This is to certify that the
dissertation entitled

The Effect of Weighting in Kernel Equating Using Counter-Balanced Designs

presented by

Yanxuan Qu

has been accepted towards fulfillment
of the requirements for the

Ph. D. degree in Counseling, Educational Psychology, and Special Education

Mark D. Reckase
Major Professor's Signature

6/7/07
Date

THE EFFECT OF WEIGHTING IN KERNEL EQUATING USING COUNTER-BALANCED DESIGNS

By

Yanxuan Qu

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Counseling, Educational Psychology, and Special Education

2007

ABSTRACT

THE EFFECT OF WEIGHTING IN KERNEL EQUATING
USING COUNTER-BALANCED DESIGNS

By

Yanxuan Qu

The Counter-Balanced (CB) design for test equating is often used in pilot studies for testing programs when sample size is limited. When a CB design is used to conduct equating, data are usually treated as an Equivalent Group design or a Single Group design (Kolen & Brennan, 2004). On the other hand, von Davier, Holland and Thayer (2004) proposed a new approach under the Kernel Equating (KE) framework which treats data as a weighted synthesized mixture of data from the two groups. This new approach is named as the two independent Single Group approach (2SG approach).

This study investigates the performance of the 2SG approach in comparison to other data treatment approaches under different sample sizes and order effect situations. Both linear and equipercentile equating methods under KE and traditional equating frameworks were applied to two real datasets and six simulated datasets. The results from traditional equipercentile equating on each simulated population data were considered as the benchmark to which all the other equating methods were compared. Standard Errors of Equating (SEE), Root Mean Square Error (RMSE), equating bias, and Standard Error of Equating Difference (SEED) were reported for each equating of the simulated data. The standard Error of Equating and Root Mean Square Error were reported for equating of the real data samples.

The results indicated the 2SG approach unifies the Equivalent Group approach

and the Single Group approach into its flexible framework. The weighting mechanism in the 2SG approach seemed to be sensitive to different order effects. Possible criteria for selecting the best weights are discussed.

DEDICATION

To my dear parents, my husband, and my little brother

ACKNOWLEDGEMENTS

This dissertation work is completed under the help of many people. First, I am deeply indebted to Professor Mark D. Reckase for his guidance and encouragement in this dissertation work. Without his constant and unconditional support, this work would not have been possible. I learned from him not only his knowledge, but also his dedication to work and his peaceful and respectful attitude to people.

I would like to thank Dr. Alina von Davier for her generous help and guidance. She is enthusiastic, upbeat and proactive. Thanks also go to Dr. Richard Houang, and Dr. Sharif Shakrani for their insightful comments on this dissertation; Dr. Ning Han and Dr. Henry Chen for their assistance in the KE software and Dr. Linda Chard for her assistance in editing the early version of my dissertation.

Meanwhile, I am very grateful to Dr. Betsy Becker, Dr. Mary Kennedy, and Dr. Edward Wolfe. Working with them on different projects broadened my scope of knowledge. Their financial support made me concentrate on my study and made me feel the family-like atmosphere.

Finally, my deep gratitude goes to my husband Lixiong Gu for his love and support, and to my parents and my brother, for their understanding and encouragement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

NOTATION

Symbol	Explanation
X, Y	Names of two test forms to be equated
$\boldsymbol{X}$, $\boldsymbol{Y}$	Scores on X and Y, random variables
P	Population of examinees
T	Target population of examinees on which the equating of X and Y takes place
CB	Counter-Balanced data collection design
EG	Equivalent Group Design
SG	Single Group Design
DF	Design Function
X_1	Test X that is taken first
X_2	Test X that is taken second
Y_1	Test Y that is taken first
Y_2	Test Y that is taken second
$F(x)$	Cumulative distribution of variable $\boldsymbol{X}$
$G(y)$	Cumulative distribution of variable $\boldsymbol{Y}$
J	Number of possible $\boldsymbol{X}$ scores
K	Number of possible $\boldsymbol{Y}$ scores
x_j	A possible score value for $\boldsymbol{X}$, j is from 1 to J
y_k	A possible score value for $\boldsymbol{Y}$, k is from 1 to K
$\boldsymbol{R}$	Generic symbol for the population probability of $\boldsymbol{X}$ after pre-smoothing for all designs
$\boldsymbol{S}$	Generic symbol for the population probability of $\boldsymbol{Y}$ after pre-smoothing for all designs
$\boldsymbol{r}$	Estimated probabilities on target population T, transformed by DF from $\boldsymbol{R}$ into $\boldsymbol{r}$
$\boldsymbol{s}$	Estimated probabilities on target population T, transformed by DF from $\boldsymbol{S}$ into $\boldsymbol{s}$
$\hat{e}_Y(x)$	Estimated score x on Form X equated to Form Y
$\hat{e}_X(y)$	Estimated score y on Form Y equated to Form X
$\hat{r}_j$	An estimated specific value of $\boldsymbol{r}$
$\hat{s}_k$	An estimated specific value of $\boldsymbol{s}$
$\hat{p}_j$	Estimated probability of getting a score x_j on X
$\hat{p}_k$	Estimated probability of getting a score y_k on Y
$\hat{p}_{jk}$	Estimated joint probability of getting a score x_j on X and a score y_k on Y over the target population, T.
$\hat{p}_{(12)jk}$	Estimated population probability of getting a score x_j on test X_1 which

