University of Alberta

INVESTIGATION OF IRT-BASED EQUATING METHODS IN THE PRESENCE OF OUTLIERS

by

HUIQIN HU

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of Doctor of Philosophy

in

Measurement, Evaluation, and Cognition

Department of Educational Psychology

Edmonton, Alberta

Fall 2004

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Library and Archives Canada

Published Heritage Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et Archives Canada

Direction du Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 0-612-95947-3
Our file Notre référence
ISBN: 0-612-95947-3

The author has granted a non-exclusive license allowing the Library and Archives Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou aturement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

Canada

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Acknowledgements

Many people helped me to complete this thesis. I would like to first thank the members of my thesis committee for their willingness to provide valuable and constructive suggestions and criticism. I am especially grateful to Dr. W. Todd Rogers, who guided me to the educational testing area, has been very patient while answering all the questions I had during my Ph. D. studies, provided generous help in improving my written English, and gave helpful comments on the organization and all the details of this thesis. I would like to extend my gratitude to Dr. Mark J. Gierl, who introduced me to item response theory, thus making it possible for me to conduct research using modern test theory. In addition, I would like to thank Dr. Mike D. Carbonaro, who introduced me to the world of computer programming. The basic programming knowledge I learned from him enabled me to complete this thesis and will have an influence on my future career.

Special thanks goes to Dr. Zarko Vukmirovic, who gave me a chance to apply what I have learned in school to educational testing practice and provided me with generous help and encouragement in pursuing and completing this dissertation topic. A special acknowledgement also goes to Dr. Judy C. Turner, whose encouragement, enthusiasm, and perseverance changed my life.

Finally, I own my husband and family a sincere thanks. Their patience and unconditional support throughout my Ph. D. studies have been greatly appreciated. They are the inspiration behind all that I do and that I hope to become.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Table of Contents

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

List of Tables

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

List of Figures

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

CHAPTER 1: INTRODUCTION

Background of the Problem

Students' achievement test scores are often used as a piece of information in making educational decisions, such as selecting students for a particular program and evaluating educational progress. The making of decisions in many contexts requires a test be administered on multiple occasions. For example, some large-scale achievement tests are given over years to track educational trends over time. For security reasons, different parallel forms of the test are often administered on different dates. However, the use of "parallel" forms leads to the concerns that the test forms may differ somewhat in difficulty or the two groups of examinees may differ in ability, thereby confounding the educational decisions to be made. For example, if two students each complete one of two parallel college entrance examinations on different dates, and the second student's raw score is two points higher than the first student's score, does the higher reported score reflect a higher achievement level or a less difficult test? In order to answer this question, different equating designs and procedures have been proposed to adjust for differences in difficulty among test forms that are built to be similar in difficulty and content (e.g., Angoff, 1971, 1984; Kolen & Brennan, 1995; Lord, 1980).

Kolen and Brennan (1995) described three basic equating designs: common-item non-equivalent groups design, single group design, and random groups design (e.g.,). In the common-item non-equivalent groups design, one test form with unique items is given to one group of examinees, an alternative form with another set of unique items is given to a second group, and an internal or external anchor test (common items) is given to both groups. In the single group design, all examinees are administered two test forms, often at

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

two different times and counter-balanced for order. In the random groups design, two test forms are administered typically by distributing the test forms alternatively at a test site. Among these designs, the common-item non-equivalent groups design is frequently used in many testing programs, because it has certain advantages over the others. For example, in the common-item non-equivalent groups design, only one test form needs to be administered at a given test date and each examinee takes only a single exam. No assumption is made about the equivalence of the groups on the latent trait being measured. Any differences of group ability or test difficulty can be identified and controlled by the common items. Consequently, this design will be the only design used for addressing the research questions in the current study.

Many equating methods have been proposed for the common-item non-equivalent groups design (e.g., Kolen & Brennan, 1995; Loyd & Hoover, 1980; Stocking & Lord, 1983). These methods can be categorized as classical equating methods, which typically refer to linear equating and equipercentile equating, and item response theory (IRT) based equating methods. Researchers (e.g., Han, Kolen, & Pohlmann, 1997; Hills, Subhiyah, & Hirsch, 1988; Kolen & Whitney, 1982; Lord & Wingersky, 1984; Modu, 1982; Yang & Houang, 1996) have compared the effectiveness of classical equating methods and IRT-based equating methods. Overall, most studies have shown that the two types of equating methods lead to comparable results. Kolen and Brennan (1995) presented a detailed list of the characteristics of equating situations for which each of the equating methods is most appropriate. However, in the case of equating large-scale achievement tests, no sound evidence has shown that the classical equating methods are more precise than the IRT equating methods or vice versa. For example, Hills et al.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

(1988) compared linear equating, Rasch model equating, three-parameter item response theory (3PL) - concurrent calibration method, 3PL - fixed common item parameters (FCIP) method, and 3PL – separate calibration with the linear transformation method. They found that all these equating methods yielded similar results.

However, Kolen and Brennan (1995) pointed out that since many large-scale testing programs use unidimensional IRT models to develop tests, the use of IRT-based equating methods often seems natural. Hambleton, Swaminathan, and Rogers (1991) claimed that IRT-based equating methods are particularly useful for test developers to manage test forms, maintain the security of tests, and compare students across test forms. Thus, IRT-based equating methods have become more and more attractive to large-scale testing practitioners. A plethora of studies have been conducted on applying IRT to test equating (e.g., Baker, 1992; Cook & Eignor, 1983; Haebara, 1980; Han et al., 1997; Kolen & Brennan, 1995; Lord, 1982; Loyd & Hoover, 1980; Stocking & Lord, 1983). However, there are still some questions, especially related to the assumptions of IRT, open to researchers and the practitioners in the field.

The basic assumptions of the commonly used IRT models are unidimensionality, local independence, and nonspeededness (Hambleton & Murray, 1983; Lord, 1980). Unidimensionality means that only one dominant latent trait accounts for examinees' performance on a test. Local independence follows the assumption of unidimensionality. It means that once the major latent trait influencing examinees' test performance is held constant, examinees' responses to any pair of items are statistically independent. In other words, the latent trait or ability is the only factor influencing examinees' responses to test items. The assumption of nonspeededness is based on the same belief that there is only

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

one latent trait. Speededness will not influence examinees' performance. When a given IRT model fits the test data of interest, several features are obtained: (1) examinee ability estimates are not test dependent; (2) item parameter estimates are not group dependent; (3) ability estimates obtained from different sets of items will be the same except for measurement error; and (4) item parameter estimates obtained in different groups of examinees will be the same except for measurement error.

It is important to choose an IRT model that fits the test data of interest while applying this model. Hambleton and Murray (1983) suggested three ways to check the model data fit: (1) evaluating the assumptions (for example, unidimensionality) of an IRT model with the given test data; (2) evaluating if the expected advantages derived from the use of an IRT model (for example, invariant item parameter and ability estimates) are obtained; and (3) evaluating the closeness of fit between predictions and observable outcomes (for example, test score distribution) using the parameter estimates and the test data.

It is also a crucial aspect of IRT applications to study the robustness of the models to violations of the assumptions and expected advantages (Kolen & Brennan, 1995). Several studies have been conducted to explore the effects of violating the assumptions (for example, dimensionality and local dependence) of IRT models on the equating results under a variety of conditions (e.g., Bogan & Yen, 1983; Bolt, 1999; De Champlain, 1996; Dorans & Kingston, 1985; Lee, Kolen, Frisbie, & Ankenmann, 2001; Modu, 1982; Skaggs & Lissitz, 1986a; Yen, 1984). However, only a few studies involved the issue of item parameter invariance that is specific to the common-item non-equivalent groups design (e.g., Bejar & Wingersky, 1981; Cook, Eignor, & Hutton, 1979; Linn,

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Levine, Hastings, & Wardrop, 1980; Stocking & Lord, 1983; Vukmirovic, Hu, & Turner, 2003).

In the common-item non-equivalent groups design, the IRT-based equating methods are typically applied following a three-step process (Kolen & Brennan, 1995). First, item parameters of the reference and equated tests are calibrated separately. Second, item parameter estimates from the equated test are scaled onto the scale of the parameter estimates for the reference test using a linear transformation method (also referred to as formula methods), which involves using methods such as mean/mean (Loyd & Hoover, 1980), mean/sigma (Marco, 1977), and characteristic curve (Haebara, 1980; Stocking & Lord, 1983) to estimate the transformation coefficients. Alternatively, the parameters of common items are held constant in the calibration of the equated test using the parameters estimated in the reference test (referred to as fixed common item parameters method, FCIP); consequently, the estimation of unique items is constrained by the scale of the common items (Hills et al., 1988; Li, Lissitz, & Yang, 1999; Zenisky, 2001; Vukmirovic et al., 2003). An alternative procedure for the first two steps is IRT concurrent calibration. Examinees' responses from the two tests to be equated are combined as one data file, and the parameters are estimated simultaneously from one computer run; thus, they are put onto one scale. In the third step, if number-correct scores are reported instead of the estimated theta, the number correct scores on the equated test are converted to the number-correct scores on the reference test using true score equating (Kolen, 1981; Lord, 1980) or observed score equating (Lord & Wingersky, 1984).

In the first step, one may expect that item parameters, such as the discrimination (*a*-parameter) and difficulty (*b*-parameter) parameters, of the common items are the

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

same, within sampling error, if they are estimated separately from two randomly equivalent groups; and they are different but have a linear relation if they are estimated from two non-equivalent groups. The guessing parameter (*c*-parameter), if it is specified in the model, will remain the same regardless of the form of the group equivalence (Hambleton & Murray, 1983).

In the case of common-item non-equivalent groups, as indicated, the *a*- and *b*-parameters of common items calibrated separately from two groups may be different. These differences are due to the indeterminacy of the estimation. Since both item parameters and ability parameters are not known in an IRT model, in order to conduct the estimation, the mean of the ability scale is often defined as 0 and the standard deviation as 1. That is, although the two groups may differ in ability, the abilities for each group are scaled to have a mean of 0 and a standard deviation of 1. As a result, the common item parameters of two non-equivalent groups are expected to have a linear relation and become the same once they are put on the same scale. If a scatter plot is made based on the *a*- or *b*- parameters estimated from two groups, all the points are expected to be located along a straight line (Hambleton & Murray, 1983) or within a narrow band around the least square fitting straight line. However, some points may be located outside the band. These "outliers" may be due to the estimation errors, curriculum changes, disclosure of common items, or sampling fluctuation. Regardless of the type of outlier, how should they be handled?

Researchers have realized that poorly estimated item difficulties may have a serious impact on the computation of sample moments, thus, producing a linear transformation that does not fit most of the estimated items' difficulties (Stocking &

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Lord, 1983). Various methods have been developed to reduce the influence of outliers when using the mean/mean and mean/sigma methods (e.g., Cook et al., 1979; Bejar & Wingersky, 1981). For example, Cook et al. (1979) proposed to restrict the range of the difficulties used in computing moments. Linn et al. (1980) tried to reduce the influence of outliers by using weighted moments where the weights are inversely proportional to the estimated standard error of the estimates of the item difficulties. Bejar and Wingersky (1981) proposed robust methods that give smaller weights to outliers. Stocking and Lord (1983) combined Linn et al.'s (1980) and Bejar et al.'s (1981) methods in an attempt to solve the outlier problem.

Outliers also adversely affect the FCIP procedure. Vukmirovic et al. (2003) compared the equating results between fixing and not fixing outlier common item parameters when estimating the equated test parameters using the FCIP. The difference between the two equating results produced by the two procedures increased as a function of the number of outliers and, especially, when the outliers were located on one side of the straight line in the scatter plot.

Questions that remain are that (1) do IRT-based equating methods that consider outliers produce a better result? (2) Which IRT-based equating method best reduces the influence of outliers? and (3) What conclusions drawn from existing comparison studies of IRT-based equating methods still hold in the presence of outliers?

Purpose of Study

The purpose of this study, therefore, was to investigate the comparability of IRT - separate calibration with mean/sigma transformation and test characteristic curve transformation methods, IRT - FCIP calibration method, and IRT - concurrent calibration

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

method in the presence of outliers with inconsistent *b*-parameter estimates. The group ability differences, number/score-point of outliers, and the types of outliers were manipulated. The IRT - concurrent calibration method and the IRT - separate calibration with test characteristic curve transformation method were conducted where the data to be analyzed included outliers and did not include outliers; the IRT – separate calibration with mean/sigma transformation method was conducted with data with no outliers, outliers, and outliers with weights; and the IRT - FCIP calibration method was conducted in which the outliers were fixed, not fixed, and removed. These ten variations of IRT-based methods were compared under a variety of conditions to answer the research questions: (1) Do the IRT-based equating methods that consider the influence of outliers produce a better result than the IRT-based equating methods that do not consider the influence of outliers? (2) Is the effect found in Question 1, if any, confounded by factors such as the characteristics of outliers and the group ability differences? (3) Which of the IRT-based equating methods produces a better result, especially among the IRT-based equating methods that consider the influence of outliers? (4) Is the effect found in Question 3, if any, confounded by factors such as the characteristics of outliers and the group ability differences?

Definition of Terms

Equating: A statistical procedure that is used to adjust the differences in difficulty among test forms that are built to be similar in difficulty and content. The scores on different test forms can be interpreted interchangeable after equating. The following properties need to be satisfied for equating two or more test forms:

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Symmetry Property: the function used to transform a score on Form X to the Form Y scale is the inverse of the function used to transform a score on Form Y to the Form X scale.

Same Specification Property: the test forms must be built to the same content and statistical specifications.

Equity Properties: the examinees with a given true score have identical observed score means, standard deviations, and distribution shapes of converted scores on Form X and scores on Form Y.

Equal Observed Score Distributions Property: the characteristics of score distributions are set equal for a specified population of examinees.

Group Invariance Property: the equating relationship is the same regardless of the groups of examinees used to conduct the equating.

Common-item Non-equivalent Groups Design: Test form X that contains a set of unique items and Form Y that contains another set of unique but randomly parallel items are administered to two non-equivalent groups on different occasions. Further, Form Z that contains items representing the content and statistical characteristics of Forms X and Y is administered to both groups. Thus, scores on Form X are equated onto the scores on Form Y through the common items in Form Z.

Reference and Equated Tests: Reference test is also referred to as base test to which another form is to be equated. The latter is referred to as the equated form. In the present study, Form Y is the reference test and Form X is the equated test.

Anchor Test: In the present study, Form Z that contains a set of common items is called the anchor test.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Internal Anchor Test: When the scores on the set of common items contribute to the examinees' scores on the test, the set of common items is referred to as the internal anchor test.

External Anchor Test: When the scores on the set of common items do not contribute to the examinees' scores on the test form, the set of common items is referred to as the external anchor test. Typically, the external anchor test is administered as a separately timed section.

Outliers: When the *b*-parameters of common items are estimated from the data sets of two non-equivalent groups, the two sets of *b*-parameters are supposed to be distributed along a least squares fit straight line and have a linear relationship. In other words, if two perpendicular straight lines are drawn from this item's X-axis and Y-axis position, the intersection point of these two perpendicular lines is supposed to be on the least squares fit line. However, if the intersection point is not on the least squares fit line, and the distance between the intersection point and its presumed position is equal to or more than two score points, which is based on the scale with a mean of 0 and standard deviation of 1, then this item is defined as an outlier. For example, in the left panel of Figure 1, there are no outliers. In contrast, in the right panel, Item i1 is an outlier.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

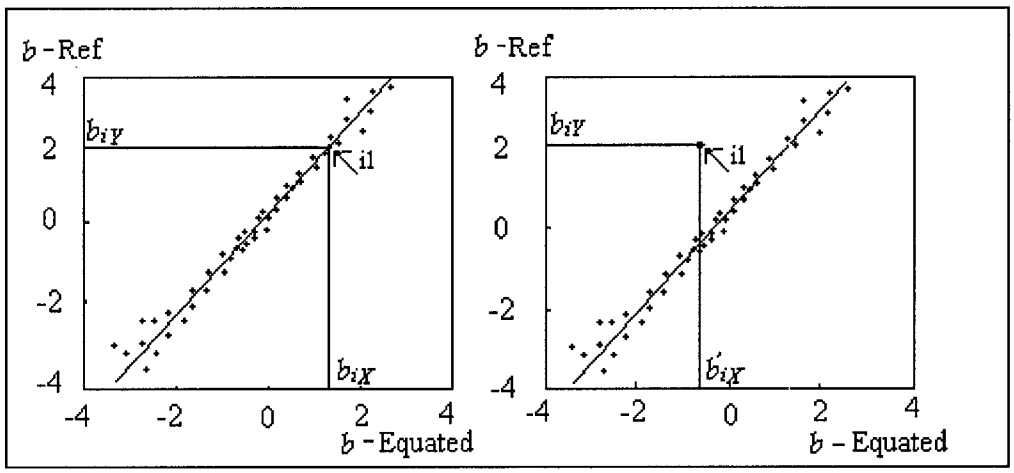

Figure 1. Illustration of the definition of outliers.

Organization of the Dissertation

Chapter 2 contains a review of the IRT models for dichotomous and polytomous items, the common-item non-equivalent groups equating design, IRT-based equating methods, criteria to evaluate IRT-based equating methods, comparison studies on the IRT-based equating methods, and studies on outliers. In Chapter 3, the equating design, controlled and manipulated conditions, computer simulation, and evaluation of the IRT-based equating methods are presented. This is followed by the presentation and discussions of results in Chapter 4. The summary of the results and methods, limitations of the current study, conclusions, implications for future practice, and recommendations for future research are presented in Chapter 5.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

CHAPTER 2: LITERATURE REVIEW

The equating of two parallel tests has been described in various ways using classical test theory (CTT) since the early 1950s (e.g., Flannangan, 1951; Lord, 1950, 1955). Typically, these traditional equating methods are conducted based on examinees' total scores. However, the total score may not reflect the latent trait (or ability) precisely because it is confounded with factors such as item difficulty, item discrimination, and/or guessing. For example, if two examinees get the same number-correct scores, it may not be that the two examinees have the same ability level because one examinee may get most of the difficult items right and the other may answer only the easier items correctly. In the framework of IRT, more factors are considered. More specifically, the probability (P) for an examinee to answer one question correctly not only depends on the examinee's ability (θ) but also on the item difficulty (b), item discrimination (a), and /or guessing chance (c). Many IRT models have described the mathematical function between P and θ, b, a, and/or c. Along with the development of IRT models, parameter estimation methods, and corresponding computer programs, IRT-based equating methods are now being used in many large-scale testing programs.

IRT models for dichotomously and polytomously scored items are described in the first section of this chapter. Then, equating designs, IRT-based equating methods, and the criteria to evaluate IRT-based equating methods will be reviewed. This is followed by a review of comparison studies of IRT-based equating methods. Finally, the studies on the influence of outliers will be presented.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

IRT Models

IRT consists of a family of probabilistic models that hypothesize the relationship between an examinee's latent trait and a correct response to an item. The basic belief of these models is that the factors that determine an examinee's performance on an item could be the characteristics of the item including item difficulty, item discrimination, and guessing chance; a latent trait; and/or a number of latent traits. More complicated IRT models have been developed to capture the factors that may influence examinees' performance on a specific test (Van der Linden, & Hambleton, 1997). For example, Reckase (1985) proposed a compensatory multidimensional IRT model that describes the non-linear logistical regression relation between the probability of a correct response of a person to a specific item and the individual's multidimensional latent traits. The latent traits in this model are supposed to be additive, which means that being high on one or more traits can compensate for being low on another trait. Although some equating procedures have been developed for multidimensional IRT (Li & Lissitz, 2000), one of the difficulties of applying these procedures to large-scale achievement tests is that there is no sound evidence that shows that the underlying latent traits of achievement tests have the relations indicated in the existing models. Further, providing meaning for each of the latent traits is difficult (Li & Lissitz, 2000). Cognitive psychologists have proposed IRT models (e.g., Emberson, 1984, 1985; Whitely, 1980) based on the cognitive study of a latent construct of interest. However, the application of these models to achievement tests is less than ideal due to the construct complexity of achievement in different subject areas.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Compared to the multidimensional or cognitive IRT models, unidimensional IRT models have been and continue to be widely used in the large-scale testing programs. Although the assumption of unidimensionality has been criticized, factor analysis has revealed that large-scale achievement tests usually have one major factor that is large relative to the remaining factors. Consequently, good ability and item parameter estimates can be obtained using unidimensional IRT models (Reckase, 1979). Hence, the following will focus on the review of the unidimensional IRT models.

IRT Models for Dichotomously Scored Items

The 1PL, 2PL, and 3PL IRT models are commonly used for dichotomously scored items. In the 1PL model, which is also called the Rasch model (Rasch, 1960), the probability of correct response, P_{ij}, of individual i on item j depends only on the ability of person i, θ_i, and the difficulty of item j, b_j. This relationship is expressed by:

$$P_{ij}(\theta_i) = \frac{\exp(\theta_i - b_j)}{1 + \exp(\theta_i - b_j)} , \qquad (1)$$

where "exp" is the natural logarithm exponent.

In the 2PL model, the probability of answering a question correctly is not only related to an examinee's ability and the item difficulty, but also to the item discrimination, a_j. In this case, the logistic function equation is (Lord, 1980):

$$P_{ij}(\theta_i) = \frac{\exp(D a_j(\theta_i - b_j))}{1 + \exp(D a_j(\theta_i - b_j))} , \qquad (2)$$

where D is the constant 1.7, which is used to make the logistic model similar to the normal ogive model (Hambleton et al., 1991). Comparing Equations 1 and 2, it can be seen that the 1PL model is a special case of the 2PL model, where "Da_i" equals one. This

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

in turn means that, in the case of 1PL model, the items are assumed to be equally discriminating.

In the 3PL model, the pseudo-guessing parameter or lower asymptote, c_j, is allowed to be different for each item. However, in the 1PL and 2PL models, no guessing is assumed. In other words, the 1PL and 2PL models are special cases of the 3PL model when the pseudo-guessing parameter equals 0. In the 3PL model, the probability of a correct response to the item j for the examinee i with ability θ_i is given by (Lord, 1980):

$$P_{ij}(\theta_i) = c_j + (1 - c_j)\frac{\exp(D\,a_j(\theta_i - b_j))}{1 + \exp(D\,a_j(\theta_i - b_j))}.$$

An item characteristic curve (ICC) is often plotted to show the relation between the probability of correct response to an item and the ability of an examinee. Figures 2 to 4 illustrate the ICCs of two items for the above models, respectively. The X-axis represents the "true" ability or the latent trait. It cannot be directly observed and has to be estimated based on the observed item responses of an examinee. Theoretically, the range of the ability parameter is from -∞ to +∞; however, in practice, the ability parameter is often located within a limited range, say -4 to +4. The Y-axis is the probability of a correct response to an item. It is assumed that this probability increases as ability increases.

The pseudo-guessing parameter, c_j, is "the probability that a person completely lacking in ability ($\theta = -\infty$) will answer the item correctly" (Lord, 1980, p.12). It is also called the pseudo-chance score level. When the items cannot be answered correctly by guessing, the c-parameter equals zero. In this case, it is appropriate to use the 1PL model or the 2PL model. As shown in Figures 2 to 4, the ICCs in the 1PL and 2PL models start from the points that correspond to probabilities (Y-axis) equal zero. The ICCs in the 3PL

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

model start from the points along the Y-axis at 0.12 and 0.20, which represent the probabilities of low ability examinees correctly answering the items.

