core Team (2007). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Reckase, M. (1985). The difficulty of items that measure more than one ability. *Applied Psychological Measurement*, *9*(5), 401–412.

Reckase, M. (1995). *A linear logistic multidimensional model for dichotomous item response data*, (pp. 271–286). New York: Springer.

Reckase, M. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement*, *21*, 25–36.

Reckase, M. & Hirsh, T. (1991). Interpretation of number correct scores when the true number of dimensions assessed by a test is greater than two. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Reckase, M. & Martineau, J. A. (2004). *The vertical scaling of science achievement tests*. Unpublished Report. Michigan State University.

Reise, S. P. & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement, 27,* 133 –144.

Roussos, L., Stout, W., & Marden, J. (1998). Using new proximity measures with hierarchical cluster analysis to detect multidimensionality. *Journal of Educational Measurement, 35,* 1–30.

Schönemann, P. (1966). A generalized solution of the orthogonal procrustes problem. *Psychometrika, 31,* 1–10.

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. (1988). Effect of examinee ability on test equating invariance. *Applied Psychological Measurement, 12,* 69–82.

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

Sympson, J. (1978). *A model for testing multidimensional items*, (pp. 82–98). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Tabachnick, B. G. & Fidell, L. S. (2007). *Using multivariate statistics* (Fifth ed.). Boston: Allyn and Bacon.

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

Thompson, T., Nering, M., & Davey, T. (1997). Multidimensional IRT scale linking. Paper presented at the annual meeting of the Psychometric Society, Gatlinburg, TN, June.

Tsai, T., Hanson, B., Kolen, M., & Forsyth, R. (2001). A comparison of bootstrap standard errors of IRT equating methods for the common-item nonequivalent groups design. *Applied Measurement In Education, 13*, 17–30.

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

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

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

Whitely, S. (1991). Measuring aptitude processes with multicomponent latent trait models. Technical report, Lawrence: University of Kansas.

Wingersky, M. & Lord, F. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. *Applied Psychological Measurement*, *8*(3), 347–364.

Wood, R., Wilson, D., Gibbons, R., Schilling, S., Muraki, E., & Bock, D. (1987). *TESTFACT*. Mooresville, IN: Scientific Software.

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

Yon, H. (2006). *Multidimensional Item Response Theory (MIRT) approaches to vertical scaling*. PhD thesis, Michigan State University.

Zimowski, M., Muraki, E., Mislevy, R., & Bock, R. (1996). *BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items*. Computer program.

Appendix   **A**

Appendix

A.1   Population Item Parameters

Table A.1. Item parameters of 20 common items.

Discrimination 1[1]	Discrimination 2[1]	Difficulty	Guessing
1.25	0.00	0.27	0.16
1.37	0.00	0.27	0.16
1.46	0.00	0.53	0.31
1.07	0.00	1.99	0.34
1.34	0.00	0.00	0.28
0.69	0.00	1.08	0.24
1.14	0.00	1.61	0.11
0.85	0.00	1.25	0.26
1.10	0.00	2.17	0.22
0.77	0.00	1.15	0.22
0.00	1.13	1.95	0.32
0.00	1.00	0.95	0.27
0.00	1.02	1.99	0.20
0.00	1.01	1.22	0.32
0.00	0.91	1.70	0.20
0.00	0.66	0.84	0.22
0.00	0.75	1.47	0.25
0.00	0.68	1.26	0.25
0.00	0.76	0.62	0.22
0.00	0.92	0.94	0.28

[1]1 indicates dimension 1; 2 indicates dimension 2.

Table A.2. Item parameters of unique items for form X.

Discrimination 1[1]	Discrimination 2[1]	Difficulty	Guessing
1.07	0.00	0.20	0.29
0.82	0.00	0.26	0.22
0.65	0.00	1.22	0.16
0.90	0.00	−0.67	0.20
0.95	0.00	0.40	0.31
0.76	0.00	0.84	0.17
0.89	0.00	0.53	0.34
0.53	0.00	0.51	0.18
0.78	0.00	1.51	0.21
0.86	0.00	0.81	0.24
0.00	0.62	−0.17	0.26
0.00	0.62	1.21	0.16
0.00	1.02	0.87	0.29
0.00	0.68	0.67	0.23
0.00	0.93	1.78	0.18
0.00	0.87	1.10	0.23
0.00	0.99	1.30	0.25
0.00	0.67	0.92	0.19
0.00	0.86	0.56	0.27
0.00	0.67	0.84	0.20

[1]1 indicates dimension 1; 2 indicates dimension 2.

Table A.3. Item parameters of unique items for form Y.

Discrimination 1^1	Discrimination 2^1	Difficulty	Guessing
0.99	0.00	0.77	0.22
0.95	0.00	1.25	0.29
0.78	0.00	0.95	0.09
1.17	0.00	-0.00	0.12
0.80	0.00	0.93	0.21
0.97	0.00	0.49	0.39
1.03	0.00	1.75	0.43
0.81	0.00	0.93	0.09
0.75	0.00	1.46	0.24
0.71	0.00	0.89	0.24
0.00	0.97	0.83	0.26
0.00	1.13	1.35	0.24
0.00	0.65	-0.03	0.21
0.00	1.12	2.13	0.27
0.00	1.13	2.25	0.24
0.00	0.53	0.89	0.21
0.00	1.35	2.41	0.14
0.00	0.87	0.86	0.24
0.00	0.70	-0.10	0.26
0.00	0.89	1.48	0.35

[1]1 indicates dimension 1; 2 indicates dimension 2.

Appendix **B**

Appendix

B.1   Tables of means and standard deviations for RMSE

Table B.1. Mean RMSE (Root Mean Squared Error) for $a1$ with test length of 40 items.

Mean Diff (SD)[2]	r	Sample size	Min	Direct	TCF[1]	ICF[1]	Concurrent
0.00	0.00	500	0.15	0.15	0.15	0.15	0.13
0.00	0.00	1000	0.11	0.11	0.11	0.11	0.09
0.00	0.00	3000	0.07	0.07	0.07	0.07	0.06
0.00	0.50	500	0.18	0.18	0.18	0.18	0.15
0.00	0.50	1000	0.12	0.12	0.12	0.12	0.11
0.00	0.50	3000	0.08	0.08	0.08	0.08	0.07
0.00	0.80	500	0.26	0.27	0.26	0.26	0.22
0.00	0.80	1000	0.18	0.19	0.18	0.18	0.16
0.00	0.80	3000	0.11	0.11	0.11	0.11	0.09
0.50	0.00	500	0.20	0.20	0.20	0.20	0.15
0.50	0.00	1000	0.13	0.13	0.13	0.13	0.11
0.50	0.00	3000	0.08	0.08	0.08	0.08	0.07
0.50	0.50	500	0.21	0.21	0.21	0.21	0.18
0.50	0.50	1000	0.15	0.15	0.15	0.15	0.12
0.50	0.50	3000	0.09	0.09	0.09	0.09	0.07
0.50	0.80	500	0.35	0.33	0.35	0.34	0.28
0.50	0.80	1000	0.22	0.23	0.22	0.22	0.19
0.50	0.80	3000	0.13	0.13	0.13	0.13	0.11
1.00	0.00	500	0.28	0.28	0.28	0.28	0.24
1.00	0.00	1000	0.18	0.18	0.18	0.18	0.14
1.00	0.00	3000	0.10	0.10	0.10	0.10	0.10
1.00	0.50	500	0.33	0.34	0.34	0.34	0.25
1.00	0.50	1000	0.18	0.18	0.18	0.18	0.14
1.00	0.50	3000	0.11	0.11	0.11	0.11	0.09
1.00	0.80	500	0.46	0.45	0.46	0.46	0.37
1.00	0.80	1000	0.29	0.28	0.29	0.29	0.24
1.00	0.80	3000	0.17	0.17	0.17	0.17	0.14

[1]TCF is the test characteristic function method; ICF is the item characteristic function method. [2]Diff (SD) means difference between two groups in standard deviation unit.

Table B.2. Standard deviation of RMSE (Root Mean Squared Error) for a1 with test length of 40 items.

Mean Diff (SD)[2]	r	Sample size	Min	Direct	TCF[1]	ICF[1]	Concurrent
0.00	0.00	500	0.04	0.04	0.04	0.04	0.03
0.00	0.00	1000	0.01	0.01	0.01	0.01	0.01
0.00	0.00	3000	0.01	0.01	0.01	0.01	0.01
0.00	0.50	500	0.03	0.03	0.03	0.03	0.02
0.00	0.50	1000	0.02	0.02	0.02	0.02	0.02
0.00	0.50	3000	0.01	0.01	0.01	0.01	0.01
0.00	0.80	500	0.05	0.05	0.05	0.05	0.03
0.00	0.80	1000	0.02	0.02	0.02	0.02	0.02
0.00	0.80	3000	0.01	0.01	0.01	0.01	0.01
0.50	0.00	500	0.07	0.07	0.08	0.08	0.03
0.50	0.00	1000	0.02	0.02	0.02	0.02	0.02
0.50	0.00	3000	0.01	0.01	0.01	0.01	0.01
0.50	0.50	500	0.04	0.04	0.04	0.04	0.05
0.50	0.50	1000	0.02	0.02	0.02	0.02	0.02
0.50	0.50	3000	0.01	0.01	0.01	0.01	0.01
0.50	0.80	500	0.10	0.08	0.11	0.10	0.08
0.50	0.80	1000	0.03	0.03	0.03	0.03	0.03
0.50	0.80	3000	0.02	0.02	0.02	0.02	0.01
1.00	0.00	500	0.10	0.10	0.10	0.10	0.13
1.00	0.00	1000	0.05	0.05	0.05	0.05	0.02
1.00	0.00	3000	0.01	0.01	0.01	0.01	0.01
1.00	0.50	500	0.11	0.11	0.12	0.12	0.09
1.00	0.50	1000	0.03	0.03	0.03	0.03	0.02
1.00	0.50	3000	0.02	0.02	0.02	0.02	0.01
1.00	0.80	500	0.12	0.12	0.12	0.12	0.13
1.00	0.80	1000	0.05	0.04	0.05	0.05	0.07
1.00	0.80	3000	0.02	0.02	0.02	0.03	0.02

[1]TCF is the test characteristic function method; ICF is the item characteristic function method. [2]Diff (SD) means difference between two groups in standard deviation unit.

Table B.3. Mean RMSE (Root Mean Squared Error) for $a1$ with test length of 60 items.

Mean Diff (SD)[2]	r	Sample size	Min	Direct	TCF[1]	ICF[1]	Concurrent
0.00	0.00	500	0.14	0.14	0.14	0.14	0.13
0.00	0.00	1000	0.10	0.10	0.10	0.10	0.09
0.00	0.00	3000	0.06	0.06	0.06	0.06	0.06
0.00	0.50	500	0.16	0.16	0.16	0.16	0.14
0.00	0.50	1000	0.11	0.11	0.11	0.11	0.10
0.00	0.50	3000	0.07	0.07	0.07	0.07	0.06
0.00	0.80	500	0.23	0.24	0.23	0.23	0.21
0.00	0.80	1000	0.17	0.17	0.17	0.17	0.15
0.00	0.80	3000	0.10	0.10	0.10	0.10	0.09
0.50	0.00	500	0.17	0.17	0.17	0.17	0.15
0.50	0.00	1000	0.12	0.12	0.12	0.12	0.11
0.50	0.00	3000	0.08	0.08	0.08	0.08	0.07
0.50	0.50	500	0.19	0.19	0.19	0.19	0.17
0.50	0.50	1000	0.13	0.13	0.13	0.13	0.12
0.50	0.50	3000	0.08	0.08	0.08	0.08	0.08
0.50	0.80	500	0.28	0.28	0.28	0.28	0.25
0.50	0.80	1000	0.20	0.20	0.20	0.20	0.18
0.50	0.80	3000	0.12	0.12	0.12	0.12	0.11
1.00	0.00	500	0.28	0.28	0.28	0.28	0.25
1.00	0.00	1000	0.15	0.15	0.15	0.15	0.14
1.00	0.00	3000	0.10	0.10	0.10	0.10	0.10
1.00	0.50	500	0.28	0.29	0.28	0.28	0.23
1.00	0.50	1000	0.18	0.18	0.18	0.18	0.15
1.00	0.50	3000	0.11	0.11	0.11	0.11	0.09
1.00	0.80	500	0.41	0.41	0.41	0.41	0.36
1.00	0.80	1000	0.27	0.27	0.27	0.27	0.22
1.00	0.80	3000	0.16	0.16	0.16	0.16	0.13

[1]TCF is the test characteristic function method; ICF is the item characteristic function method. [2]Diff (SD) means difference between two groups in standard deviation unit.

Table B.4. Standard deviation of RMSE (Root Mean Squared Error) for a1 with test length of 60 items.