	is taken first and a score y_k on test Y_2 which is taken second
$\hat{P}_{(21)jk}$	Estimated population probability of getting a score x_j on test X_2 which is taken second and a score y_k on test Y_1 which is taken first
h_X, h_Y	Bandwidth used to define the KE continuizations of *F(x)* and *G(y)*. They are positive numbers. Large values of the bandwidths lead to linear equating, while smaller values give more "equipercentile-like" equating functions.
$X(h_X)$	Continuized random variable for scores on Form X
$Y(h_Y)$	Continuized random variable for scores on Form Y
$J_{e_Y(\hat{r},\hat{s})}$	Jacobian matrix of the KE function, which is a function of $\hat{r}$ and $\hat{s}$
$J_{DF(\hat{R},\hat{S})}$	Jacobian matrix of the design function, which is a function of $\hat{R}$ and $\hat{S}$

Chapter I: Introduction

Test equating is an important statistical procedure in educational testing. It is used to produce scores that are comparable across different but parallel test forms, both within a year and across years. Although there have been many comparative studies investigating the accuracy of different equating methods, very few studies have been done for equating with a Counter-Balanced (CB) design. Traditionally as in Lord (1950), Angoff (1971) and Kolen and Brennan (2004), data collected by a CB design were either pooled together as a Single Group (SG) design or discarded as an Equivalent Group (EG) design. Recently, a new approach of treating data collected by a CB design was proposed by von Davier, Holland and Thayer (2004). This new approach involves weighting data before pooling them together. To evaluate the performance of this new approach, this study compared the overall equating accuracy of the two independent single group approach, abbreviated as the 2SG approach, to the other approaches of treating data collected by a CB design.

The rest of this chapter introduces the general procedure for equating using the counter-balanced design and equating approaches for a CB design including the new 2SG approach under the Kernel Equating (KE) framework, and gives a brief summary of literature on KE equating. At the end of this chapter, the research questions and research expectations of this study are presented. Chapter II describes the CB design and KE framework as well as equating errors and the evaluation of equating results. Chapter III describes the real and simulated datasets to which the equating methods were applied and the procedure of this study. Chapter IV presents the study results and Chapter V discusses the findings and limitations of this study.

1.1 Equating Procedure in General

Every equating procedure consists of two basic components: equating design and equating methods. Typical equating designs include Equivalent Group (also called random group) design, Single Group design, Counter-Balanced design, and Non-Equivalent Anchor Test (NEAT) design. Typical equating methods can be classified into the following three categories: 1) Classical observed score equating; 2) Item Response Theory (IRT) true score equating; and 3) Item Response Theory observed score equating. Classical observed score equating methods include the mean, linear, and equipercentile equating methods reported by Kolen (1988). They define the score correspondence between two forms by setting certain characteristics of observed score distributions for a specified group of examinees. Item response theory true score equating defines the score correspondence by setting the true scores of examinees to be equal (Cook & Eignor, 1991).

1.2 Counter-Balanced Design and Equating

Counterbalance or Latin Square is often used in pure experimental designs to cancel out order effects (Montogomery, 2000). In educational testing, a CB design is often used to collect data in pilot studies of testing programs. In a CB design, two independent groups of examinees usually take two parallel test forms X and Y in different order.

Various ways of dealing with data in a CB design test equating were described in Lord (1950), Angoff (1971), and Kolen and Brennan (2004). None of these approaches is satisfactory for situations when order effect cannot be cancelled out. In order to improve the equating practice for a CB design, especially when order effects cannot be cancelled

out, von Davier, Holland, and Thayer (2004) proposed a new way of treating data collected by a CB design under their Kernel Equating framework. This new way of treating data is named the *two independent single group* approach (2SG approach), which creates a synthetic target group by assigning different weights to the two tests taken in different order, and applies linear and equipercentile equating methods to the synthetic group. The significance of this approach is its weighting mechanism, which is supposed to have the potential to provide optimal equating results with the smallest equating error by using as much data information as possible. However, the effectiveness of this 2SG approach hasn't been evaluated.