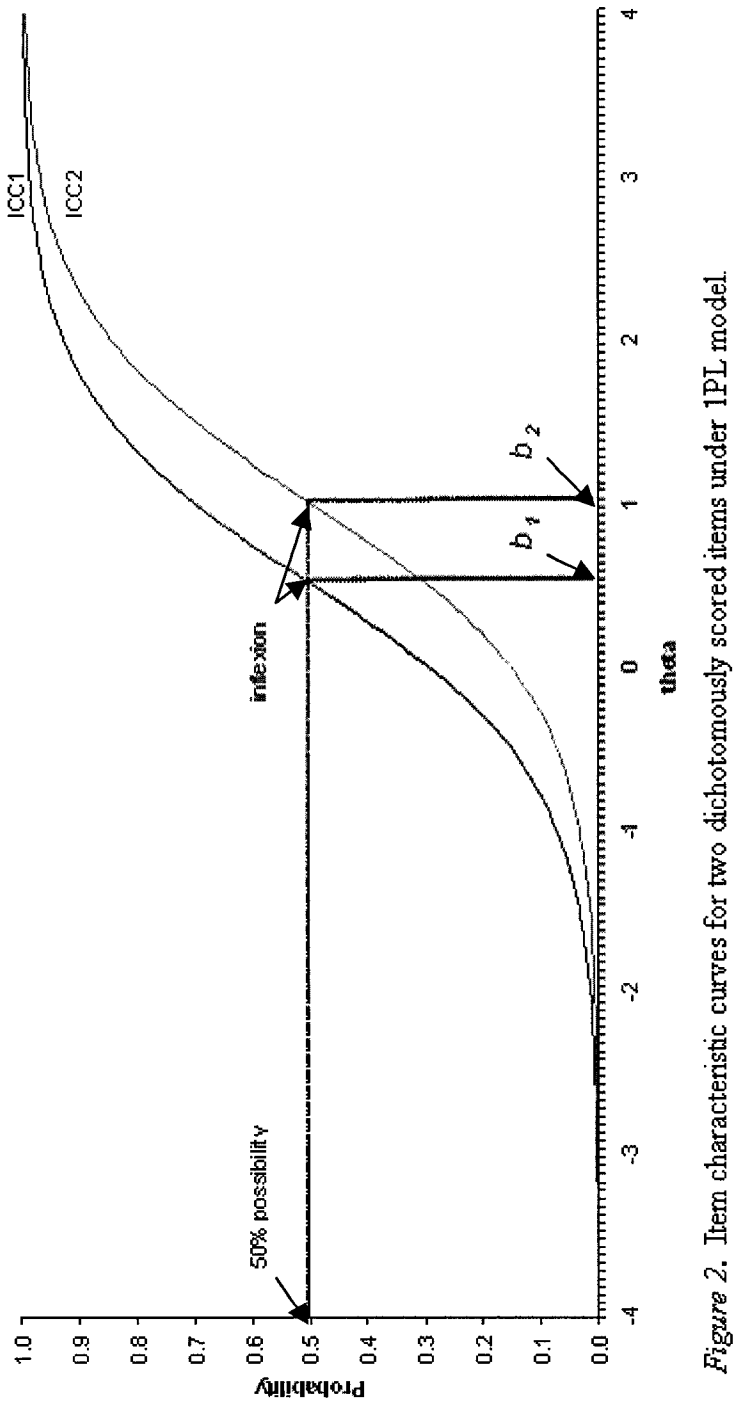

Figure 2. Item characteristic curves for two dichotomously scored items under 1PL model.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

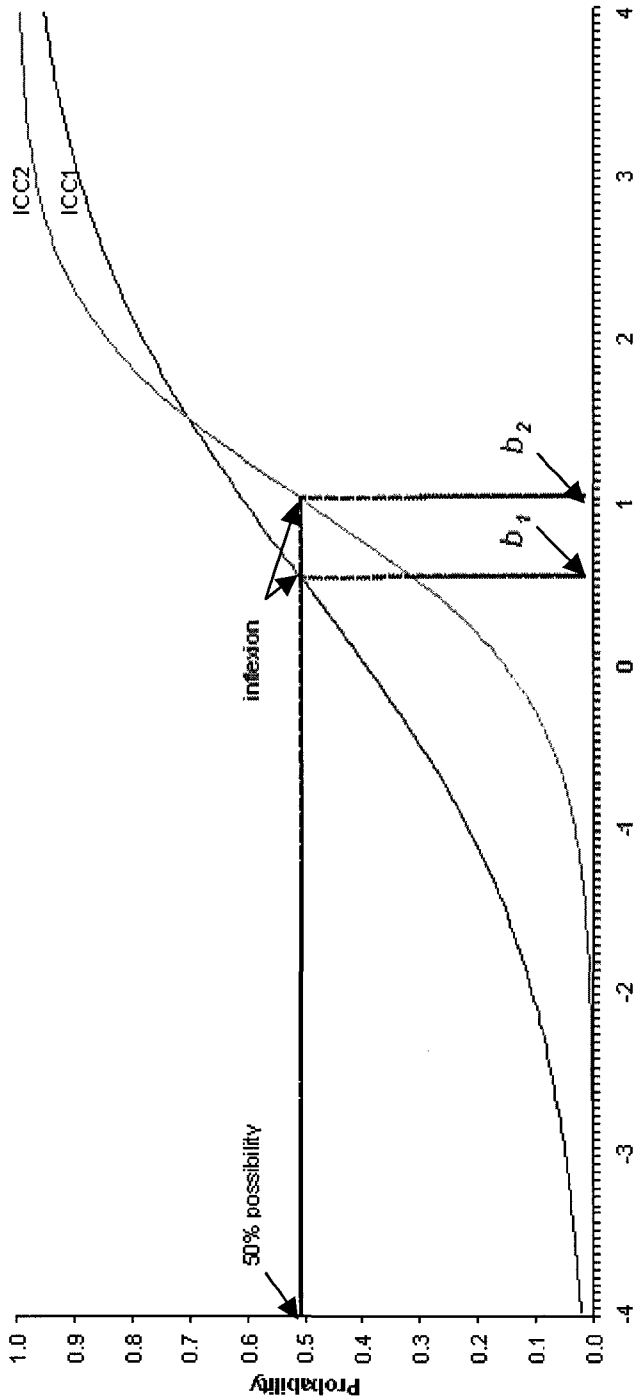

Figure 3. Item characteristic curves for two dichotomously scored items under 2PL model.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

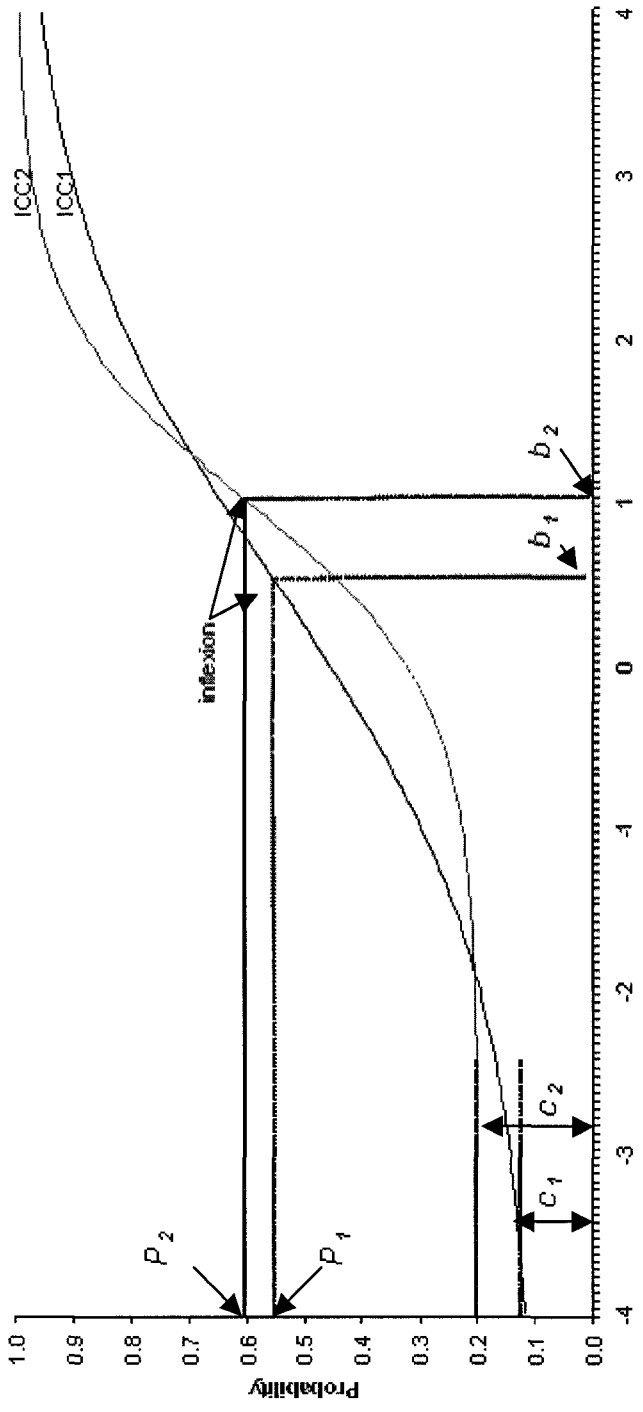

Figure 4. Item characteristic curves for two dichotomously scored items under 3PL model.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

The difficulty parameter, b_j, is the location parameter. It is located along the ability scale at the point at which the slope of the ICC is a maximum. The *b*-parameters determine the position of the curve along the ability scale. The more difficult the item, the further the curve is to the right. For example, in Figures 2 to 4, Item 2 is more difficult than Item 1. Large positive *b*-parameters indicate difficult items while large negative *b*-parameters represent easier items. In the 1PL and 2PL models, b_j is the ability level where the probability of a correct answer is 0.50; however, in the 3PL model, b_j is the ability level where the probability of a correct answer is halfway between c_j and 1.00 (Hambleton, et. al., 1991; Lord, 1980). As shown in Figure 4, the probabilities of a correct response to Item 1 and to Item 2, p_1 and p_2, are greater than 0.50 since c_1 and c_2 are greater than 0.

The discrimination parameter, a_j, indicates how well an item distinguishes between high and low ability examinees. It is proportional to the slope of the ICC at the inflexion point, i.e., at $\theta = b$ (Lord, 1980). The actual slope at $\theta = b$ is $0.425*a*(1-c)$. The steeper the slope, the higher the *a*-parameter, and the better the item discrimination power. For example, the ICCs in Figures 3 and 4 indicate that Item 2 is more discriminating than Item 1. However, under the 1PL model, as shown in Figure 2, all the items have the same discrimination. It means that the slopes of all the ICCs are similar except that they have different locations.

IRT Models for Polytomously Scored Items

A variety of IRT models are available for polytomously scored items. For example, Bock (1972) proposed the nominal response model for multi-category items. Masters (1982) developed the partial credit model (PCM) with fixed slope parameter.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Muraki (1992, 1993) developed the generalized partial credit model (GPCM) in which the slope parameter is allowed to vary across items. Muraki and Bock (1999) further extended the GPCM and Samejima's graded response model (GRM) to model items with rating scales. According to Thissen and Sternberg (1986), these models can be classified as either "difference" models or "divided-by-total" models. The best known models from the first category are the graded response model (Samejima, 1969) and its variations, and the best known models from the second category are the partial credit model (Masters, 1982) and its extensions.

Difference Models

In Samejima's (1969) GRM, person i's response to item j is categorized into one of $m_j + 1$ ordered categories. Associated with each category k of item j is a category score k, which equals 0, 1... m_j. The probability of obtaining a score k is the difference between the probability of obtaining a score of k or higher and the probability of obtaining a score of k+1 or higher, which can be written as:

$$P_{jk}(\theta_i) = P^*_{jk}(\theta_i) - P^*_{j,k+1}(\theta_i) \quad , \tag{3}$$

where

$$P^*_{jk}(\theta_i) = \frac{\exp(D\,a_j(\theta_i - b_{jk}))}{1 + \exp(D\,a_j(\theta_i - b_{jk}))} \quad , \tag{4}$$

θ_i is the latent trait,

D is the scaling constant (D = 1.7 to scale the logistic to the normal ogive metric; D = 1 to preserve the logistic metric),

a_j is the item discrimination power or common slope parameter for all response options for item j, and

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

b_{jk} is the threshold parameter for score k on item j.

Two constraints are made to Equation 3: $P^*_{j0}(\theta_i) = 1$ and $P^*_{j,m+1}(\theta_i) = 0$. That is, the probability of obtaining any score that is equal to or greater than 0 is 1, and the probability of obtaining any score that is greater than the highest score is 0. When the item is scored dichotomously, Equation 3 is simplified as Equation 2, which represents a 2PL model. In other words, the 2PL model is a special case of the GRM when an item has only two possible scores.

If the relation between $P^*_{jk}(\theta_i)$ and theta is plotted, a set of S-shaped category response curves (CRC) will be formed; and if the relation between $P_{jk}(\theta_i)$ and theta is plotted, the operating characteristic curve (OCC) will be formed. Figures 5 and 6 illustrate these curves for two items with two sets of a- and b-parameters. In the GRM, the a-parameter, a_j, is the common slope parameter for an item. It represents the steepness of the curves. The bigger the a-parameter, the steeper the curves. For example, in Figures 5 and 6, when the a-parameter decreases from 2 to 1, both the CRCs and OCCs become flatter.

b_{jk} is the category threshold parameter. It is the ability level where the probability of a correct answer to a category and its above categories is 0.50. When the b-parameter increases, the corresponding CRC(s) and OCC(s) shift to the right along the ability scale. For example, in Figures 5 and 6, when b_{j2} changes from 0 to 0.50, CRC2, OCC1, and OCC2 shift to the right. One may also notice that OCC2 becomes flatter. It indicates that the slopes of the middle OCCs are affected not only by the slope parameter, a_j, but also by the distance between adjacent categories. However, the slopes of the two extreme OCCs are affected only by the slope parameter since they are essentially the respective

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

cumulative probability functions. This phenomenon indicates that the slope parameter for the polytomous item response model is no longer the synonymy of discriminating power as it is for the dichotomous item response model.

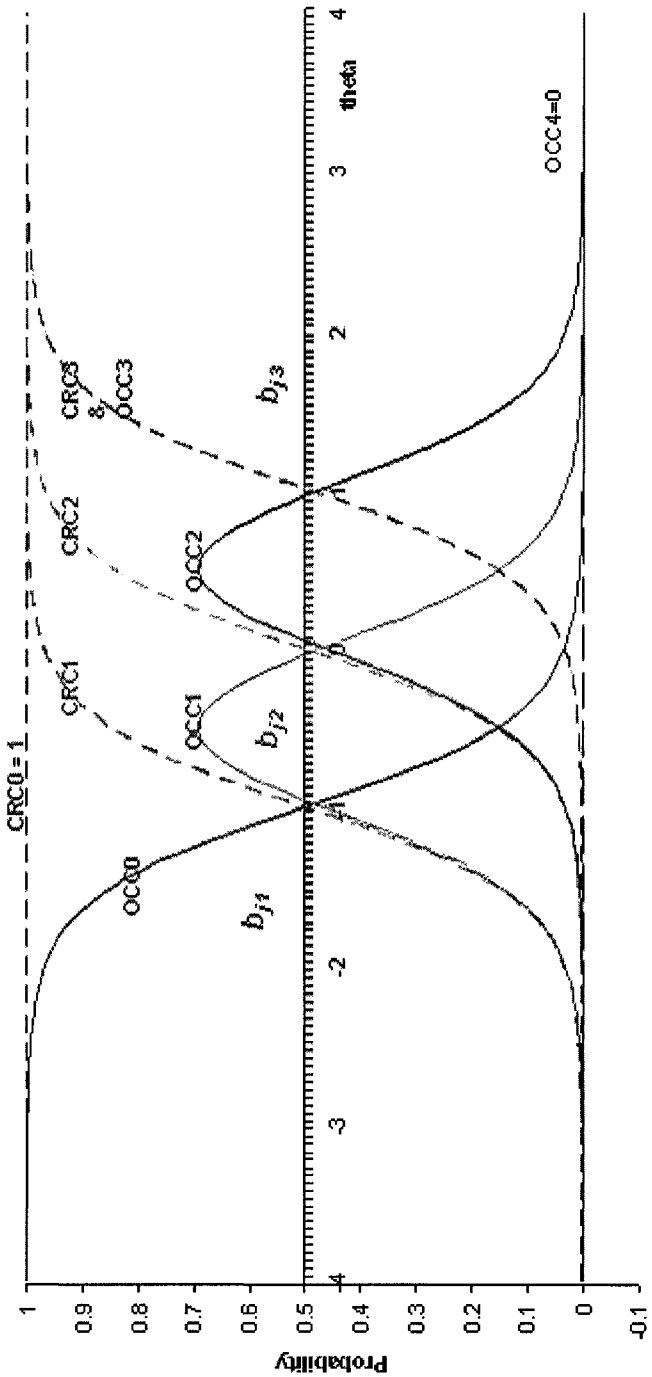

Figure 5. Category response and operating characteristic curves for a polytomously scored item under GRM with a =2, b =(-1,0,1).

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

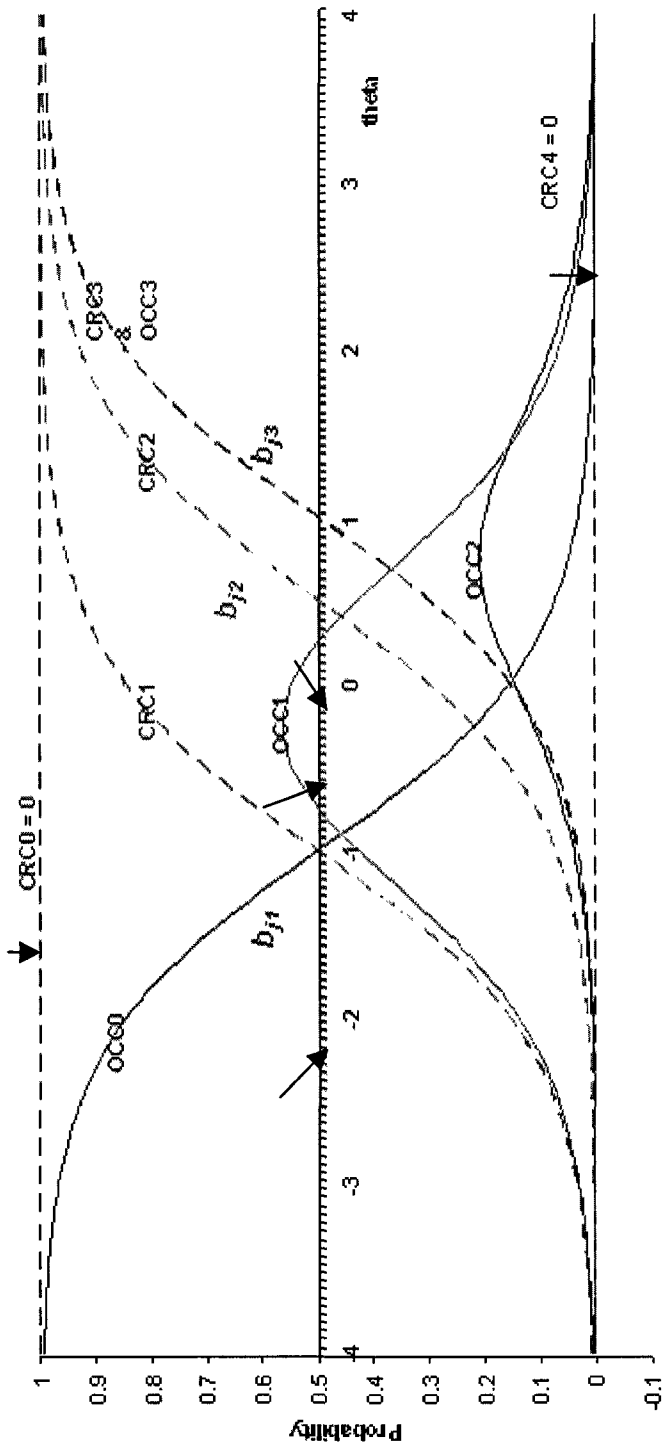

Figure 6. Category response and operating characteristic curves for a polytomously scored item under GRM with $a=1$, $b=(-1,0.5,1)$.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Muraki and Bock (1999) extended Samejima's GRM to the rating scale model for items scored in successive categories. The logistic form of this extension is:

$$P_{jk}(\theta_i) = \frac{\exp(Da_j(\theta_i - b_j + d_{jk}))}{1+\exp(Da_j(\theta_i - b_j + d_{jk}))} - \frac{\exp(Da_j(\theta_i - b_j + d_{jk+1}))}{1+\exp(Da_j(\theta_i - b_j + d_{jk+1}))} .$$

Note that item category threshold parameter, b_{jk}, in Equation 4 is resolved into an item location parameter b_j and a category parameter d_{jk}. In other words, b_{jk} equals b_j - d_{jk} (Childs & Chen, 1999). If an item has four category responses, this item would have one slope parameter (a_j), one location parameter (b_j), and three category parameters (d_{jk}).

Divided-by-Total Models

Andrich (1978) extended the Rasch model for dichotomous items to the Rasch polytomous rating response model. Masters (1980) reformulated Andrich's model to form the partial credit model. Later, Muraki (1992, 1993) generalized the partial credit model by including a slope parameter in his model. All these models are based on the assumption that the probability of choosing the kth category over the (k-1)th category is governed by the dichotomous response model. The probability of responding in a category k to an item i is expressed by the conditional probability of responding in category k, given the probability of responding in the $k-1$ and k categories. For example, Muraki's GPCM is given by:

$$P_{jk}(\theta_i) = \frac{\exp\left[\sum_{k=0}^{k} a_j(\theta_i - b_{jk})\right]}{\sum_{v=0}^{m_j} \exp\left[\sum_{k=0}^{v} a_j(\theta_i - b_{jk})\right]} , \qquad (5)$$

where $b_{j0} = 0$. When $m_j = 1$ and $k = 0, 1$, the partial credit model reduces to the 2PL model.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Operating characteristic curves for two items under the GPCM are illustrated in Figures 7 and 8. a_j is the slope parameter for item j. When the a-parameter increases, the OCCs become steeper. This is illustrated by the change of the a-parameter and the operating characteristic curves in Figures 7 and 8.

The b_{jk} in the GPCM is no longer defined as it is in the GRM. Masters (1980) named b_{jk} as the item step parameter. It is located along the ability scale at the intersection point of two adjacent characteristic curves. The magnitude of b_{jk} determines the relative difficulty of passing each step. Consequently, it is not always sequentially ordered within item j as it is in GRM because it is possible that passing one step is more difficult than passing the next step. This is illustrated in Figure 8 where b_{j2} (-1) is smaller than b_{j1} (0). When the b_{jk} increases, the corresponding $OCC_{j,\,k-1}$ and OCC_{jk} move to the right along the ability scale, which means the step from k-1 to k becomes more difficult. The property that the slope of the middle operating character curves will be affected not only by the slope parameter but also by the b-parameter found in GRM can also be found in GPCM. For example, when the distance between b_{jk} and $b_{j,k-1}$ becomes narrower, the corresponding $OCC_{j,k-1}$ become flatter. The two extreme OCCs will be influenced only by the slope parameter.

Muraki and Bock (1999) also extended the GPCM to rating scales. This extension is given by:

$$P_{jk}(\theta_i) = \frac{\exp\left[\sum_{k=0}^{k} a_j(\theta_i - b_j + d_{jk})\right]}{\sum_{v=0}^{m_j} \exp\left[\sum_{k=0}^{v} a_j(\theta_i - b_j + d_{jk})\right]}, \qquad (6)$$

where b_j is the item location parameter, and d_{jk} is the category parameter. The b_{jk} in Equation 5 equals b_j - d_{jk} in Equation 6. Andrich (1978) first introduced this separation of

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

item location and the category parameter. He claimed that the rating formulation preserves the ordering of the response categories.

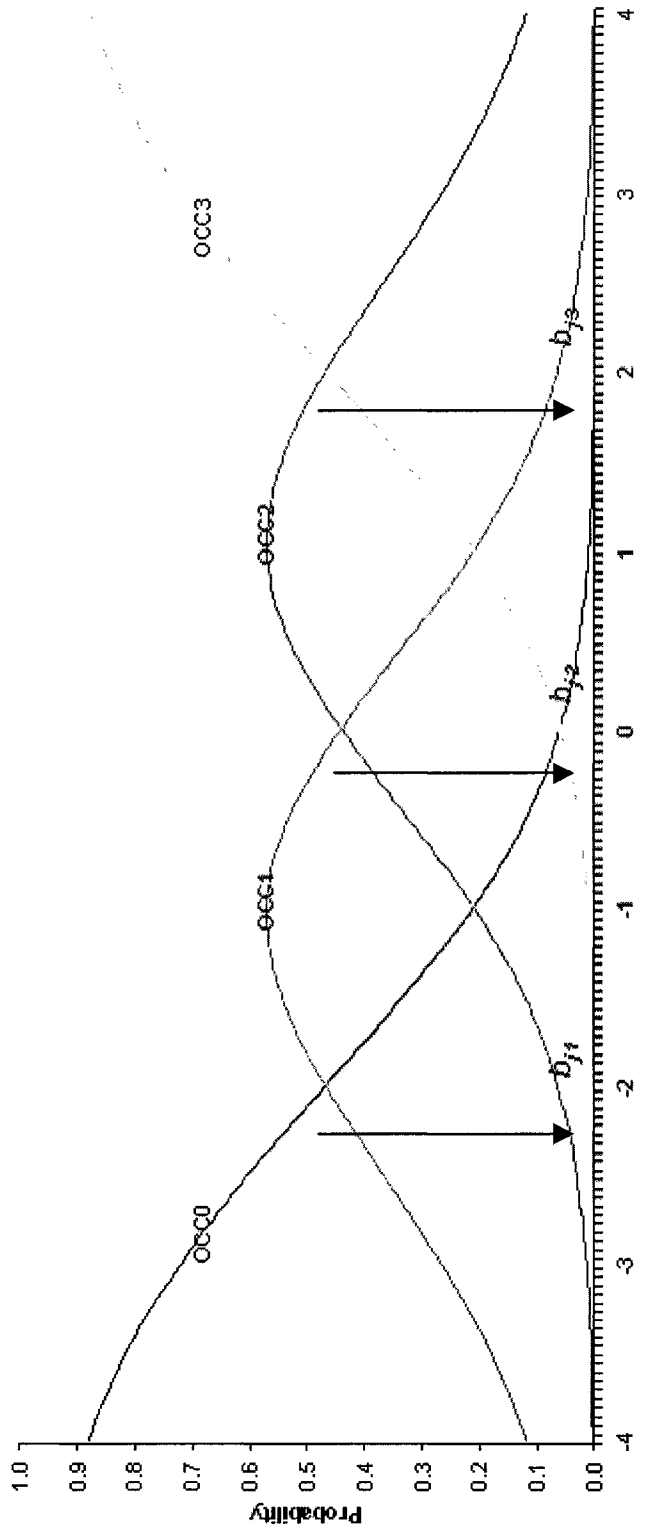

Figure 7. Operating characteristic curves for a polytomously scored item under GPCM with a=1, b=(-2,0,2).

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

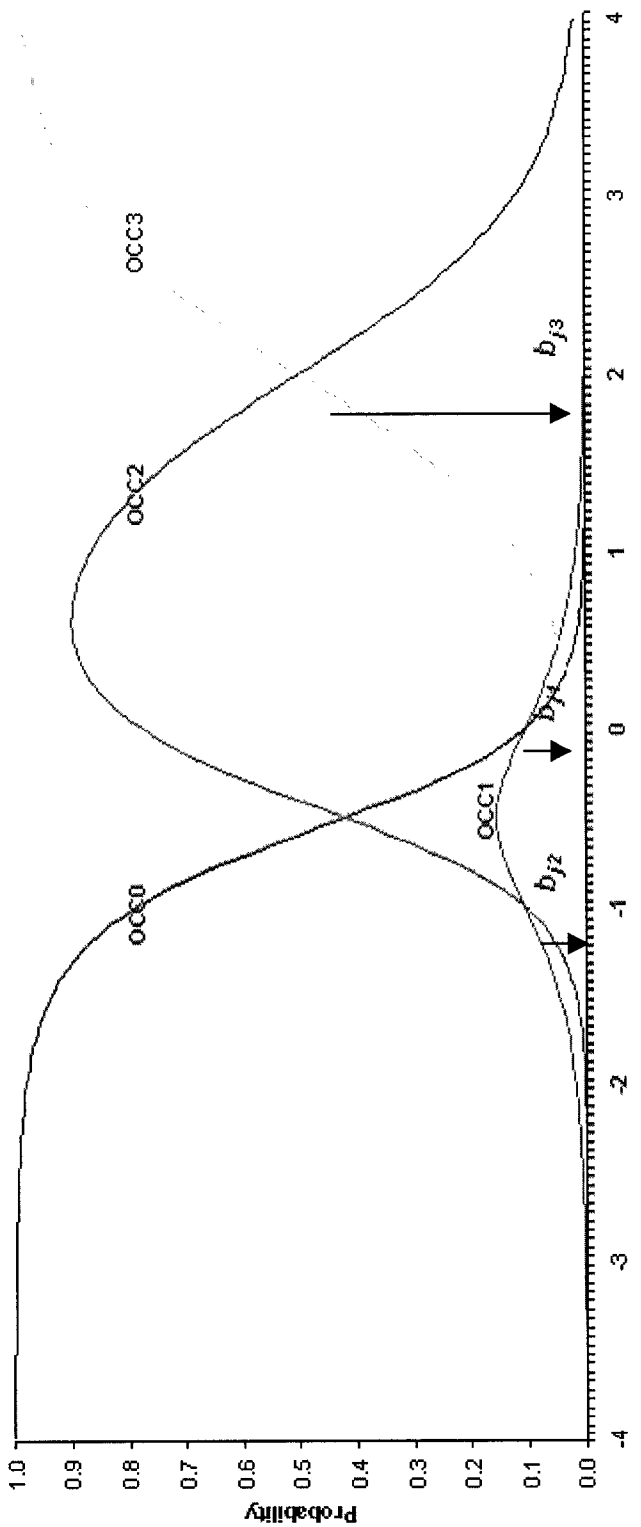

Figure 8. Operating characteristic curves for a polytomously scored item under GPCM with a=2, b=(0, -1, 2).