Mean Diff (SD)[2]	r	Sample size	Min	Direct	TCF[1]	ICF[1]	Concurrent
0.00	0.00	500	0.02	0.02	0.02	0.02	0.01
0.00	0.00	1000	0.01	0.01	0.01	0.01	0.01
0.00	0.00	3000	0.01	0.01	0.01	0.01	0.01
0.00	0.50	500	0.02	0.02	0.02	0.02	0.01
0.00	0.50	1000	0.01	0.02	0.01	0.01	0.01
0.00	0.50	3000	0.01	0.01	0.01	0.01	0.01
0.00	0.80	500	0.03	0.03	0.03	0.03	0.03
0.00	0.80	1000	0.02	0.02	0.02	0.02	0.02
0.00	0.80	3000	0.01	0.01	0.01	0.01	0.01
0.50	0.00	500	0.02	0.02	0.02	0.02	0.02
0.50	0.00	1000	0.01	0.01	0.01	0.01	0.01
0.50	0.00	3000	0.01	0.01	0.01	0.01	0.01
0.50	0.50	500	0.04	0.04	0.04	0.04	0.05
0.50	0.50	1000	0.01	0.01	0.01	0.01	0.01
0.50	0.50	3000	0.01	0.01	0.01	0.01	0.01
0.50	0.80	500	0.04	0.04	0.04	0.04	0.03
0.50	0.80	1000	0.03	0.03	0.03	0.03	0.02
0.50	0.80	3000	0.01	0.01	0.01	0.01	0.01
1.00	0.00	500	0.12	0.12	0.12	0.12	0.11
1.00	0.00	1000	0.02	0.02	0.02	0.02	0.02
1.00	0.00	3000	0.01	0.01	0.01	0.01	0.01
1.00	0.50	500	0.10	0.10	0.10	0.10	0.08
1.00	0.50	1000	0.03	0.03	0.03	0.03	0.02
1.00	0.50	3000	0.01	0.01	0.01	0.01	0.01
1.00	0.80	500	0.10	0.10	0.10	0.10	0.09
1.00	0.80	1000	0.07	0.07	0.07	0.07	0.04
1.00	0.80	3000	0.02	0.02	0.02	0.02	0.01

[1]TCF is the test characteristic function method; ICF is the item characteristic function method. [2]Diff (SD) means difference between two groups in standard deviation unit.

Table B.5. Mean RMSE (Root Mean Squared Error) for $a2$ with test length of 40 items.

Mean Diff (SD)[2]	r	Sample size	Min	Direct	TCF[1]	ICF[1]	Concurrent
0.00	0.00	500	0.16	0.16	0.16	0.16	0.14
0.00	0.00	1000	0.11	0.11	0.11	0.11	0.10
0.00	0.00	3000	0.07	0.07	0.07	0.07	0.06
0.00	0.50	500	0.18	0.18	0.18	0.18	0.15
0.00	0.50	1000	0.12	0.12	0.12	0.12	0.11
0.00	0.50	3000	0.08	0.08	0.08	0.08	0.07
0.00	0.80	500	0.27	0.27	0.27	0.27	0.23
0.00	0.80	1000	0.19	0.20	0.19	0.19	0.17
0.00	0.80	3000	0.11	0.11	0.11	0.11	0.10
0.50	0.00	500	0.21	0.21	0.21	0.21	0.17
0.50	0.00	1000	0.14	0.14	0.14	0.14	0.12
0.50	0.00	3000	0.09	0.09	0.09	0.09	0.08
0.50	0.50	500	0.22	0.22	0.22	0.22	0.19
0.50	0.50	1000	0.15	0.15	0.15	0.15	0.13
0.50	0.50	3000	0.10	0.10	0.10	0.10	0.09
0.50	0.80	500	0.36	0.36	0.37	0.36	0.30
0.50	0.80	1000	0.24	0.24	0.24	0.24	0.20
0.50	0.80	3000	0.14	0.14	0.14	0.14	0.11
1.00	0.00	500	0.32	0.32	0.32	0.32	0.26
1.00	0.00	1000	0.19	0.19	0.19	0.19	0.16
1.00	0.00	3000	0.11	0.11	0.11	0.11	0.11
1.00	0.50	500	0.34	0.35	0.35	0.35	0.27
1.00	0.50	1000	0.20	0.20	0.20	0.20	0.16
1.00	0.50	3000	0.13	0.13	0.13	0.13	0.11
1.00	0.80	500	0.46	0.47	0.47	0.47	0.40
1.00	0.80	1000	0.31	0.31	0.31	0.31	0.26
1.00	0.80	3000	0.19	0.19	0.19	0.19	0.15

[1]TCF is the test characteristic function method; ICF is the item characteristic function method. [2]Diff (SD) means difference between two groups in standard deviation unit.

Table B.6. Standard deviation of RMSE (Root Mean Squared Error) for $a2$ with test length of 40 items.

Mean Diff (SD)[2]	r	Sample size	Min	Direct	TCF[1]	ICF[1]	Concurrent
0.00	0.00	500	0.02	0.02	0.02	0.02	0.02
0.00	0.00	1000	0.02	0.02	0.02	0.02	0.02
0.00	0.00	3000	0.01	0.01	0.01	0.01	0.01
0.00	0.50	500	0.02	0.02	0.02	0.02	0.02
0.00	0.50	1000	0.02	0.02	0.02	0.02	0.02
0.00	0.50	3000	0.01	0.01	0.01	0.01	0.01
0.00	0.80	500	0.05	0.05	0.05	0.05	0.03
0.00	0.80	1000	0.02	0.02	0.02	0.02	0.02
0.00	0.80	3000	0.02	0.02	0.02	0.02	0.01
0.50	0.00	500	0.07	0.08	0.07	0.07	0.04
0.50	0.00	1000	0.02	0.02	0.02	0.02	0.02
0.50	0.00	3000	0.01	0.01	0.01	0.01	0.01
0.50	0.50	500	0.05	0.05	0.05	0.05	0.06
0.50	0.50	1000	0.02	0.02	0.02	0.02	0.02
0.50	0.50	3000	0.01	0.01	0.01	0.01	0.01
0.50	0.80	500	0.10	0.09	0.10	0.10	0.08
0.50	0.80	1000	0.03	0.03	0.03	0.03	0.03
0.50	0.80	3000	0.02	0.02	0.02	0.02	0.02
1.00	0.00	500	0.11	0.11	0.11	0.11	0.12
1.00	0.00	1000	0.05	0.05	0.05	0.05	0.03
1.00	0.00	3000	0.02	0.02	0.02	0.02	0.01
1.00	0.50	500	0.10	0.11	0.11	0.11	0.10
1.00	0.50	1000	0.03	0.03	0.03	0.03	0.03
1.00	0.50	3000	0.02	0.02	0.02	0.02	0.02
1.00	0.80	500	0.09	0.10	0.09	0.09	0.11
1.00	0.80	1000	0.06	0.05	0.06	0.06	0.07
1.00	0.80	3000	0.03	0.03	0.03	0.03	0.03

[1]TCF is the test characteristic function method; ICF is the item characteristic function method. [2]Diff (SD) means difference between two groups in standard deviation unit.

Table B.7. Mean RMSE (Root Mean Squared Error) for $a2$ with test length of 60 items.

Mean Diff (SD)[2]	r	Sample size	Min	Direct	TCF[1]	ICF[1]	Concurrent
0.00	0.00	500	0.15	0.15	0.15	0.15	0.14
0.00	0.00	1000	0.11	0.11	0.11	0.11	0.10
0.00	0.00	3000	0.08	0.08	0.08	0.08	0.07
0.00	0.50	500	0.17	0.17	0.17	0.17	0.15
0.00	0.50	1000	0.12	0.12	0.12	0.12	0.11
0.00	0.50	3000	0.08	0.08	0.08	0.08	0.07
0.00	0.80	500	0.25	0.25	0.25	0.25	0.23
0.00	0.80	1000	0.18	0.18	0.18	0.18	0.17
0.00	0.80	3000	0.11	0.11	0.11	0.11	0.10
0.50	0.00	500	0.19	0.19	0.19	0.19	0.17
0.50	0.00	1000	0.13	0.13	0.13	0.13	0.12
0.50	0.00	3000	0.09	0.09	0.09	0.09	0.08
0.50	0.50	500	0.21	0.22	0.21	0.21	0.20
0.50	0.50	1000	0.15	0.15	0.15	0.15	0.13
0.50	0.50	3000	0.09	0.09	0.09	0.09	0.08
0.50	0.80	500	0.30	0.31	0.31	0.31	0.27
0.50	0.80	1000	0.21	0.21	0.21	0.21	0.19
0.50	0.80	3000	0.13	0.13	0.13	0.13	0.12
1.00	0.00	500	0.31	0.31	0.31	0.31	0.28
1.00	0.00	1000	0.18	0.17	0.17	0.17	0.16
1.00	0.00	3000	0.11	0.11	0.11	0.11	0.11
1.00	0.50	500	0.32	0.32	0.32	0.32	0.28
1.00	0.50	1000	0.19	0.19	0.19	0.19	0.17
1.00	0.50	3000	0.12	0.12	0.12	0.12	0.11
1.00	0.80	500	0.44	0.44	0.44	0.44	0.40
1.00	0.80	1000	0.29	0.29	0.29	0.29	0.26
1.00	0.80	3000	0.17	0.17	0.17	0.17	0.15

[1]TCF is the test characteristic function method; ICF is the item characteristic function method. [2]Diff (SD) means difference between two groups in standard deviation unit.

Table B.8. Standard deviation of RMSE (Root Mean Squared Error) for $a2$ with test length of 60 items.

Mean Diff (SD)[2]	r	Sample size	Min	Direct	TCF[1]	ICF[1]	Concurrent
0.00	0.00	500	0.02	0.02	0.02	0.02	0.02
0.00	0.00	1000	0.01	0.01	0.01	0.01	0.01
0.00	0.00	3000	0.01	0.01	0.01	0.01	0.01
0.00	0.50	500	0.02	0.02	0.02	0.02	0.02
0.00	0.50	1000	0.02	0.02	0.02	0.02	0.01
0.00	0.50	3000	0.01	0.01	0.01	0.01	0.01
0.00	0.80	500	0.03	0.04	0.03	0.03	0.03
0.00	0.80	1000	0.02	0.02	0.02	0.02	0.02
0.00	0.80	3000	0.01	0.01	0.01	0.01	0.01
0.50	0.00	500	0.04	0.04	0.04	0.04	0.03
0.50	0.00	1000	0.01	0.01	0.01	0.01	0.02
0.50	0.00	3000	0.01	0.01	0.01	0.01	0.01
0.50	0.50	500	0.03	0.03	0.03	0.03	0.04
0.50	0.50	1000	0.02	0.02	0.02	0.02	0.02
0.50	0.50	3000	0.01	0.01	0.01	0.01	0.01
0.50	0.80	500	0.05	0.05	0.05	0.05	0.03
0.50	0.80	1000	0.02	0.02	0.02	0.02	0.02
0.50	0.80	3000	0.01	0.02	0.02	0.02	0.01
1.00	0.00	500	0.11	0.11	0.11	0.11	0.12
1.00	0.00	1000	0.03	0.03	0.03	0.03	0.02
1.00	0.00	3000	0.01	0.01	0.01	0.01	0.01
1.00	0.50	500	0.09	0.10	0.09	0.09	0.10
1.00	0.50	1000	0.03	0.03	0.03	0.03	0.04
1.00	0.50	3000	0.01	0.01	0.01	0.01	0.01
1.00	0.80	500	0.10	0.10	0.10	0.10	0.11
1.00	0.80	1000	0.08	0.07	0.07	0.07	0.07
1.00	0.80	3000	0.02	0.02	0.02	0.02	0.02

[1]TCF is the test characteristic function method; ICF is the item characteristic function method. [2]Diff (SD) means difference between two groups in standard deviation unit.

Table B.9. Mean RMSE (Root Mean Squared Error) for d with test length of 40 items.

Mean Diff (SD)[2]	r	Sample size	Min	Direct	TCF[1]	ICF[1]	Concurrent
0.00	0.00	500	0.17	0.17	0.17	0.17	0.15
0.00	0.00	1000	0.12	0.12	0.12	0.12	0.11
0.00	0.00	3000	0.09	0.09	0.09	0.09	0.08
0.00	0.50	500	0.17	0.17	0.17	0.17	0.14
0.00	0.50	1000	0.12	0.12	0.12	0.12	0.11
0.00	0.50	3000	0.08	0.08	0.08	0.08	0.08
0.00	0.80	500	0.18	0.19	0.19	0.19	0.15
0.00	0.80	1000	0.12	0.12	0.12	0.13	0.11
0.00	0.80	3000	0.09	0.09	0.09	0.09	0.09
0.50	0.00	500	0.64	0.40	0.43	0.43	0.36
0.50	0.00	1000	0.49	0.29	0.29	0.29	0.29
0.50	0.00	3000	0.45	0.20	0.21	0.19	0.26
0.50	0.50	500	0.60	0.37	0.39	0.38	0.43
0.50	0.50	1000	0.49	0.28	0.29	0.28	0.28
0.50	0.50	3000	0.46	0.24	0.24	0.24	0.25
0.50	0.80	500	0.77	0.47	0.55	0.54	0.49
0.50	0.80	1000	0.47	0.22	0.25	0.25	0.25
0.50	0.80	3000	0.44	0.20	0.21	0.22	0.23
1.00	0.00	500	1.47	1.09	1.09	1.07	1.12
1.00	0.00	1000	1.00	0.64	0.64	0.62	0.58
1.00	0.00	3000	0.87	0.53	0.54	0.35	0.53
1.00	0.50	500	1.66	1.21	1.24	1.24	0.95
1.00	0.50	1000	0.95	0.55	0.57	0.54	0.55
1.00	0.50	3000	0.89	0.53	0.53	0.38	0.51
1.00	0.80	500	1.76	1.20	1.27	1.27	1.27
1.00	0.80	1000	0.94	0.48	0.56	0.52	0.58
1.00	0.80	3000	0.85	0.47	0.50	0.36	0.49

[1]TCF is the test characteristic function method; ICF is the item characteristic function method. [2]Diff (SD) means difference between two groups in standard deviation unit.