The 2SG approach, the EG approach, and the SG approach are all about data collection designs in an equating procedure. The 2SG approach is under the framework of Kernel Equating. The equating methods related to this approach are KE linear or KE equipercentile equating methods. The EG approach and SG approach can be implemented under both KE and traditional equating framework. Therefore, the equating methods related to these two approaches are the KE linear, KE equipercentile, traditional linear or traditional equipercentile equating methods (see more details in Chapter II).

1.3 Literature Review

Descriptions about equating using a CB design can be found in Lord (1950), Angoff (1971), Kolen and Brennan (2004), Zeng and Cope (1995) and von Davier, Holland, and Thayer (2004). The 2SG approach of treating data collected by a CB design was mentioned in von Davier, Holland, and Thayer (2004). The only study compared the performance of this 2SG approach with the EG and SG approach in improving equating accuracy of a CB design equating is conducted by Qu and von Davier (2006). They

compared the 2SG approach to the SG and EG approach under KE framework using a real data collected by a CB design. It was found that, when order effect can be cancelled out, the 2SG approach with equal weights produce similar equating results as the SG approach under KE framework. It is still unclear how the 2SG approach performs when order effects cannot be cancelled out. Moreover, it is not well documented in the literature how to test whether the order effects can or cannot be cancelled out.

The 2SG approach is carried out under the KE framework. KE is a unified approach to test equating based on a flexible family of equipercentile-like equating functions that contain the linear equating function as a special case. It belongs to the category of classical observed score equating. Studies comparing the KE methods with other equating procedures concluded that the KE procedure can improve or approximate the equating results of corresponding traditional equating methods.

Livingston (1993a) compared KE methods with traditional linear and equipercentile equating methods using small samples collected by a NEAT design. He evaluated the equating methods in terms of random equating error and equating bias and found that the KE methods with log-linear smoothing provided more accurate equating results, when compared to traditional equating methods without smoothing. He also found that, compared to the empirical standard error of equating, the analytic standard error of equating calculated by the delta method is larger at the lower or higher score range when sample size is less than 200.

Mao and von Davier (2005) compared Kernel Equating methods with their corresponding traditional equating methods using real data in a NEAT design and an EG design. For the NEAT design, they compared the traditional frequency estimation

equipercentile equating with KE post-stratification equating method and the Tucker method with the KE linear post-stratification equating method. They found that KE methods and their corresponding traditional equating methods have very similar equating results. Von Davier, Holland, and others (2005) did a similar study using a pseudo-test data with a NEAT design and drew the same conclusion.

Han, Li, and Hambleton (2005) compared KE with IRT true score equating methods using data collected by a NEAT design. Again, they found the KE methods provide similar equating results as those of the IRT equating methods.

1.4 Research Questions

This study intends to quantify differential order effects, to compare the 2SG equating procedures under KE framework with other traditional equating procedures, and to discover whether the weighting mechanism can enhance the equating accuracy under different order effect situations. The specific research questions are:

1) How should differential order effects in CB designs be quantified?
2) Are the KE methods better than their corresponding traditional equating methods?
3) Does the weighting in the 2SG approach provide better results under certain order effect situation?
4) What weight should be used for a 2SG approach?

Table 6 displays the 22 equating procedures compared in this dissertation. What distinguishes them from each other are the way they treat the data collected by a CB design (EG, SG or 2SG with weighting) and the equating method (linear or equipercentile) they adopted. To compare the performance of KE with traditional

equating methods, the equating results of two KE procedures are compared to the equating results of their corresponding traditional equating procedures (as listed in table 5).

1.5 Research Expectations

1) The KE equating methods and their corresponding traditional equating methods provide similar equating results.
2) As DOE increases, the weights of the 2SG approach assigned on tests taken first increases accordingly.
3) Decision on the selection of an equating function with the optimal weights may vary when using different statistical criterion to evaluate the equating results.

As presented above, the literature on any CB design equating is sparse. Since CB design is still used in research projects and in the pilot study of testing programs (Yu, 2003) when examinees are hard to find, it is useful to comprehend the 2SG approach and to evaluate how much it can enhance overall equating accuracy when compared to other methods in various order effect situations. Such a study will contribute to the general knowledge about a CB design and the methods available for equating using data collected by a CB design.

Chapter II: Theoretical Framework

This chapter first introduces the equating designs related to a CB design, the linear and equipercentile equating methods and the Kernel Equating framework, and then describes the concept of equating error and the criteria used for evaluating equating

results.