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

As described above, the mathematical formulation for the two classes of polytomously scored item response models are different. The different formulations are used to capture different response processes. The "difference" models, like the GRM, assume that the adjacent categories can be collapsed through cumulative frequencies. In contrast, in the "divided-by-total" models, like, the GPCM, the adjacent categories cannot be collapsed arbitrarily. This indicates that the response process of a respondent characterized by the first type of model is going from one category to the next category sequentially. However, the latter models reflect the process of a respondent who considers all of the response categories at once (Andrich, 1995).

IRT-based Equating

Common-item Non-equivalent Groups Design

As indicated in Chapter 1, the common-item non-equivalent groups design has some advantages over the other designs and, thus, has been widely used in practice. Although some variations of this design have been developed, the common characteristics of them are that two unique test forms are administered on two test dates respectively, and an anchor test with common items is administered on both test dates for the purpose of equating. Figure 9 illustrates a typical common-item non-equivalent groups design where test form Y with unique test U1 and anchor test C1 is administered in Year 1 and test form X with unique test U2 and anchor test C1 is administered in Year 2. These two unique tests can be equated through C1. This design may be extended over a number of different administrations.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

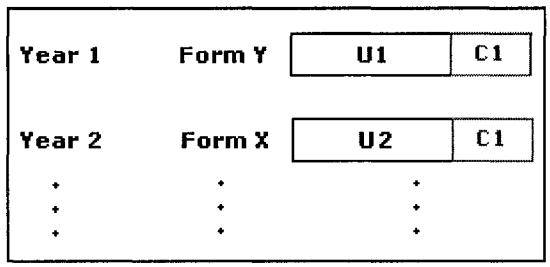

Figure 9. Illustration of common-item non-equivalent groups design.

One major concern of this design is centred on the selection or development of the common items to be included in the anchor test. Questions that are continuously asked are: (1) How many common items should be administered? (2) Should the common items represent the content and statistical characteristics of the unique test? (3) How should common items be selected when equating tests with a mixture of item formats?

Number of Common Items

The appropriate length of the anchor test has been studied under a variety of conditions (e.g., Budescu, 1985; Hills et al., 1988). Budescu (1985) revealed that large numbers of common items lead to less random equating error. Petersen, Cook, and Stocking (1983) indicated that too few items could lead to equating problems. However, Yang and Houang (1996) found that increasing the number of common items beyond 20% of the number of total items in the test led to improvements on equating accuracy that were not practically significant.

Other researchers have shown that a few common items can produce sufficient equating results. For example, Hills et al. (1988) studied the effect of anchor test length and found that ten randomly chosen anchor items were sufficient when the IRT - concurrent method was used to equate two mathematics tests. Raju, Edwards, and Osberg (1983), Raju, Bode, Larsen, and Steinhaus (1986), and Wingersky and Lord (1984)

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

suggested that as few as five or six carefully chosen items could perform well in IRT-based equating when the item parameters of both tests (i.e., U1+C1 and U2+C1) were estimated simultaneously.

If both sufficiency and efficiency are considered, two similar rules of thumb have been suggested for determining the number of common items. Angoff (1984) recommended that at least 20 items or 20% of the total number of items in a test, whichever is larger, should be included in the anchor test. Kolen and Brennan (1995) also said that the anchor test should be at least 20% of the length of a total test containing 40 or more items unless the test is very long, in which case, 30 common items might be sufficient.

Content and Statistical Representativeness of Common Items

Kolen and Brennan (1995) suggested "the number of common items to use should be considered on both content and statistical grounds" (p. 248). Petersen, Marco, and Steward (1982) investigated a variety of test form and anchor test characteristics, including content representativeness and item difficulty. They consistently found that both of these anchor test properties are crucial when the two groups differed in ability. Klein and Jarjoura (1985) found that a content representative anchor test functioned better than a longer, non-representative anchor test. Cook and Petersen (1987) reviewed several studies that considered anchor test properties. In their summary, they pointed out that the effectiveness of an anchor test depends on the extent to which the anchor test is similar to the total test, and that content and statistical representativeness is especially important when groups vary in ability. Yang (1997) found out that the accuracy of equating depended on the content representativeness of the anchor items, no matter which

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

equating method (Tucker linear and two IRT-based methods) was used to equate two test forms.

However, other researchers have shown that content and/or statistical representativeness are not important under some conditions. Kromrey, Parshall, and Yi (1998) compared the unweighted approach (using equal weights for all the items in the anchor test) and two differential weighting methods for the items in non-representative anchor tests. They found that the weighting did not perform better than no weighting, although this result may be due to the use of equivalent groups. Harris (1991) examined the effect of content and statistical non-representativeness. She found that content itself did not greatly influence equating results. However, if the anchor test was not statistically representative, a content representative anchor test may produce less equating error than a content non-representative anchor test. Budescu (1985) pointed out that the magnitude of correlation between the anchor test and the total test was the most important determinant of the efficiency of the equating process; a high correlation resulted in a better equating result. In agreement, Beguin (2002) found that the unidimensional equating procedures are fairly robust to violations of the assumption of content and statistical representativeness of the anchor test as long as the anchor test is highly correlated with the tests to be equated.

Item Format of Common Items

With the increasing use of open-ended response items in large-scale achievement tests, the issue of whether the common items should include different item types has been addressed. In practice, an anchor test consisting exclusively of MC items is frequently used instead of using mixed item types (Sykes, Hou, Hanson, & Wang, 2002). The use of

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

MC items only is based on the assumption that, when used, it is easier to represent all of the content categories covered in the total test or when there is evidence that a single dimension adequately explains the item responses in the total test. The use of an MC anchor test has its practical advantages. For example, it allows equating to take place in frequently narrow time frames demanded by the rapid turn-around of scores (Sykes et al., 2002). The use of MC items avoids the possibility that examinees will remember any open-ended response items included in the anchor test. However, Tate (2000, 2002) noted that anchor tests that were unbalanced with respect to item type (i.e., exclusively MC items) underestimated the simulated increase in abilities relative to anchor tests that were balanced across item type, and that the exclusive use of MC items in an anchor test failed to capture the large change in the mean ability attributable to the inclusion of open-ended response items. Since different types of items are believed to assess somewhat different constructs or cognitive processes (Bennett & Ward, 1993), it is reasonable to construct an anchor test that includes the item types employed in the total test.

However, in the simple common-item non-equivalent groups design, there are practical concerns about constructing a single anchor test with enough content, statistical, and item format representative items. For example, if all these factors are considered, it may lead to a long anchor test. However, a long internal anchor test may not be allowed in practice due to limited test administration time. Some large-scale achievement testing programs employ an external anchor test to solve this problem. However, other factors, such as lack of motivation, may influence examinees' observed performance on the external anchor test. Consequently, the scores may under represent the latent trait. Further, as indicated previously, it is difficult to guarantee the security of the common

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

items using this simple single common-item anchor non-equivalent groups design. For example, in the design with an internal anchor test, if an examinee took Form Y on one date and Form X on the next test date, it is very possible that the examinee may remember the answers for some of the common items.

A variant of the simple common-item non-equivalent groups design has been employed in some studies and large-scale testing programs, especially when the number of examinees is quite large (e.g., greater than ten thousand) (e.g., Zenisky, 2001; Vukmirovic, et al., 2003). In this design (see Figure 10), known as the common-item non-equivalent groups matrix design, the tests to be equated are administered on two different test dates. On one test date, say Year 1, multiple test forms are administered to different students at the same time. For example, in Year 1, a sub-form, FormY_1, of the test form Y could be administered to 3000 students, the FormY_2 could be administered to a different 3000 students, and so on. On one test date, different test forms include exactly the same unique items but different sets of common or equating items. For example, in Year 1, all the test forms include one set of unique items, U1, but different sets of common items: C_1, C_2, C_3, and C_4. The unique items administered in Year 2 are different from those administered in Year 1. However, the same sets of common items are used. Thus, examinees scores on U1 and U2 can be equated through common items in C_1, C_2, C_3, and C_4. Note that the common items are not scored as part of examinees' final scores although they are administered with the unique items at the same time. Further, these different sets of common items are assumed to be statistically similar and with the same item format. Together, the full set of common items represents the content and statistical characteristics of the unique items.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

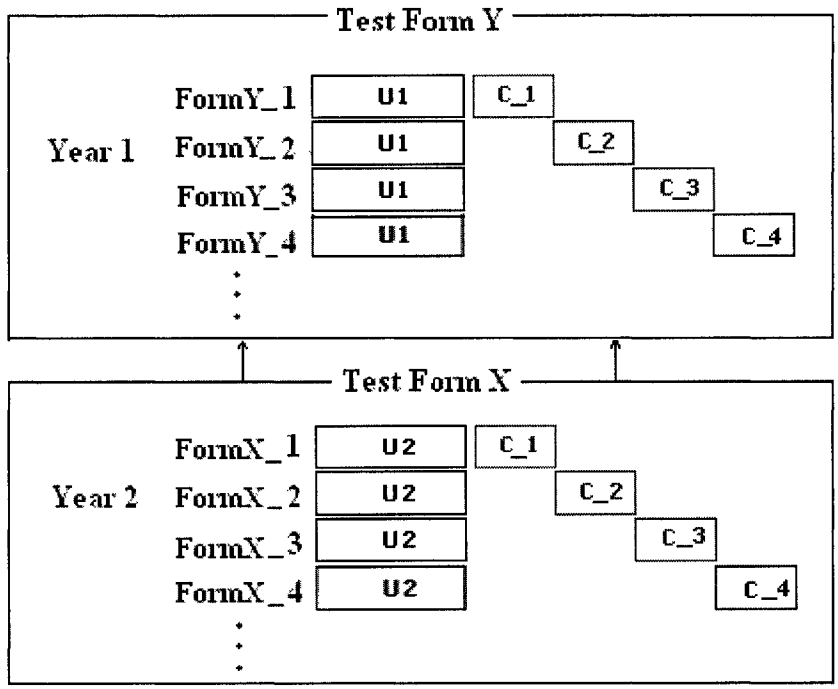

Figure 10. Illustration of common-item non-equivalent groups matrix design.

The advantage of the common-item non-equivalent groups matrix design over the simple common-item non-equivalent groups design (see Figure 9) is that part of the common items can be imbedded into each test form without excessively prolonging the test administration time. Meanwhile, examinees are equally motivated to take both unique and common items. If some examinees take the test next year, the possibility that they will take the same common items they did in the previous year is low. In other words, the common items are more secure in this design. As a result, it is possible to include open-ended response items in the anchor test without risking test security. Finally, the total number of common items shared between two consecutive years could be far more than the minimum number suggested by the rules of thumb provided by Angoff (1984) and Kolen and Brennan (1995). Because of these advantages, the common-item non-equivalent matrix groups design will be considered in the present study.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

IRT Calibration/Transformation

As indicated in Chapter 1, the estimated parameters for the two non-equivalent equating groups are often not on the same scale and, therefore, need to be equated. Three IRT calibration/transformation methods are frequently used to perform the calibration and transformation: separate calibration/linear transformation, concurrent calibration, and calibration with fixed common item parameters.

Separate Calibration/Linear Transformation

In practice, test forms from different years are often calibrated separately. In the random groups design, because the groups are randomly equivalent, and the abilities are scaled to the same metric with the same mean and standard deviation in both groups, no further transformation is needed. In contrast, for the common-item non-equivalent groups design, the two groups are assumed to be drawn from different populations. The estimated parameters for the two groups will not be in the same metric. In this case, a transformation procedure is needed to convert the parameter estimates of the two test forms onto a common scale. After the transformation, the parameters of the common items estimated from the two groups are expected to be the same within sampling error.

A linear relationship exists between the IRT scales for the two test forms. In other words, if an IRT model fits a set of data, then any linear transformation of the θ scale also fits the set of data, given the item parameters also are transformed (Kolen & Brennan, 1995; Lord, 1982). For example, in the 3PL model, the theta values and the three parameters for the X and Y scales are related as follows (Kolen & Brennan, 1995, p.163):

$$\theta_{Yi} = A\theta_{Xi} + B$$
$$a_{Yj} = a_{Xj} / A$$
$$b_{Yj} = Ab_{Xj} + B$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

$c_{Yj} = c_{Xi}$,

where

A and B are referred as to equating coefficients,

θ_{Yi} and θ_{Xi} are the theta values for individual i on Scale Y and Scale X,

a_{Xj}, b_{Xj}, and c_{Xi} are the item parameters for item j on Scale X, and

a_{Yj}, b_{Yj}, and c_{Yj} are the item parameters for item j on Scale Y.

For the polytomously scored IRT models, the following linear relation exists (Li, et al., 1999, p.8):

$$\theta_{Yi} = A\theta_{Xi} + B$$
$$a_{Yj} = a_{Xj} / A$$
$$_{Yj} = Ab_{Xj} + B$$
$$b_{Yjk} = Ab_{Xjk} + B$$
$$d_{Yjk} = Ad_{Xik} \text{ ,}$$

where

b_{Yjk} is equal to b_{Yj} - d_{Yjk},

b_{Xjk} is equal to b_{Xj} - d_{Xjk} ,

b_{Xjk}, b_{Xj}, and d_{Xjk} are the category threshold, location, and category parameters on Scale X, and b_{Yjk}, b_{Yj}, and d_{Yjk} are the category threshold, location, and category parameters on Scale Y.

Many procedures have been proposed to estimate the coefficients A and B (e.g., Haebara, 1980; Lin, et al., 1980; Loyd & Hoover, 1980; Marco, 1977; Stocking & Lord, 1983). One procedure is to use moment statistics in terms of groups of items and/or persons. For example,

$$A = \frac{\sigma(b_Y)}{\sigma(b_X)} \qquad (7)$$

$$= \frac{\mu(a_X)}{\mu(a_Y)} \qquad (8)$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

$$= \frac{\sigma(\theta_Y)}{\sigma(\theta_X)}$$

$$B = \mu(b_Y) - A\mu(b_X) \qquad (9)$$
$$= \mu(\theta_Y) - A\mu(\theta_X),$$

where

$\mu(b_Y)$, $\mu(b_X)$, $\mu(a_Y)$, and $\mu(a_X)$ are the means of common items' b- and a-parameters on Scale Y and Scale X,

$\sigma(b_Y)$ and $\sigma(b_X)$ are the standard deviations of common items' b-parameters on Scale Y and Scale X,

$\mu(\theta_Y)$, $\mu(\theta_X)$, $\sigma(\theta_Y)$, and $\sigma(\theta_X)$ are the means and standard deviations of examinees' abilities on Scale Y and Scale X.

Marco (1977) described a mean/sigma method that uses the mean and standard deviation of b-parameters to estimate the coefficients A and B (Equations 7 and 9). Loyd and Hoover (1980) proposed to use the means of a- and b-parameters to estimate the equating coefficients A and B (Equations 8 and 9), and this method is referred as to mean/mean method. However, these methods are sensitive to deviate values (outliers), especially poorly estimated item difficulties (Cohen & Kim, 1998; Stocking & Lord, 1983). Cook et al. (1979), Bejar and Wingersky (1981), Linn et al. (1980), and Stocking and Lord (1983) developed procedures to overcome this problem. For example, Linn et al. (1980) modified the mean/sigma method by using weighted item difficulty estimates where the weights are inversely proportional to the estimated standard errors of the item difficulties. Cohen and Kim (1998) extended the mean/sigma method and Linn et al.'s modification to the graded response model by treating each categorical threshold parameter as an individual b-parameter.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

It is theoretically correct to use the mean/mean and mean/sigma methods. However, Stocking and Lord (1983) pointed out that a potential drawback of these methods is that the *a*-parameter and/or the *b*-parameter are separately used to estimate the equating coefficients. The above methods may lose the information that can be obtained from considering all the item parameters simultaneously. Thus, Stocking and Lord (1983) proposed the test characteristic curve method. Baker (1992) and Li et al. (1999) extended this procedure to the polytomously-scored items. In their approach, equating coefficients are obtained by minimizing the quadratic loss function:

$$F = \frac{1}{N}\sum_{i=1}^{N}(t_{iY} - t_i^{*}{}_{X \to Y})^2 ,$$

where

N is the number of "arbitrary" points along the reference test θ scale.

t_{iY} is the expected number-correct true scores for the reference test,

and $t_i^{*}{}_{X \to Y}$ is the transformed number-correct true scores for the equated test.

t_{iY} and $t_i^{*}{}_{X \to Y}$ are defined, respectively, as

$$t_{iY} = \sum_{j=1}^{n_c}\sum_{k=0}^{m_j} u_{jk}P_{jkY}(\theta_{iY}; a_{jY}, b_{jY}, c_{jY}) \qquad (10)$$

and

$$t_i^{*}{}_{X \to Y} = \sum_{j=1}^{n_c}\sum_{k=0}^{m_j} u_{jk}P_{jk}^{*}{}_{X \to Y}(A\theta_{iX} + B; \frac{a_{jX}}{A}, Ab_{jX} + B, c_{jX}) , \qquad (11)$$

where

n_c is the number of common items,

m_j is the number of categories minus 1 for item j,

u_{jk} is the weight allocated to the response category k for item j, ranging from 0 to 1 for a dichotomous scored item or 0 to 3 for a four-category scored item,

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

P_{jkY} is the observed probability that an examinee answers item j's category k correctly given the item parameters and ability parameters are calibrated on Scale Y,

and $P_{jk}{}^{*}{}_{X \rightarrow Y}$ is the probability that an examinee answers item j's category k correctly given the item parameters and ability parameters on Scale X have been transformed onto Scale Y.

Note that θ_{Yi} in Equations 10 and 11 is in the reference test metric because the goal is to convert the scores on the equated test to the reference test metric. The function F, which is a function of both A and B, will be minimized when $\partial F / \partial A = 0$ and $\partial F / \partial B = 0$. An iterative multivariate procedure has been proposed to find the two equating coefficients that will minimize F (Baker, 1992; Li et al., 1999; Stocking & Lord, 1983).

No matter which linear transformation method is used, the item parameter estimates of common items in the reference test metric are often not the same as those transformed from the equated test due to sampling error, parameter estimation error, and transformation error. Hambleton and Swaminathan (1985) suggested averaging these two sets of common item parameter estimates. However, Kim and Cohen (1998) argued that when the item parameter estimates are changed, the subsequent ability distribution of the reference group may no longer be N (0, 1). This problem is overcome by the following two calibration methods.

Concurrent Calibration

Concurrent calibration means the parameter estimation for the test forms to be equated is completed in one analysis. The parameter estimates, therefore, are automatically calibrated onto the same scale and no further transformation is needed

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

(Hambleton et al., 1991; Hanson & Belguin, 2002; Mislevy & Bock, 1990). Hambleton et al. (1991) summarised the general procedure of concurrent calibration:

1. Examinees taking the parallel test forms are treated as one sample and the data from different groups are combined;
2. The scores for the items that are not responded by some group(s) are coded as "not presented";
3. The ability parameters for all the examinees and the item parameters for all the items are calibrated in a single analysis.

Theoretically, the concurrent calibration method is expected to yield more stable equating results than the separate calibration method because more examinees are available to take the common items and more information is available for the parameter estimation. Further, the equating errors produced by calibration and inaccurate transformation functions may be reduced or, perhaps, removed (Li, Griffth, & Tam, 1997; Li et al., 1999). Li et al. (1997) indicated that the greatest potential benefit of using concurrent calibration is that concurrent calibration may minimize the impact of sampling fluctuations on estimating the guessing parameters. The increased number of low ability examinees that comes about by combining the different samples is especially informative in the estimation of the guessing parameters (Stocking, 1990). In contrast, when there is a small number of examinees, the guessing parameter estimates will be unstable (Van der Linden & Hambleton, 1997). This uncertainty may, in turn, cause uncertainty in the difficulty parameter estimates, especially for the easy items (Thissen & Wainer, 1982). Since the parameter estimation of common items may improve due to the increased

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

number of examinees, the estimation of examinees' ability parameters may profit from this improvement.

However, when the concurrent calibration is conducted using the marginal maximum likelihood estimation (MMLE) procedure, there needs to be a single underlying ability distribution for the total or combined sample. If the ability distribution of the target group is different from that of the reference group, the marginalization of the likelihood function under the assumption of a single ability distribution may not be correctly specified, which, in turn, will lead to the problems in estimating the parameters (Kim & Cohen, 1997).

There are also some practical difficulties that can occur when the concurrent calibration method is used to equate test forms. For example, Li et al. (1997) pointed out that coding the "not reached items" can be very tedious when there are more than two test forms. When the two groups are not equivalent, the ability scale derived from the concurrent calibration will be located somewhere between the ability scales of the two groups. This may lead to the difficulty of scale explanation when more test forms are equated over time (Vukmirovic et al., 2003).

FCIP Calibration

Mislevy and Bock (1990) combined the separate and concurrent calibration procedures to produce what they called the fixed common item parameters (FCIP) method. First, the reference and equated tests are calibrated separately. Second, the parameters of common items are held constant in the calibration of the equated test using the values of the parameter estimates from the reference test (Li et al., 1997). As a result, the equated test items are calibrated onto the existing reference scale without changing

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

the parameter values for the reference test.

Like concurrent calibration, FCIP calibration is expected to produce more stable equating results than separate calibration because it removes the equating errors caused by inaccurate transformation functions, and it takes the *c*-parameter into account for metric conversion. Further, if multiple groups take the common items, the common item parameter estimates are expected to be more stable, which, in turn, leads to a more accurate equating results. FCIP can be adapted to a variety of data collection methods such as computer adaptive testing and online item linking (Vale, 1986; Li et al., 1999).

However, there are some potential problems with FCIP when the two equated groups are extremely different in terms of location and variability of ability. For example, it may produce extreme item parameter estimates. It may not converge when some common items are fixed. To overcome this latter problem, these item parameters have to be freed (Li et al., 1997).

IRT True Score Equating

IRT true score equating is used to equate number-correct true scores on Form X and Form Y (Kolen & Brennan, 1995). It is completed via a given ability parameter θ. This θ is associated with the number-correct true scores on the two test forms in the following way (Kolen & Brennan, 1995, p.175-176):

$$\tau_X(\theta i) = \sum_{j=1}^{nx} \sum_{k=0}^{mj} u_{jk} P_{jk}(\theta_i)$$

and

$$\tau_Y(\theta i) = \sum_{j=1}^{nY} \sum_{k=0}^{mj} u_{jk} P_{jk}(\theta_i),$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

where

n_X and n_Y are the number of items in the Form X and Form Y,

m_j is the number of categories for item j minus 1,

u_{jk} is the weight allocated to the response category k for item j,

$P_{jk}(\theta_i)$ is the item response function that could be defined in any of the IRT models.

In the case of 3PL, $\tau_X(\theta_i)$ and $\tau_Y(\theta_i)$ are constrained by the sum of c-parameters and the total number of items and categories in Form X and Form Y as defined below:

$$\sum_{j=1}^{nX}\sum_{k=0}^{mj} c_{jk} < \tau_X < \sum_{j=1}^{nX}\sum_{k=0}^{mj} u_{jk} \quad \text{and} \quad \sum_{j=1}^{ny}\sum_{k}^{mi} c_{jk} < \tau_Y < \sum_{j=1}^{ny}\sum_{k}^{mi} u_{jk}$$

Kolen and Brennan (1995) described the procedure of conducting true score equating: First, specify a number-correct true score on Form X. Second, find the θ_i that corresponds to this number-correct true score. Third, find the number-correct true score on Form Y that corresponds to the θ_i. The second step requires the solution of a nonlinear equation using an iterative process. In this case, the Newton-Raphson method is used for finding the roots of the nonlinear function:

$$\theta^{+} = \theta^{-} - \frac{\tau_X - \sum_{j=1}^{nX}\sum_{k=0}^{mj} u_{jk}P_{jk}(\theta_i)}{-\sum_{j=1}^{nX}\sum_{k=0}^{mj} u_{jk}P'_{jk}(\theta_i)},$$

where

θ^{+} is the new value calculated after the initial value θ^{-} is chosen.

This new θ^{+} is typically closer to the root of the equation than θ^{-}. In the next cycle, θ^{-} is replaced by the θ^{+} calculated from the previous run. This process continues

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

until a criterion such as $\theta^{+} - \theta^{-} \leq 0.001$ is obtained. Then θ^{+} from the last run is the θ corresponding to the true score assigned on Form X.

$P'_{jk}(\theta_i)$ is the first derivative of $P_{jk}(\theta_i)$ with respect to θ_i. It is defined (Lord, 1980) as:

$$P'_{jk}(\theta_i) = \frac{1.7a_j(1-P_{jk}(\theta_i))(P_{jk}(\theta_i)-c_j)}{1-c_j}.$$

True score equating has been criticized. For example, Kolen and Brennan (1995) pointed out that there is no theoretical basis to apply the relationship between the estimated number-correct true scores to that of the observed number-correct scores (Kolen & Brennan, 1995). Besides, the true scores are not available in practice. Thus, observed-score equating was proposed to overcome this theoretical disadvantage. However, comparison studies did not show consistent results about the comparability of the IRT observed-score equating and the IRT true-score equating. For example, Kolen's (1981) study revealed that the two methods produced different equating results across the 1PL, 2PL, and 3PL models using cross-validation criterion and the random groups design. The IRT observed-score equating showed better stability. However, Lord and Wingersky (1984) found no difference between these two methods when the circular chain equating criterion was used to evaluate equating stability using a common-item non-equivalent groups design. Han et al. (1997) further examined these two methods and the traditional equipercentile equating. They found that the IRT true-score equating produced more stable equating results than the IRT observed-score equating. However, the mean difference in equating stability was statistically insignificant. Despite these concerns, the IRT true score equating has been more widely used than the IRT observed-

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

score equating in practice because the IRT true score equating is sample independent and easy to compute (Kolen & Brennan, 1995).