Table B.10. Standard deviation of RMSE (Root Mean Squared Error) for d with test length of 40 items.

Mean Diff (SD)[2]	r	Sample size	Min	Direct	TCF[1]	ICF[1]	Concurrent
0.00	0.00	500	0.07	0.07	0.06	0.07	0.06
0.00	0.00	1000	0.02	0.02	0.02	0.02	0.02
0.00	0.00	3000	0.01	0.01	0.01	0.01	0.01
0.00	0.50	500	0.03	0.02	0.02	0.02	0.02
0.00	0.50	1000	0.02	0.02	0.02	0.02	0.02
0.00	0.50	3000	0.01	0.01	0.01	0.01	0.01
0.00	0.80	500	0.09	0.09	0.09	0.09	0.04
0.00	0.80	1000	0.02	0.02	0.02	0.03	0.02
0.00	0.80	3000	0.01	0.01	0.01	0.02	0.01
0.50	0.00	500	0.37	0.19	0.34	0.34	0.10
0.50	0.00	1000	0.03	0.04	0.04	0.05	0.03
0.50	0.00	3000	0.02	0.08	0.08	0.08	0.01
0.50	0.50	500	0.20	0.16	0.19	0.19	0.66
0.50	0.50	1000	0.03	0.05	0.04	0.06	0.02
0.50	0.50	3000	0.02	0.06	0.06	0.06	0.02
0.50	0.80	500	0.57	0.57	0.60	0.57	0.67
0.50	0.80	1000	0.03	0.07	0.06	0.05	0.03
0.50	0.80	3000	0.02	0.07	0.06	0.05	0.02
1.00	0.00	500	0.76	0.81	0.81	0.82	1.08
1.00	0.00	1000	0.40	0.41	0.41	0.42	0.07
1.00	0.00	3000	0.03	0.06	0.02	0.21	0.02
1.00	0.50	500	1.00	0.94	0.96	0.96	0.85
1.00	0.50	1000	0.06	0.10	0.04	0.12	0.04
1.00	0.50	3000	0.03	0.06	0.02	0.21	0.02
1.00	0.80	500	1.13	1.13	1.02	1.03	1.27
1.00	0.80	1000	0.08	0.15	0.07	0.12	0.31
1.00	0.80	3000	0.03	0.12	0.02	0.18	0.02

[1]TCF is the test characteristic function method; ICF is the item characteristic function method. [2]Diff (SD) means difference between two groups in standard deviation unit.

Table B.11. Mean RMSE (Root Mean Squared Error) for d with test length of 60 items.

Mean Diff (SD)[2]	r	Sample size	Min	Direct	TCF[1]	ICF[1]	Concurrent
0.00	0.00	500	0.16	0.16	0.16	0.16	0.15
0.00	0.00	1000	0.12	0.12	0.12	0.12	0.11
0.00	0.00	3000	0.10	0.10	0.10	0.10	0.09
0.00	0.50	500	0.16	0.16	0.16	0.16	0.15
0.00	0.50	1000	0.12	0.12	0.12	0.12	0.12
0.00	0.50	3000	0.09	0.09	0.09	0.09	0.09
0.00	0.80	500	0.18	0.18	0.18	0.18	0.16
0.00	0.80	1000	0.12	0.13	0.13	0.14	0.12
0.00	0.80	3000	0.09	0.10	0.10	0.10	0.09
0.50	0.00	500	0.53	0.34	0.34	0.34	0.35
0.50	0.00	1000	0.46	0.27	0.27	0.27	0.28
0.50	0.00	3000	0.43	0.18	0.19	0.16	0.25
0.50	0.50	500	0.55	0.34	0.34	0.33	0.35
0.50	0.50	1000	0.47	0.26	0.26	0.26	0.27
0.50	0.50	3000	0.44	0.20	0.22	0.21	0.24
0.50	0.80	500	0.54	0.29	0.33	0.32	0.31
0.50	0.80	1000	0.45	0.21	0.24	0.23	0.25
0.50	0.80	3000	0.41	0.18	0.20	0.19	0.22
1.00	0.00	500	1.55	1.13	1.14	1.14	1.09
1.00	0.00	1000	0.90	0.55	0.55	0.52	0.57
1.00	0.00	3000	0.83	0.49	0.52	0.32	0.52
1.00	0.50	500	1.31	0.83	0.89	0.86	0.86
1.00	0.50	1000	0.92	0.53	0.56	0.52	0.55
1.00	0.50	3000	0.84	0.49	0.51	0.27	0.49
1.00	0.80	500	1.56	0.99	1.08	1.05	1.11
1.00	0.80	1000	0.95	0.57	0.60	0.54	0.54
1.00	0.80	3000	0.80	0.45	0.47	0.32	0.46

[1]TCF is the test characteristic function method; ICF is the item characteristic function method. [2]Diff (SD) means difference between two groups in standard deviation unit.

Table B.12. Standard deviation of RMSE (Root Mean Squared Error) for d with test length of 60 items.

Mean Diff (SD)[2]	r	Sample size	Min	Direct	TCF[1]	ICF[1]	Concurrent
0.00	0.00	500	0.03	0.03	0.02	0.02	0.03
0.00	0.00	1000	0.01	0.01	0.01	0.01	0.01
0.00	0.00	3000	0.01	0.01	0.01	0.01	0.01
0.00	0.50	500	0.02	0.03	0.02	0.03	0.02
0.00	0.50	1000	0.02	0.02	0.02	0.02	0.01
0.00	0.50	3000	0.01	0.01	0.01	0.01	0.01
0.00	0.80	500	0.04	0.04	0.04	0.04	0.04
0.00	0.80	1000	0.02	0.02	0.03	0.03	0.02
0.00	0.80	3000	0.01	0.01	0.01	0.01	0.01
0.50	0.00	500	0.09	0.09	0.09	0.09	0.09
0.50	0.00	1000	0.03	0.04	0.04	0.04	0.02
0.50	0.00	3000	0.02	0.08	0.07	0.07	0.01
0.50	0.50	500	0.13	0.14	0.13	0.13	0.17
0.50	0.50	1000	0.04	0.05	0.05	0.05	0.03
0.50	0.50	3000	0.02	0.07	0.06	0.06	0.01
0.50	0.80	500	0.17	0.18	0.17	0.17	0.05
0.50	0.80	1000	0.03	0.07	0.05	0.05	0.03
0.50	0.80	3000	0.02	0.06	0.05	0.05	0.01
1.00	0.00	500	0.93	0.89	0.89	0.89	0.81
1.00	0.00	1000	0.05	0.04	0.04	0.11	0.04
1.00	0.00	3000	0.02	0.10	0.01	0.20	0.02
1.00	0.50	500	0.63	0.50	0.58	0.60	0.60
1.00	0.50	1000	0.05	0.10	0.04	0.13	0.06
1.00	0.50	3000	0.02	0.08	0.02	0.19	0.02
1.00	0.80	500	0.87	0.75	0.72	0.74	1.00
1.00	0.80	1000	0.34	0.41	0.39	0.42	0.12
1.00	0.80	3000	0.02	0.09	0.02	0.17	0.02

[1]TCF is the test characteristic function method; ICF is the item characteristic function method. [2]Diff (SD) means difference between two groups in standard deviation unit.

Appendix   **C**

Appendix

C.1   Tables of means and standard deviations for BIAS

Table C.1. Mean bias for $a1$ with test length of 40 items.

Mean Diff (SD)[2]	r	Sample size	Min	Direct	TCF[1]	ICF[1]	Concurrent
0.00	0.00	500	−0.03	−0.03	−0.03	−0.03	−0.02
0.00	0.00	1000	−0.01	−0.01	−0.01	−0.01	−0.01
0.00	0.00	3000	−0.01	−0.01	−0.01	−0.01	−0.01
0.00	0.50	500	−0.03	−0.02	−0.03	−0.03	−0.02
0.00	0.50	1000	−0.01	−0.01	−0.01	−0.01	−0.01
0.00	0.50	3000	−0.00	−0.00	−0.00	−0.00	−0.00
0.00	0.80	500	−0.05	−0.02	−0.05	−0.05	−0.03
0.00	0.80	1000	−0.02	0.00	−0.01	−0.01	−0.01
0.00	0.80	3000	0.00	0.01	0.01	0.01	0.01
0.50	0.00	500	−0.05	−0.05	−0.05	−0.05	−0.02
0.50	0.00	1000	−0.02	−0.02	−0.02	−0.02	−0.01
0.50	0.00	3000	−0.01	−0.01	−0.01	−0.01	0.00
0.50	0.50	500	−0.04	−0.04	−0.04	−0.04	−0.03
0.50	0.50	1000	−0.02	−0.02	−0.02	−0.02	−0.01
0.50	0.50	3000	−0.01	−0.01	−0.01	−0.01	−0.00
0.50	0.80	500	−0.11	−0.04	−0.10	−0.10	−0.06
0.50	0.80	1000	−0.03	−0.01	−0.03	−0.03	−0.02
0.50	0.80	3000	−0.00	0.00	0.00	0.00	0.00
1.00	0.00	500	−0.10	−0.10	−0.10	−0.11	−0.05
1.00	0.00	1000	−0.04	−0.04	−0.04	−0.04	0.01
1.00	0.00	3000	−0.01	−0.01	−0.01	−0.02	0.02
1.00	0.50	500	−0.11	−0.12	−0.12	−0.12	−0.06
1.00	0.50	1000	−0.03	−0.03	−0.03	−0.03	−0.01
1.00	0.50	3000	−0.01	−0.01	−0.01	−0.01	0.00
1.00	0.80	500	−0.19	−0.15	−0.19	−0.19	−0.13
1.00	0.80	1000	−0.06	−0.04	−0.07	−0.06	−0.03
1.00	0.80	3000	−0.01	−0.02	−0.02	−0.01	0.01

[1]TCF is the test characteristic function method; ICF is the item characteristic function method. [2]Diff (SD) means difference between two groups in standard deviation unit.

Table C.2. Standard deviation of bias for $a1$ with test length of 40 items.

Mean Diff (SD)[2]	r	Sample size	Min	Direct	TCF[1]	ICF[1]	Concurrent
0.00	0.00	500	0.03	0.03	0.02	0.03	0.02
0.00	0.00	1000	0.01	0.01	0.01	0.01	0.01
0.00	0.00	3000	0.01	0.01	0.01	0.01	0.01
0.00	0.50	500	0.02	0.02	0.02	0.02	0.01
0.00	0.50	1000	0.01	0.01	0.01	0.01	0.01
0.00	0.50	3000	0.00	0.00	0.00	0.00	0.00
0.00	0.80	500	0.04	0.05	0.04	0.05	0.02
0.00	0.80	1000	0.01	0.02	0.02	0.02	0.01
0.00	0.80	3000	0.01	0.01	0.01	0.01	0.01
0.50	0.00	500	0.05	0.04	0.05	0.05	0.02
0.50	0.00	1000	0.01	0.01	0.01	0.01	0.01
0.50	0.00	3000	0.01	0.01	0.01	0.01	0.01
0.50	0.50	500	0.03	0.02	0.03	0.03	0.03
0.50	0.50	1000	0.01	0.01	0.01	0.01	0.01
0.50	0.50	3000	0.00	0.01	0.01	0.01	0.00
0.50	0.80	500	0.09	0.08	0.09	0.09	0.07
0.50	0.80	1000	0.02	0.03	0.02	0.02	0.02
0.50	0.80	3000	0.01	0.02	0.01	0.01	0.01
1.00	0.00	500	0.08	0.08	0.08	0.08	0.11
1.00	0.00	1000	0.04	0.04	0.04	0.04	0.01
1.00	0.00	3000	0.01	0.01	0.01	0.01	0.01
1.00	0.50	500	0.10	0.10	0.10	0.10	0.06
1.00	0.50	1000	0.01	0.01	0.01	0.01	0.01
1.00	0.50	3000	0.01	0.01	0.01	0.01	0.01
1.00	0.80	500	0.12	0.13	0.12	0.12	0.12
1.00	0.80	1000	0.04	0.05	0.05	0.04	0.06
1.00	0.80	3000	0.01	0.01	0.01	0.02	0.01

[1]TCF is the test characteristic function method; ICF is the item characteristic function method. [2]Diff (SD) means difference between two groups in standard deviation unit.

Table C.3. Mean bias for $a1$ with test length of 60 items.