2.1 Counter-Balanced Design

A CB design is often used in practice when administering two forms to examinees where it is difficult to obtain sufficiently large group of examinees (Kolen & Brennan, 2004). To explain the CB design in more detail, a brief description about EG design and SG design is necessary:

Equivalent Group Design

TABLE 1. *Equivalent-Groups design*

Population	Sample	*X*	*Y*
P	1	√	
P	2		√

In an EG design, two independent random samples are drawn from a common population of examinees, P. Each group of examinees is randomly assigned to take one of the two parallel forms *X* and *Y* as shown in Table 1.

Single Group Design

TABLE 2. *Single-Group design*

Population	Sample	X	Y
P	1	√	√

In a SG design, only one random sample of examinees is selected from population P, and all the examinees take the two test forms X and Y in one administration as shown in Table 2. Because the two test forms are parallel and they are taken by the same examinee, it is almost certain that the examinee's performance on the second form will be affected by their performance on the first form. The effect may be a "practice/learning effect," or "fatigue effect." If familiarity with the test increased performance, then Form Y could appear to be easier than Form X. On the other hand, if fatigue is a factor in examinee performance, then Form Y could appear relatively more difficult than Form X because examinees would be tired when administered Form Y (Kolen & Brennan, 2004). For simplicity, all such possible effects will be named as "order effect" (Lord, 1950). If the two test forms are administered in the same order to all examinees, as in a SG design, it is impossible to obtain any estimate of the amount of order effect. Consequently, to control for the order effect, it is usual to counterbalance the order of administration by dividing the group in a SG design into two random halves and giving two test forms to each group but in different order. This design is what is often called a CB design.

TABLE 3. *Counter-Balanced design*

Population	Sample	X_1	Y_1	X_2	Y_2
P	1	√			√
P	2		√	√	

*The subscripts of X and Y indicate the order. Eg., X_1 means take test X first, Y_2 means take test Y second.

Table 3 illustrates a CB design, in which, two samples of examinees were

randomly chosen from a same population *P* and were randomly assigned as sample 1 and sample 2. Sample 1 takes test X first (denoted as X_1), test Y second (denoted as Y_2), and sample 2 takes test Y first (denoted as Y_1) and test X second (denoted as X_2). The purpose of counterbalancing the order of testing is to ensure any order effects are present equally in the scores obtained for both test forms X and Y such that the order effects on Form X and Form Y can be cancelled out.

Theoretically, if random selection and random assignment of the examinees are carried out strictly in operation, the purpose of canceling out "order effect" can be accomplished by collecting data using a CB design. However, in practice, the assumption of random selection is often violated. Usually, random sampling is replaced by random cluster sampling. The violation of these two assumptions leads to the interaction between group abilities and form difficulties, which is the reason why the order effects often cannot be cancelled out. For example, some group of people might do better on the second test after practicing on the first test, while the other groups might do worse.

There have been different definitions for order effects in literature. Lord (1950) and Angoff (1971) defined the order effect on Form X as $K_X = X_2 - X_1 = C\sigma_{X_1} = C\sigma_{X_2} = C\sigma_X$, and the order effect on test Form Y as $K_Y = Y_2 - Y_1 = C\sigma_{Y_1} = C\sigma_{Y_2} = C\sigma_Y$ (where *C* is a constant). They assumed that order effects are constant for all examinees and are proportional to the standard deviations. Kolen and Brennan (2004) explained order effects without assuming they are constant for each examinee. They defined Differential Order Effect (DOE) as $(\bar{X}_1 - \bar{Y}_1) - (\bar{X}_2 - \bar{Y}_2)$ and suggested that a significant DOE would indicate that order

effects cannot be cancelled out in a CB design. However, there is not a significance test described in their book. In chapter III, this dissertation adopted their definition of DOE, described a hypothesis testing for the statistical significance of DOE and suggested using the effect size statistics for the magnitude of DOE.

2.2 Equating Using Counter-Balanced Designs

Like every equating procedure, equating using a CB design has two parts: data collection design and equating methods.

2.2.1 Approaches to Treating Data in a CB Design: The nature of CB design leads to different ways of dealing with data. Comparing tables 1, 2 and 3 we see that CB design actually contains both EG and SG designs. For example, there are two (dependent) EG designs, one for X_1 and Y_1, and the other for X_2 and Y_2. In addition, there are two (independent) SG designs, one for X_1 and Y_2, and the other for X_2 and Y_1. Finally, the two groups of examinees can be pooled together and all the data from X_1, Y_2, X_2 and Y_1 can be treated as a pooled SG design.