Criteria for Evaluating IRT-based Equating Methods

As described above, there are many equating designs and equating methods available for conducting equating. One may have to answer the question about whether a specific option of these designs and methods leads to an acceptable equating result. A variety of criteria have been used to evaluate the equating results (Harris & Crouse, 1993). What are called the indices and the standard error are the two most commonly used criteria to compare and evaluate the precision of equating methods.

Indices

Overall summary indices are often used to compare the equating results across methods. These indices typically include root mean square difference (*RMSD*) (also referred to as root mean square error), mean square error (*MSE*) or total error, mean absolute difference (*MAD*), and mean signed difference (*MSD*). These four indices can be used with both weighting and no weighting. The weighted indices are defined as:

$$RMSD_w = \sqrt{\frac{\sum_{i=1}^{n} f_i(V_{i1} - V_{i2})^2}{\sum_{i=1}^{n} f_i}} \tag{12}$$

$$MSE_w = \frac{\sum_{i=1}^{n} f_i(V_{i1} - V_{i2})^2}{\sum_{i=1}^{n} f_i} \tag{13}$$

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

$$MAD_w = \frac{\sum_{i=1}^{n} f_i |V_{i1} - V_{i2}|}{\sum_{i=1}^{n} f_i} \quad (14)$$

$$MSD_w = \frac{\sum_{i=1}^{n} f_i (V_{i1} - V_{i2})}{\sum_{i=1}^{n} f_i}, \quad (15)$$

where

V_{i1} and V_{i2} typically are the raw scores of examinee i on the equated test that are equivalent to the raw scores on the reference test using equating methods 1 and 2 when the purpose of using these indices is to evaluate the consistency between different equating methods. In the computer simulation studies, V_{i1} often refers to the estimated values, and V_{i2} refers to as the true values.

f_i is the frequency of the raw score of examinee i on the equated test, and i runs over the possible raw score range. If the indices are not weighted, the f_i in Equations 12-15 is set to one.

n is the number of values of V.

The weighted indices and unweighted indices sometimes produce different results. For example, Skaggs and Lissitz (1986b) found that although the weighted indices appeared acceptable in some instances, the unweighted indices appeared relatively large. The selection of any of these indices depends on the researcher's emphasis on the differences. Weighted indices allow relatively more emphasis to be given to score point differences that occur frequently, and lower emphasis to differences in the extremes of the score range where few or no examinees score. However, sometimes it is also important to know the differences throughout the score scale. For

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

example, Jaeger (1981) found that substantial equating error might be considered serious, even if it only occurs for score points that have small frequency. In this case, the use of unweighted indices is more appropriate.

In computer simulation studies, the comparison of differences are usually replicated R times. Gifford and Swaminathan (1990) and Hanson and Belguin (2002) demonstrated that the mean squared difference for each comparison across $r = 1, 2, \ldots, R$ replications, can be decomposed into the variance of the difference across replications (also referred to as random error) and bias (referred to as systematic error) squared. For example, the formula for the MSE of b-parameters across R replications, when decomposed, becomes:

$$MSE_b = \frac{\sum_{r=1}^{R} (b_{jr}^{*} - \bar{b}_j^{*})^2}{R} + (\bar{b}_j^{*} - b_j)^2 , \qquad (16)$$

where

b_{jr}^{*} is the b-parameter for item j in the equated test, say Test X, that has been equated on the reference test, say Test Y, for replication r,

$\bar{b}_j^{*}$ is the mean of b_{jr}^{*} across r replications, and

b_j is the true value of the b-parameter for item j.

In Equation 16, the first term on the right of the equation is the random error and the second part is the bias squared. Only the bias term (systematic error, denoted by MSE_{b_SE}) reflects of accuracy of equating. Thus, it is used to calculate the MSE of b-parameters across items. For example:

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

$$MSE_{b\_SE} = \frac{\sum_{j=1}^{n} (\bar{b}_j^* - b_j)^2}{n}$$

where n is the number (or score-points) of items in the test.

Although a majority of equating studies have employed one or more of these indices to evaluate the consistency of equating methods, there are some concerns about these indices. For example, Harris and Crouse (1993) pointed out that it was not clear what amount of difference between two indices is significant, or what amount of differences indicates a satisfactory equating. Further, it is not clear that which equating method should be considered as the standard against which other equating methods are compared. Even when the differences between two methods are found, it is not clear which method is better.

However, the accuracy of equating can be evaluated in simulation studies because when one generates the data, one knows the true values one is trying to recover. For example, examinees' true ability, the true item parameters, and the true equating coefficients are known in a computer simulation study. Thus, the estimated values using of an equating method can be compared to the true values.

The use of these indices has been abundant in the literature, especially in computer simulation studies. For example, Li et al. (1997) calculated the *MSD* and *RMSD* for the difference between the true parameters and the estimated parameters to examine the robustness of FCIP under the situation of large standard errors in the item difficulty and guessing parameters. Kim and Cohen (1998) used *RMSD* for the *a*-parameter and *b*-parameter to compare the methods of separate calibration with the characteristic curve transformation and the method of concurrent calibration. Kaskowitz

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

and De Ayala (2001) examined the effect of error in item parameter estimates on the test characteristic curve method using *RMSD* and *MSD* for the equating coefficients *A* and *B*. However, no evidence has been presented that indicates which index is the best to use.

Standard Error

Equating error may occur randomly or systematically (Kolen & Brennan, 1995). Random error is due to random sampling of examinees from the population of interest. Thus, increasing the sample size can decrease the random error. Systematic error may result from the estimation error of the equating relationship, violation of the assumptions of an equating method, or improper implementation of an equating design. For example, in the common-item non-equivalent groups design, systematic error may arise when the common items are not representative of the total test in terms of the content and statistical characteristics. Systematic error may also occur when the common items function differently from one test administration to another.

The delta method or analytical procedure (Kolen & Brennan, 1995) has been used to develop formulas for calculating the random error of IRT-based equating. For example, Thissen and Wainer (1982) developed a mathematical expression to examine how the sample size, the shape of examinees' ability distribution, and the characteristics of test items cause differences in the errors of parameter estimates. Li et al. (1999) extended this formula to the general partial credit model. Ogasawara (2001) derived asymptotic standard errors of the IRT-based equating coefficient estimates for the 2PL and 3PL models. However, due to the lack of simple computation, the analytical procedure is not widely used in evaluating the IRT-based equating methods in the literature.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Another procedure to estimate the random equating error is referred to as the "bootstrap" (Kolen & Brennan, 1995). It can be used when the analytical procedure is not available or its assumptions are questionable. In this procedure, many samples, for example, 25 to 200 (Efron & Tibshirani, 1993), are drawn from the sample data and the equating is conducted for each replication. Then, the standard deviations are computed over all the replications to obtain the standard error of equated scores at all raw score points for the equated test form (Kolen & Brennan, 1995; Tsai, Hanson, Kolen, & Forsyth, 2001). However, this procedure is time consuming in the context of IRT-based equating; therefore, it is not widely used.

Some empirical suggestions have been made on how to control the random error as well as the systematic error. For example, Kolen and Brennan (1995) recommended that a sample size of 400 for the Rasch model or 1500 for the 3PL model should be used when using IRT-based equating. Kolen and Brennan (1995) also suggested that the proportion of common items over the total test items should be at least 20% in the common-item non-equivalent groups design.

Comparison Studies on the IRT-based Equating Methods

Because of the lack of definite criteria to evaluate an equating method, many studies have focused on comparing the different equating methods (e.g., Tsai et al., 2001). For example, Kolen (1981) compared the traditional linear, equipercentile, modified 1PL estimated true score, modified 1PL estimated observed score, 2PL estimated true score, 2PL estimated observed score, 3PL estimated true score, and 3PL estimated observed score equating methods. The equating data was collected under the randomly equivalent groups design. The results from the equating methods were

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

compared using cross-validation criterion. He found four major results: (1) the 1PL method results were inadequate for equating tests differing in difficulty; (2) the 3PL methods produced the most stable cross-validation results; (3) the linear method was not satisfactory when equating tests of unequal difficulty; and (4) the equipercentile method produced reasonably adequate results. Kolen and Whitney (1982) compared the equipercentile, linear, 1PL, and 3PL equating methods when the sample size was 200. Cross-validation analyses were used to evaluate the equating results. They found that the 3PL equating method produced unacceptable equating results. The 1PL method produced results that were as stable as those from traditional methods.

Petersen et al. (1983) compared the equipercentile, linear, and IRT -concurrent calibration methods in terms of scale drift using a sample of approximately 2670 cases. Overall, they found the IRT - concurrent calibration method appeared to be the best equating method for reducing scale drift over time. Cook and Eignor (1983) investigated the feasibility of applying IRT-based equating methods by comparing these methods with traditional equating methods. Scale drift was used to evaluate the equating results. If the results of equating test form A directly to test form D are not the same as that obtained by equating test form A to test form D through forms B and C, then scale drift was thought to occur. They found that the IRT - concurrent calibration and the IRT - separate calibration with characteristic curve transformation methods were both feasible and that the two methods produced similar results.

Hills et al. (1988) compared the linear equating, 1PL, 3PL-concurrent, 3PL-fixed parameter, and 3PL-linear transformation methods. He reported that the different methods produced similar results in the situation in which the tests were made parallel in

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

difficulty and content, and the groups were equivalent over two years. Further, an anchor test with 10 items provided equating as effective as an anchor test with 30 items using the IRT - concurrent calibration method. However, Kim and Cohen (1998) found some differences between the IRT separate and concurrent equating methods. In their study, a test of 50 items and four sets of common items (5, 10, 25, and 50) were administered to two groups with 500 examinees. Three IRT-based equating methods were compared: separate calibration with Stocking-Lord characteristic curve method, concurrent calibration via marginal maximum "a posteriori" estimation, and concurrent calibration via marginal maximum likelihood estimation. They found that separate calibration yielded smaller RMSD for both item discrimination and difficulty parameters for a smaller number of common items than the other two methods. For a larger number of common items, the three methods yielded essentially the same results.

Hanson and Belguin (2002) compared IRT - concurrent calibration and IRT - separate calibration with mean/mean, mean/sigma, Stocking-Lord characteristic curve, and Haebara characteristic curve methods. Four factors were considered in their study: program (MULTILOG versus BILOG-MG), sample size per form (3000 versus 1000), number of common items (20 versus 10), and equivalent and non-equivalent groups (no mean difference and a mean difference of 1 standard deviation). They found that, overall, the concurrent calibration resulted in less *MSE* than the separate calibration/linear transformation.

Tsai et al. (2001) employed bootstrap error as their criterion to compare the IRT - separate and IRT - concurrent calibration methods. In the separate calibration, Stocking –Lord's characteristic curve method was followed to estimate the equating coefficients.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

In both calibrations, the true score equating and the observed score equating were examined separately. Overall, five methods were compared. The standard deviation was computed over 500 bootstrap replications to obtain the standard error of IRT-based equating at each raw score point for the equated test form. They found that the concurrent calibration methods produced smaller standard errors than the separate calibration method did.

Cohen and Kim (1998) extended the comparison studies to tests with only polytomously scored items. Simulated data were used to investigate the IRT - separate calibration with different transformation methods: Stocking-Lord characteristic curve method, mean/mean, mean/sigma, and weighted mean/sigma. Underlying ability distribution (N (0, 1) - N (0, 1) versus N (0, 1) - N (1, 1)), sample size (300 versus 1000), and number of common items (5, 10, 20, and 30) were manipulated. The results indicate that differences in the equating coefficients estimated by these methods were small. *RMSD* of ability parameters were very small under most conditions. The methods yielded similar results for the longer common-item designs with larger sample size.

It is difficult to summarize these comparison studies because (1) almost every study compared different IRT-based equating methods under a variety of conditions, which sometimes lead to inconsistency of the results, and (2) different criteria were used to evaluate the equating results. The only common conclusion one may draw from these previous studies is that all the IRT-based equating methods yield more similar results when their assumptions were not violated.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Studies on Outliers

One of the major benefits of using IRT in large-scale achievement tests is that the estimation of item parameters is group independent. That is to say, when a set of items are administered to two groups, the item parameters (e.g., *a*-, *b*-, *c*-parameters) estimated from these groups are expected to be linearly related (i.e., located along a straight line if a scatter plot is plotted). However, some observed item parameters of the two groups to be equated have been found that are located far away from the straight line in the scatter plots (e.g., Stocking & Lord, 1983; Vukmirovic et al., 2003).

Many researchers have been aware of the effects of outliers on IRT-based equating, especially when the common-item non-equivalent groups design is employed (e.g., Bejar & Wingersky, 1981; Cohen & Kim, 1998; Cook et al, 1979; Hanson & Feinstein, 1997; Linn et al., 1980; Stocking & Lord, 1983). For example, Stocking and Lord (1983) pointed out that the poorly estimated item difficulties might negatively affect the estimation of the equating coefficients *A* and *B* when the mean/mean and mean/sigma transformation methods were used. Vukmirovic et al. (2003) found that fixing and not fixing the item parameters with inconsistent *b*-parameters might lead to different equating results when the FCIP was employed.

To overcome this problem, procedures have been proposed to modify the mean/mean and mean/sigma transformation methods. For example, Cook et al. (1979) tried to restrict the range of the difficulties used in computing moments. Bejar and Wingersky (1981) suggested giving smaller weights to the outliers that are used to estimate moments. Linn et al. (1980) used weighted item difficulties where the weights were the inverse of the squared standard errors. Stocking and Lord (1983) proposed an

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

iterative procedure that employed both Linn et al.'s (1980) method and Bejar and Wingersky's (1981) robust method. Cohen and Kim (1998) extended Linn et al.'s (1980) procedure to calculate the equating coefficients for polytomously scored items. However, it wasn't clear how much these procedures improved the equating results. For example, Cook et al.'s (1979) method may lead to the deletion of some outliers. Although this may eliminate the negative effect of outliers, it may lead to a non representative sample of common items. The other modified methods may be useful when the outliers are due to the item parameter estimation errors. However, other reasons, such as disclosure of some items, may also produce outliers. In this case, it is not clear whether these modified methods will result in a better equating result.

Few studies have addressed the issue of outliers when other IRT-based equating methods such as the separate-calibration/characteristic curve transformation, FCIP, and concurrent calibration are employed. Vukmirovic et al. (2003) explored the effects of fixing and not fixing random outliers using FCIP. However, how to deal with outliers when they appeared non-randomly was not clear. Theoretically, one may suggest removing outliers when no harm would be done to the balance of content in the set of common items (Hanson & Feinstein, 1997). However, further systematic study needs to be conducted to obtain a more clear understanding of the effect of the presence of outliers on the results obtained using IRT-based equating methods.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

CHAPTER 3: METHODOLOGY

In the current study, simulated data were generated for the common-item non-equivalent groups matrix design that reflected the characteristics of outliers and group ability differences. Ten IRT-based equating methods were used to equate the simulated data. The equating results were compared and evaluated; as a result, the performance of the IRT-based equating methods in the presence of outliers was investigated. The methodology involved in the above procedures is described in the following four sections.

Equating Design

The results of the current study are intended to be generalized to equating large-scale achievement tests with mixed item formats using IRT-based methods. As indicated in Chapter 2, since the common-item non-equivalent groups matrix design has many advantages in the context of large-scale achievement tests, a simple form of this design (see Figure 11) was employed. As shown in Figure 11, test forms Y and X were administered in two years and needed to be equated. There were three sub forms in each of these two test forms defined by different sets of the equating items. For example, test form Y included sub forms FormY_1, FormY_2, and FormY_3. In Year 1, all the test forms included one set of unique items, U1, but different sets of common items: C_1, C_2, and C_3. The unique items, U2, administered in Year 2 were different from those administered in Year 1. However, the same sets of common items were used. Thus, examinees scores on U1 and U2 were equated through common items in C_1, C_2, and C_3.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

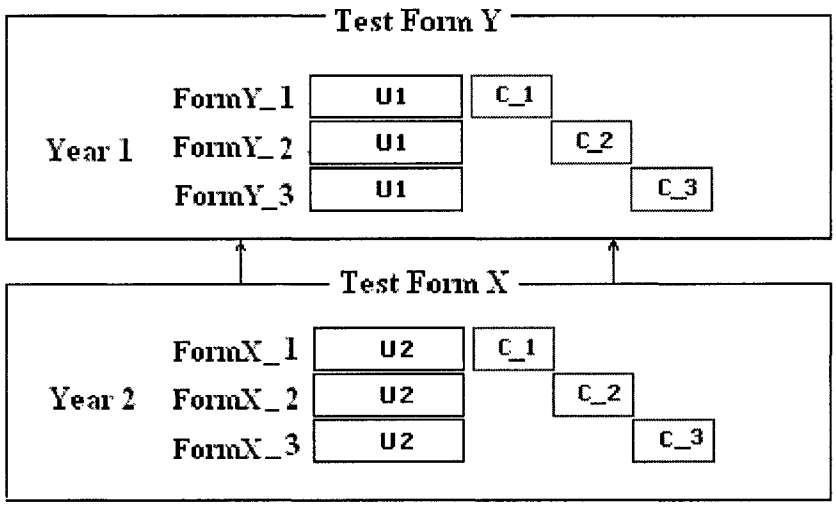

Figure11. Illustration of the equating design employed in the current study.

The number of items and the types of items considered in the current study are intended to best simulate a state-wide large-scale mathematics achievement test. Based on the available item characteristics from the mathematics test and the basic considerations for test development (for example, content representativeness and test administration time), 72 unique items and 30 common items were used to measure examinee's achievement in five mathematics content areas. As shown in Figure 12, there were 36 unique items in each sub test of Form Y. Of these, 26 were multiple-choice items (MC) with two score categories, 5 were short-answer items (SA) with two score categories, and 5 were open-ended response items (OR) with five score categories. Each sub form contained a different set of 8 common MC items, 1 common SA item, and 1 common OR item. The 30 common items together represented the statistical, content, and item format characteristics of the unique items. The same structure was used for Form X but with a different set of 36 unique items. Using the 30 common items, the unique items in the test form Y and X were equated onto the same scale. The common items were not scored as part of the examinees' final scores.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

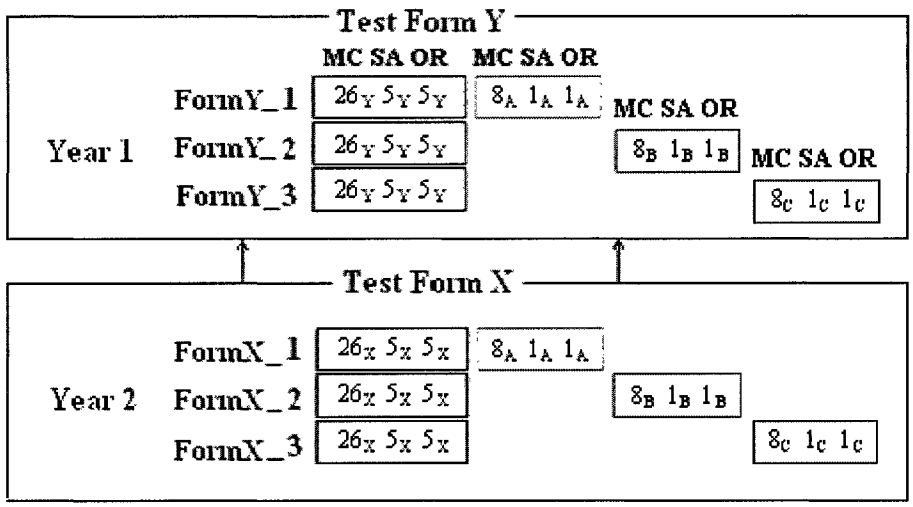

Figure 12. The number and the types of items in each test form.

The numbers of different types of items are presented for each content area in Table 1. For example, in the content area of number sense, there are 10 MC, 1 SA, and 1 OR unique items and 9 MC, 2 SA, and 2 OR common items. The number of items varies across content areas.

Table 1

The Number of Unique and Common Items in Each Content Area

Content Area	Item	MC	SA	OR	Score-points
Number Sense	Unique Items	10	1	1	15
	Common Items	9	2	2	19
Patterns, Relations, and Functions	Unique Items	6	1	1	11
	Common Items	5	0	1	9
Statistics and Probability	Unique Items	5	1	1	10
	Common Items	5	0	0	5
Geometry	Unique Items	3	1	1	8
	Common Items	3	0	0	3
Measurement	Unique Items	2	1	1	7
	Common Items	2	1	0	3

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Fixed and Manipulated Factors

The accuracy of IRT-based equating for the common-item non-equivalent groups matrix design might be influenced by factors such as sample size, the IRT model used for the estimation of parameters, the computer program used for parameter estimation, the number/score-points and the representativeness of common items, the characteristics of outliers, group ability differences, and equating method. In the current study, the first three factors were fixed and, therefore, were not examined. The remaining factors were manipulated.

Sample Size

The sample size for each sub test form was controlled at 2000, which has been suggested as large enough to produce stable parameter estimates (e.g., Kolen & Brennan, 1995; Zeng, 1991). Consequently, the total sample size for each of the test forms X and Y was 6000 (2000 x 3 = 6000), respectively.

IRT Models

Three IRT models, 3PL, 2PL, and GRM, were chosen for modelling the data generated for the test form X and test form Y. The 3PL model was used for modelling the multiple-choice items. This is based on the observation that it is always possible for an examinee to answer a multiple-choice item correctly by guessing since the answer is provided with the alternatives. The 2PL model was used for modeling the short-answer items. The use of 2PL model is based on the belief that the possibility of answer a short-answer question correctly by guessing is close to zero, and it is reasonable to assume the discrimination power for all the items are different. The rationale for choosing the extended GRM to model the

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

polytomously scored items is that (1) the scores for the math open-ended items are ordered; (2) the adjacent scores, for example, 2, 3, and 4, can be collapsed as one category, if necessary; and (3) it is meaningful to know the possibility of getting a higher score over getting a lower score.

Computer Programs

Several computer programs have been developed for parameter estimation. For example, once the data for two equating groups are collected, the parameters can be estimated through IRT software packages such as LOGIST (Wingersky, Barton, & Lord, 1982), BILOG (Mislevy & Bock, 1990), BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996), MULTILOG (Thissen, 1991), or PARSCALE (Muraki & Bock, 1999). The last two programs can be used for polytomously scored items. These programs were thought as one factor that might influence the equating results (e.g., Hanson & Beguin, 2002). However, after comparing the performance of MULTILOG and BILOG-MG in the context of concurrent calibration and separate/linear transformation, Hanson and Beguin (1999) concluded that the two programs tended to perform similarly. Childs and Chen (1999) found that although the MULTILOG and PARSCALE parameterize the polytomous IRT models differently (for example, MULTILOG uses GRM while PARSCALE uses the extended GRM as described in Chapter 2), similar parameter estimates were obtained. Thus, only PARSCALE was used to estimate the parameters.

The PARSCALE control files were altered for the concurrent, separate, and FCIP calibrations. The sample control files for each of these methods were presented in Appendixes A, B, and C.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

The Number/Score-points and the Representativeness of Common Items

The number and the representativeness of common items were originally well controlled. Li et al. (1999) and Tate (2000) suggested that while determining the number of common items, a polytomously scored item can be treated as several dichotomously scored items that are equal to the number of score categories. Since the tests simulated in the current study have open-ended response items, the score-points rather than the number of items was considered. However, the number of items was also reported as a reference. The ratios expressed as percentage of common items over the unique items based on the score-points and the number of items are listed in Table 2. As shown in Table 2, the total percentage of the score-points of the common items over that of the unique items was about 76.5%; and the percentage of the number of common items over the number of unique items was about 83.3%. In terms of content areas, the smallest percentage based on the score-points was 37.5% for the content of geometry. Table 2 indicates that the number or the score-points of the common items was originally large enough to represent the unique items.

Table 2

Percentages of the Common Items over the Unique Items

Content Area	No. of Score-points of the Common Items / No. of Score-points of the Unique items	No. of the Common Items / No. of the Unique Items
Number Sense	126.7%	108.3%
Patterns, Relations, and Functions	81.8%	75.0%
Statistics and Probability	50.0%	71.4%
Geometry	37.5%	60.0%
Measurement	42.9%	75.0%
Total	76.5%	83.3%

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

The means and the standard deviations of the *b*-parameters of the unique and common items based on the score-points and the number of items are listed in Table 3. As indicated in the table, the *b*-parameters of the common items (or anchor test) represented the *b*-parameters of the two unique tests. The mean difficulties of the three sets of common items in each test form were similar, which indicated that the difficulties of each sub test form were similar.

Table 3

The Descriptive Statistics of the b-parameters of the Unique and the Common Items

Item	Based on the Score-points			Based on the Number of Items		
	Total Score-points	Mean	S.D.	Total Number of Items	Mean	S.D.
U1	51	-0.0595	1.1130	36	-0.3074	0.6902
U2	51	-0.0460	1.1648	36	-0.3061	0.7348
Common	39	-0.1220	0.8599	30	-0.2942	0.7525
C_1	13	-0.0963	0.8631	10	-0.2919	0.6388
C_2	13	-0.1302	0.8622	10	-0.2888	0.8373
C_3	13	-0.1395	0.9231	10	-0.3019	0.8459

However, these properties were changed when the outliers and the IRT-based equating methods were manipulated. For example, if the outliers were from one content area and they were removed while conducting equating, then the content representativeness of common items would be violated.

Characteristics of Outliers

The presence of outliers may be due to a change in the instructional emphasis of a certain content area, the revelation of some common items, and item parameter estimation error (especially for items with extreme *b*-values). Thus, the manipulation of outliers reflected these possible reasons. For example, some outliers were from one content area to reflect the first possible reason, which meant that the content

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

representativeness was altered in this manipulation. That is to say, by excluding all the outliers from one content area, the content representativeness was violated. Based on the characteristics of the data used in the current study, the following six conditions were examined:

1. There was no outlier in the common items.
2. The outliers were 3 MC items with 3 score-points. These items were from one content area.
3. The outliers were 3 MC items with 3 score-points. These items were randomly selected from any of the five content areas.
4. The outliers were 5 MC items and 1 OR items with 9 score-points. These items were from one content area.
5. The outliers were 5 MC items and 1 OR items with 9 score-points. These items were randomly selected from any of the five content areas.
6. The outliers were 3 MC items with 3 score-points. These items could be either very difficult or very easy. However, in the current study, all of them were very easy given the available properties of the *b*-parameters.