Mean Diff (SD)[2]	r	Sample size	Min	Direct	TCF[1]	ICF[1]	Concurrent
0.00	0.00	500	−0.02	−0.02	−0.02	−0.02	−0.02
0.00	0.00	1000	−0.01	−0.01	−0.01	−0.01	−0.01
0.00	0.00	3000	−0.00	−0.01	−0.00	−0.00	−0.00
0.00	0.50	500	−0.02	−0.02	−0.02	−0.02	−0.02
0.00	0.50	1000	−0.01	−0.01	−0.01	−0.01	−0.01
0.00	0.50	3000	−0.00	−0.00	−0.00	−0.00	−0.00
0.00	0.80	500	−0.04	−0.01	−0.03	−0.02	−0.03
0.00	0.80	1000	−0.01	0.00	−0.01	−0.00	−0.01
0.00	0.80	3000	0.00	0.00	0.00	0.00	0.00
0.50	0.00	500	−0.03	−0.03	−0.03	−0.03	−0.02
0.50	0.00	1000	−0.02	−0.02	−0.02	−0.02	−0.01
0.50	0.00	3000	−0.01	−0.01	−0.01	−0.01	0.00
0.50	0.50	500	−0.03	−0.03	−0.03	−0.03	−0.03
0.50	0.50	1000	−0.01	−0.01	−0.01	−0.01	−0.01
0.50	0.50	3000	−0.01	−0.01	−0.01	−0.01	−0.00
0.50	0.80	500	−0.06	−0.02	−0.06	−0.05	−0.04
0.50	0.80	1000	−0.02	−0.01	−0.02	−0.02	−0.01
0.50	0.80	3000	−0.00	−0.00	−0.00	−0.00	0.00
1.00	0.00	500	−0.11	−0.11	−0.10	−0.10	−0.05
1.00	0.00	1000	−0.03	−0.03	−0.03	−0.03	0.00
1.00	0.00	3000	−0.01	−0.02	−0.01	−0.02	0.01
1.00	0.50	500	−0.08	−0.08	−0.08	−0.08	−0.05
1.00	0.50	1000	−0.03	−0.03	−0.03	−0.03	−0.01
1.00	0.50	3000	−0.01	−0.01	−0.01	−0.01	−0.00
1.00	0.80	500	−0.15	−0.13	−0.16	−0.15	−0.11
1.00	0.80	1000	−0.06	−0.05	−0.06	−0.05	−0.03
1.00	0.80	3000	−0.01	−0.01	−0.01	−0.01	0.00

[1]TCF is the test characteristic function method; ICF is the item characteristic function method. [2]Diff (SD) means difference between two groups in standard deviation unit.

Table C.4. Standard deviation of bias for $a1$ with test length of 60 items.

Mean Diff (SD)[2]	r	Sample size	Min	Direct	TCF[1]	ICF[1]	Concurrent
0.00	0.00	500	0.02	0.02	0.02	0.02	0.01
0.00	0.00	1000	0.01	0.01	0.01	0.01	0.01
0.00	0.00	3000	0.01	0.01	0.01	0.01	0.00
0.00	0.50	500	0.01	0.01	0.01	0.01	0.01
0.00	0.50	1000	0.01	0.01	0.01	0.01	0.01
0.00	0.50	3000	0.00	0.00	0.00	0.00	0.00
0.00	0.80	500	0.02	0.03	0.02	0.03	0.02
0.00	0.80	1000	0.01	0.02	0.02	0.02	0.01
0.00	0.80	3000	0.01	0.01	0.01	0.01	0.01
0.50	0.00	500	0.02	0.02	0.02	0.02	0.01
0.50	0.00	1000	0.01	0.01	0.01	0.01	0.01
0.50	0.00	3000	0.01	0.01	0.01	0.01	0.01
0.50	0.50	500	0.03	0.03	0.03	0.03	0.03
0.50	0.50	1000	0.01	0.01	0.01	0.01	0.01
0.50	0.50	3000	0.00	0.00	0.00	0.00	0.00
0.50	0.80	500	0.03	0.05	0.03	0.04	0.02
0.50	0.80	1000	0.01	0.02	0.01	0.02	0.01
0.50	0.80	3000	0.01	0.01	0.01	0.01	0.01
1.00	0.00	500	0.10	0.09	0.09	0.09	0.09
1.00	0.00	1000	0.02	0.02	0.02	0.02	0.01
1.00	0.00	3000	0.01	0.01	0.01	0.01	0.01
1.00	0.50	500	0.08	0.08	0.08	0.09	0.06
1.00	0.50	1000	0.01	0.02	0.01	0.02	0.01
1.00	0.50	3000	0.00	0.01	0.00	0.01	0.00
1.00	0.80	500	0.10	0.12	0.11	0.11	0.09
1.00	0.80	1000	0.06	0.07	0.06	0.07	0.03
1.00	0.80	3000	0.01	0.01	0.01	0.01	0.01

[1]TCF is the test characteristic function method; ICF is the item characteristic function method. [2]Diff (SD) means difference between two groups in standard deviation unit.

Table C.5. Mean bias for $a2$ with test length of 40 items.

Mean Diff (SD)[2]	r	Sample size	Min	Direct	TCF[1]	ICF[1]	Concurrent
0.00	0.00	500	−0.03	−0.03	−0.03	−0.02	−0.02
0.00	0.00	1000	−0.01	−0.01	−0.01	−0.01	−0.01
0.00	0.00	3000	−0.00	−0.00	−0.00	−0.00	−0.00
0.00	0.50	500	−0.03	−0.02	−0.03	−0.02	−0.02
0.00	0.50	1000	−0.01	−0.01	−0.01	−0.01	−0.01
0.00	0.50	3000	−0.00	−0.00	−0.00	−0.00	−0.00
0.00	0.80	500	−0.06	−0.03	−0.06	−0.05	−0.04
0.00	0.80	1000	−0.02	−0.01	−0.02	−0.02	−0.01
0.00	0.80	3000	0.00	0.01	0.01	0.01	0.01
0.50	0.00	500	−0.05	−0.05	−0.05	−0.05	−0.03
0.50	0.00	1000	−0.02	−0.02	−0.02	−0.02	−0.01
0.50	0.00	3000	−0.00	−0.01	−0.01	−0.01	0.01
0.50	0.50	500	−0.05	−0.04	−0.05	−0.05	−0.03
0.50	0.50	1000	−0.02	−0.02	−0.02	−0.02	−0.01
0.50	0.50	3000	−0.00	−0.00	−0.00	−0.00	0.00
0.50	0.80	500	−0.12	−0.07	−0.12	−0.12	−0.08
0.50	0.80	1000	−0.04	−0.03	−0.04	−0.04	−0.02
0.50	0.80	3000	−0.00	0.00	0.00	−0.00	0.01
1.00	0.00	500	−0.13	−0.13	−0.13	−0.13	−0.08
1.00	0.00	1000	−0.04	−0.04	−0.04	−0.04	0.00
1.00	0.00	3000	−0.01	−0.01	−0.01	−0.01	0.02
1.00	0.50	500	−0.13	−0.14	−0.14	−0.14	−0.08
1.00	0.50	1000	−0.03	−0.04	−0.03	−0.03	−0.01
1.00	0.50	3000	−0.01	−0.01	−0.01	−0.01	0.00
1.00	0.80	500	−0.21	−0.19	−0.23	−0.23	−0.16
1.00	0.80	1000	−0.09	−0.07	−0.09	−0.08	−0.06
1.00	0.80	3000	−0.02	−0.02	−0.02	−0.02	−0.00

[1]TCF is the test characteristic function method; ICF is the item characteristic function method. [2]Diff (SD) means difference between two groups in standard deviation unit.

Table C.6. Standard deviation of bias for $a2$ with test length of 40 items.

Mean Diff (SD)[2]	r	Sample size	Min	Direct	TCF[1]	ICF[1]	Concurrent
0.00	0.00	500	0.01	0.02	0.01	0.01	0.01
0.00	0.00	1000	0.01	0.01	0.01	0.01	0.01
0.00	0.00	3000	0.01	0.01	0.01	0.01	0.01
0.00	0.50	500	0.01	0.02	0.01	0.01	0.01
0.00	0.50	1000	0.01	0.01	0.01	0.01	0.01
0.00	0.50	3000	0.00	0.00	0.00	0.00	0.00
0.00	0.80	500	0.04	0.06	0.04	0.04	0.02
0.00	0.80	1000	0.01	0.02	0.02	0.02	0.01
0.00	0.80	3000	0.01	0.01	0.01	0.01	0.01
0.50	0.00	500	0.06	0.07	0.06	0.06	0.03
0.50	0.00	1000	0.01	0.01	0.01	0.01	0.01
0.50	0.00	3000	0.01	0.01	0.01	0.01	0.01
0.50	0.50	500	0.04	0.03	0.04	0.04	0.05
0.50	0.50	1000	0.01	0.01	0.01	0.01	0.01
0.50	0.50	3000	0.00	0.01	0.01	0.01	0.00
0.50	0.80	500	0.10	0.09	0.10	0.10	0.08
0.50	0.80	1000	0.02	0.03	0.02	0.02	0.02
0.50	0.80	3000	0.01	0.02	0.01	0.01	0.01
1.00	0.00	500	0.10	0.10	0.10	0.09	0.11
1.00	0.00	1000	0.04	0.04	0.04	0.04	0.02
1.00	0.00	3000	0.01	0.01	0.01	0.01	0.01
1.00	0.50	500	0.10	0.10	0.10	0.10	0.09
1.00	0.50	1000	0.01	0.01	0.01	0.02	0.01
1.00	0.50	3000	0.01	0.01	0.01	0.01	0.01
1.00	0.80	500	0.11	0.12	0.10	0.10	0.12
1.00	0.80	1000	0.05	0.05	0.05	0.05	0.07
1.00	0.80	3000	0.01	0.01	0.01	0.02	0.01

[1]TCF is the test characteristic function method; ICF is the item characteristic function method. [2]Diff (SD) means difference between two groups in standard deviation unit.

Table C.7. Mean bias for $a2$ with test length of 60 items.

Mean Diff (SD)[2]	r	Sample size	Min	Direct	TCF[1]	ICF[1]	Concurrent
0.00	0.00	500	−0.02	−0.02	−0.02	−0.02	−0.02
0.00	0.00	1000	−0.01	−0.01	−0.01	−0.01	−0.01
0.00	0.00	3000	−0.00	−0.00	−0.00	−0.00	−0.00
0.00	0.50	500	−0.02	−0.02	−0.02	−0.02	−0.02
0.00	0.50	1000	−0.01	−0.00	−0.01	−0.01	−0.00
0.00	0.50	3000	0.00	0.00	0.00	0.00	0.00
0.00	0.80	500	−0.04	−0.01	−0.04	−0.03	−0.04
0.00	0.80	1000	−0.02	0.00	−0.01	−0.01	−0.01
0.00	0.80	3000	0.01	0.01	0.01	0.01	0.01
0.50	0.00	500	−0.04	−0.04	−0.04	−0.04	−0.03
0.50	0.00	1000	−0.02	−0.02	−0.02	−0.02	−0.00
0.50	0.00	3000	−0.00	−0.00	−0.00	−0.00	0.01
0.50	0.50	500	−0.04	−0.04	−0.04	−0.04	−0.03
0.50	0.50	1000	−0.01	−0.01	−0.01	−0.01	−0.01
0.50	0.50	3000	−0.00	−0.00	−0.00	−0.00	0.00
0.50	0.80	500	−0.07	−0.04	−0.08	−0.07	−0.06
0.50	0.80	1000	−0.03	−0.01	−0.03	−0.02	−0.02
0.50	0.80	3000	0.00	0.00	−0.00	−0.00	0.00
1.00	0.00	500	−0.12	−0.12	−0.12	−0.12	−0.08
1.00	0.00	1000	−0.03	−0.03	−0.03	−0.03	0.00
1.00	0.00	3000	−0.01	−0.01	−0.01	−0.01	0.02
1.00	0.50	500	−0.10	−0.11	−0.11	−0.10	−0.08
1.00	0.50	1000	−0.03	−0.03	−0.03	−0.03	−0.02
1.00	0.50	3000	−0.01	−0.01	−0.01	−0.01	0.00
1.00	0.80	500	−0.19	−0.18	−0.20	−0.19	−0.16
1.00	0.80	1000	−0.07	−0.07	−0.08	−0.07	−0.05
1.00	0.80	3000	−0.01	−0.02	−0.01	−0.01	0.00

[1]TCF is the test characteristic function method; ICF is the item characteristic function method. [2]Diff (SD) means difference between two groups in standard deviation unit.

Table C.8. Standard deviation of bias for $a2$ with test length of 60 items.