Because of these different ways of considering data in a CB design, several data treatment approaches have been used to equate test forms X and Y. Lord (1950) and Angoff (1971) described a linear equating method that actually treated the data as pooled single group design. They assume constant order effect and bivariate normal distributions of test X and Y in the population. By constant order effect, they mean that order effects are the same for all examinees and are proportional to the relevant standard deviations. Kolen and Brennan (2004) did not assume constant order effects across examinees. They suggested using the pooled SG approach when order effects can be cancelled out. Otherwise, only the EG approach with X_1 and Y_1 should be used, since it is perhaps the

only unbiased way of treating data in a CB design.

Nonetheless, each of these two approaches for treating data has its own weaknesses. Although *The EG approach using X_1 and Y_1 only* is unbiased, it throws away half of the data and makes no use of the correlation between X and Y, which is implicit in the SG aspects of the CB design. *The pooled SG approach* is considered problematic when order effects cannot be cancelled out because it is hard to interpret the pooled distribution of X_1 and X_2 (or Y_1 and Y_2) when they each have a different distribution (von Davier, Holland, & Thayer, 2004).

In an attempt to find a better way of using data collected by a CB design, von Davier, Holland, and Thayer (2004) proposed *the 2SG approach*, a new approach using all data information as much as possible and more flexibly. It is expected to be able to unify the other three approaches into one single approach and provide an optimal equating solution while taking into account different sizes of order effects. Section 2.3 explains this approach under the KE framework in detail.

Table 4 summarizes different ways of dealing with data in a CB design discussed in literature review.

TABLE 4. *Ways of treating data in a CB design appearing in the literature*

EG design for X_1 and Y_1 only	Explanation	Use data from X_1 and Y_1 only
	Assumptions	Random selection from a single population & random assignment Suggested when DOE is significant
	Advantage/Disadvantage	Unbiased/loss of half data
	Source	Kolen and Brennan (2004), von Davier et al. (2004)
EG design for X_2 and Y_2 only	Explanation	Use data from X_2 and Y_2 only
	Assumptions	Random selection from a single population; random assignment Definitely not when DOE is significant
	Advantage/Disadvantage	/biased; loss of half data
	Source	Kolen and Brennan (2004)
EG pooling approach	Explanation	Average two EG equating functions
	Assumptions	Random selection; random assignment DOE is not significant
	Advantage/Disadvantage	Use full data information/ignore dependency between two equating functions
	Source	Von Davier et al. (2004)
SG design for X_1 and Y_2 only	Explanation	Use data from X_1 and Y_2 only
	Assumptions	Random selection DOE is not significant
	Advantage/Disadvantage	/loss of data information
	Source	Kolen and Brennan (2004)
SG design for X_2 and Y_1 only	Explanation	Use data from X_2 and Y_1 only
	Assumptions	Random selection DOE is not significant
	Advantage/Disadvantage	/loss of data information
	Source	Kolen and Brennan (2004)
Pooled SG approach	Explanation	Use all data from X_1, Y_1, X_2 and Y_2 equally when order effect can be cancelled out
	Assumptions	Random selection; random assignment DOE is not significant
	Advantage/Disadvantage	Use full data information/not applicable when DOE is significant
	Source	Kolen and Brennan (2004), Lord (1950), von Davier et al. (2004)
2SG approach	Explanation	Use all data information unequally when different order effects present
	Assumptions	Random selection & random assignment All kinds of DOE
	Advantage/Disadvantage	Use full data information/
	Source	Von Davier et al. (2004)

* Approaches 2, 3, 4, 5 are possible ways of treating data in a CB design but are of no interest to this study

2.2.2 Equating Methods for a CB Design: Linear or equipercentile equating

methods following KE or traditional equating procedure are the equating methods related to a CB design found in literature.

Every equating method defines a target population *T*, on which scores on the two test forms are to be made equivalent (for the population as a whole, not necessarily for every individual in the population) (Livingston, 2004; von Davier, Holland, & Thayer 2004; etc.). The target population depends on the data collection design. This study focuses on the CB, EG, and SG designs where there is only one population *P* of test takers from which particular samples are drawn. For these designs the target population *T* is assumed to be the same as the underlying population *P* (von Davier, Holland, & Thayer, 2004). The linear equating method is appropriate when tests X and Y have the same distribution on the target population while the equipercentile equating method adjusts for the differences in the distribution.