The first condition served as the baseline of comparison. Conditions 2 to 4 reflected the increase of the number/score-points of outliers and the first two possible reasons for the presence of outliers. Since the most possible result of the instruction emphasis on one content area and the exposure of some common items is that the corresponding items become easier when they are administered in the second year, only the outliers located on the left side of the straight line on the scatter plot of *b*-parameters

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

were examined in the current study. The sixth condition represents the situation that estimation error leads to the presence of outliers.

Group Ability Differences

Group ability difference was the second factor manipulated in the current study. Theoretically, in the common-item non-equivalent groups matrix design, equating is only needed when the groups taking the two tests are non-equivalent. However, for the purpose of comparison, equating was conducted for the situations with equivalent and non-equivalent groups in the current study. Samples for the item responses for test form Y were generated by sampling the latent trait (θ) from a normal distribution with mean zero and standard deviation one (N (0, 1)). Two sets of item responses were generated for test form X by sampling θ from an N (0, 1) distribution and an N (1, 1) distribution. The samples with N (0, 1) for test form Y and test form X were used to examine the case that the two groups were equivalent. The samples with N (0, 1) for test form Y and with N (1, 1) for test form X were used to examine the case where the two groups were not equivalent. This factor was fully crossed with the six conditions defined by the outliers. Thus, a total of 12 (2 x 6 = 12) response conditions were produced.

IRT-based Equating Methods

The IRT-based equating method was the third factor manipulated in the current study. The following ten equating methods were conducted for each of the 12 response conditions:

1. concurrent calibration with outliers included (i.e., this method did not consider the influence of outliers by including the responses to outliers in the data file used for the concurrent calibration; i.e., the outliers were ignored)

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

2. concurrent calibration with outliers excluded (i.e., this method considered the influence of outliers by excluding the responses to outliers from the data file used for the concurrent calibration)
3. TCC transformation with outliers included
4. TCC transformation with outliers excluded
5. M/S transformation with outliers included
6. M/S transformation with outliers excluded
7. M/S transformation with outliers weighted
8. FCIP calibration with outliers fixed
9. FCIP calibration with outliers not fixed
10. FCIP calibration with outliers excluded

Note that for the response conditions with no outliers, the above ten methods reduced to four methods: concurrent calibration, TCC transformation, M/S transformation, and FCIP calibration. As a result, a total of 108 (1 x 2 x 4 ("0" outlier by group ability differences by equating method) + 5 x 2 x 10 (5 combinations of outliers' characteristics by group ability differences by equating method) = 108) conditions were examined in the current study.

Computer Simulation

Simulation studies are actually statistically sampling experiments with an underlying model whose results are used to address research questions (Robinstein, 1981). Using simulated data to study statistical problems can be found in the early 1900s. With the development of high-speed computers, the computer-based simulation became a popular and a formal research method for solving statistical problems (Harwell, Stone,

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Hsu, & Kirisci, 1996). Computer simulations have played an important role for studying the properties of the IRT models and their applications. For example, many studies have employed this method to investigate IRT-based equating methods (e.g., Baker, 1992, 1996; Bolt, 1999; Cohen & Kim, 1998; Hanson & Beguin, 2002; Li, et al. 1997; Li, et al. 1999; Wang, Hanson, & Harris, 2000).

Lehman and Bailey (1968) pointed out that a computer simulation might be conducted when an experiment study in the real world is too costly or impossible. In the publication policy of *Psychometrika* (Psychometric Society, 1979), it is pointed that simulation studies should be employed only if the information cannot reasonably be obtained in other ways, for example, in an analytical way. These reasons support the use of computer simulation in the current study. It is difficult for a researcher to collect real data to represent the outliers of interest. More importantly, it is almost impossible to pursue a research question such as which IRT-based equating method produces a more accurate result using real data due to the lack of definite evaluation criteria. A simulation study can solve these problems. For example, the different conditions of interest can be reflected in the simulated data or implied in the simulation process. The equating results of the IRT-based equating methods can be compared with the true scores that are known before the simulated data are generated. Therefore, the accuracy of the performance of the IRT-based equating methods can be compared.

In the current study, computer simulation was used. The steps followed were:

1. For test form Y, an item response sample was generated for each of the 6 outlier conditions that had an underlying theta distribution of N (0, 1).

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

2. For test form X, an item response sample was generated for each of the 6 outlier conditions with an underlying theta distribution of N (0, 1); another item response sample was generated for each of the 6 outlier conditions with an underlying theta distribution of N (1, 1). The samples for the test form X were paired with the samples for the test form Y to represent the 12 response conditions.
3. The response samples for the two test forms were calibrated and/or equated by the four IRT-based equating methods or their ten variations listed on Pages 66 and 67. Each of these methods was followed by the IRT true score equating.
4. This process was replicated 50 times, which is thought to be sufficient to compare the results obtained from Step 3 (Hanson & Belguin, 2002; Harwell et al., 1996).

Evaluation of the IRT-based Equating Methods

The purpose of the current study was to compare the IRT-based equating methods in the presence of outliers. As described in Chapter 1, four specific research questions were studied: (1) Do the IRT-based equating methods that consider the influence of outliers produce a better result than the IRT-based equating methods that do not consider the influence of outliers? (2) Is the effect found in Question 1, if any, confounded by factors such as the characteristics of outliers and the group ability differences? (3) Which of the IRT-based equating methods produces a better result, especially among the IRT-based equating methods that consider the influence of outliers? (4) Is the effect found in Question 3, if any, confounded by factors such as the characteristics of outliers and the group ability differences?

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

To answer these questions, the unweighted mean square error for the *b*-parameters (MSE_b) and the unweighted mean square error for the number-correct true scores (MSE_t) were calculated for the 108 conditions:

$$MSE_b = \frac{\sum_{r=1}^{50}\sum_{j=1}^{36}\sum_{k=0}^{mj}(b^*_{jkr} - b_{jk})^2}{50\times 51} \tag{17}$$

and

$$MSE_t = \frac{\sum_{r=1}^{50}\sum_{s=0}^{51}(t^*_{sr} - \tau_s)^2}{50\times 52} \tag{18}$$

where

m_j is the number of categories minus 1 for item j,

b^*_{jkr} is the *b*-parameter for the unique item j category k in the equated test, say Test X, that has been equated on the reference test, say Test Y, for replication r,

b_{jk} is the true values of the *b*-parameter for item j category k,

t^*_{sr} is the number-correct true score at score point s in the equated test, say Test X, that has been equated on the reference test, say Test Y, for replication r,

and τ_s is the true number-correct true score at score point s.

The mean square errors were further decomposed into systematic errors (MSE_b_SE and MSE_t_SE) and random errors (MSE_b_RE and MSE_t_RE). The systematic errors were calculated by:

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

$$MSE_{b\_SE} = \frac{\sum_{j=1}^{36}\sum_{k=0}^{mj}(\bar{b}^{*}_{jk} - b_{jk})^2}{51}$$

and

$$MSE_{t\_SE} = \frac{\sum_{s=0}^{51}(\bar{t}^{*}_{s} - \tau_s)^2}{52} \quad .$$

The random errors were calculated by:

$$MSE_{b\_RE} = \frac{\sum_{r=1}^{50}\sum_{j=1}^{36}\sum_{k=0}^{mj}(b^{*}_{jkr} - \bar{b}^{*}_{jk})^2}{50\times 51}$$

and

$$MSE_{t\_RE} = \frac{\sum_{r=1}^{50}\sum_{s=0}^{51}(t^{*}_{sr} - \bar{t}^{*}_{s})^2}{50\times 52} \quad ,$$

where

$\bar{b}^{*}_{jk}$ is the mean of the b-parameters for the unique item j category k in the equated test that have been equated on the reference test across 50 replications,

$\bar{t}^{*}_{s}$ is the mean of the number-correct true scores at score point s in the equated test that have been equated on the reference test across 50 replications,

and the other symbols have the same meaning as those in Equations 17 and 18.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

In the current study, the systematic errors for the *b*-parameters and the number-correct-true scores under the 108 conditions were compared to answer the research questions. The justification for using the systematic errors is described in detail in Chapter 4.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

CHAPTER 4: RESULTS

The purpose of this chapter is to present and discuss the results of the analyses conducted to answer the four research questions addressed in this study:

(1) Do the IRT-based equating methods that consider the influence of outliers produce a better result than the IRT-based equating methods that do not consider the influence of outliers?

(2) Is the effect found in Question 1, if any, confounded by factors such as the characteristics of outliers and the group ability differences?

(3) Which of the IRT-based equating methods produces a better result, especially among the IRT-based equating methods that consider the influence of outliers?

(4) Is the effect found in Question 3, if any, confounded by factors such as the characteristics of outliers and the group ability differences?

As described in Chapter 3, three factors were manipulated. These included (a) characteristics of outliers, which were six combination of number/score-points of outliers (0 outliers with score-points 0, 3 MC items with score-points 3, and 5MC items and 1 OR item with score-points 9) and types of outliers (outliers from one content area, randomly from any content area, and with extreme values applied to the 3 score-points condition and the first two types of outliers applied to the 9 score-points condition), (b) group ability differences (equivalent groups (N (0, 1) vs. N (0, 1)) and non-equivalent groups (N (0, 1) vs. N (1, 1)), (b), and (c) IRT-based equating methods (four methods with no outliers present and ten variations in the presence of outliers). The evaluation criteria used to compare the IRT-based equating methods were the systematic errors of the *b*-parameters and the number-correct true scores (MSE_b_SE and MSE_t_SE).

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

The error values under various conditions are summarized in Tables 4 to 9. It was found that when the MSE_b and MSE_t values increased greatly, the corresponding MSE_b_SE and MSE_t_SE values also increased greatly. In contrast, the corresponding MSE_b_RE and MSE_t_RE values did not change too much. For example, when no outliers present in the data set, the MSE_b, MSE_b_SE, and MSE_b_RE values for the concurrent calibration under the equivalent groups condition were 0.0157, 0.0108, and 0.0049 respectively (see left side, Panel A, Table 4, p. 76); the MSE_b, MSE_b_SE, and MSE_b_RE values for the concurrent calibration under the non-equivalent groups condition were 0.6907, 0.6836, and 0.0071 (see left side, Panel B, Table 4). This example indicated that the change of mean square errors was mainly due to the change of systematic errors. Besides, theoretically, systematic errors reflect the magnitudes of the bias introduced by specific equating methods and do not decrease as the sample size increases (Kolen & Brennan, 1995). Thus, the MSE_b_SE and MSE_t_SE were used to (a) evaluate the accuracy of the equating methods and (b) to make comparisons among the methods.

In most equating simulation studies reported in the literature, the relative magnitudes of error values were compared to determine the relative accuracy of the equating methods considered. However, the use of relative magnitudes of systematic errors in the current study was problematic. For example:

(1) The MSE_t_SE values for the M/S transformation and FCIP calibration in the presence of no outliers and with equivalent groups were 0.0859 and 0.0858 respectively (see Panel A, Table 4). Based on the relative criteria, one may conclude that the FCIP calibration performed better than the M/S

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

transformation since the former method had a smaller MSE_t_SE value than the latter one. However, should both the MSE_t_SE values of 0.0859 and 0.0858 be treated as small and the difference of 0.0001 be ignored?

(2) The MSE_t_SE values for the TCC transformation with outliers included and concurrent calibration with outliers included under the condition of 3 outliers with extreme values and equivalent groups were 3.7795 and 0.6026 respectively (see Panel A, Table 7, p.89). Based on the relative criteria, one may conclude that the concurrent calibration with outliers included performed better than the TCC transformation with outliers included since the former method had a smaller MSE_t_SE value than the latter one. While the conclusion may sound reasonable for this case, it likely makes no sense in the following case.

(3) The MSE_t_SE values for the M/S transformation with outliers included and the TCC transformation with outliers included under the condition of outliers with 9 score-points from one content area and equivalent groups were 19.5356 and 16.2440 respectively (see Panel A, Table 8, p.91). Based on the relative criteria, one may conclude that the TCC transformation with outliers included performed better than the M/S transformation with outliers included. However, should one care about the relative performance of these methods? The answer is probably no given the MSE_t_SE values were large, perhaps too large. Consequently, neither the TCC nor the M/S transformations with outliers included would be recommended in this situation.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

These examples lead to the question: what should the size of MSE_b_SE and MSE_t_SE be to claim the systematic error is small, moderate, or large? To answer this question and to make the discussion consistent, absolute rules for interpreting the sizes of the systematic errors for the *b*-parameters and the number-correct true scores were developed.

Rules for MSE_b_SE and MSE_t_SE

The development of rules was based on the magnitude of the square root of systematic errors (referred to as bias), which represents the difference between the observed *b*-parameters or the number-correct true scores and their corresponding true *b*-parameters or number-correct true scores. For the *b*-parameter, the bias values of 0.2500, one-fourth of the standard deviation of the distribution of the *b*-parameters, and 0.5000, one-half of the standard deviation of the distribution of the *b*-parameters, were adopted as the cut-off scores. These values correspond to 0.0625 and 0.2500 in the metric of mean square errors. Consequently, the rules for the MSE_b_SE are: (a) $MSE_b\_SE \leq 0.06$ is considered as small; (b) $0.06 < MSE_b\_SE \leq 0.25$ is considered as moderate; (c) $MSE_b\_SE > 0.25$ is considered as large. The MSE_b_SE values were rounded to two decimal points to avoid the situations when a MSE_b_SE value is placed in a higher category due to a small difference from a cut-off value. The rules for the MSE_t_SE are: (a) $MSE_t\_SE \leq 2.25$ is considered as small; (b) $2.25 < MSE_t\_SE \leq 6.25$ is considered as moderate; (c) $MSE_t\_SE > 6.25$ is considered as large. As for the case of MSE_b_SE values, two decimal points were used in judging the size of MSE_t_SE values.

In the following sections, the systematic errors for the conditions with no outliers are reported and discussed first. Although these results do not answer any of the research

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

questions, they serve as the baseline for the comparison of IRT-based equating methods in the presence of outliers. The results for the conditions with outliers are then presented and discussed.

No Outliers Present in the Data Set

The results of the current study revealed that there were differences between the results of the four IRT-based equating methods when there was no group ability difference and the results when the two equating groups differed by one standard deviation of ability. Hence, the results for the two equivalent equating groups and the two non-equivalent equating groups are discussed separately. For each type of group, the systematic errors for the *b*-parameters are reported first, and then the systematic errors for the number-correct true scores are presented.

Equivalent Equating Groups

Item Difficulty b

As shown in Panel A, Table 4, the MSE_b_SE values for each of the four equating methods were small ($MSE_b\_SE \leq 0.06$) when the two equating groups were equivalent. Further, rounding each value to two decimal points yielded the same result, 0.01, which indicated that the four methods were equally accurate as determined by the criterion of MSE_b_SE.

Number-correct True Score t

Turning to the number-correct true scores, the MSE_t_SE values for each of the four equating methods were small (see Panel A, Table 4, $MSE_t\_SE \leq 2.25$). This result indicated that the four equating methods performed equally accurate as determined by the criterion of MSE_t_SE.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Table 4

Mean Square Total, Systematic, and Random Errors of the four IRT-based Equating Methods

Method	Item Difficulty *b*				Number-correct True Score *t*			
	MSE_b	MSE_b_SE		MSE_b_RE	MSE_t	MSE_t_SE		MSE_t_RE
A. Equivalent Groups (N(0,1) vs. N(0,1))								
Concurrent	0.0157	0.0108	(S)	0.0049	0.1550	0.0459	(S)	0.1091
TCC	0.0189	0.0108	(S)	0.0081	0.2285	0.0646	(S)	0.1638
M/S	0.0176	0.0110	(S)	0.0067	0.2754	0.0859	(S)	0.1895
FCIP	0.0244	0.0118	(S)	0.0127	0.2277	0.0858	(S)	0.1419
B. Non-equivalent Groups (N(0,1) vs. N(1,1))								
Concurrent	0.6907	0.6836	(L)	0.0071	2.7538	2.6989	(M)	0.0549
TCC	0.0699	0.0562	(S)	0.0136	1.6142	1.3888	(S)	0.2255
M/S	0.0288	0.0191	(S)	0.0097	0.3127	0.0960	(S)	0.2167
FCIP	0.1385	0.1302	(M)	0.0083	2.7546	2.7021	(M)	0.0525

Note: S, M, and L represent the size of systematic errors. S refers to small; M refers to moderate; and L refers to large.

Non-equivalent Equating Groups

Item Difficulty b

When the two equating groups differed by one standard deviation of ability, the MSE_b_SE values varied (see Panel B, Table 4). The TCC and M/S transformations had small MSE_b_SE values; the FCIP calibration had a moderate MSE_b_SE value; and the concurrent calibration had a large MSE_b_SE value.

Compared to the results for the equivalent groups (Panel A, Table 4), there was more variability among the systematic errors across the four equating methods when the two equating groups differed by one standard deviation of ability. The MSE_b_SE values for the TCC and M/S transformations were small regardless of the group equivalence. For the FCIP calibration, the MSE_b_SE value increased to moderate when the group ability difference increased to one standard deviation. The MSE_b_SE value for the concurrent

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

calibration was small under the equivalent groups condition but large under the non-equivalent groups condition.

Number-correct True Score t

As shown in Panel B, Table 4, the values for MSE_t_SE varied. The TCC and M/S transformations had small MSE_t_SE values. The concurrent and FCIP calibrations had moderate MSE_t_SE values.

Compared to the results for the equivalent groups (Panel A, Table 4), there was more variability among the MSE_t_SE values across the four equating methods when the two equating groups differed by one standard deviation of ability. The TCC and M/S transformations had small MSE_t_SE values regardless of the equivalence of the equating groups. The MSE_t_SE values for the concurrent and FCIP calibrations were small for the equivalent groups but moderate for the non-equivalent groups.

To summarize, while the four equating methods produced comparable estimates of the item difficulties and number-correct true scores when the two equating groups were equivalent, the same cannot be said when the two equating groups were not equivalent. The four methods were sensitive, but not equally, to the presence of non-equivalent groups.

Presence of Outliers

When outliers are present in the data set, the four IRT-based equating methods can still be used to equate the two tests by simply ignoring the influence of outliers. However, six variations of the four methods that take into account the presence of outliers had been proposed in the literature in an attempt to reduce the adverse influence of outliers. To emphasize the presence of outliers and the difference between the methods

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

that considered the influence of outliers and that did not consider the influence of outliers, the four basic equating methods were renamed as concurrent calibration with outliers included, TCC transformation with outliers included, mean/sigma transformation with outliers included, and FCIP calibration with outliers fixed. The corresponding methods that considered the influence of outliers were named as concurrent calibration with outliers excluded, TCC transformation with outliers excluded, mean/sigma transformation with outliers excluded, mean/sigma transformation with outliers weighted, FCIP calibration with outliers not fixed, and FCIP calibration with outliers excluded. The systematic errors for the ten methods are summarized in Tables 5 to 7 for the conditions involving outliers with 3 score-points and in Tables 8 and 9 for the conditions involving outliers with 9 score-points.

As for the case when there were no outliers present in the data, there were differences between the results of the ten IRT-based equating methods when there were no group ability differences and the results when the groups' ability differed by one standard deviation under the condition of outliers with 3 score-points from one content area (see Table 5, p.80). Consequently, the results for the equivalent groups and non-equivalent groups are discussed separately, with the results presented first for the *b*-parameters and then for the number-correct true scores.

Further, the results for the condition in which the outliers with 3 score-points were from one content area (Table 5), were randomly from any content area (Table 6, p.88), and had extreme values (Table 7, p.89) were similar. However, these results differed from the results for the condition of outliers with 9-score points from one content area (Table 8, p.91) and randomly from any content area (Table 9, p.101), which were similar

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

to each other. Therefore, the results are discussed in detail for the condition of outliers with 3 score-points from one content area, followed by a brief discussion of the results for the conditions of outliers with 3 score-points randomly from any content area and with extreme values. Likewise, the results are discussed in detail for the condition of outliers with 9 score-points from one content area, followed by a brief discussion of the results for the condition with 9 score-points randomly from any content area.

Outliers with 3 Score-points from One Content Area

Equivalent Equating Groups

Item Difficulty b

As shown in Panel A1, Table 5, the concurrent calibration with outliers included and the FCIP calibration with outliers fixed had small MSE_b_SE values; the TCC transformation with outliers included and the M/S transformation with outliers included had moderate MSE_b_SE values. In contrast, all the MSE_b_SE values for the methods that considered the influence of outliers in Panel A2, Table 5 were small.

Comparison of the MSE_b_SE values in Panels A1 and A2, Table 5. It was found that the methods that did not consider the influence of outliers had either small or moderate MSE_b_SE values. In contrast, all the methods that considered the influence of outliers had small MSE_b_SE values. More specifically, for the concurrent calibration, including outliers and excluding outliers produced small MSE_b_SE values. Likewise, for the FCIP calibration, fixing, not fixing, and excluding outliers produced small MSE_b_SE values. However, for the TCC and M/S transformations, without considering the influence of outliers produced moderate MSE_b_SE values while considering the influence of outliers produced small MSE_b_SE values.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Table 5

Mean Square Total, Systematic, and Random Errors for the Ten IRT-based Equating Methods under the Condition of Outliers with 3 Score-points and From One Content Area

Method	Item Difficulty *b*				Number-correct True Score *t*			
	MSE_b	MSE_b_SE		MSE_b_RE	MSE_t	MSE_t_SE		MSE_t_RE
A. Equivalent Groups (N(0,1) vs. N(0,1))								
A1. Methods that Did Not Consider the Influence of Outliers								
C+include	0.0194	0.0154	(S)	0.0041	0.8131	0.7366	(S)	0.0765
TCC+include	0.1008	0.0907	(M)	0.0101	3.9423	3.6257	(M)	0.3166
M/S+include	0.1241	0.1179	(M)	0.0062	10.2838	9.5457	(L)	0.7381
FCIP+fixed	0.0345	0.0303	(S)	0.0042	0.9192	0.8197	(S)	0.0995
A2. Methods that Considered the Influence of Outliers								
C+exclude	0.0150	0.0106	(S)	0.0044	0.2042	0.0648	(S)	0.1394
TCC+exclude	0.0203	0.0104	(S)	0.0099	0.3023	0.0687	(S)	0.2336
M/S+exclude	0.0163	0.0105	(S)	0.0057	0.3285	0.1103	(S)	0.2182
M/S+weight	0.0170	0.0104	(S)	0.0066	0.4073	0.0944	(S)	0.3129
FCIP+nofixed	0.0166	0.0105	(S)	0.0061	0.2142	0.0733	(S)	0.1409
FCIP+exclude	0.0167	0.0107	(S)	0.0061	0.1992	0.0660	(S)	0.1332
B. Non-equivalent Groups (N(0,1) vs. N(1,1))								
B1. Methods that Did Not Consider the Influence of Outliers								
C+include	0.5735	0.5673	(L)	0.0062	1.2162	1.0825	(S)	0.1337
TCC+include	0.1814	0.1697	(M)	0.0117	5.9277	5.6282	(M)	0.2995
M/S+include	0.1048	0.0901	(M)	0.0147	7.2565	6.2887	(L)	0.9679
FCIP+fixed	0.0790	0.0718	(M)	0.0072	1.7454	1.6267	(S)	0.1187
B2. Methods that Considered the Influence of Outliers								
C+exclude	0.7170	0.7102	(L)	0.0068	3.4654	3.3607	(M)	0.1047
TCC+exclude	0.0635	0.0497	(S)	0.0137	1.5154	1.2560	(S)	0.2594
M/S+exclude	0.0259	0.0168	(S)	0.0091	0.2354	0.0726	(S)	0.1628
M/S+weight	0.4591	0.4545	(L)	0.0046	35.0008	34.5878	(L)	0.4130
FCIP+nofixed	0.1552	0.1473	(M)	0.0078	3.4173	3.3135	(M)	0.1038
FCIP+exclude	0.1576	0.1497	(M)	0.0078	3.5069	3.4004	(M)	0.1065

Note: C+include—concurrent calibration with outliers included; TCC+include—TCC transformation with outliers included; M/S include—M/S transformation with outliers included; FCIP+fixed—FCIP calibration with outliers fixed; C+exclude—concurrent calibration with outliers excluded; TCC+exclude—TCC transformation with outliers excluded; M/S+exclude—M/S transformation with outliers excluded; M/S+weight—M/S transformation with outliers weighted; FCIP+nofixed—FCIP calibration with outliers not fixed; and FCIP+exclude—FCIP calibration with outliers excluded.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Comparison of the MSE_b_SE values in Panels A1 and A2, Table 5 and in Panel A, Table 4. When the systematic errors for the *b*-parameters in Panel A1, Table 5 were compared with the systematic errors in Panel A, Table 4, it was found that the MSE_b_SE values for the concurrent calibration with outliers included and the FCIP calibration with outliers fixed and for the two corresponding methods under the condition of no outliers were small. In contrast, while the TCC and M/S transformations with outliers included had moderate MSE_b_SE values, they had small MSE_b_SE values under the condition of no outliers. The MSE_b_SE values for the methods that considered the influence of outliers (Panel A2, Table 5) and the four methods under the condition of no outliers (Panel A, Table 4) were all small.

Number-correct True Score t

As shown in Panel A1, Table 5, the concurrent calibration with outliers included and FCIP calibration with outliers fixed had small MSE_t_SE values; the TCC transformation with outliers included had a moderate MSE_t_SE value; and the M/S transformation with outliers included had a large MSE_t_SE value. In contrast, all the methods in Panel A2, Table 5 that considered the influence of outliers had small MSE_t_SE values.

Comparison of the MSE_t_SE values in Panels A1 and A2, Table 5. While the MSE_t_SE values for the methods that did not consider the influence of outliers varied from large to small, they were all small for the methods that considered the influence of outliers. More specifically, for the concurrent calibration, both including and excluding outliers produced small MSE_t_SE values. Likewise, for the FCIP calibration, fixing, not

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

fixing, and excluding outliers produced small MSE_t_SE values. However, for the TCC transformation, including outliers produced a moderate MSE_t_SE value while excluding outliers produced a small MSE_t_SE value. For the M/S transformation, including outliers produced a large MSE_t_SE value while excluding outliers and weighting outliers produced a small MSE_t_SE value.