Mean Diff (SD)	R	Sample size	Min	Direct	TCF	ICF	Concurrent
0.00	0.00	500	0.02	0.02	0.02	0.02	0.01
0.00	0.00	1000	0.01	0.01	0.01	0.01	0.01
0.00	0.00	3000	0.01	0.00	0.01	0.01	0.00
0.00	0.50	500	0.01	0.01	0.01	0.01	0.01
0.00	0.50	1000	0.01	0.01	0.01	0.01	0.01
0.00	0.50	3000	0.00	0.00	0.00	0.00	0.00
0.00	0.80	500	0.02	0.04	0.03	0.03	0.02
0.00	0.80	1000	0.01	0.02	0.02	0.02	0.01
0.00	0.80	3000	0.00	0.01	0.01	0.01	0.01
0.50	0.00	500	0.03	0.03	0.03	0.03	0.02
0.50	0.00	1000	0.01	0.01	0.01	0.01	0.01
0.50	0.00	3000	0.01	0.01	0.01	0.01	0.01
0.50	0.50	500	0.02	0.03	0.02	0.02	0.02
0.50	0.50	1000	0.01	0.01	0.01	0.01	0.01
0.50	0.50	3000	0.00	0.01	0.00	0.00	0.00
0.50	0.80	500	0.04	0.06	0.05	0.05	0.03
0.50	0.80	1000	0.01	0.03	0.01	0.02	0.01
0.50	0.80	3000	0.01	0.01	0.01	0.01	0.01
1.00	0.00	500	0.09	0.09	0.09	0.09	0.10
1.00	0.00	1000	0.02	0.02	0.02	0.02	0.01
1.00	0.00	3000	0.01	0.01	0.01	0.01	0.01
1.00	0.50	500	0.08	0.08	0.08	0.08	0.09
1.00	0.50	1000	0.02	0.02	0.02	0.02	0.02
1.00	0.50	3000	0.00	0.01	0.00	0.01	0.00
1.00	0.80	500	0.11	0.12	0.11	0.12	0.11
1.00	0.80	1000	0.07	0.06	0.06	0.07	0.05
1.00	0.80	3000	0.01	0.01	0.01	0.01	0.01

[1]TCF is the test characteristic function method; ICF is the item characteristic function method. [2]Diff (SD) means difference between two groups in standard deviation unit.

Table C.9. Mean bias for d with test length of 40 items.

Mean Diff (SD)[2]	r	Sample size	Min	Direct	TCF[1]	ICF[1]	Concurrent
0.00	0.00	500	0.00	−0.00	−0.00	−0.00	−0.01
0.00	0.00	1000	−0.01	−0.02	−0.02	−0.02	−0.02
0.00	0.00	3000	−0.03	−0.03	−0.03	−0.03	−0.03
0.00	0.50	500	0.00	−0.00	−0.00	−0.01	−0.01
0.00	0.50	1000	−0.01	−0.01	−0.01	−0.01	−0.02
0.00	0.50	3000	−0.02	−0.03	−0.03	−0.03	−0.03
0.00	0.80	500	−0.01	−0.00	−0.01	−0.02	−0.02
0.00	0.80	1000	−0.03	−0.04	−0.04	−0.05	−0.04
0.00	0.80	3000	−0.04	−0.04	−0.04	−0.05	−0.04
0.50	0.00	500	0.49	0.25	0.28	0.27	0.27
0.50	0.00	1000	0.45	0.24	0.24	0.24	0.25
0.50	0.00	3000	0.42	0.14	0.14	0.11	0.23
0.50	0.50	500	0.51	0.25	0.28	0.27	0.28
0.50	0.50	1000	0.45	0.23	0.24	0.22	0.24
0.50	0.50	3000	0.43	0.21	0.19	0.18	0.22
0.50	0.80	500	0.53	0.16	0.25	0.24	0.28
0.50	0.80	1000	0.43	0.13	0.18	0.18	0.22
0.50	0.80	3000	0.41	0.15	0.15	0.16	0.21
1.00	0.00	500	1.06	0.62	0.62	0.60	0.67
1.00	0.00	1000	0.89	0.54	0.54	0.50	0.53
1.00	0.00	3000	0.83	0.50	0.51	0.25	0.50
1.00	0.50	500	1.09	0.61	0.65	0.65	0.61
1.00	0.50	1000	0.89	0.50	0.52	0.48	0.51
1.00	0.50	3000	0.85	0.49	0.50	0.30	0.48
1.00	0.80	500	1.14	0.49	0.61	0.61	0.66
1.00	0.80	1000	0.87	0.38	0.49	0.42	0.49
1.00	0.80	3000	0.81	0.43	0.47	0.26	0.46

[1]TCF is the test characteristic function method; ICF is the item characteristic function method. [2]Diff (SD) means difference between two groups in standard deviation unit.

Table C.10. Standard deviation of bias for d with test length of 40 items.

Mean Diff (SD)[2]	r	Sample size	Min	Direct	TCF[1]	ICF[1]	Concurrent
0.00	0.00	500	0.04	0.04	0.03	0.03	0.03
0.00	0.00	1000	0.03	0.02	0.02	0.02	0.02
0.00	0.00	3000	0.01	0.01	0.01	0.01	0.01
0.00	0.50	500	0.05	0.04	0.04	0.04	0.04
0.00	0.50	1000	0.04	0.03	0.03	0.02	0.02
0.00	0.50	3000	0.02	0.02	0.02	0.02	0.01
0.00	0.80	500	0.05	0.06	0.05	0.06	0.04
0.00	0.80	1000	0.03	0.03	0.03	0.04	0.02
0.00	0.80	3000	0.02	0.02	0.02	0.02	0.02
0.50	0.00	500	0.08	0.11	0.06	0.07	0.04
0.50	0.00	1000	0.03	0.05	0.06	0.07	0.02
0.50	0.00	3000	0.02	0.11	0.13	0.14	0.01
0.50	0.50	500	0.08	0.09	0.07	0.08	0.12
0.50	0.50	1000	0.03	0.08	0.06	0.10	0.02
0.50	0.50	3000	0.02	0.08	0.11	0.12	0.02
0.50	0.80	500	0.13	0.19	0.18	0.18	0.12
0.50	0.80	1000	0.03	0.13	0.11	0.12	0.03
0.50	0.80	3000	0.02	0.11	0.12	0.11	0.02
1.00	0.00	500	0.20	0.17	0.17	0.20	0.22
1.00	0.00	1000	0.09	0.09	0.09	0.15	0.04
1.00	0.00	3000	0.02	0.06	0.02	0.28	0.02
1.00	0.50	500	0.22	0.19	0.16	0.16	0.17
1.00	0.50	1000	0.05	0.11	0.03	0.15	0.04
1.00	0.50	3000	0.03	0.06	0.02	0.27	0.02
1.00	0.80	500	0.29	0.42	0.22	0.23	0.24
1.00	0.80	1000	0.06	0.22	0.03	0.19	0.06
1.00	0.80	3000	0.02	0.12	0.02	0.27	0.02

[1]TCF is the test characteristic function method; ICF is the item characteristic function method. [2]Diff (SD) means difference between two groups in standard deviation unit.

Table C.11. Mean bias for d with test length of 60 items.

Mean Diff (SD)[2]	r	Sample size	Min	Direct	TCF[1]	ICF[1]	Concurrent
0.00	0.00	500	−0.01	−0.01	−0.01	−0.01	−0.02
0.00	0.00	1000	−0.02	−0.03	−0.03	−0.03	−0.03
0.00	0.00	3000	−0.04	−0.04	−0.04	−0.04	−0.04
0.00	0.50	500	−0.02	−0.01	−0.02	−0.02	−0.02
0.00	0.50	1000	−0.02	−0.03	−0.03	−0.03	−0.04
0.00	0.50	3000	−0.03	−0.04	−0.04	−0.04	−0.04
0.00	0.80	500	−0.01	−0.01	−0.01	−0.02	−0.02
0.00	0.80	1000	−0.03	−0.04	−0.05	−0.06	−0.04
0.00	0.80	3000	−0.04	−0.05	−0.05	−0.05	−0.05
0.50	0.00	500	0.45	0.25	0.25	0.25	0.27
0.50	0.00	1000	0.42	0.22	0.22	0.22	0.24
0.50	0.00	3000	0.40	0.11	0.10	0.04	0.23
0.50	0.50	500	0.46	0.22	0.23	0.21	0.26
0.50	0.50	1000	0.43	0.21	0.21	0.20	0.23
0.50	0.50	3000	0.41	0.16	0.16	0.14	0.22
0.50	0.80	500	0.45	0.12	0.20	0.17	0.24
0.50	0.80	1000	0.41	0.11	0.16	0.12	0.21
0.50	0.80	3000	0.38	0.11	0.13	0.10	0.19
1.00	0.00	500	1.03	0.58	0.59	0.59	0.65
1.00	0.00	1000	0.85	0.50	0.50	0.46	0.52
1.00	0.00	3000	0.80	0.46	0.49	0.23	0.50
1.00	0.50	500	0.99	0.51	0.56	0.51	0.58
1.00	0.50	1000	0.86	0.48	0.51	0.45	0.50
1.00	0.50	3000	0.81	0.46	0.48	0.14	0.47
1.00	0.80	500	1.03	0.46	0.58	0.53	0.60
1.00	0.80	1000	0.84	0.44	0.49	0.39	0.47
1.00	0.80	3000	0.77	0.42	0.45	0.20	0.44

[1]TCF is the test characteristic function method; ICF is the item characteristic function method. [2]Diff (SD) means difference between two groups in standard deviation unit.

Table C.12. Standard deviation of bias for d with test length of 60 items.

Mean Diff (SD)[2]	r	Sample size	Min	Direct	TCF[1]	ICF[1]	Concurrent
0.00	0.00	500	0.03	0.03	0.03	0.03	0.03
0.00	0.00	1000	0.03	0.02	0.02	0.02	0.02
0.00	0.00	3000	0.01	0.01	0.01	0.01	0.01
0.00	0.50	500	0.05	0.05	0.05	0.05	0.04
0.00	0.50	1000	0.03	0.03	0.03	0.03	0.02
0.00	0.50	3000	0.02	0.02	0.02	0.02	0.01
0.00	0.80	500	0.05	0.05	0.04	0.05	0.04
0.00	0.80	1000	0.03	0.04	0.04	0.04	0.03
0.00	0.80	3000	0.02	0.02	0.01	0.02	0.01
0.50	0.00	500	0.05	0.07	0.06	0.07	0.05
0.50	0.00	1000	0.02	0.06	0.07	0.07	0.02
0.50	0.00	3000	0.02	0.12	0.14	0.13	0.01
0.50	0.50	500	0.06	0.12	0.10	0.11	0.05
0.50	0.50	1000	0.04	0.07	0.09	0.10	0.03
0.50	0.50	3000	0.02	0.10	0.11	0.13	0.01
0.50	0.80	500	0.06	0.16	0.11	0.15	0.04
0.50	0.80	1000	0.03	0.13	0.13	0.15	0.03
0.50	0.80	3000	0.02	0.11	0.11	0.13	0.01
1.00	0.00	500	0.20	0.16	0.13	0.13	0.14
1.00	0.00	1000	0.04	0.03	0.03	0.15	0.03
1.00	0.00	3000	0.02	0.11	0.01	0.27	0.01
1.00	0.50	500	0.15	0.17	0.10	0.20	0.11
1.00	0.50	1000	0.04	0.11	0.03	0.17	0.03
1.00	0.50	3000	0.02	0.09	0.02	0.26	0.02
1.00	0.80	500	0.18	0.30	0.15	0.23	0.16
1.00	0.80	1000	0.08	0.15	0.07	0.24	0.04
1.00	0.80	3000	0.02	0.10	0.02	0.27	0.02

[1]TCF is the test characteristic function method; ICF is the item characteristic function method. [2]Diff (SD) means difference between two groups in standard deviation unit.

Appendix **D**

Appendix

D.1   Tables of means of correlation between estimates and generating parameters

Table D.1. Mean untransformed correlation between estimates and population parameter for $a1$ with test length of 40 items.

Mean Diff (SD)[2]	r	Sample size	Min	Direct	TCF[1]	ICF[1]	Concurrent
0.00	0.00	500	0.96	0.96	0.96	0.96	0.97
0.00	0.00	1000	0.98	0.98	0.98	0.98	0.98
0.00	0.00	3000	0.99	0.99	0.99	0.99	0.99
0.00	0.50	500	0.94	0.94	0.94	0.94	0.96
0.00	0.50	1000	0.97	0.97	0.97	0.97	0.98
0.00	0.50	3000	0.99	0.99	0.99	0.99	0.99
0.00	0.80	500	0.87	0.87	0.87	0.87	0.91
0.00	0.80	1000	0.94	0.94	0.94	0.94	0.95
0.00	0.80	3000	0.98	0.98	0.98	0.98	0.98
0.50	0.00	500	0.93	0.93	0.93	0.93	0.96
0.50	0.00	1000	0.97	0.97	0.97	0.97	0.98
0.50	0.00	3000	0.99	0.99	0.99	0.99	0.99
0.50	0.50	500	0.92	0.92	0.92	0.92	0.94
0.50	0.50	1000	0.96	0.96	0.96	0.96	0.97
0.50	0.50	3000	0.99	0.99	0.99	0.99	0.99
0.50	0.80	500	0.78	0.77	0.75	0.78	0.86
0.50	0.80	1000	0.91	0.91	0.91	0.91	0.93
0.50	0.80	3000	0.97	0.97	0.97	0.97	0.98
1.00	0.00	500	0.87	0.86	0.86	0.87	0.88
1.00	0.00	1000	0.94	0.94	0.94	0.94	0.96
1.00	0.00	3000	0.98	0.98	0.98	0.98	0.98
1.00	0.50	500	0.80	0.80	0.80	0.80	0.88
1.00	0.50	1000	0.94	0.94	0.94	0.94	0.96
1.00	0.50	3000	0.98	0.98	0.98	0.98	0.99
1.00	0.80	500	0.61	0.60	0.62	0.62	0.74
1.00	0.80	1000	0.85	0.85	0.85	0.85	0.89
1.00	0.80	3000	0.94	0.94	0.94	0.94	0.96

[1]TCF is the test characteristic function method; ICF is the item characteristic function method. [2]Diff (SD) means difference between two groups in standard deviation unit.