Linear equating defines the equating relationship as the equivalence of *Z*-scores, whereas equipercentile equating method defines equating relationship as the equivalence of cumulative distribution functions of ***X*** and ***Y*** in the population. Equation (1) and equation (2) define the equating relationship for linear and equipercentile equating when equating ***X*** onto ***Y***, which means each of the raw scores, x_j is transformed to $e_Y(x_j)$ or y by these equating functions, i.e., a raw score of x_j on test X is interchangeable with a raw score of $e_Y(x_j)$ or y on test Y.

$$\frac{x-\mu_X}{\sigma_X}=\frac{y-\mu_Y}{\sigma_Y} \Rightarrow y=\mu_Y+\frac{\sigma_Y}{\sigma_X}(x-\mu_X) \quad (1)$$

$$G(y)=F(x) \Rightarrow y=G^{-1}(F(x)) \quad (2)$$

Equation 2 holds only when ***X*** and ***Y*** are continuous. KE applies the Gaussian

Kernel continuization procedure (von Davier, Holland, & Thayer, 2004). While the traditional equipercentile equating in this study uses linear interpolation to continuize score distributions.

2.3 Equating with a CB Design under the Kernel Equating Framework

The KE framework accommodates both linear and equipercentile equating procedures with pre-smoothing and continuization. Pre-smoothing is the log-linear smoothing before scores are equated. Continuization is used to convert discrete score distributions to continuous distributions by using a normal (Gaussian) "kernel" (Holland & Thayer, 1989; von Davier, Holland, & Thayer, 2004). In the case of a CB design, the KE framework incorporates three different ways of treating data -- *the EG approach, the pooled SG approach, and the 2SG approach*. Both linear and equipercentile equating methods are available to each of the three ways of treating data. The following section introduces the five steps of the KE framework particularly for a CB design and presents how the three approaches differ with respect to each of these five steps.

2.3.1 Step 1. Log-linear Pre-smoothing

In pre-smoothing, the empirical score distributions are smoothed. Smoothing can remove irregularity in the empirical score distributions and make them as smooth as the population score distribution relationship. Smoothing is necessary, especially when sample size is small (Livingston, 1993). KE conducts pre-smoothing using a log-linear method. Compared to the other pre-smoothing methods, the log-linear method has the flexibility of accommodating many distributions and is well-behaved and relatively easy to estimate. Because the log-linear models are a part of the exponential families, the

estimated distribution can match the sample distribution by as many moments as possible (Holland & Thayer, 2000; Kolen & Brennan, 2004).

In this step, a log-linear model with best fit is selected to fit the sample data and to estimate discrete score probabilities. The fit of the log-linear models can be evaluated by examining changes in the likelihood ratio chi-square index over different models and conditional Freeman-Tukey residual plots. The Freeman-Tukey residual plot displays the deviation between $e_y(X)$ and Y or between $e_x(Y)$ and X. A log-linear model with good fit will have conditional Freeman-Tukey residuals randomly distributed within 3 units above or below the zero line. In addition, the fit of a log-linear model can be somehow reflected by the Standard Error of Equating introduced in step 5. A bad model fit could lead to large SEE.

Let J and K denotes the total number of possible scores on Form X and Form Y respectively, x_j represents a possible score value for test X, j=1 to J on X; y_k represents a possible score value for test Y, k = 1 to K on test Y; p_{jk} =Prob{X= x_j , Y= y_k | T }=the bivariate score probability of X= x_j and Y= y_k over the target population T; let β's be the slope parameters that will be estimated by maximum likelihood method, α and α^* are the normalizing constants selected to make the sum of population score probabilities equal to one; let T_X and T_Y denote the number of moments matched between the fitted probabilities and the observed score probabilities; and let I and L denote the number of cross moments matched between the fitted and the observed score probabilities. Then,

A univariate log-linear model takes the form of:

$$\log(p_j) = \alpha + \sum_{i=1}^{I} \beta_i (x_j)^i \tag{3}$$

A bivariate log-linear model takes the form of:

$$\log(p_{jk}) = \alpha^* + \sum_{i=1}^{T_X} \beta_X^i (x_j)^i + \sum_{i=1}^{T_Y} \beta_Y^i (y_k)^i + \sum_{i=1}^{I} \sum_{l=1}^{L} \beta_{il} x_j^i y_k^l \tag{4}$$

For the *SG KE method*, one single bivariate log-linear model is fit to the pooled data to get the probability of an examinee getting a score of j on Form X and a score of k on Form Y (that is $\hat{p}_{jk}$).

For the *2SG KE method*, two separate bivariate log-linear models are fit to two groups of data to get two sets of probability estimates $\hat{p}_{(12)jk}$ and $\hat{p}_{(21)jk}$, where $\hat{p}_{(12)jk}$ is the estimated population probability of getting a score x_j on test X_1, which is taken first and a score y_k on test Y_2 which is taken second; $\hat{p}_{(21)jk}$ is the estimated population probability of getting a score x_j on test X_2 which is taken second and a score y_k on test Y_1 which is taken first.