Comparison of the MSE_t_SE values in Panels A1 and A2, Table 5 and in Panel A, Table 4. When the systematic errors for the number-correct true scores in Panel A1, Table 5 were compared with the corresponding systematic errors in Panel A, Table 4, it was found that the concurrent and FCIP calibrations had small MSE_t_SE values in both tables. The TCC and M/S transformations with outliers included had moderate and large MSE_t_SE values respectively but small MSE_t_SE values under the condition of no outliers. In contrast, the MSE_t_SE values in Panel A2, Table 5 for each method that considered the influence of outliers were small as were the corresponding values in Panel A, Table 4.

Non-equivalent Equating Groups

Item Difficulty b

As shown in Panel B1, Table 5, the TCC transformation with outliers included, the M/S transformation with outliers included, and the FCIP calibration with outliers fixed had moderate MSE_b_SE values; and the concurrent calibration with outliers included had a large MSE_b_SE value. Examination of the MSE_b_SE values in Panel B2, Table 5 revealed that: the MSE_b_SE values for the TCC transformation with outliers excluded and the M/S transformation with outliers excluded were small; the MSE_b_SE

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

value for the FCIP calibration with outliers not fixed and excluded were moderate; and the MSE_b_SE values for the concurrent calibration with outlier excluded and M/S transformation with outliers weighted were large.

Comparison of the MSE_b_SE values in Panels B1 and B2, Table 5. Comparison of these two sets of results revealed that: for the TCC transformation, including outliers produced a moderate MSE_b_SE value while excluding outliers produced a small MSE_b_SE value. For the M/S transformation, including outliers produced a moderate MSE_b_SE value. However, excluding outliers produced a small MSE_b_SE value and weighting outliers produced a large MSE_b_SE value for this transformation. For the FCIP calibration, fixing, without fixing, and excluding outliers produced moderate MSE_b_SE values. Lastly, for the concurrent calibration, including and excluding outliers produced large MSE_b_SE values.

Comparison of the MSE_b_SE values in Panels B1 and B2, Table 5 and in Panel B, Table 4. When the systematic errors for the *b*-parameters in Panel B1, Table 5 were compared with the systematic errors in Panel B, Table 4, it was found that for the TCC and M/S transformations, including outliers produced moderate MSE_b_SE values. However, the MSE_b_SE values for these two methods were small under the condition of no outliers. The concurrent and FCIP calibrations without considering the influence of outliers resulted in large and moderate MSE_b_SE values. These two calibrations also had large and moderate MSE_b_SE values under the condition of no outliers. In contrast, with the exception of the M/S transformation, the MSE_b_SE values for the other three methods that considered the influence of outliers in Panel B2, Table 5 were the same size

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

as those in Panel B, Table 4. For the M/S transformation, excluding outliers produced as small MSE_b_SE value as that under the condition of no outliers present in the data set. However, weighting outliers produced a large rather than a small MSE_b_SE value.

Comparison of the MSE_b_SE values in Panels A and B, Table 5. For the methods that did not consider the influence of outliers, the FCIP calibration with outliers fixed had a small MSE_b_SE value under the equivalent groups condition but a moderate MSE_b_SE value under the non-equivalent groups condition; the concurrent calibration had a small MSE_b_SE value under the equivalent groups condition but a large MSE_b_SE value under the non-equivalent groups; and the TCC and M/S transformations with outliers included had moderate MSE_b_SE values regardless of the equivalence of the groups. All the methods that considered the influence of outliers had small MSE_b_SE values under the equivalent groups condition. However, their MSE_b_SE values varied under the non-equivalent groups condition: the MSE_b_SE values for the TCC and M/S transformations with outliers excluded were small; the MSE_b_SE values for the FCIP calibration with outliers not fixed and excluded were moderate; and the MSE_b_SE values for the concurrent calibration with outliers excluded and M/S with outliers weighted were large.

Number-correct True Score t

As shown in Panel B1, Table 5, the concurrent calibration with outliers included and FCIP calibration with outliers fixed had small MSE_t_SE values; the TCC transformation with outliers included had a moderate MSE_t_SE value; and the M/S transformation with outliers included had a large MSE_t_SE value. For the methods that

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

considered the influence of outliers (see Panel B2, Table 5), the TCC and M/S transformations with outliers excluded had small MSE_t_SE values, however, the M/S transformation with outliers weighted had a large MSE_t_SE value; and the concurrent calibration with outliers excluded and FCIP calibration with outliers not fixed and excluded had moderate MSE_t_SE values.

Comparison of the MSE_t_SE values in Panels B1 and B2, Table 5. The concurrent and FCIP calibrations, without considering the influence of outliers, produced small MSE_t_SE values while considering the influence of outliers produced moderate MSE_t_SE values. For the TCC transformation, including outliers produced a moderate MSE_t_SE value while excluding outliers produced a small MSE_t_SE value. Lastly, for the M/S transformation, including and weighting outliers produced large MSE_t_SE values, while excluding outliers produced a small MSE_t_SE value.

Comparison of the MSE_t_SE values in Panels B1 and B2, Table 5 and in Panel B, Table 4. When the systematic errors for the number-correct true scores in Panel B1, Table 5 were compared with the corresponding systematic errors in Panel B, Table 4, it was found that, for the concurrent and FCIP calibrations, the MSE_t_SE values were small when the influence of outliers were not considered while the corresponding values were moderate under the condition of no outliers. In contrast, for the TCC and M/S transformations, their MSE_t_SE values were moderate and large when the influence of outliers was not considered but small under the condition of no outliers. When the systematic errors for the number-correct true scores in Panel B2, Table 5 were compared with the corresponding systematic errors in Panel B, Table 4, it was found that, with the

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

exception of the M/S transformation, the MSE_t_SE values for the other three methods that considered the influence of outliers were the same size as those under the condition of no outliers. For the M/S transformation, the MSE_t_SE values were small when the outliers were excluded and no outliers were present but large when the outliers were weighted.

Comparison of the MSE_t_SE values in Panels A and B, Table 5. The MSE_t_SE values for the methods that did not consider the influence of outliers under the equivalent groups condition were the same size as those under the non-equivalent groups condition. In contrast, with the exception of the TCC and M/S transformations with outliers excluded, while the methods that considered the influence of outliers under the equivalent groups condition had small MSE_t_SE values, they had either large or moderate MSE_t_SE values under the non-equivalent groups condition. For the TCC and M/S transformations with outliers excluded, the MSE_t_SE values were small regardless of the equivalence of equating groups.

Outliers with 3 Score-points

Randomly from Any Content Area and with Extreme Values

As mentioned previously, the results for the condition in which the outliers with 3 score-points were from one content area, were randomly from any content area, and had extreme values were similar, which indicated that the performance of the IRT-based equating methods was not confounded by the types of outliers (see Tables 5, 6, and 7). Hence, all the observations made for the condition of outliers with 3 score-points from one content area can be applied to the conditions of outliers with 3 score-points randomly from any content area and with extreme values.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Summary

When the equating groups were equivalent, whether the methods that considered the influence of outliers performed better, as determined by the MSE_b_SE and MSE_t_SE values, than the methods that did not consider the influence of outliers depended on the specific method. For the TCC and M/S transformations, considering the influence of outliers resulted in small systematic errors while without considering the influence of outliers resulted in moderate or large systematic errors. However, for the concurrent and FCIP calibrations, the systematic errors were small no matter the influence of outliers were considered or not. All the methods that considered the influence of outliers had small systematic errors, which indicated that they all performed equally well.

When the equating groups were not equivalent, not all of the systematic errors for the methods that considered the influence of outliers were smaller than the corresponding values for the methods that did not consider the influence of outliers. Thus, caution needs to be taken while deciding whether the methods that considered the influence of outliers performed better than the methods that did not consider the influence of outliers when the equating groups were not equivalent. Among the methods that considered the influence of outliers, only the TCC and M/S transformations with outliers excluded produced small systematic errors. The findings are applicable to the conditions involving outliers with 3 score-points a) from one content area, b) randomly from any content area, and c) with extreme values.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Table 6

Mean Square Total, Systematic, and Random Errors for the Ten IRT-based Equating Methods under the Condition of Outliers with 3 Score-points and Randomly From Any Content Area

Method	Item Difficulty *b*				Number-correct True Score *t*			
	MSE_b	MSE_b_SE		MSE_b_RE	MSE_t	MSE_t_SE		MSE_t_RE
A. Equivalent Groups (N(0,1) vs. N(0,1))								
A1. Methods that Did Not Consider the Influence of Outliers								
C+include	0.0211	0.0164	(S)	0.0047	0.8285	0.7601	(S)	0.0683
TCC+include	0.0836	0.0753	(M)	0.0082	3.0077	2.7929	(M)	0.2148
M/S+include	0.1077	0.1037	(M)	0.0040	7.7596	7.5302	(L)	0.2294
FCIP+fixed	0.0399	0.0346	(S)	0.0053	1.1577	1.0358	(S)	0.1218
A2. Methods that Considered the Influence of Outliers								
C+exclude	0.0165	0.0116	(S)	0.0049	0.1335	0.0413	(S)	0.0922
TCC+exclude	0.0215	0.0118	(S)	0.0096	0.3050	0.0927	(S)	0.2123
M/S+exclude	0.0178	0.0117	(S)	0.0061	0.2238	0.0727	(S)	0.1512
M/S+weight	0.0198	0.0117	(S)	0.0081	0.4381	0.0819	(S)	0.3561
FCIP+nofixed	0.0267	0.0105	(S)	0.0162	0.2742	0.0803	(S)	0.1940
FCIP+exclude	0.0213	0.0118	(S)	0.0094	0.1777	0.0625	(S)	0.1152
B. Non-equivalent Groups (N(0,1) vs. N(1,1))								
B1. Methods that Did Not Consider the Influence of Outliers								
C+include	0.5648	0.5583	(L)	0.0065	1.1700	1.0259	(S)	0.1441
TCC+include	0.2062	0.1923	(M)	0.0139	6.4558	6.0737	(M)	0.3820
M/S+include	0.1412	0.1334	(M)	0.0078	10.3915	9.6805	(L)	0.7111
FCIP+fixed	0.0765	0.0691	(M)	0.0074	1.7081	1.5823	(S)	0.1258
B2. Methods that Considered the Influence of Outliers								
C+exclude	0.7172	0.7099	(L)	0.0072	3.4527	3.3384	(M)	0.1143
TCC+exclude	0.0774	0.0642	(S)	0.0132	1.8550	1.5888	(S)	0.2662
M/S+exclude	0.0272	0.0181	(S)	0.0090	0.2742	0.0824	(S)	0.1919
M/S+weight	0.4563	0.4508	(L)	0.0055	35.5981	35.1571	(L)	0.4409
FCIP+nofixed	0.1556	0.1473	(M)	0.0083	3.4400	3.3267	(M)	0.1133
FCIP+exclude	0.1568	0.1485	(M)	0.0083	3.4712	3.3558	(M)	0.1154

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Table 7

Mean Square Total, Systematic, and Random Errors for the Ten IRT-based Equating Methods under the Condition of Outliers with 3 Score-points and with Extreme Values

Method	Item Difficulty *b*				Number-correct True Score *t*			
	MSE_b	MSE_b_SE		MSE_b_RE	MSE_t	MSE_t_SE		MSE_t_RE
A. Equivalent Groups (N(0,1) vs. N(0,1))								
A1. Methods that Did Not Consider the Influence of Outliers								
C+include	0.0182	0.0137	(S)	0.0045	0.6381	0.6026	(S)	0.0355
TCC+include	0.0909	0.0805	(M)	0.0104	4.1023	3.7795	(M)	0.3228
M/S+include	0.1707	0.1658	(M)	0.0049	16.9815	16.4099	(L)	0.5717
FCIP+fixed	0.0238	0.0192	(S)	0.0046	0.6573	0.6189	(S)	0.0384
A2. Methods that Considered the Influence of Outliers								
C+exclude	0.0168	0.0120	(S)	0.0048	0.1615	0.0538	(S)	0.1078
TCC+exclude	0.0213	0.0120	(S)	0.0093	0.2954	0.0962	(S)	0.1992
M/S+exclude	0.0191	0.0120	(S)	0.0072	0.3092	0.0683	(S)	0.2410
M/S+weight	0.0190	0.0122	(S)	0.0068	0.3654	0.0958	(S)	0.2696
FCIP+nofixed	0.0213	0.0108	(S)	0.0106	0.2054	0.0711	(S)	0.1343
FCIP+exclude	0.0171	0.0122	(S)	0.0049	0.1846	0.0681	(S)	0.1165
B. Non-equivalent Groups (N(0,1) vs. N(1,1))								
B1. Methods that Did Not Consider the Influence of Outliers								
C+include	0.6310	0.6242	(L)	0.0068	2.2896	2.2091	(S)	0.0805
TCC+include	0.1107	0.0909	(M)	0.0198	4.4081	3.7821	(M)	0.6260
M/S+include	0.1405	0.1218	(M)	0.0186	12.4942	10.8171	(L)	1.6771
FCIP+fixed	0.1081	0.1001	(M)	0.0080	2.3669	2.2923	(M)	0.0746
B2. Methods that Considered the Influence of Outliers								
C+exclude	0.7122	0.7047	(L)	0.0074	3.2681	3.1658	(M)	0.1022
TCC+exclude	0.0721	0.0601	(S)	0.0120	1.7085	1.4682	(S)	0.2402
M/S+exclude	0.0286	0.0182	(S)	0.0104	0.3177	0.0860	(S)	0.2316
M/S+weight	0.4590	0.4521	(L)	0.0068	34.9881	34.4702	(L)	0.5179
FCIP+nofixed	0.1480	0.1394	(M)	0.0086	3.1204	3.0315	(M)	0.0889
FCIP+exclude	0.1514	0.1428	(M)	0.0086	3.2292	3.1325	(M)	0.0967

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Outliers with 9 Score-points from One Content Area

The error values for the conditions of outliers with 9 score-points are summarized in Tables 8 and 9. As mentioned previously, the results for the condition of outliers with 9-score points from one content area (Table 8) and randomly from any content area (Table 9, p.101) were similar to each other. Therefore, the results are discussed in detail for the former condition, followed by a brief discussion of the results for the latter condition.

Equivalent Equating Groups

Item Difficulty b

As shown in Panel A1, Table 8, the concurrent calibration with outliers included had a small MSE_b_SE value; the FCIP calibration with outliers fixed had a moderate MSE_b_SE value; and the TCC and M/S transformations with outliers included had large MSE_b_SE values. In contrast, with the exception of the M/S transformation with outliers weighted that had a moderate MSE_b_SE value, the methods that considered the influence of outliers had small MSE_b_SE values (see Panel A2, Table 8).

Comparison of the MSE_b_SE values in Panels A1 and A2, Table 8. The methods that did not consider the influence of outliers had small, moderate, or large MSE_b_SE values. In contrast, with the exception of the M/S transformation with outliers weighted, all the methods that considered the influence of outliers had small MSE_b_SE values. More specifically, for the concurrent calibration, including and excluding outliers produced small MSE_b_SE values. For the TCC transformation, including outliers produced a large MSE_b_SE value while excluding outliers produced a small MSE_b_SE value. For the M/S transformation, including outliers produced a large MSE_b_SE value,

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

excluding outliers produced a small MSE_b_SE value, and weighting outliers produced a moderate MSE_b_SE value. Lastly, for the FCIP calibration, fixing outliers produced a moderate MSE_b_SE value while not fixing and excluding outliers produced small MSE_b_SE values.

Table 8

Mean Square Total, Systematic, and Random Errors for the Ten IRT-based Equating Methods under the Condition of Outliers with 9 Scored-points and From One Content Area

Method	Item Difficulty *b*				Number-correct True Score *t*			
	MSE_b	MSE_b_SE		MSE_b_RE	MSE_t	MSE_t_SE		MSE_t_RE
A. Equivalent Groups (N(0,1) vs. N(0,1))								
A1. Methods that Did Not Consider the Influence of Outliers								
C+include	0.0504	0.0454	(S)	0.0050	4.6742	4.6003	(M)	0.0740
TCC+include	0.4837	0.4750	(L)	0.0087	16.4719	16.2440	(L)	0.2279
M/S+include	0.3230	0.3198	(L)	0.0032	19.8008	19.5356	(L)	0.2652
FCIP+fixed	0.1421	0.1364	(M)	0.0058	4.5481	4.4692	(M)	0.0789
A2. Methods that Considered the Influence of Outliers								
C+exclude	0.0186	0.0107	(S)	0.0079	0.1462	0.0440	(S)	0.1022
TCC+exclude	0.0219	0.0107	(S)	0.0111	0.2354	0.0445	(S)	0.1909
M/S+exclude	0.0174	0.0107	(S)	0.0067	0.3231	0.0750	(S)	0.2481
M/S+weight	0.2022	0.2004	(M)	0.0018	39.0992	38.3385	(L)	0.7607
FCIP+nofixed	0.0399	0.0106	(S)	0.0293	0.2527	0.0591	(S)	0.1936
FCIP+exclude	0.0163	0.0109	(S)	0.0054	0.1777	0.0586	(S)	0.1190
B. Non-equivalent Groups (N(0,1) vs. N(1,1))								
B1. Methods that Did Not Consider the Influence of Outliers								
C+include	0.3777	0.3709	(L)	0.0068	0.4662	0.3283	(S)	0.1379
TCC+include	0.6389	0.6227	(L)	0.0162	20.7792	20.3331	(L)	0.4461
M/S+include	0.2777	0.2703	(L)	0.0075	14.8481	14.3375	(L)	0.5106
FCIP+fixed	0.0339	0.0252	(S)	0.0086	0.4173	0.3161	(S)	0.1012
B2. Methods that Considered the Influence of Outliers								
C+exclude	0.7601	0.7521	(L)	0.0080	4.6354	4.5402	(M)	0.0952
TCC+exclude	0.0802	0.0633	(S)	0.0170	1.9827	1.6172	(S)	0.3655
M/S+exclude	0.0280	0.0174	(S)	0.0106	0.3054	0.0855	(S)	0.2199
M/S+weight	0.4280	0.4229	(L)	0.0051	37.6123	37.1089	(L)	0.5034
FCIP+nofixed	0.1928	0.1839	(M)	0.0089	4.7385	4.6466	(M)	0.0918
FCIP+exclude	0.1947	0.1857	(M)	0.0089	4.7823	4.6911	(M)	0.0912

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Comparison of the MSE_b_SE values in Panels A1 and A2, Tables 8 and 5 and in Panel A, Table 4. When the systematic errors for the *b*-parameters in Panel A1, Table 8 were compared with the systematic errors in Panel A1, Table 5 and in Panel A, Table 4, it was found that when the score-points of outliers increased, the MSE_b_SE values for the concurrent calibration with outliers included remained small. However, the MSE_b_SE values for the TCC and M/S transformations with outliers included changed from small to large with the increase in score-points of outliers. For the FCIP calibration with outliers fixed, although the MSE_b_SE values were small under the conditions of no outliers and outliers with 3 score-points, they became moderate under the condition of outliers with 9 score-points. In contrast, the comparison of the MSE_b_SE values in Panel A2, Table 8 with the corresponding values in Panel A2, Table 5 and in Panel A, Table 4 showed that, with the exception of the M/S transformation with outliers weighted, the methods that considered the influence of outliers had small MSE_b_SE values as do the corresponding methods under the conditions of no outliers and outliers with 3 score-points. The M/S transformation with outliers weighted had a moderate MSE_b_SE value under the condition of outliers with 9 score-points but a small MSE_b_SE value under the conditions of no outliers and outliers with 3 score-points.

Number-correct True Score t

As shown in Panel A1, Table 8, the concurrent calibration with outliers included and FCIP calibration with outliers fixed had moderate MSE_t_SE values; and the TCC and M/S transformations with outliers included had large MSE_t_SE values. In contrast, with the exception of the M/S transformation with outliers weighted that had a large

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

MSE_t_SE value, the remaining methods that considered the influence of outliers had small MSE_t_SE values (see Panel A2, Table 8).

Comparison of the MSE_t_SE values in Panels A1 and A2, Table 8. The MSE_t_SE values for the methods that did not consider the influence of outliers were either moderate or large. In contrast, with the exception of the M/S transformation with outliers weighted, the methods that considered the influence of outliers had small MSE_t_SE values. More specifically, for the concurrent calibration, including outliers produced a moderate MSE_t_SE value while excluding outliers produced a small MSE_t_SE value; likewise, for the FCIP calibration, fixing outliers produced a moderate MSE_t_SE value while not fixing and excluding outliers produced small MSE_t_SE values; for the TCC transformation, including outliers produced a large MSE_t_SE value while excluding outliers produced a small MSE_t_SE value; and for the M/S transformation, including outliers produced a large MSE_t_SE value, excluding outliers produced a small MSE_t_SE value, and weighting outliers produced a large MSE_t_SE value.

Comparison of the MSE_t_SE values in Panels A1 and A2, Tables 8 and 5 and in Panel A, Table 4. When the systematic errors for the number-correct true scores in Panel A1, Table 8 were compared with the corresponding systematic errors in Panel A1, Table 5 and Panel A, Table 4, it was found that with the increase of the score-points of outliers, the MSE_t_SE values for the methods that did not consider the influence of outliers also increased. More specifically, for the concurrent and FCIP calibrations, the MSE_t_SE values changed from small to moderate when the score-points of outliers increased from

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

3 to 9. The MSE_t_SE values for the TCC transformation changed from small to moderate and to large. And for the M/S transformation, the MSE_t_SE values changed from small to large when the score-points of outliers changed from 0 to 3. In contrast, the comparison of the MSE_t_SE values in Panel A2, Table 8, in Panel A2, Table 5, and in Panel A, Table 4 showed that, with the exception of the M/S transformation with outliers weighted, the MSE_t_SE values for the methods that considered the influence of outliers remained small as the score-points of outliers increased.

Non-equivalent Equating Groups

Item Difficulty b

As shown in Panel B1, Table 8, the FCIP calibration with outlier fixed had a small MSE_b_SE value. In contrast, the concurrent calibration with outliers included and the TCC and M/S transformations with outliers included had large MSE_b_SE values. Examination of the MSE_b_SE values in Panel B2, Table 8 showed that the MSE_b_SE values for the TCC and M/S transformations with outliers excluded were small; the MSE_b_SE values for the FCIP calibration with outliers not fixed and excluded were moderate; and the MSE_b_SE values for the concurrent calibration with outlier excluded and M/S transformation with outliers weighted were large.

Comparison of the MSE_b_SE values in Panels B1 and B2, Table 8. It was found that whether the methods that considered the influence of outliers had smaller MSE_b_SE values than the methods that did not consider the influence of outliers depended on the specific method used. The concurrent calibration produced large MSE_b_SE values regardless of whether the outliers were included or excluded. For the TCC

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

transformation, including outliers produced a large MSE_b_SE value while excluding outliers produced a small MSE_b_SE value. For the M/S transformation, including and weighting outliers produced large MSE_b_SE values while excluding outliers produced a small MSE_b_SE value. Lastly, for the FCIP calibration, fixing outliers produced a small MSE_b_SE value while without fixing and excluding outliers produced moderate MSE_b_SE values.

Comparison of the MSE_b_SE values in Panels B1 and B2, Tables 8 and 5 and in Panel B, Table 4. When the systematic errors for the *b*-parameters in Panel B1, Table 8 were compared with the systematic errors in Panel B1, Table 5 and in Panel B, Table 4, it was found that the MSE_b_SE values for the concurrent calibration were large across the three tables. The MSE_b_SE values for the TCC and M/S transformations increased from small to moderate, and to large as the score-points of outliers increased. The MSE_b_SE values for the FCIP calibration were moderate under the conditions of no outliers and outliers with 3 score-points but small under the condition of outliers with 9 score-points. In contrast, the MSE_b_SE values for the methods that considered the influence of outliers were the same size across Tables 5 and 8. More specifically, when the equating groups were not equivalent, the MSE_b_SE values for the TCC and M/S transformations with outliers excluded were both small; the MSE_b_SE values for the FCIP calibration with outliers not fixed and excluded were both moderate; and the MSE_b_SE values for the concurrent calibration with outliers excluded and M/S transformation with outliers weighted were both large. With the exception of the M/S transformation with outliers

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

weighted, the MSE_b_SE vales for the remaining methods in Panel B2, Table 5 and 8 were the same size as the corresponding values in Panel B, Table 4.

Comparison of the MSE_b_SE values in Panels A and B, Table 8. For the methods that did not consider the influence of outliers, the concurrent calibration with outliers included had a small MSE_b_SE value under the equivalent groups condition but a large value under the non-equivalent groups condition; the TCC and M/S transformations with outliers included had large MSE_b_SE values regardless of the equivalence of the groups; and the FCIP calibration with outliers fixed had a moderate MSE_b_SE value under the equivalent groups condition but a small value under the non-equivalent groups condition. The methods that considered the influence of outliers, with the exception of the M/S transformation with outliers weighted, had small MSE_b_SE values under the equivalent groups condition. However, the MSE_b_SE values for these methods varied under the non-equivalent groups condition: the MSE_b_SE values for the TCC and M/S transformations with outliers excluded were small; the MSE_b_SE values for the FCIP calibration with outliers not fixed and excluded were moderate; and the MSE_b_SE values for concurrent calibration with outliers excluded and M/S with outliers weighted were large.

Number-correct True Score t

As shown in Panel B1, Table 8, the MSE_t_SE values for the concurrent calibration with outliers included and FCIP calibration with outliers fixed were small, while for the TCC and M/S transformations with outliers included, they were large. For the methods that considered the influence of outliers (see Panel B2, Table 8), the TCC and M/S transformations with outliers excluded had small MSE_t_SE values. The

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

concurrent calibration with outliers excluded and the FCIP calibration without fixing outliers and with outliers excluded had moderate MSE_t_SE values. The M/S transformation with outliers weighted had a large MSE_t_SE value.

Comparison of the MSE_t_SE values in Panels B1 and B2, Table 8. It was found that whether the methods that considered the influence of outliers had smaller MSE_b_SE values than the methods that did not consider the influence of outliers depended on the specific method used. For the concurrent and FCIP calibrations, without considering the influence of outliers produced small MSE_t_SE values while considering the influence of outliers produced moderate MSE_t_SE values. For the TCC transformation, including outliers produced a large MSE_t_SE value while excluding outliers produced a small MSE_t_SE value. For the M/S transformation, including and weighting outliers produced large MSE_t_SE values while excluding outliers produced a small MSE_t_SE value.