Table D.2. Mean untransformed correlation between estimates and population parameter for $a1$ with test length of 60 items.

Mean Diff (SD)[2]	r	Sample size	Min	Direct	TCF[1]	ICF[1]	Concurrent
0.00	0.00	500	0.96	0.96	0.96	0.96	0.97
0.00	0.00	1000	0.98	0.98	0.98	0.98	0.98
0.00	0.00	3000	0.99	0.99	0.99	0.99	0.99
0.00	0.50	500	0.95	0.95	0.95	0.95	0.96
0.00	0.50	1000	0.97	0.97	0.97	0.97	0.98
0.00	0.50	3000	0.99	0.99	0.99	0.99	0.99
0.00	0.80	500	0.89	0.89	0.89	0.89	0.91
0.00	0.80	1000	0.94	0.94	0.94	0.94	0.95
0.00	0.80	3000	0.98	0.98	0.98	0.98	0.98
0.50	0.00	500	0.95	0.95	0.95	0.95	0.96
0.50	0.00	1000	0.97	0.97	0.97	0.97	0.98
0.50	0.00	3000	0.99	0.99	0.99	0.99	0.99
0.50	0.50	500	0.93	0.93	0.93	0.93	0.94
0.50	0.50	1000	0.97	0.97	0.97	0.97	0.97
0.50	0.50	3000	0.99	0.99	0.99	0.99	0.99
0.50	0.80	500	0.84	0.83	0.84	0.84	0.88
0.50	0.80	1000	0.92	0.92	0.92	0.92	0.94
0.50	0.80	3000	0.97	0.97	0.97	0.97	0.98
1.00	0.00	500	0.84	0.84	0.84	0.84	0.87
1.00	0.00	1000	0.96	0.96	0.96	0.96	0.96
1.00	0.00	3000	0.98	0.98	0.98	0.98	0.98
1.00	0.50	500	0.84	0.83	0.84	0.84	0.89
1.00	0.50	1000	0.94	0.94	0.94	0.94	0.95
1.00	0.50	3000	0.98	0.98	0.98	0.98	0.98
1.00	0.80	500	0.66	0.66	0.67	0.67	0.75
1.00	0.80	1000	0.85	0.85	0.85	0.85	0.90
1.00	0.80	3000	0.95	0.95	0.95	0.95	0.96

[1]TCF is the test characteristic function method; ICF is the item characteristic function method. [2]Diff (SD) means difference between two groups in standard deviation unit.

Table D.3. Mean untransformed correlation between estimates and population parameter for $a2$ with test length of 40 items.

Mean Diff (SD)[2]	r	Sample size	Min	Direct	TCF[1]	ICF[1]	Concurrent
0.00	0.00	500	0.95	0.95	0.95	0.95	0.96
0.00	0.00	1000	0.97	0.97	0.97	0.97	0.98
0.00	0.00	3000	0.99	0.99	0.99	0.99	0.99
0.00	0.50	500	0.93	0.93	0.93	0.93	0.95
0.00	0.50	1000	0.97	0.97	0.97	0.97	0.97
0.00	0.50	3000	0.99	0.99	0.99	0.99	0.99
0.00	0.80	500	0.84	0.83	0.84	0.84	0.88
0.00	0.80	1000	0.91	0.91	0.91	0.91	0.93
0.00	0.80	3000	0.97	0.97	0.97	0.97	0.98
0.50	0.00	500	0.90	0.89	0.90	0.90	0.93
0.50	0.00	1000	0.96	0.96	0.96	0.96	0.97
0.50	0.00	3000	0.98	0.98	0.98	0.98	0.99
0.50	0.50	500	0.89	0.89	0.89	0.89	0.91
0.50	0.50	1000	0.95	0.95	0.95	0.95	0.96
0.50	0.50	3000	0.98	0.98	0.98	0.98	0.98
0.50	0.80	500	0.70	0.67	0.70	0.70	0.79
0.50	0.80	1000	0.87	0.86	0.87	0.87	0.91
0.50	0.80	3000	0.96	0.96	0.96	0.96	0.97
1.00	0.00	500	0.79	0.79	0.79	0.79	0.84
1.00	0.00	1000	0.92	0.92	0.92	0.92	0.94
1.00	0.00	3000	0.97	0.97	0.97	0.97	0.98
1.00	0.50	500	0.75	0.75	0.75	0.75	0.83
1.00	0.50	1000	0.91	0.91	0.91	0.91	0.94
1.00	0.50	3000	0.96	0.96	0.96	0.96	0.97
1.00	0.80	500	0.53	0.47	0.54	0.54	0.66
1.00	0.80	1000	0.79	0.78	0.79	0.79	0.84
1.00	0.80	3000	0.92	0.92	0.92	0.92	0.94

[1]TCF is the test characteristic function method; ICF is the item characteristic function method. [2]Diff (SD) means difference between two groups in standard deviation unit.

Table D.4.   Mean untransformed correlation between estimates and population parameter for $a2$ with test length of 60 items.

Mean Diff (SD)[2]	r	Sample size	Min	Direct	TCF[1]	ICF[1]	Concurrent
0.00	0.00	500	0.95	0.95	0.95	0.95	0.96
0.00	0.00	1000	0.97	0.97	0.97	0.97	0.98
0.00	0.00	3000	0.99	0.99	0.99	0.99	0.99
0.00	0.50	500	0.94	0.94	0.94	0.94	0.95
0.00	0.50	1000	0.97	0.97	0.97	0.97	0.97
0.00	0.50	3000	0.99	0.99	0.99	0.99	0.99
0.00	0.80	500	0.86	0.86	0.86	0.86	0.89
0.00	0.80	1000	0.93	0.93	0.93	0.93	0.94
0.00	0.80	3000	0.98	0.98	0.98	0.98	0.98
0.50	0.00	500	0.93	0.93	0.93	0.93	0.94
0.50	0.00	1000	0.96	0.96	0.96	0.96	0.97
0.50	0.00	3000	0.98	0.98	0.98	0.98	0.99
0.50	0.50	500	0.90	0.90	0.90	0.90	0.91
0.50	0.50	1000	0.95	0.95	0.95	0.95	0.96
0.50	0.50	3000	0.98	0.98	0.98	0.98	0.98
0.50	0.80	500	0.80	0.78	0.80	0.80	0.84
0.50	0.80	1000	0.90	0.90	0.90	0.90	0.92
0.50	0.80	3000	0.96	0.96	0.96	0.96	0.97
1.00	0.00	500	0.80	0.80	0.80	0.80	0.82
1.00	0.00	1000	0.94	0.94	0.94	0.94	0.94
1.00	0.00	3000	0.97	0.97	0.97	0.97	0.98
1.00	0.50	500	0.79	0.77	0.79	0.79	0.82
1.00	0.50	1000	0.92	0.92	0.92	0.92	0.93
1.00	0.50	3000	0.97	0.97	0.97	0.97	0.98
1.00	0.80	500	0.59	0.58	0.60	0.60	0.66
1.00	0.80	1000	0.81	0.81	0.81	0.81	0.85
1.00	0.80	3000	0.93	0.93	0.93	0.93	0.95

[1]TCF is the test characteristic function method; ICF is the item characteristic function method. [2]Diff (SD) means difference between two groups in standard deviation unit.

Table D.5. Mean untransformed correlation between estimates and population parameter for d with test length of 40 items.

Mean Diff (SD)[2]	r	Sample size	Min	Direct	TCF[1]	ICF[1]	Concurrent
0.00	0.00	500	0.97	0.97	0.97	0.97	0.97
0.00	0.00	1000	0.99	0.99	0.99	0.99	0.99
0.00	0.00	3000	0.99	0.99	0.99	0.99	1.00
0.00	0.50	500	0.97	0.97	0.97	0.97	0.98
0.00	0.50	1000	0.99	0.99	0.99	0.99	0.99
0.00	0.50	3000	0.99	0.99	0.99	0.99	1.00
0.00	0.80	500	0.97	0.97	0.96	0.96	0.98
0.00	0.80	1000	0.99	0.99	0.98	0.99	0.99
0.00	0.80	3000	0.99	0.99	0.99	0.99	1.00
0.50	0.00	500	0.89	0.92	0.92	0.92	0.95
0.50	0.00	1000	0.95	0.97	0.97	0.97	0.98
0.50	0.00	3000	0.97	0.99	0.99	0.99	0.99
0.50	0.50	500	0.90	0.94	0.93	0.93	0.94
0.50	0.50	1000	0.96	0.98	0.97	0.98	0.98
0.50	0.50	3000	0.98	0.99	0.99	0.99	0.99
0.50	0.80	500	0.86	0.90	0.88	0.88	0.92
0.50	0.80	1000	0.96	0.98	0.98	0.98	0.98
0.50	0.80	3000	0.98	0.99	0.99	0.99	0.99
1.00	0.00	500	0.72	0.78	0.78	0.78	0.81
1.00	0.00	1000	0.86	0.92	0.92	0.92	0.93
1.00	0.00	3000	0.92	0.96	0.96	0.97	0.97
1.00	0.50	500	0.68	0.75	0.74	0.74	0.83
1.00	0.50	1000	0.88	0.94	0.94	0.94	0.95
1.00	0.50	3000	0.92	0.96	0.96	0.97	0.97
1.00	0.80	500	0.67	0.74	0.73	0.73	0.78
1.00	0.80	1000	0.87	0.94	0.93	0.93	0.93
1.00	0.80	3000	0.92	0.97	0.97	0.98	0.97

[1]TCF is the test characteristic function method; ICF is the item characteristic function method. [2]Diff (SD) means difference between two groups in standard deviation unit.

Table D.6.   Mean untransformed correlation between estimates and population parameter for d with test length of 60 items.

Mean Diff (SD)[2]	r	Sample size	Min	Direct	TCF[1]	ICF[1]	Concurrent
0.00	0.00	500	0.97	0.97	0.97	0.97	0.98
0.00	0.00	1000	0.99	0.99	0.99	0.99	0.99
0.00	0.00	3000	0.99	0.99	0.99	0.99	1.00
0.00	0.50	500	0.97	0.97	0.97	0.97	0.98
0.00	0.50	1000	0.99	0.99	0.99	0.99	0.99
0.00	0.50	3000	0.99	0.99	0.99	0.99	1.00
0.00	0.80	500	0.97	0.97	0.97	0.97	0.97
0.00	0.80	1000	0.99	0.99	0.99	0.99	0.99
0.00	0.80	3000	1.00	1.00	1.00	1.00	1.00
0.50	0.00	500	0.93	0.95	0.95	0.95	0.95
0.50	0.00	1000	0.96	0.98	0.98	0.98	0.98
0.50	0.00	3000	0.98	0.99	0.99	0.99	0.99
0.50	0.50	500	0.92	0.94	0.94	0.94	0.94
0.50	0.50	1000	0.96	0.98	0.98	0.98	0.98
0.50	0.50	3000	0.98	0.99	0.99	0.99	0.99
0.50	0.80	500	0.92	0.94	0.94	0.94	0.96
0.50	0.80	1000	0.96	0.98	0.98	0.98	0.98
0.50	0.80	3000	0.98	0.99	0.99	0.99	0.99
1.00	0.00	500	0.72	0.77	0.77	0.77	0.79
1.00	0.00	1000	0.90	0.94	0.94	0.95	0.95
1.00	0.00	3000	0.94	0.97	0.97	0.98	0.97
1.00	0.50	500	0.77	0.83	0.82	0.82	0.84
1.00	0.50	1000	0.90	0.94	0.94	0.95	0.95
1.00	0.50	3000	0.94	0.97	0.97	0.98	0.98
1.00	0.80	500	0.71	0.77	0.76	0.76	0.79
1.00	0.80	1000	0.88	0.92	0.92	0.93	0.93
1.00	0.80	3000	0.94	0.98	0.97	0.98	0.98

[1]TCF is the test characteristic function method; ICF is the item characteristic function method. [2]Diff (SD) means difference between two groups in standard deviation unit.