For the EG approach, data is fit by two univariate log-linear models. Alternatively, the *EG with X_1 and Y_1 only KE method* can be considered as a special case of the *2SG KE method* with weights of (1, 1).

2.3.2 Step 2. Estimating Score Probabilities on the Target Population

In this step, a Design Function (DF), either linear or non-linear, is applied to map the estimated population score probabilities from step 1 into the estimated score

probabilities for X and Y on the target population T, denoted as $\hat{r}_j$ and $\hat{s}_k$.

In the KE method of *EG with X_1 and Y_1 only*, the DF is an identity function, i.e., the estimated probabilities on target population T ($\hat{r}_j$ or $\hat{s}_k$) is identical to the estimated population probabilities, $\hat{p}_j$ or $\hat{p}_k$. For both *pooled SG* and *2SG KE methods*, a non-identity DF is needed to transform the estimated population probabilities from step 1, which is relevant to the data design, into the estimated probabilities over target population T. For *the pooled SG KE method*, $\hat{r}_j^*$ and $\hat{s}_k^*$ is the sum of the joint probabilities over k and j respectively. For *the 2SG KE method*, $\hat{r}_j$ or $\hat{s}_k$ is the weighted average of the two sets of estimates from the two groups.

$$\hat{r}_j^* = \sum_k \hat{p}_{jk}, \tag{5}$$

$$\hat{s}_k^* = \sum_j \hat{p}_{jk}, \tag{6}$$

$$\hat{r}_j = w_x \sum_k \hat{p}_{(12)jk} + (1 - w_x) \sum_k \hat{p}_{(21)jk}, \tag{7}$$

$$\hat{s}_k = w_y \sum_j \hat{p}_{(21)jk} + (1 - w_y) \sum_j \hat{p}_{(12)jk}, \tag{8}$$

Where w_x and w_y indicate the weights placed on the test forms taken first. Depending on the size of DOE, they can be adjusted somewhere between 0.5 and 1 to emphasize information collected from tests taken first. When both w_x and w_y are set to be 1, data from test forms taken second are completely discarded. Thus the *2SG approach*

becomes the *EG approach with X_1 and Y_1 only*. On the other hand, when both w_x and w_y are set to be 0.5, the *2SG approach* approximates the *SG approach* by treating the data equally from tests taken first and second.

2.3.3 Step 3. Continuization

Livingston (1993) clearly explained this step. In all equipercentile equatings, score x on Form X and score y on Form Y are equated in a population of test-takers if and only if they have the same percentile rank in that population. In the real world of educational testing, since the observed test scores are discrete, it is rare to find a score on Form Y that has exactly the same percentile rank in the test-taker population as score x on Form X. In order to do equipercentile equating, discrete percentile rank score distribution has to be continuized. In the KE framework, this "continuization" of the distribution is accomplished when it replaces the frequency at each discrete score value with a continuous frequency distribution centered at that value. In contrast, the traditional equipercentile method uses linear interpolation to continuize discrete score distributions.

By adding a continuous random variable V distributed as $N(0, 1)$, the discrete random variables X and Y are transformed into continuous variables $X(h_X)$ and $Y(h_Y)$ in KE:

$$X(h_X) = a_X(X + h_X V) + (1 - a_X)\mu_X \tag{9}$$

$$Y(h_Y) = a_Y(X + h_Y V) + (1 - a_Y)\mu_Y \tag{10}$$

In the above formula, h_X and h_Y can be any positive number. They are the

bandwidth of the replaced normal distributions for each discrete score; μ_X and σ_X^2 denote the mean and variance of variable X over target population T,

$$\mu_X = \sum_j x_j r_j, \sigma_X^2 = \sum_j (x_j - \mu_X)^2 r_j \; ; \; a_X^2 = \frac{\sigma_X^2}{\sigma_X^2 + h_X^2} \text{ is an adjusting constant.}$$

Since variable V has a continuous normal distribution, it is obvious that $X + h_X V$ will be continuous and so does $X(h_X)$. It can be proved that the transformed continuous variable $X(h_X)$ and $Y(h_Y)$ has the same mean and standard deviation as the discrete variables X and Y respectively.

The selection of h_X (or h_Y) determines the equating method. The KE Optimal (simply as "KE" in Table 6) equating method selects h_X (or h_Y) automatically by minimizing the difference between the probability distributions of X (or Y) before and after continuization $\sum_j (\hat{r}_j - \hat{f}_{h_X}(x_j))^2$, where f_{h_X} is the density of $X(h_X)$). While the KE_Linear (linear) equating method can be approximated by using a large "bandwidth" value which is usually larger than 10 times of the standard deviation of an observed score distribution.