Comparison of the MSE_t_SE values in Panels B1 and B2, Tables 8 and 5 and in Panel B, Table 4. When the systematic errors for the number-correct true scores in Panel B1, Table 8 were compared with the corresponding systematic errors in Panel B1, Table 5 and in Panel B, Table 4, it was found that, for the concurrent and FCIP calibrations, the MSE_t_SE values were small when outliers were present but not considered, while the corresponding values were moderate under the condition of no outliers. For the TCC and M/S transformations, the MSE_t_SE values increased as the score-points of outliers increased. In contrast, when the systematic errors for the number-correct true scores in Panel B2, Table 8 were compared with the corresponding systematic errors in Panel B2, Table 5 and in Panel B, Table 4, it was found that the MSE_t_SE values for the methods

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

that considered the influence of outliers were the same size across Tables 5 and 8. More specifically, when the equating groups were not equivalent, the MSE_t_SE values for the TCC and M/S transformations with outliers excluded were small; the MSE_t_SE values for the concurrent calibration with outliers excluded and the FCIP calibration with outliers not fixed and excluded were moderate; and the MSE_t_SE values for the M/S transformation with outliers weighted were large. With the exception of the M/S transformation with outliers weighted, the MSE_t_SE values were the same size as the corresponding MSE_t_SE values under the condition of no outliers.

Comparison of the MSE_t_SE values in Panels A and B, Table 8. For the methods that did not consider the influence of outliers, the MSE_t_SE values for the concurrent calibration with outliers included and FCIP calibration with outliers fixed were moderate under the equivalent groups condition while these values were small under the non-equivalent groups condition. The MSE_t_SE values for the TCC and M/S transformations with outliers included were large regardless of the group equivalence. In contrast, all the methods that considered the influence of outliers, except the M/S transformation with outliers weighted, had small MSE_t_SE values under the equivalent groups condition while, with the exception of the TCC and M/S transformations with outliers excluded, they had either large or moderate MSE_t_SE values under the condition of non-equivalent groups. It is worth to point out that the TCC and M/S transformations with outliers excluded had small MSE_t_SE values regardless of the equivalence of equating groups.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Outliers with 9 Score-points Randomly from Any Content Area

As indicated previously, the results for the condition in which the outliers with 9 score-points were from one content area and were randomly from any content area were similar, which again indicated that the performance of the IRT-based equating methods was not confounded by the types of outliers. Hence, all the observations found under the condition of outliers with 9 score-points were from one content area can be applied to the condition of outliers with 9 score-points were randomly from any content area.

Summary

When the equating groups were equivalent, the methods that did not consider the influence of outliers tended to have greater systematic errors than the methods that did consider the influence of outliers, which indicated that the latter methods performed better, as determined by the MSE_b_SE and MSE_t_SE values, than the former methods. Among the methods that considered the influence of outliers, with the exception of the M/S transformation with outliers weighted, the remaining methods produced small systematic errors, which indicated these methods performed equally well under the condition of equivalent groups.

When the equating groups were not equivalent, not all the systematic errors for the methods that did not consider the influence of outliers were greater than the corresponding values for the methods that did consider the influence of outliers. Thus, as was the case for the outliers with 3 score-points, caution needs to be taken when one draws conclusion on whether the methods that considered the influence of outliers performed better than the methods that did not consider the influence of outliers when the equating groups were not equivalent. Among the methods that considered the influence

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

of outliers, only the TCC and M/S transformations with outliers excluded produced small systematic errors. It is worth to note that the M/S transformation with outliers weighted produced large systematic errors under the conditions of outliers with 3 and 9 score-points and non-equivalent equating groups. This is probably due to the use of weight in this method. As described in Chapter 2, this method uses the weighted item difficulties to calculate the equating coefficients. The weights are inversely proportional to the standard errors of the item difficulty estimates. Under the non-equivalent groups condition, one group has an ability distribution with mean 1 and standard deviation 1, which means the standard errors of the item difficulty estimate are large when the item responses from this group are used. Unfortunately, this method uses these large standard errors to weight the item difficulty, which in turn leads to the large systematic errors.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Table 9

Mean Square Total, Systematic, and Random Errors for the Ten IRT-based Equating Methods under the Condition of Outliers with 9 Score-points and Randomly From Any Content Area

Method	Item Difficulty *b*				Number-correct True Score *t*			
	MSE_b	MSE_b_SE		MSE_b_RE	MSE_t	MSE_t_SE		MSE_t_RE
A. Equivalent Groups (N(0,1) vs. N(0,1))								
A1. Methods that Did Not Consider the Influence of Outliers								
C+include	0.0461	0.0420	(S)	0.0041	4.6465	4.5843	(M)	0.0623
TCC+include	0.4073	0.3966	(L)	0.0107	16.4731	16.1420	(L)	0.3311
M/S+include	0.3304	0.3277	(L)	0.0027	27.1038	26.8693	(L)	0.2346
FCIP+fixed	0.1415	0.1372	(M)	0.0043	4.8023	4.7382	(M)	0.0641
A2. Methods that Considered the Influence of Outliers								
C+exclude	0.0148	0.0102	(S)	0.0046	0.1285	0.0372	(S)	0.0913
TCC+exclude	0.0211	0.0102	(S)	0.0109	0.3004	0.0660	(S)	0.2344
M/S+exclude	0.0166	0.0102	(S)	0.0065	0.2731	0.0847	(S)	0.1884
M/S+weight	0.2463	0.2448	(M)	0.0015	32.2223	31.9253	(L)	0.2970
FCIP+nofixed	0.0192	0.0090	(S)	0.0103	0.1808	0.0541	(S)	0.1267
FCIP+exclude	0.0227	0.0104	(S)	0.0123	0.1954	0.0524	(S)	0.1430
B. Non-equivalent Groups (N(0,1) vs. N(1,1))								
B1. Methods that Did Not Consider the Influence of Outliers								
C+include	0.4069	0.4005	(L)	0.0063	0.4915	0.3633	(S)	0.1283
TCC+include	0.4889	0.4727	(L)	0.0162	19.4662	18.8192	(L)	0.6470
M/S+include	0.3514	0.3476	(L)	0.0038	28.1300	27.6667	(L)	0.4633
FCIP+fixed	0.0362	0.0284	(S)	0.0077	0.5435	0.4353	(S)	0.1082
B2. Methods that Considered the Influence of Outliers								
C+exclude	0.7676	0.7597	(L)	0.0079	4.6465	4.5644	(M)	0.0821
TCC+exclude	0.0654	0.0507	(S)	0.0148	1.4692	1.1761	(S)	0.2931
M/S+exclude	0.0282	0.0185	(S)	0.0097	0.2850	0.0770	(S)	0.2080
M/S+weight	0.9206	0.9049	(L)	0.0156	33.2908	32.7707	(L)	0.5201
FCIP+nofixed	0.1929	0.1839	(M)	0.0090	4.6912	4.6068	(M)	0.0844
FCIP+exclude	0.1963	0.1873	(M)	0.0090	4.7627	4.6822	(M)	0.0805

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

CHAPTER 5: SUMMARY AND CONCLUSIONS

In this chapter, a summary of the research questions and methods used to answer the research questions is presented first. The key findings are then summarized. This is followed by the discussion of the limitations of the study. Conclusions, implications for practice, and recommendations for future research are then made.

Summary of Research Questions and Methods

A potential benefit of using IRT models is that the item parameter estimates (e.g., *a*-, *b*-, or *c*-parameters) are not group dependent. However, in the context of test equating, it has been found that the estimates of item parameters are not necessarily consistent when they are estimated from two equating groups and have been transformed onto the same scale. Outliers with large inconsistent item parameter estimates may have a serious impact on IRT-based equating results. Variations of the basic IRT-based equating methods have been used to remove the possible adverse influence of outliers. Do the IRT-based equating methods that consider the influence of outliers produce better results than the methods that do not consider the influence of outliers? Which IRT-based equating method best reduces the influence of outliers? Thus, the primary purpose of the current study was to investigate the comparability of ten IRT-based equating methods in the presence of outlier items with inconsistent *b*-parameter estimates. The ten methods were concurrent calibration with outliers included, TCC transformation with outliers included, M/S transformation with outliers included, FCIP calibration with outliers fixed, concurrent calibration with outliers excluded, TCC transformation with outliers excluded, M/S transformation with outliers excluded, M/S transformation with outliers weighted, FCIP calibration with outliers not fixed, and FCIP calibration with outliers excluded.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Simulated data were generated based on the real item parameters of a large-scale state mathematics achievement test. The equating design was the common-item non-equivalent matrix groups design. In this design, two test forms need to be equated. There were three sub forms in each of these two test forms. Each sub form contained 26 unique MC items (with two score categories), 5 unique SA items (with two score categories), 5 unique OR items (with five score categories), 8 common MC items, 1 common SA item, and 1 common OR item. The corresponding IRT models used were the 3PL, 2PL, and GRM models. The factors manipulated were group ability differences (equivalent groups (N (0, 1) vs. N (0, 1)) and non-equivalent groups (N (0, 1) vs. N (1, 1)), number/score-points of outliers (0 outliers with score-points 0, 3 MC items with score-points 3, and 5MC items and 1 OR item with score-points 9), types of outliers (outliers from one content area, randomly from any content area, and with extreme values applied to the 3 score-points condition and the first two types of outliers applied to the 9 score-points condition), and IRT-based equating methods (four basic methods with no outliers present and ten variations in the presence of outliers). As a result, a total of 108 conditions (1 x 2 x 4 ("0" outlier by group ability differences by equating method) + 5 x 2 x 10 (5 combinations of outliers' characteristics by group ability differences by equating method) = 108) were examined to answer the following research questions:

(1) Do the IRT-based equating methods that consider the influence of outliers produce a better result than those that do not consider the influence of outliers?

(2) Is the effect found in Question 1, if any, confounded by factors such as the group ability differences and the characteristics of outliers?

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

(3) Which of the IRT-based equating methods produces a better result, especially among the IRT-based equating methods that consider the influence of outliers?

(4) Is the effect found in Question 3, if any, confounded by factors such as the group ability differences and the characteristics of outliers?

To answer the research questions, the unweighted mean square error for the *b*-parameters (MSE_b) and the unweighted mean square error for the number-correct true scores (MSE_t) were calculated for the 108 conditions. The mean square errors were further decomposed into systematic errors (MSE_b_SE and MSE_t_SE) and random errors (MSE_b_RE and MSE_t_RE). Since the change of mean square errors was mainly due to the change of systematic errors and systematic error reflect the magnitudes of the bias introduced by specific equating methods, the MSE_b_SE and MSE_t_SE were used to (a) evaluate the accuracy of the equating methods and (b) to make comparisons among the methods.

To make the comparisons consistent, the following absolute rules were used: (a) $MSE_b\_SE \leq 0.06$ was considered as small, (b) $0.06 < MSE_b\_SE \leq 0.25$ was considered as moderate, and (c) $MSE_b\_SE > 0.25$ was considered as large for the *b*-parameters; and (a) $MSE_t\_SE \leq 2.25$ was considered as small, (b) $2.25 < MSE_t\_SE \leq 6.25$ was considered as moderate, and (c) $MSE_t\_SE > 6.25$ was considered as large for the number-correct true scores.

Summary of Findings

The key findings of the current study are:

1. When outliers were not present in the data set and the two equating groups were equivalent, the MSE_b_SE and MSE_t_SE values for the methods of concurrent

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

calibration, TCC transformation, M/S transformation, and FCIP calibration were small. However, when the difference between groups mean abilities was one standard deviation, the MSE_b_SE and MSE_t_SE values for the four methods varied: the TCC and M/S transformations produced small MSE_b_SE and MSE_t_SE values; the FCIP produced moderate MSE_b_SE and MSE_t_SE values; and the concurrent calibration produced large MSE_b_SE and MSE_t_SE values.

2. When outliers were present in the data, the IRT-based equating methods that did not consider the influence of outliers tended to produce different MSE_b_SE and MSE_t_SE values from the IRT-based equating methods that considered the influence of outliers. However, whether the latter methods produced smaller MSE_b_SE and MSE_t_SE values than the former methods depended on the specific condition.

2a. When the equating groups were equivalent, the methods that did not consider the influence of outliers tended to have greater MSE_b_SE and MSE_t_SE values than the methods that considered the influence of outliers. When the number/score-points of outliers increased, with the exception of the MSE_b_SE values for the concurrent calibration with outliers included, the MSE_b_SE and MSE_t_SE values for the methods that did not consider the influence of outliers tended to increase. In contrast, among the methods that considered the influence of outliers, with the exception of M/S transformation with outliers weighted, the

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

remaining methods produced small MSE_b_SE and MSE_t_SE values, and these values did not increase as the number/score-points of outliers increased.

2b. When the equating groups were not equivalent, not all the MSE_b_SE and MSE_t_SE values for the methods that did not consider the influence of outliers were greater than the corresponding values for the methods that did consider the influence of outliers. When the number/score-points of outliers increased, the change of the MSE_b_SE and MSE_t_SE values for each method that did not consider the influence of outliers were not consistent, however, the MSE_b_SE and MSE_t_SE values for the methods that considered the influence of outliers did not change.

3. Among the IRT-based equating methods that considered the influence of outliers, the TCC and M/S transformations with outliers excluded consistently produced small MSE_b_SE and MSE_t_SE values regardless of the group equivalence. In contrast, the M/S transformation with outliers weighted produced large MSE_b_SE and MSE_t_SE values especially under the non-equivalent groups condition. The FCIP calibration with outliers not fixed and excluded produced small MSE_b_SE and MSE_t_SE values under the equivalent groups condition and moderate MSE_b_SE and MSE_t_SE values under the non-equivalent groups condition. The concurrent calibration with outliers included produced small MSE_b_SE and MSE_t_SE values under the equivalent groups condition but large MSE_b_SE values and moderate MSE_t_SE values under the non-equivalent groups condition.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Limitations of the Study

The present study was limited to the investigation of IRT-based equating methods in the presence of outliers with inconsistent *b*-parameter estimates. The influence of outliers with inconsistent *a*- and/or c-parameter estimates was not considered. The selection of the *b*-parameter was based on the observations that poorly estimated item difficulties had a serious impact on the equating results (Stocking & Lord, 1983) and that the *a*- and *c*-parameter estimates are not as stable as the *b*-parameter estimates (Ironson, 1983). However, theoretically, the exclusive use of *b*-parameters did not cover all the item characteristic information.

Furthermore, only the outliers located on the left side of the straight line on the scatter plot of *b*-parameters (see Figure 1, p. 11) were examined. This was based on the assumption that the most plausible result of placing instructional emphasis on one content area and/or the revelation of common items is that the corresponding items will become easier when they are administrated in the second year. However, it is possible that, in real data, some outliers may be located on both sides of the straight line.

Conclusions

The results of the current study revealed that when outliers were not present in the data set and the equating groups were equivalent, the methods of concurrent calibration, TCC transformation, M/S transformation, and FCIP calibration performed equally well. However, the same cannot be said when the two equating groups were not equivalent. The four methods were sensitive, but not equally, to the presence of non-equivalent groups. These findings about the concurrent calibration and the TCC and M/S transformations are consistent with the previous research (e.g., Hanson & Belguin, 2002; Kim & Kohen, 1998).

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

When outliers were present in the data set, the performance of the ten IRT-based equating methods was confounded by the group ability differences and the number/score-points of outliers. However, the type of outliers had little impact on the performance of the ten IRT-based equating methods.

When the two equating groups were equivalent, the performance of the methods that did not consider the influence of outliers was sensitive to the number/score-points of outliers. When the number/score-points of outliers were small, the concurrent calibration with outliers included and the FCIP calibration with outliers fixed performed surprisingly well. However, the performance of the methods that did not consider the influence of outliers tended to become worse when the number/score-points of outliers increased. In contrast, with the exception of the M/S transformation with outliers weighted, the methods that considered the influence of outliers performed equally well regardless of the characteristics of outliers. The M/S transformation with outliers weighted performed well as suggested by Lin et al. (1980), but only when fewer outliers present in the data.

When the equating groups were not equivalent, overall, the performance of the methods that did not consider the influence of outliers was not ideal. Among the methods that considered the influence of outliers, the M/S and TCC transformations with outliers excluded performed consistently well regardless of the characteristics of outliers. The remaining methods, especially the M/S transformation with outliers weighted, did not perform well, in contrast to when the equating groups were equivalent.

Taken together, the TCC and M/S transformations with outliers excluded consistently performed well regardless of the group equivalence and the characteristics of outliers.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Implications for Future Practice

The results of the current study reveal that if outliers with inconsistent *b*-parameter estimates are detected in the common items, the influence of such outliers should be removed. The methods that can be considered are the TCC and M/S transformations with outliers excluded. If the two equating groups have similar ability distributions, the concurrent calibration with outlier excluded and the FCIP calibration with outliers excluded or not fixed can be considered too. When the two equating groups have large ability distribution difference, the M/S transformation with outliers weighted and the concurrent calibration with outliers excluded are not recommended.

Recommendations for Future Research

Further research is needed to determine the influence of outlier items defined by their values for the *a*-parameters, *c*-parameters, and item characteristic curves that represent the interaction of item parameters. Investigation of the influence of outliers located on both sides of the straight line in the scatter plot of *b*-parameters is also needed.

In the current study, it was relatively easy to detect the outliers since this was a computer simulation study and the outliers with inconsistent *b*-parameter estimates were pre-designed. However, how to detect outliers in a real data set needs to be investigated. Detection procedures have been indicated in the literature. For example, Linn et al. (1980) indicated that outliers might be due to the large standard error of the item difficulty estimate and proposed the method of M/S transformation with outliers weighted. However, as revealed in this study, the M/S transformation with outliers weighted did not work when equating groups were not equivalent and a large number of outliers were present in the

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

data. This result may indicate that using standard error of the item difficulty estimate to determine if an item as outlier may not be appropriate. The perpendicular distance between a point, which represents an item, and the straight line of best fit (see Figure 1, p.11) could be used to detect outliers based on Stocking and Lord (1983). However, since the true location of the straight line of best fit is unknown in the real data, it is impossible to determine an outlier based on an accurate perpendicular distance. Thus, a graphic procedure (Z. Vukmirovic, personal communication, August 23, 2002) was proposed. In this procedure, outliers are detected by looking at the relative location of an item compared to the location of the most of the items (e.g., Figure 1, p.11). If it is relatively far from the straight line that represents the most of the items, this item may be considered as an outlier. However, this is a very subjective procedure. A more objective procedure needs to be developed.

Absolute rules were proposed in the current study to distinguish small, moderate, and large systematic errors of *b*-parameters and number correct true scores. The development of these rules is somewhat subjective. More research is needed to investigate whether these rules will hold over the other studies.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43,* 561-573.

Andrich, D. (1995). Distinctive and incompatible properties of two common classes of IRT models for graded responses. *Applied Psychological Measurement, 19,* 101-119.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thirndike (Ed.), *Educational measurement* (2nd). Wanshinton, DC: American Council on Education.

Angoff, W. H. (1984). *Scales, norms, and equivalent scores.* Princeton, NJ: Educational Testing Service.

Baker, F. B. (1992). Equating tests under the graded response model. *Applied Psychological Measurement, 16(1),* 87-96.

Baker, F. B. (1996). An investigation of the sampling distributions of equating coefficients. *Applied Psychological Measurement, 20(1),* 45-57.

Beguin, A. A. (2002). *Robustness of IRT test equating to violations of the representativeness of the common items in a non-equivalent groups design.* Paper presented to the annual meeting of the National Council on Measurement in Education, New Orleans.

Bejar, I., & Wingersky, M. S. (1981). *An application of item response theory to equating the Test of Standard Written English* (College Board Report No. 81-8). Princeton NJ: Educational Testing Service, 1981. (ETS No. 81-35)

Bennett, R. E., & Ward, W. C. (1993). *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment.* New Jersey: Hillsdale.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Bock, R. D. (1972). Estimating item parameters and latent ability when the responses are scored in two or more nominal categories. *Psychometrika, 37,* 29-51.

Bogan, E. D., & Yen, W. M. (1983). *Detecting multidimensionality and examine its effects on vertical equating with the three-parameter logistical model.* Paper presented at the Annual Meeting of the American Educational Research Association, Montreal.

Bolt, D. M. (1999). Evaluating the effects of multidimensionality on IRT true-score equating. *Applied Measurement in Education, 12(3),* 383-407.

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

Childs, R. A., & Chen, Wen-Hung (1999). Obtaining comparable item parameter estimates in MULTILOG and PARSCALE for two polytomous IRT models. *Applied Psychological Measurement, 23(4),* 371-379.

Cohen, A. S., & Kim, Seock-Ho, (1998). An investigation of linking methods under the graded response model. *Applied Psychological Measurement. 22(2),* 116-130.

Cook, L. L., & Eignor, D. R. (1983). *An investigation of the feasibility of applying item response theory to equate achievement tests.* Paper presented at the Annual Meeting of the American Educational Research Association, Montreal.

Cook, L. L., Eignor, D. R., & Hutton, L. R. (1979). *Considerations in the application of latent trait theory to objective-based criterion-referenced tests.* Paper presented at the meeting of the American Educational Research Association, San Francisco.

Cook, L. L., & Petersen, N. S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. *Applied Psychological Measurement, 11,* 225-244.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

De Champlain, A. F. (1996). The effect of multidimensionality on IRT true-score equating for subgroups of examinees. *Journal of Educational Measurement, 33(2),* 181-201.

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

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap* (Monographs on Statistics and Applied Probalility 57). New York: Chapman & Hall.

Embreson, S. E. (1984). A general component latent trait model for response process, *Psychometrika, 49,* 175-186.

Embretson, S. E. (1985). *Test design: Developments in psychology and psychometrics.* New York: Academic Press.

Flanagan, J. C. (1951). Units, scores and norms. In E. F. Lindquist (Ed.), *Educational Measurement.* (2nd ed.) Washington, D.C.: American Council on Education, 695-763.

Gifford, J. A., & Swaminathan, H. (1990). Bias and the effect of priors in Bayesian estimation of parameters of item response models. *Applied Psychological Measurement, 14,* 33-43.

Harris, D. J. (1991). *Equating with nonrepresentative common item sets and non-equivalent groups.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Harris, D. J., & Crouse, J. D. (1993). A Study of Criteria Used in Equating. *Applied Measurement in Education, 6(3),* 195-240.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

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

Hambleton, R. K., Murray, L. (1983). Some goodness of fit investigations for item response models. In R. K. Hambleton, *Applications of item response theory* (pp.71-94). British Columbia: Educational Research Institute of British Columbia.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Item response theory.* Newbury Park, CA: Sage.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: Kluwer Nijhoff Publishing.

Han, Tianqi, Kolen, M., & Pohlmann, J. (1997). A comparison among IRT true- and observed-score equatings and traditional equipercentile equating. *Applied Measurement in Education, 10(2),* 105-121.

Hanson, B. A., & Beguin, A. A. (2002). Obtaining a Common Scale for Item Response Theory Item Parameters Using Separate versus Concurrent Estimation in the Common-Item Equating Design. *Applied Psychological Measurement, 26(1),* 3-24.

Hanson, B. A., & Feinstein, Z. S. (1997). *Application of a polynomial loglinear model to assessing differential item functioning for common items in the common-item equating design* (ACT Research Report Series 97-1). Iowa City, IA: American College Testing.

Harwell, M., Stone, C. A., Hsu, Tse-Chi, & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement, 20(2),* 101-125.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Hills, J. R., Subhiyah, R. G., & Hirsch, T. M. (1988). Equating minimum-competency tests: comparison of methods. *Journal of Educational Measurement, 25(3),* 221-231.

Ironson, G. H. (1983). Using item response theory to measure bias. In R. K. Hambleton, *Applications of item response theory* (pp.155-174). British Columbia: Educational Research Institute of British Columbia.

Jaeger, R. M. (1981). Some exploratory indices for selection of a test equating method. *Journal of Educational Measurement, 18,* 23-38.

Kaskowitz, G. S., & De Ayala, R. J. (2001). The effect of error in item parameter estimates on the test response function method of linking. *Applied Psychological Measurement, 25(1),* 39-52.

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

Kim, Seock-Ho., & Cohen, A. S.(1997). *A comparison of linking and concurrent calibration under the graded response model.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

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

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

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Kolen, M. J., & Brennan, R. L. (1995). *Test equating: methods and practices.* New York: Springer.

Kolen, M. J. & Whitney, D. R. (1982). Comparison of four procedures for equating the tests of General Educational Development. *Journal of Educational Measurement. 19(4),* 279-293.

Kromrey, J. D., Parshall, C. G., & Yi, Q. (1998). *The effects of content representativeness and differential weighting on test equating: a Monte Carlo study.* Paper presented at the Annual Meeting of the American Educational Research Association, San Diego.

Lee, G. Kolen, M. J. Frisbie, D. A., & Ankenmann, R. D. (2001). Comparison of dichotomous and polytomous item response models in equating scores from tests composed of testlets. *Applied Psychological Measurement, 25(4),* 357-372.

Lehman, R. S., & Bailey, D. E. (1968). *Digital computing: Fortran IV and its applications in behavioural science.* New York: John Wiley and Sons.

Li, Yuan-H, Griffith, W. D., & Tam, H. P. (1997). *Equating multiple tests via an IRT linking design: Utilizing a single set of anchor items with fixed common item parameters during the calibration process.* Paper presented at the Psychometric Society Meeting, Knoxville.

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

Li, Yuan H., Lissitz, R. W., & Yang, Yu-Nu (1999). *Estimating IRT equating coefficients for tests with polytomously and dichotomously scored items.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Montreal.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. (1980). *An investigation of item bias in a test of reading comprehension* (Technical Report No. 163). Urbana IL: Center for the Study of Reading, University of Illinois.

Lord, F. M. (1950). *Notes on comparable scales for test scores.* (ETS RB 50-48). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1955). Equating test scores – a maximum likelihood solution. *Psychometrika,* 20, 193-200.

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

Lord, F. M. (1982). Item response theory and equating – A technical summary. In P. W. Holland and D. B. Rubin (Eds.), *Test equating* (pp. 141-148). New York: Academic.

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

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

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149-174.

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

Mislevy, R. J., & Bock, R. D. (1990). *BILOG-3 (2nd ed): Item Analysis and test scoring with binary logistic models.* Mooresvilk: Scientific Software.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Modu, C. C. (1982). *The robustness of latent trait model for achievement test score equating*. Paper presented at the Annual Meeting of the American Educational Research Association, New York.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16,* 159-176.

Muraki, E. (1993). Information functions of the generalized partial credit model. *Applied Psychological Measurement, 17,* 351-363.