Appendix   **E**

Appendicx

E.1   TESTFACT codes

E.1.1   **Concurrent calibration with 40 items, equivalent groups, r=0**

```
>TITLE
trial.TSF - TWO FACTOR ANALYSIS OF FORM xy with 60 items
                    .
>PROBLEM        NITEMS=60, RESPONSE=4, NOTPRES;
>COMMENTS
                Data.
>NAMES          ITEM00,
ITEM01,
ITEM02,
ITEM03,
ITEM04,
ITEM05,
ITEM06,
ITEM07,
ITEM08,
ITEM09,
ITEM10,
ITEM11,
ITEM12,
ITEM13,
ITEM14,
ITEM15,
ITEM16,
ITEM17,
ITEM18,
ITEM19,
ITEM20,
ITEM21,
ITEM22,
ITEM23,
ITEM24,
ITEM25,
ITEM26,
ITEM27,
ITEM28,
ITEM29,
ITEM30,
ITEM31,
ITEM32,
ITEM33,
ITEM34,
```

```
                 ITEM35,
                 ITEM36,
                 ITEM37,
                 ITEM38,
                 ITEM39,
                 ITEM40,
                 ITEM41,
                 ITEM42,
                 ITEM43,
                 ITEM44,
                 ITEM45,
                 ITEM46,
                 ITEM47,
                 ITEM48,
                 ITEM49,
                 ITEM50,
                 ITEM51,
                 ITEM52,
                 ITEM53,
                 ITEM54,
                 ITEM55,
                 ITEM56,
                 ITEM57,
                 ITEM58,
                 ITEM59;
                 >RESPONSE      ' ','0','1','9';
                 >KEY           11111111111111111111111111111111111111111111111111111111;
                 >TETRACHORIC   NDEC=3, LIST=100;
                 >FACTOR        NFAC=2, NROOT=3, ROTATE=VARIMAX;
                 >FULL          CYCLES=100,
                                CPARMS=(0.16391,
0.16496,
0.30673,
0.33949,
0.27745,
0.23524,
0.10611,
0.25955,
0.21558,
0.22214,
0.29469,
0.22486,
```

0.16185,
0.19772,
0.31149,
0.1705,
0.33842,
0.18315,
0.2066,
0.23594,
0.31831,
0.26934,
0.1997,
0.31912,
0.19825,
0.22094,
0.25254,
0.25341,
0.2183,
0.27818,
0.25736,
0.15542,
0.28935,
0.2317,
0.18,
0.22764,
0.2457,
0.18623,
0.26719,
0.19852,
0.21742,
0.28848,
0.09435,
0.11762,
0.20715,
0.38821,
0.42594,
0.08751,
0.23631,
0.23809,
0.26064,
0.24422,
0.20548,
0.27176,

```
0.24351,
0.20616,
0.14228,
0.24154,
0.26222,
0.34531);
>SCORE        CHANCE LIST=20;
>SAVE          PARM FSCORE ROTATE;
>INPUT         NIDCHAR=5, SCORES, FILE='res_con_c01_fxy060_s01_r001.txt';
(5A1,60A1)
>STOP
```

E.1.2 Separate calibration with 40 items, non-equivalent groups (.5SD), r=0.8

```
>TITLE
trial.TSF - TWO FACTOR ANALYSIS OF FORM y with 40 items
              .
>PROBLEM       NITEMS=40, RESPONSE=4, NOTPRES;
>COMMENTS
              Data.
>NAMES         ITEM00,
ITEM01,
ITEM02,
ITEM03,
ITEM04,
ITEM05,
ITEM06,
ITEM07,
ITEM08,
ITEM09,
ITEM10,
ITEM11,
ITEM12,
ITEM13,
ITEM14,
ITEM15,
ITEM16,
ITEM17,
ITEM18,
ITEM19,
ITEM20,
ITEM21,
ITEM22,
ITEM23,
ITEM24,
ITEM25,
ITEM26,
ITEM27,
ITEM28,
ITEM29,
ITEM30,
ITEM31,
ITEM32,
ITEM33,
ITEM34,
```

```
ITEM35,
ITEM36,
ITEM37,
ITEM38,
ITEM39;
>RESPONSE      ' ','0','1','9';
>KEY           1111111111111111111111111111111111111111;
>TETRACHORIC   NDEC=3, LIST=100;
>FACTOR        NFAC=2, NROOT=3, ROTATE=(PROMAX, 2, 4);
>FULL          CYCLES=100,
               CPARMS=(0.16391,
0.16496,
0.30673,
0.33949,
0.27745,
0.23524,
0.10611,
0.25955,
0.21558,
0.22214,
0.21742,
0.28848,
0.09435,
0.11762,
0.20715,
0.38821,
0.42594,
0.08751,
0.23631,
0.23809,
0.31831,
0.26934,
0.1997,
0.31912,
0.19825,
0.22094,
0.25254,
0.25341,
0.2183,
0.27818,
0.26064,
0.24422,
```

```
0.20548,
0.27176,
0.24351,
0.20616,
0.14228,
0.24154,
0.26222,
0.34531);
>SCORE        CHANCE LIST=20;
>SAVE          PARM FSCORE ROTATE;
>INPUT         NIDCHAR=5, SCORES, FILE='res_sep_c19_fy040_s01_r001.txt';
(5A1,40A1)
>STOP
```

Appendix   **F**

Appendix

F.1   R codes for equating

F.1.1   R code for equating

```
MAXTRIAL=20;
MAXSSE=.2;


# Implementation of the ICF and TCF equating methods
#
# OD:  Oshima's direct equating
# ICF: sum_{i} sum_{q1} sum_{q2} (P_i(q1,q2)-Pstar_i(q1,s2))^2
#      q1: -3, -2, ..., 2, 3;  q2: -3, -2, ..., 2, 3
# TCF: sum_{q1} sum_{q2} [ sum_{I}(P_i(q1,q2)) - sum_{i}(Pstar_i(q1,q2)) ]^2
#      q1, q2 as in case of ICF (based on Oshima 2000)
#


#source("R/integral.r");
source("R/bfgs.r");


#-------------------------------------------------------------------
# Model function
#
# The probability of correct response from a specific student
# (with theta ability vector) to an item (with parameters a, c, d).
irprob=function(c, a, d, theta){
   c+(1-c)*exp(a%*%theta+d)/(1+exp(a%*%theta+d));
}



#-------------------------------------------------------------------
# Objective functions
#
# NOTATION:
# values:  values for the variable to be optimized
#    values[1:4]: the rotation matrix
#                 A=matrix(values[1,4], nrow=2, ncol=2, byrow=F)
#    values[5:6]: the translation vector (m)
# c:        guessing paramter
# aBase:    a1 and a2 (slopes) for the base form
# dBase:    d (difficulty) for the base form
# aEq:      a1 and a2 (slopes) for the form to be brought to Base-scale
# dEq
# A:        rotation matrix
# m:        translation vector
```

```
####
# objective function of Test Characteristic Function
#
# thetas are sampled from (-3, 3) along each dimension for a total of 49
# points.  Based on Oshima 2000.
objectiveTCF=function(values, c, aBase, dBase, aEq, dEq){
   A=matrix(values[1:4], nrow=2, ncol=2, byrow=F);
   m=values[5:6];

   aEqStar=aEq%*%A;
   dEqStar=dEq+aEq%*%A%*%m;

   theta1s=seq(-3,3,by=1);
   theta2s=seq(-3,3,by=1);

   sumQ=0;
   for(theta1idx in 1:length(theta1s)){
      theta1=theta1s[theta1idx];
      for(theta2idx in 1:length(theta2s)){
         theta2=theta2s[theta2idx];
         irpBase=irprob(c, aBase, dBase, c(theta1, theta2));
         irpEqStar=irprob(c, aEqStar, dEqStar, c(theta1, theta2));
         sumQ=sumQ+((sum(irpEqStar)-sum(irpBase))^2);
      }
   }
   sumQ/49/length(c);
}

# objective function for Item Characteristic Function
#
# theta is sampled from (-3, 3), 7 samples along each dimension
# based on Oshima 2000
objectiveICF=function(values, c, aBase, dBase, aEq, dEq){
   A=matrix(values[1:4], nrow=2, ncol=2, byrow=F);
   B=values[5:6];

   aEqStar=aEq%*%A;
   dEqStar=dEq+aEq%*%A%*%B;

   theta1s=seq(-3,3,by=1);
   theta2s=seq(-3,3,by=1);
```

```
    sumQ=0;
    for(theta1idx in 1:length(theta1s)){
        theta1=theta1s[theta1idx];
        for(theta2idx in 1:length(theta2s)){
            theta2=theta2s[theta2idx];
            irpBase=irprob(c, aBase, dBase, c(theta1, theta2));
            irpEqStar=irprob(c, aEqStar, dEqStar, c(theta1, theta2));
            sumQ=sumQ+sum((irpEqStar-irpBase)^2);
        }
    }
    sumQ/(49*length(c));
}


####
# objective function for Oshima's direct equating
#
objectiveOD=function(values, c, aBase, dBase, aEq, dEq){
    A=matrix(values[1:4], nrow=2, ncol=2, byrow=F);
    m=values[5:6];

    # print(A); print(m);

    # aYstar=aEq%*%A;
    # dYstar=dEq+aEq%*%A%*%m;
    #
    # sum((aYstar-aBase)^2)+sum((dYstar-dBase)^2)

    sumQ=0;
    for(i in 1:length(c)){
        sumQ=sumQ + ((aBase[i,]-aEq[i,]%*%A)^2 %*%c(1,1));
        sumQ=sumQ + ((dBase[i]-dEq[i]-aBase[i,]%*%A%*%m)^2 );
    }
    sumQ;
}


####
# objective for translation in Min's method
#
# values=m (translation vector); rotation matrix A is given
objectiveMin=function(values, dBase, aEq, dEq, A){
    m=values[1:2];
    dEqStar=dBase+aEq%*%A%*%m;
```

```
    sum((dEqStar-dBase)^2)
}


#-----------------------------------------------------------------
# Non-optimization based equating
#
# Min's method
rotation_min=function(aBase, aEq){

    S=t(aEq)%*%aBase;

    SST=S%*%t(S);
    sst=svd(SST);

    STS=t(S)%*%S;
    sts=svd(STS);

    T=sst$u %*% t(sts$v);
    T
}

dilation_min=function(aBase, aEq, T, center=F){
    B=aBase; A=aEq;

    if(center==TRUE){
        cat(c("    Using centering for dilation in Min's method\n"));
        B=aBase-apply(aBase, 2, FUN=mean);
        A=aEq-apply(aEq, 2, FUN=mean);
    }

    K=diag(diag(t(B)%*%A%*%T))%*%solve(diag(diag(t(T)%*%t(A)%*%A%*%T)));
    # Estimate of B: A %*% T %*% K


    K
}



translation_min=function(dBase, aEq, dEq, T){

    # Initial estimate of m
    m=rep((sum(dBase)-sum(dEq))/(sum(T[,1])+sum(T[,2])),2)
```

```
   # Newton-Rapson
   m=minimizeNR(m, objectiveMin, dBase=dBase, aEq=aEq, dEq=dEq, A=T);

   m
}

equate_min=function(aBase, dBase, aEq, dEq, center=F){
   T=rotation_min(aBase, aEq);
   K=dilation_min(aBase, aEq, T, center);
   m=translation_min(dBase, aEq, dEq, T);
   result=list();
   result$T=T;
   result$K=K;
   result$m=m;
   result$method="min";
   result
}


#---------------------------------------------------------------------
# Equating functions
#
# base: matrix for the common items(!) in the base form
#        columns: 1: ID, 2: c (guess), 3: d (difficult) 4,5: a1,a2 (slopes)
# eq:   matrix for the common items for the form to be equated
# A:     initial rotation matrix
# m:     initial translation matrix
#

transform=function(eq, res){
   aEq=as.matrix(eq[,4:5]);
   dEq=eq[,3];
   c=eq[,2];
   ids=eq[,1];

   ystar=matrix(NA, ncol=5, nrow=length(c));
   ystar[,1]=ids;
   ystar[,2]=c;
   ystar[,3]=dEq+aEq%*%res$T%*%res$m;
   ystar[,4:5]=aEq%*%res$T%*%res$K;

   ystar
```

```
}

evaluate=function(base, eq, res){
   aBase=as.matrix(base[,4:5]);
   dBase=base[,3];
   c=base[,2];

   aEq=as.matrix(eq[,4:5]);
   dEq=eq[,3];

   aYStar=aEq%*%res$T%*%res$K;
   dYStar=dEq+aEq%*%res$T%*%res$m;

   ssea1=sum((aYStar[,1]-aBase[,1])^2);
   ssea2=sum((aYStar[,2]-aBase[,2])^2);
   ssed=sum((dYStar-dBase)^2);

   print(ssea1+ssea2+ssed);

   cat(c("----- Method=", res$method, " SSE=", ssea1+ssea2+ssed, "RMSE=",
   sqrt((ssea1+ssea2+ssed)/3*length(c)), "\n"));
   cat(c("RMSE(a1)=", sqrt(ssea1/length(c)), "RMSE(a2)=", sqrt(ssea2/length(c)),
   "RMSE(d)=", sqrt(ssed/length(c)), "\n"));
}

equate=function(base, eq, method, maxsse=MAXSSE, maxtrial=MAXTRIAL,
center=FALSE){

   cat(c("--------------------------------------------------- Method: ",
   method, "\n"));

   # Process the data matrices
   aBase=as.matrix(base[,4:5]);
   dBase=base[,3];
   c=base[,2];
   aEq=as.matrix(eq[,4:5]);
   dEq=eq[,3];

   # Methods other than Min's require reasonable initial values
   if(method != "min"){
      # Let the matrix of PC vectors for aBase be Rx ('prx' in the code)
      #                                      aEq be Ry   ('pry' in the code)
```

```
# Assuming that the PC vectors are similar for aBase and aEq,
# for any vector x in aBase and any vector y in aEq,
# x Rx = y Ry T
#
# Setting x and y to unit vectors (e.g. x=I (identity) )
# T = Ry^{-1} %*% Rx = Ry %*% Rx
#

prx=prcomp(aBase)$rotation;
pry=prcomp(aEq)$rotation;

# Problem:
# - the order of the PC vectors may not be right
#      aprx=abs(acos(prx));
# if(aprx[1,1]+aprx[2,2]>aprx[1,2]+aprx[2,1]){
   # cat("Swapping the principal vectors in x\n");
   # prx=cbind(prx[,2], prx[,1]);
# }
#      apry=abs(acos(pry));
# if(apry[1,1]+apry[2,2]>apry[1,2]+apry[2,1]){
   # cat("Swapping the principal vectors in y\n");
   # pry=cbind(pry[,2], pry[,1]);
# }

# - the signs of the columns of Rx and Ry are arbitrary
signMatrices=list();
signMatrices[[1]]=diag(c(1,1));   # No change to the signs
signMatrices[[2]]=diag(c(-1,1)); # Change the sign of the first PC vector
signMatrices[[3]]=diag(c(1,-1)); #                              second
signMatrices[[4]]=diag(c(-1,-1));#                              both PC vectors

bestT0=pry%*%prx;                      # The best PCA-based rotation matrix
bestsse0=sum((aBase-aEq%*%bestT0)^2); # and its SSE
for(first in 1:2){      # first=c(3,4) are duplicates
   for(second in 1:4){
      T0=pry%*%signMatrices[[first]]%*%prx%*%signMatrices[[second]];
      sse0=sum((aBase-aEq%*%T0)^2);
      if(sse0<bestsse0) { bestsse0=sse0; bestT0=T0; }
      #   print(T0);
      #   cat(c("----- sse0=", sse0, " bestsse0=", bestsse0, "\n"));
   }
}
```

```
    # Find the initial translation vector
    m0=rep((mean(dBase)-mean(dEq))/sum(c(1,1)%*%bestT0),2);
    cat(c("Initial rotation matrix based on PCA:\n"));
    print(bestT0);
    cat(c("SSE of the initial rotation matrix: ", bestsse0, "\n"));

    # Create a list of initial matrices
    # -- PCA-based rotation matrix
    # -- Min's orthogonal rotation matrix
    # -- Identity matrix
    # -- Matrix that swaps the columns of aEq
    V0s=list(); # List of initial matrices: 7 structured, some random
    V0s[[1]]=c(bestT0,m0);
    V0s[[2]]=c(bestT0,c(0,0));
    V0s[[3]]=c(rotation_min(aBase, aEq), m0);
    V0s[[4]]=c(rotation_min(aBase, aEq), c(0,0));
    V0s[[5]]=c(diag(c(1,1)), c(0,0)); # 1 0 0 1
    V0s[[6]]=c(diag(c(1,1)), c(1,1));
    V0s[[7]]=c(0, 1, 1, 0, 0, 0);  # Swaps the columns of aEq
    for(i in 8:maxtrial){
       V0s[[i]]=c(runif(6, -1, 1));
    }
}

# Do the equating
if(method!="min"){
   cat(c("Equating using ", method, " method\n"));
   bestresult=NULL; bestsse=10000000;
   for(trial in 1:maxtrial){
      cat(c("========= Trial ", trial, "\n"));
      values=c(V0s[[trial]]);
      print(values);

      func=NULL;
      if(method=="TCF"){ func=objectiveTCF; }
         else if(method=="ICF") { func=objectiveICF; }
            else if(method=="OD") { func=objectiveOD; }
      if(is.null(func)){
         cat(c("Invalid method specified ", method, ".
         Valid methods: Min, TCF, ICF, OD\n"));
         return;
```

```
    }

    # Try to optimize using Newton-Raphson
    values=minimizeNR(values, func, c=c, aBase=aBase,
    dBase=dBase, aEq=aEq, dEq=dEq);
    result=list();
    result$T=matrix(values[1:4], nrow=2, ncol=2, byrow=F);
    result$m=values[5:6];
    result$K=diag(c(1,1));
    result$method=method;

    aYStar=aEq%*%result$T;
    dYStar=dEq+aEq%*%result$T%*%result$m;

    sse=sum((aYStar-aBase)^2)+sum((dYStar-dBase)^2);
    if(sse<bestsse) {
        bestsse=sse;
        bestresult=result;
    }
    cat(c("======== SSE=", sse, "Best SSE=", bestsse, "\n"));

    # Try to optimize using BFGS
    values=minimizeBFGS(values, func, c=c, aBase=aBase,
    dBase=dBase, aEq=aEq, dEq=dEq);
    result=list();
    result$T=matrix(values[1:4], nrow=2, ncol=2, byrow=F);
    result$m=values[5:6];
    result$K=diag(c(1,1));
    result$method=method;

    aYStar=aEq%*%result$T;
    dYStar=dEq+aEq%*%result$T%*%result$m;

    sse=sum((aYStar-aBase)^2)+sum((dYStar-dBase)^2);
    if(sse<bestsse) {
        bestsse=sse;
        bestresult=result;
    }
    cat(c("======== SSE=", sse, "Best SSE=", bestsse, "\n"));

    if(bestsse<maxsse) { return(bestresult); }
}
```

```
    if(bestsse0<bestsse) {
        cat(c("WARNING: failed to improve on the initial estimates\n"));
        result0=list();
        result0$T=bestT0;
        result0$m=m0;
        result0$method="PCA";
        result0$K=diag(c(1,1));
        return(result0);
    }
    return(bestresult);
} else {
    cat("Equating using Min's method\n");
    return(equate_min(aBase, dBase, aEq, dEq, center));
}


}



#----------------------------------------------------------------------
# Example
#
#commonItems=c(1:10,31:40);
#baseFull=read.table("../data/
#EstimatedParameter/par_sep_c01_fx060_s01_r001.txt", skip=1);
#eqFull=read.table("../data/
#EstimatedParameter/par_sep_c01_fy060_s01_r001.txt", skip=1);
#base=baseFull[commonItems,c(2:6)];
#eq=eqFull[commonItems,c(2:6)];

#baseFull=read.table("../par_est_sep_010103.txt", header=T);
#eqFull=read.table("../par_est_sep_010203.txt", header=T);
#base=baseFull[commonItems,]
#eq=eqFull[commonItems,]
#aBase=base[,4:5];
#aEq=eq[,4:5];

#res.tcf=equate(base, eq, method="TCF", maxsse=.2, maxtrial=3)
#res.icf=equate(base, eq, method="ICF", maxsse=.2, maxtrial=5)
#res.od=equate(base, eq, method="OD", maxsse=.2, maxtrial=7);
#res.min=equate(base, eq, method="min");
#evaluate(base, eq, res.tcf);
```

```
#evaluate(base, eq, res.min);
#evaluate(base, eq, res.icf);
#evaluate(base, eq, res.od);
```

F.1.2   R code for Newton-Raphson and Broyden-Fletcher-Goldfarb-Shanno (BFGS) methods

```
DELTA=.00001;
EPS=.0001;
INFTY=1000000000; # Obj function value>INFTY indicates divergence


# Computes the first derivative of func 'fun' w.r.t. variable 'var1'
# using the centered difference formula
# The values of variable var1 is 'values[var1]'
derive1=function(values, var1, fun, ...){
   valuesp=values;
   valuesp[var1]=valuesp[var1]+DELTA;
   valuesm=values;
   valuesm[var1]=valuesm[var1]-DELTA;
   (fun(valuesp, ...)-fun(valuesm, ...))/(2*DELTA)
}


# Second order partial derivatives
derive2=function(values, var1, var2, fun, ...){
   valuesp=values;
   valuesp[var2]=valuesp[var2]+DELTA;
   valuesm=values;
   valuesm[var2]=valuesm[var2]-DELTA;
   (derive1(valuesp, var1, fun, ...)-derive1(valuesm, var1, fun, ...))/(2*DELTA)
}


# Minimizes using the BFGS algorithm.
# values   contains the initial values for the variables to be optimized
# func     the objective function to be optimized
# ...      additional parameters (not subject to optimization) to be
#          passed onto 'func' (e.g. constants, matrices, etc)
minimizeBFGS=function(values, func, ...){
   nvars=length(values);

      # Initial Hessian
   H=diag(rep(1,nvars));

   for(iter in 1:99){

      val=func(values,...);
      cat(c("BFGS Iteration ", iter, " value=", val,"\n"));
      #        print(c(values));
```

```
    f=rep(NA, nvars); # gardient of f(values)
    for(i in 1:nvars){ # First order derivatives
        f[i]=derive1(values, i, func, ...);
    }
#          print(f);

    # Stopping condition
    if(is.nan(val) || val > INFTY){
        cat(c("**** DIVERGENCE DETECTED. QUITTING.\n"));
        return(values);
    }
    done=1;
    for(i in 1:nvars){
        if(abs(f[i])>EPS) { done=0; break; }
    }
    if(done) { break; }

    # update
    ds=-solve(H)%*%f;
    # print(ds);
    values_next=values+ds;
    # print(values_next);

    f_next=rep(NA, nvars); # gradient of f(values_next)
    for(i in 1:nvars){ # First order derivatives for values_next
        f_next[i]=derive1(values_next, i, func, ...);
    }
#          print(f_next);

    y=f_next-f;
    # print(y);
    temp1=(t(y)%*%ds)[1][1]; # Transform 1x1 matrix into number
    temp2=(t(ds)%*%H%*%ds)[1][1];
    H=H+(y%*%t(y))/temp1-(H%*%ds%*%t(ds)%*%H)/temp2;
    # print(H);

    values=values_next;
    }
    values
}
```

```r
# Minimizes using the Newton-Raphson algorithm.
# values   contains the initial values for the variables to be optimized
# func     the objective function to be optimized
# ...      additional parameters (not subject to optimization) to be
#          passed onto 'func' (e.g. constants, matrices, etc)
minimizeNR=function(values, func, ...){
   nvars=length(values);

   for(iter in 1:99){

      val=func(values, ...);
      cat(c("NR iteration ", iter, " value=", val, "\n"));

      H=matrix(NA, nrow=nvars, ncol=nvars); # Hessian
      f=rep(NA, nvars);  # first-order derivative
      for(i in 1:nvars){
         f[i]=derive1(values, i, func, ...);
         for(j in 1:nvars){
            H[i, j]=derive2(values, i, j, func, ...);
         }
      }

      if(is.nan(val) || val>INFTY) {
         cat(c("**** DIVERGENCE DETECTED. QUITTING.\n"));
         return(values);
      }
      done=1;
      for(i in 1:nvars){
         if(abs(f[i])>EPS) { done=0; break; }
      }
      if(done) {
         cat(c("NR: done.  Returning: val=", func(values,...), "\n"));
         return(values);
      }

      ds=-solve(H)%*%f;
      values=values+ds*.1; # instead of ds, can use ds*.1
   }
   cat(c("**** NR: max iteration reached. val=", func(values,...), "\n"));
   values;
}
```

```
# Examle
#
# Consider function func1(x,y)=ax^2+by^2.  Minimze func1 in x and y
# for given a, b.
#
# func1=function(values, a, b) { # values = c(x, y)
#    a*values[1]^2+b*values[2]^2;
# }
#
# Minimize .5x^2+2.5y^2 (that is func1(x,y,a=.5, b=2.5)) in x and y.
# Let the initial guess be x=5 and y=1.  'values' is then c(5,1).
# vals=minimizeBFGS(c(5,1), func1, a=.5, b=2.5);
# cat(c("x=", vals[1], "y=", vals[2], "\n"));
#vals=minimizeBFGS(c(5,1), func1, a=.5, b=2.5);
#cat(c("x=", vals[1], "y=", vals[2], "\n"));
```

F.1.3   R code to convert item parameters of form Y onto the scale of form X

```r
args=list();
for(e in commandArgs()) {
    ta=strsplit(e, "=");
    cat(c(ta[[1]][1], "=", ta[[1]][2], "\n"));
    arg=ta[[1]][1];
    args[substr(arg,2,nchar(arg))]=ta[[1]][2]
}
#load equating code
source("/Users/mayukokanada/Research/R/equating.r");


baseFile=args$base;        # base parameter file
eqFile=args$eq;        # linked parameter file
nitem=as.numeric(args$nitem);
eqmethod=args$eqmethod;

X=read.table(baseFile, header=T); #base
Y=read.table(eqFile, header=T); #equted
z = matrix(X,nitem,5)

Ya1 = matrix(Y$a1,nitem,1);
Ya2 = matrix(Y$a2,nitem,1);
Ya=cbind(Ya1,Ya2);
dim(Ya);
Yd = matrix(Y$d,nitem,1);
# matrix only with common items;
uni1 =nitem/2 +1;
uni2 = nitem/2 + 10;
X.par.com= rbind(X[1:10,],X[uni1:uni2,])
Y.par.com= rbind(Y[1:10,],Y[uni1:uni2,])

if(eqmethod =="min"){
    res=equate(X.par.com, Y.par.com, method="min")
}
if(eqmethod =="od"){
    res=equate(X.par.com, Y.par.com, method="OD", maxsse=.2, maxtrial=10)
}
if(eqmethod =="tcf"){
    res=equate(X.par.com, Y.par.com, method="TCF", maxsse=.2, maxtrial=10)
}
if(eqmethod =="icf"){
```

```
    res=equate(X.par.com, Y.par.com, method="ICF", maxsse=.2, maxtrial=10)
}

res$T
res$K
res$m
res$method

# convert all items of Y using the equating parameters;
id = Y[,1];
c = Y[,2];
slope=Ya%*%res$T%*%res$K;
a1 = slope[,1];
a2 = slope[,2];
diff =Yd+Ya%*%res$T%*%res$m;
Ystar = cbind(id, c, diff, a1, a2);
head(Ystar)

write("item c d  a1 a2 ", file=args$ystar);
write(t(Ystar),file=args$ystar,append = T, ncolumns=5, sep = " ");
```