Muraki, E., & Bock, R. D. (1999). *PARSCALE: IRT based test scoring and item analysis for graded open-ended exercises and performance tasks (Version 3)* [Computer software]. Chicago: Scentific Software.

Ogasawara, H. (2001). Standard errors of item response theory equating/linking by response function methods. *Applied Psychological Measurement, 25(1),* 53-67.

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

Petersen, N. C., Marco, G. L., & Stewart, E. E. (1982). A test of the adequacy of linear score equating models. In P. W. Holland and D. B. Rubin (Eds.), *Test equating.* New York: Academic Press, 71-135.

Psychometric Society. (1979). Publication policy regarding Monto Carlo studies. *Psychometrika, 44,* 133-134.

Raju, N. S., Edwards, J. E., & Osgerg, D. W. (1983). *The effect of anchor test size in vertical equating with the Rasch and three-parameter models*. Paper presented at

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

the Annual Meeting of the National Council on Measurement in Education, Montreal.

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

Raju, N. S., Bode, R. K., Larsen, V. S., & Steinhaus, S. (1986). *Anchor-test size and horizontal equating with the Rasch and three-parameter models.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: results and implications. *Journal of Educational Statistics, 4(3),* 207-230.

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

Rubinstein, R. V. (1981). *Simulation and the Monte Carlo method.* New York: Wiley.

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

Skaggs, G., & Lissitz, R. W. (1986a). An exploration of the robustness of four test equating models. *Applied Psychological Measurement, 10,* 303-317.

Skaggs, G., & Lissitz, R. W. (1986b). *The effects of examinee ability for equating invariance.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Stocking, M. L. (1990). Specifying Optimum Examinees for Item Parameter Estimation in Item Response Theory. *Psychometrika, 55(3),* 461-475.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

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

Sykes, R. C., Hou, L., Hanson, B., & Wang, Z. (2002). *Multidimensionality and the equating of a mixed-format math examination.* Paper presented to the annual meeting of the National Council on Measurement in Education, New Orleans.

Tate, R. (2000). Performance of a proposed method for the linking of mixed format tests with constructed response and multiple choice items. *Journal of Educational Measurement, 37(4),* 329-346.

Tate, R., & Kamata, A. (2002). *The performance of a method for the long-term equating of a mixed format assessment (Phase I).* Paper presented to the annual meeting of the National Council on Measurement in Education, New Orleans.

Thissen, D. (1991). *MULTILOG user's guide: Multiple, categorical item analysis and test scoring using item response theory (Version 6.0).* New York: Springer.

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

Thissen, D., & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika, 47,* 397-412.

Tsai, Tsung-Hsun, Hanson, B. A., Kolen, M. J., & Forsyth, R. A. (2001). A comparison of bootstrap standard errors of IRT equating methods for the common-item nonequivalent groups design. *Applied Measurement in Education, 14(1),* 17-30.

Vale, C. D. (1986). Linking item parameters onto a common scale. *Applied Psychological Measurement, 10,* 333-344.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Van der Linden, W.J., & Hambleton, R.K. (1997). *Handbook of Modern Item Response Theory.* New York: Springer.

Vukmirovic, Z., Hu, H., & Turner, J. C. (2003). *The effects of outliers on IRT equating with fixed common item parameters.* Paper presented at the National Council on Measurement in Education, Chicago.

Wang, Tian-You, Hanson, B. A., & Harris, D. J. (2000). The effectiveness of circular equating as a criterion for evaluating equating. *Applied Psychological Measurement, 24(3),* 195-210.

Whitely, S. E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika, 45,* 479-494.

Wingersky, M. S., Barton, M. A., & Lord, F. M. (1982). *LOGIST user's guide.* Princeton NJ: Educational Testing Service.

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

Yang, W. (1997). *The effects of content mix and equating method on the accuracy of test equating using anchor-item design.* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.

Yang, W., & Houang, R. T. (1996). *The effect of anchor length and equating method on the accuracy of test equating: comparisons of linear and IRT-based equating using an anchor-item design.* Paper presented at the Annual Meeting of the American Educational Research Association, New York.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8(2),* 125-145.

Zenisky, A. L. (2001). *Investigating the accumulation of equating error in fixed common item parameter linking: A simulation study.* Paper presented to the annual meeting of the Northeastern Educational Research Association, Kerhonkson, NY.

Zeng, L. (1991). *Standard errors of linear equating for the single-group design* (ACT Research Report 91-4). Iowa City, IA: American College Testing.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BILOG-MG: Multiple group IRT Analysis and Test Maintenance for Binary Items* [Computer program]. Chicago: Scientific Software International.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Appendix A

Concurrent Calibration

This appendix gives the sample PARSCALE control file used in the FCIP calibrations for the conditions that outliers (with 3 and 9 score-points) were included and outliers (with 3 score-points and from one content area) were excluded in the data file.

Record File for the Concurrent Calibration with Outliers Included

```
Concurrent calibration of reference and equated tests.
>COMMENT ;
This analysis is based on 102 items. There are:
26 operational MC items in the reference test, 3PL model,
 5 operational SA items in the reference test, 2PL Model,
 5 operational OR items in the reference test, GR model, 5 categories,
24 common MC items, 3PL model,
 3 common SA items, 2PL Model,
 3 common OR items, GR model, 5 categories,
26 operational MC items in the equated test, 3PL model,
 5 operational SA items in the equated test, 2PL Model,
 5 operational OR items in the equated test, GR model, 5 categories,;
>FILE DFNAME='C:\Psl3\conc.DAT',
    NFNAME='C:\Psl3\conc.KEY',
    SAVE;
>SAVE PARM='C:\Psl3\conc.PAR';
>INPUT NIDW=5,NTOTAL=102,NTEST=1,LENGTH=(102), SAMPLE=100;
 (5A1,102A1)
>TEST1 TNAME=concurrentin,ITEM=(1(1)102),NBLOCK=102;
>BLOCK1 BNAME=MC-UNI,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
  GUESSING=(2,ESTIMATE),GPARM=(.25),REPEAT=26;
>BLOCK2 BNAME=SA-UNI,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
  REPEAT=5;
>BLOCK3 BNAME=OR-UNI,NITEMS=1,NCAT=5,ORIGINAL=(0,1,2,3,4),
  REPEAT=5;
>BLOCK4 BNAME=MC-EQ1,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
  GUESSING=(2,ESTIMATE),GPARM=(.25),REPEAT=8;
>BLOCK5 BNAME=SA-EQ1,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
  REPEAT=1;
>BLOCK6 BNAME=OR-EQ1,NITEMS=1,NCAT=5,ORIGINAL=(0,1,2,3,4),
  REPEAT=1;
>BLOCK7 BNAME=MC-EQ2,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
  GUESSING=(2,ESTIMATE),GPARM=(.25),REPEAT=8;
>BLOCK8 BNAME=SA-EQ2,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
```

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

```
 REPEAT=1;
>BLOCK9 BNAME=OR-EQ2,NITEMS=1,NCAT=5,ORIGINAL=(0,1,2,3,4),
 REPEAT=1;
>BLOCK10 BNAME=MC-EQ3,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
 GUESSING=(2,ESTIMATE),GPARM=(.25),REPEAT=8;
>BLOCK11 BNAME=SA-EQ3,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
 REPEAT=1;
>BLOCK12 BNAME=OR-EQ3,NITEMS=1,NCAT=5,ORIGINAL=(0,1,2,3,4),
 REPEAT=1;
>BLOCK13 BNAME=MC-EQ3,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
 GUESSING=(2,ESTIMATE),GPARM=(.25),REPEAT=26;
>BLOCK14 BNAME=SA-EQ3,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
 REPEAT=5;
>BLOCK15 BNAME=OR-EQ3,NITEMS=1,NCAT=5,ORIGINAL=(0,1,2,3,4),
 REPEAT=5;
>CAL GRADED,LOGISTIC,NQPTS=30,CYCLE=(50,2,2,2,2),
   GPRIOR,TPRIOR,SPRIOR,NEWTON=0,CRIT=0.001,DIAGNOS=1;
>SCORE NOSCORE;
```

Record File for the Concurrent Calibration with Outliers Excluded

```
Concurrent calibration of reference and equated tests with outliers excluded.
>COMMENT ;
This analysis is based on 99 items. There are:
26 operational MC items in the reference test, 3PL model,
 5 operational SA items in the reference test, 2PL Model,
 5 operational OR items in the reference test, GR model, 5 categories,
21 common MC items, 3PL model,
 3 common SA items, 2PL Model,
 3 common OR items, GR model, 5 categories,
26 operational MC items in the equated test, 3PL model,
 5 operational SA items in the equated test, 2PL Model,
 5 operational OR items in the equated test, GR model, 5 categories,;
>FILE DFNAME='C:\Psl3\conc.DAT',
    NFNAME='C:\Psl3\conc.KEY',
    SAVE;
>SAVE PARM='C:\Psl3\conc.PAR';
>INPUT NIDW=5,NTOTAL=102,NTEST=1,LENGTH=(99);
 (5A1,102A1)
>TEST1 TNAME=concEX,ITEM=(1(1)43,45(1)52,54(1)62,64(1)102),
 NBLOCK=99;
>BLOCK1 BNAME=MC-UNI,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
 GUESSING=(2,ESTIMATE),GPARM=(.25),REPEAT=26;
>BLOCK2 BNAME=SA-UNI,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
 REPEAT=5;
>BLOCK3 BNAME=OR-UNI,NITEMS=1,NCAT=5,ORIGINAL=(0,1,2,3,4),
```

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

```
 REPEAT=5;
>BLOCK4 BNAME=MC-EQ1,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
 GUESSING=(2,ESTIMATE),GPARM=(.25),REPEAT=7;
>BLOCK5 BNAME=SA-EQ1,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
 REPEAT=1;
>BLOCK6 BNAME=OR-EQ1,NITEMS=1,NCAT=5,ORIGINAL=(0,1,2,3,4),
 REPEAT=1;
>BLOCK7 BNAME=MC-EQ2,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
 GUESSING=(2,ESTIMATE),GPARM=(.25),REPEAT=7;
>BLOCK8 BNAME=SA-EQ2,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
 REPEAT=1;
>BLOCK9 BNAME=OR-EQ2,NITEMS=1,NCAT=5,ORIGINAL=(0,1,2,3,4),
 REPEAT=1;
>BLOCK10 BNAME=MC-EQ3,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
 GUESSING=(2,ESTIMATE),GPARM=(.25),REPEAT=7;
>BLOCK11 BNAME=SA-EQ3,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
 REPEAT=1;
>BLOCK12 BNAME=OR-EQ3,NITEMS=1,NCAT=5,ORIGINAL=(0,1,2,3,4),
 REPEAT=1;
>BLOCK13 BNAME=MC-EQ,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
 GUESSING=(2,ESTIMATE),GPARM=(.25),REPEAT=26;
>BLOCK14 BNAME=SA-EQ3,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
 REPEAT=5;
>BLOCK15 BNAME=OR-EQ3,NITEMS=1,NCAT=5,ORIGINAL=(0,1,2,3,4),
 REPEAT=5;
>CAL GRADED,LOGISTIC,NQPTS=30,CYCLE=(50,2,2,2,2),
   GPRIOR,TPRIOR,SPRIOR,NEWTON=0,CRIT=0.001,DIAGNOS=1;
>SCORE NOSCORE;
```

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Appendix B

Separate Calibration

This appendix gives the sample PARSCALE control files used in the separate calibrations with TCC and M/S transformations. Note that the procedure of transformation rather than the calibration involves the consideration of outliers.

Record File for the Separate Calibration of Reference Tests

```
Calibration of the reference test.
>COMMENT ;
This analysis is based on 66 items. There are:
26 operational MC items in the reference test, 3PL model,
 5 operational SA items in the reference test, 2PL Model,
 5 operational OR items in the reference test, GR model, 5 categories,
24 common MC items, 3PL model,
 3 common SA items, 2PL Model,
 3 common OR items, GR model, 5 categories,
>FILE DFNAME='C:\Psl3\ref.DAT',
    NFNAME='C:\Psl3\ref.KEY',
    SAVE;
>SAVE PARM='C:\Psl3\ref.PAR';
>INPUT NIDW=4,NTOTAL=66,NTEST=1,LENGTH=(66), SAMPLE=100;
 (4A1,66A1)
>TEST1 TNAME=ref,ITEM=(1(1)66),NBLOCK=66;
>BLOCK1 BNAME=MC-UNI,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
  GUESSING=(2,ESTIMATE),GPARM=(.25),REPEAT=26;
>BLOCK2 BNAME=SA-UNI,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
  REPEAT=5;
>BLOCK3 BNAME=OR-UNI,NITEMS=1,NCAT=5,ORIGINAL=(0,1,2,3,4),
  REPEAT=5;
>BLOCK4 BNAME=MC-EQ1,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
  GUESSING=(2,ESTIMATE),GPARM=(.25),REPEAT=8;
>BLOCK5 BNAME=SA-EQ1,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
  REPEAT=1;
>BLOCK6 BNAME=OR-EQ1,NITEMS=1,NCAT=5,ORIGINAL=(0,1,2,3,4),
  REPEAT=1;
>BLOCK7 BNAME=MC-EQ2,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
  GUESSING=(2,ESTIMATE),GPARM=(.25),REPEAT=8;
>BLOCK8 BNAME=SA-EQ2,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
  REPEAT=1;
>BLOCK9 BNAME=OR-EQ2,NITEMS=1,NCAT=5,ORIGINAL=(0,1,2,3,4),
  REPEAT=1;
```

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

```
>BLOCK10 BNAME=MC-EQ3,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
 GUESSING=(2,ESTIMATE),GPARM=(.25),REPEAT=8;
>BLOCK11 BNAME=SA-EQ3,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
 REPEAT=1;
>BLOCK12 BNAME=OR-EQ3,NITEMS=1,NCAT=5,ORIGINAL=(0,1,2,3,4),
 REPEAT=1;
>CAL GRADED,LOGISTIC,NQPTS=30,CYCLE=(50,2,2,2,2),
   GPRIOR,TPRIOR,SPRIOR,NEWTON=0,CRIT=0.001,DIAGNOS=1;
>SCORE NOSCORE;
```

Record File for the Separate Calibration of Equated Tests

```
Calibration of the equated test.
>COMMENT ;
This analysis is based on 66 items. There are:
26 operational MC items in the reference test, 3PL model,
 5 operational SA items in the reference test, 2PL Model,
 5 operational OR items in the reference test, GR model, 5 categories,
24 common MC items, 3PL model,
 3 common SA items, 2PL Model,
 3 common OR items, GR model, 5 categories,
>FILE DFNAME='C:\Psl3\new.DAT',
    NFNAME='C:\Psl3\new.KEY',
    SAVE;
>SAVE PARM='C:\Psl3\new.PAR';
>INPUT NIDW=4,NTOTAL=66,NTEST=1,LENGTH=(66);
 (4A1,66A1)
>TEST1 TNAME=new,ITEM=(1(1)66),NBLOCK=66;
>BLOCK1 BNAME=MC-UNI,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
 GUESSING=(2,ESTIMATE),GPARM=(.25),REPEAT=26;
>BLOCK2 BNAME=SA-UNI,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
 REPEAT=5;
>BLOCK3 BNAME=OR-UNI,NITEMS=1,NCAT=5,ORIGINAL=(0,1,2,3,4),
 REPEAT=5;
>BLOCK4 BNAME=MC-EQ1,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
 GUESSING=(2,ESTIMATE),GPARM=(.25),REPEAT=8;
>BLOCK5 BNAME=SA-EQ1,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
 REPEAT=1;
>BLOCK6 BNAME=OR-EQ1,NITEMS=1,NCAT=5,ORIGINAL=(0,1,2,3,4),
 REPEAT=1;
>BLOCK7 BNAME=MC-EQ2,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
 GUESSING=(2,ESTIMATE),GPARM=(.25),REPEAT=8;
>BLOCK8 BNAME=SA-EQ2,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
 REPEAT=1;
>BLOCK9 BNAME=OR-EQ2,NITEMS=1,NCAT=5,ORIGINAL=(0,1,2,3,4),
 REPEAT=1;
```

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

```
>BLOCK10 BNAME=MC-EQ3,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
 GUESSING=(2,ESTIMATE),GPARM=(.25),REPEAT=8;
>BLOCK11 BNAME=SA-EQ3,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
 REPEAT=1;
>BLOCK12 BNAME=OR-EQ3,NITEMS=1,NCAT=5,ORIGINAL=(0,1,2,3,4),
 REPEAT=1;
>CAL GRADED,LOGISTIC,NQPTS=30,CYCLE=(50,2,2,2,2),
   GPRIOR,TPRIOR,SPRIOR,NEWTON=0,CRIT=0.001,DIAGNOS=1;
>SCORE NOSCORE;
```

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

Appendix C

FCIP Calibration

This appendix gives the sample PARSCALE control files used in the FCIP calibrations for the conditions that outliers (with 3 and 9 score-points) were fixed, outliers (with 3 score-points and from one content area) were not fixed, outliers (with 3 score-points and from one content area) were excluded in the data file. The file ref.IFL was renamed from the file ref.PAR, which was generated from the calibration of reference tests.

Record File for the FCIP Calibration with Outliers Fixed

```
FCIP Calibration with outliers fixed.
>COMMENT ;
This analysis is based on 66 items. There are:
26 operational MC items in the reference test, 3PL model,
 5 operational SA items in the reference test, 2PL Model,
 5 operational OR items in the reference test, GR model, 5 categories,
24 common MC items, 3PL model,
 3 common SA items, 2PL Model,
 3 common OR items, GR model, 5 categories,
>FILE DFNAME='c:\Psl3\new.DAT',
    NFNAME='c:\Psl3\new.KEY',
    IFNAME='c:\Psl3\ref.IFL',
    SAVE;
>SAVE PARM='c:\Psl3\new.PAR';
>INPUT NIDW=5,NTOTAL=66,NTEST=1,LENGTH=(66), SAMPLE=100;
 (5A1,66A1)
>TEST1 TNAME=ref,ITEM=(1(1)66),NBLOCK=66;
>BLOCK1 BNAME=MC-UNI,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
  GUESSING=(2,ESTIMATE),GPARM=(.25),REPEAT=26;
>BLOCK2 BNAME=SA-UNI,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
  REPEAT=5;
>BLOCK3 BNAME=OR-UNI,NITEMS=1,NCAT=5,ORIGINAL=(0,1,2,3,4),
  REPEAT=5;
>BLOCK4 BNAME=MC-EQ1,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
  GUESSING=(2,ESTIMATE),GPARM=(.25),REPEAT=8,SKIP;
>BLOCK5 BNAME=SA-EQ1,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
  REPEAT=1,SKIP;
>BLOCK6 BNAME=OR-EQ1,NITEMS=1,NCAT=5,ORIGINAL=(0,1,2,3,4),
```

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

```
 REPEAT=1,SKIP;
>BLOCK7 BNAME=MC-EQ2,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
 GUESSING=(2,ESTIMATE),GPARM=(.25),REPEAT=8,SKIP;
>BLOCK8 BNAME=SA-EQ2,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
 REPEAT=1,SKIP;
>BLOCK9 BNAME=OR-EQ2,NITEMS=1,NCAT=5,ORIGINAL=(0,1,2,3,4),
 REPEAT=1,SKIP;
>BLOCK10 BNAME=MC-EQ3,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
 GUESSING=(2,ESTIMATE),GPARM=(.25),REPEAT=8,SKIP;
>BLOCK11 BNAME=SA-EQ3,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
 REPEAT=1,SKIP;
>BLOCK12 BNAME=OR-EQ3,NITEMS=1,NCAT=5,ORIGINAL=(0,1,2,3,4),
 REPEAT=1,SKIP;
>CAL GRADED,LOGISTIC,NQPTS=30,CYCLE=(50,2,2,2,2),
   GPRIOR,TPRIOR,SPRIOR,NEWTON=0,CRIT=0.001,DIAGNOS=1;
>SCORE NOSCORE;
```

Record File for the FCIP Calibration with Outliers Not Fixed

```
FCIP Calibration with outliers NOT fixed.
>COMMENT ;
This analysis is based on 66 items. There are:
26 operational MC items in the reference test, 3PL model,
 5 operational SA items in the reference test, 2PL Model,
 5 operational OR items in the reference test, GR model, 5 categories,
24 common MC items, 3PL model,
 3 common SA items, 2PL Model,
 3 common OR items, GR model, 5 categories,
>FILE DFNAME='c:\Psl3\new.DAT',
    NFNAME='c:\Psl3\new.KEY',
    IFNAME='c:\Psl3\ref.IFL',
    SAVE;
>SAVE PARM='c:\Psl3\new.PAR';
>INPUT NIDW=4,NTOTAL=66,NTEST=1,LENGTH=(66);
 (4A1,66A1)
>TEST1 TNAME=fcipNF,ITEM=(1(1)66),NBLOCK=66;
>BLOCK1 BNAME=MC-UNI,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
 GUESSING=(2,ESTIMATE),GPARM=(.25),REPEAT=26;
>BLOCK2 BNAME=SA-UNI,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
 REPEAT=5;
>BLOCK3 BNAME=OR-UNI,NITEMS=1,NCAT=5,ORIGINAL=(0,1,2,3,4),
 REPEAT=5;
>BLOCK4 BNAME=MC-EQ1,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
 GUESSING=(2,ESTIMATE),GPARM=(.25),REPEAT=7,SKIP;
>BLOCK5 BNAME=MC-EQ1,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
 GUESSING=(2,ESTIMATE),GPARM=(.25),REPEAT=1;
```

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

```
>BLOCK6 BNAME=SA-EQ1,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
 REPEAT=1,SKIP;
>BLOCK7 BNAME=OR-EQ1,NITEMS=1,NCAT=5,ORIGINAL=(0,1,2,3,4),
 REPEAT=1,SKIP;
>BLOCK8 BNAME=MC-EQ2,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
 GUESSING=(2,ESTIMATE),GPARM=(.25),REPEAT=6,SKIP;
>BLOCK9 BNAME=MC-EQ1,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
 GUESSING=(2,ESTIMATE),GPARM=(.25),REPEAT=1;
>BLOCK10 BNAME=MC-EQ1,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
 GUESSING=(2,ESTIMATE),GPARM=(.25),REPEAT=1,SKIP;
>BLOCK11 BNAME=SA-EQ2,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
 REPEAT=1,SKIP;
>BLOCK12 BNAME=OR-EQ2,NITEMS=1,NCAT=5,ORIGINAL=(0,1,2,3,4),
 REPEAT=1,SKIP;
>BLOCK13 BNAME=MC-EQ3,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
 GUESSING=(2,ESTIMATE),GPARM=(.25),REPEAT=6,SKIP;
>BLOCK14 BNAME=MC-EQ1,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
 GUESSING=(2,ESTIMATE),GPARM=(.25),REPEAT=1;
>BLOCK15 BNAME=MC-EQ1,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
 GUESSING=(2,ESTIMATE),GPARM=(.25),REPEAT=1,SKIP;
>BLOCK16 BNAME=SA-EQ3,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
 REPEAT=1,SKIP;
>BLOCK17 BNAME=OR-EQ3,NITEMS=1,NCAT=5,ORIGINAL=(0,1,2,3,4),
 REPEAT=1,SKIP;
>CAL GRADED,LOGISTIC,NQPTS=30,CYCLE=(50,2,2,2,2),
   GPRIOR,TPRIOR,SPRIOR,NEWTON=0,CRIT=0.001,DIAGNOS=1;
>SCORE NOSCORE;
```

Record File for the FCIP Calibration with Outliers Excluded

```
 FCIP Calibration with outliers excluded.
>COMMENT ;
This analysis is based on 63 items. There are:
26 operational MC items in the equated test, 3PL model,
 5 operational SA items in the equated test, 2PL Model,
 5 operational OR items in the equated test, GR model, 5 categories,;
21 common MC items, 3PL model,
 3 common SA items, 2PL Model,
 3 common OR items, GR model, 5 categories,
>FILE DFNAME='C:\Psl3\new.DAT',
    NFNAME='C:\Psl3\new.KEY',
    IFNAME='c:\Psl3\ref.IFL',
    SAVE;
>SAVE PARM='C:\Psl3\new.PAR';
>INPUT NIDW=4,NTOTAL=66,NTEST=1,LENGTH=(63);
 (4A1,66A1)
```

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

```
>TEST1 TNAME=fcipEX,ITEM=(1(1)43,45(1)52,54(1)62,64(1)66),
 NBLOCK=63;
>BLOCK1 BNAME=MC-UNI,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
 GUESSING=(2,ESTIMATE),GPARM=(.25),REPEAT=26;
>BLOCK2 BNAME=SA-UNI,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
 REPEAT=5;
>BLOCK3 BNAME=OR-UNI,NITEMS=1,NCAT=5,ORIGINAL=(0,1,2,3,4),
 REPEAT=5;
>BLOCK4 BNAME=MC-EQ1,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
 GUESSING=(2,ESTIMATE),GPARM=(.25),REPEAT=7,SKIP;
>BLOCK5 BNAME=SA-EQ1,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
 REPEAT=1,SKIP;
>BLOCK6 BNAME=OR-EQ1,NITEMS=1,NCAT=5,ORIGINAL=(0,1,2,3,4),
 REPEAT=1,SKIP;
>BLOCK7 BNAME=MC-EQ2,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
 GUESSING=(2,ESTIMATE),GPARM=(.25),REPEAT=7,SKIP;
>BLOCK8 BNAME=SA-EQ2,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
 REPEAT=1,SKIP;
>BLOCK9 BNAME=OR-EQ2,NITEMS=1,NCAT=5,ORIGINAL=(0,1,2,3,4),
 REPEAT=1,SKIP;
>BLOCK10 BNAME=MC-EQ3,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
 GUESSING=(2,ESTIMATE),GPARM=(.25),REPEAT=7,SKIP;
>BLOCK11 BNAME=SA-EQ3,NITEMS=1,NCAT=2,ORIGINAL=(0,1),
 REPEAT=1,SKIP;
>BLOCK12 BNAME=OR-EQ3,NITEMS=1,NCAT=5,ORIGINAL=(0,1,2,3,4),
 REPEAT=1,SKIP;
>CAL GRADED,LOGISTIC,NQPTS=30,CYCLE=(50,2,2,2,2),
   GPRIOR,TPRIOR,SPRIOR,NEWTON=0,CRIT=0.001,DIAGNOS=1;
>SCORE NOSCORE;
```

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.