2.3.4 Step 4. Equating

KE defines the equating relationship as the equivalence between the continuized cumulative distributions of $X(h_X)$ and $Y(h_Y)$. For example, the equating function for equating X to Y on target population T is given by:

$$G_{h_Y}(y;\hat{s}) = F_{h_X}(x;\hat{r}) \Rightarrow \hat{y} = G_{h_Y}^{-1}(F_{h_X}(x;\hat{r});\hat{s}) \Leftrightarrow \hat{e}_y(x) = \hat{G}_{h_Y}^{-1}(\hat{F}_{h_X}(x))$$

(11)

Where $F(h_X)$ and $G(h_Y)$ represent cumulative density functions of $X(h_X)$ and $Y(h_Y)$ respectively. The linear equating method is considered as a special case in KE framework.

2.3.5 Step 5. Calculating Standard Error of Equating (SEE) and Standard Error of Equating Difference (SEED)

KE provides a formula for calculating SEE derived from the delta method (see von Davier, Holland, and Thayer, 2004):

$$SEE(\hat{e}_Y(x)) = SEE(e_Y(x;\hat{r},\hat{s})) = \sqrt{J_{e_Y(\hat{r},\hat{s})} J_{DF(\hat{R},\hat{S})} \Sigma_{\hat{R},\hat{S}} J^{t}_{e_Y(\hat{R},\hat{S})} J^{t}_{DF(\hat{r},\hat{s})}} \quad (12)$$

Here $\hat{R}$ and $\hat{S}$ are used as generic names over all the designs for the population score probabilities of $\boldsymbol{X}$ and $\boldsymbol{Y}$ estimated by the log-linear pre-smoothing model in step 1, like $\hat{p}_j$, $\hat{p}_k$, $\hat{p}_{(12)jk}$, and $\hat{p}_{(21)jk}$ etc. When sample size is large, $\begin{pmatrix} \hat{R} \\ \hat{S} \end{pmatrix}$ is asymptotically normally distributed with mean of $\begin{pmatrix} R \\ S \end{pmatrix}$ and variance matrix of $\Sigma_{\hat{R},\hat{S}}$ with dimension $((JK+JK)\times(JK+JK))$; $\hat{r}$ and $\hat{s}$ are the estimated population score probabilities of $\boldsymbol{X}$ and $\boldsymbol{Y}$ over target population T; $\Sigma_{\hat{R},\hat{S}}$ is the covariance matrix of $\hat{R}$ and $\hat{S}$. The estimated equating function is a composition of $\hat{e}_Y$ and DF ($\hat{e}_Y(x) = e_Y(x;\hat{r},\hat{s}) = G^{-1}(\hat{F}(x))$); the design function (DF) is a function of

$\hat{R}$ and $\hat{S}$; $J_{e_Y(\hat{r},\hat{s})}$ and $J_{DF(\hat{R},\hat{S})}$ are Jacobian matrices (in formula 13 and 14) related to the equating function and the design function respectively. $J_{e_Y(\hat{r},\hat{s})}$ is a $(1\times(J+K))$-row vector of the first derivatives of the estimated equating function with respect to each estimated score probabilities $\hat{r}$ and $\hat{s}$ over target population T, and $J_{DF(\hat{R},\hat{S})}$ is a $((J+K)\times(JK+JK))$ - matrix of the first derivatives of the DF with respect to each of the output variables from the pre-smoothing procedure:

$$J_{e_Y(\hat{r},\hat{s})} = \left(\frac{\partial \hat{e}_Y}{\partial \hat{r}}, \frac{\partial \hat{e}_Y}{\partial \hat{s}}\right)_{(1\times(J+K))} \tag{13}$$

$$J_{DF(\hat{R},\hat{S})} = \begin{pmatrix} \frac{\partial \hat{r}}{\partial \hat{R}}, \frac{\partial \hat{r}}{\partial \hat{S}} \\ \frac{\partial \hat{s}}{\partial \hat{R}}, \frac{\partial \hat{s}}{\partial \hat{S}} \end{pmatrix}_{((J+K)\times(JK+JK))} \tag{14}$$

Kernel Equating provides an analytic tool to calculate standard error of equating. It is known as the delta method (also known as Taylor Series method) and provides a statistical procedure widely used to estimate the variance or standard error of a function of some statistical estimates with known asymptotic distributions (Kolen & Brennan, 2004; von Davier, Holland, & Thayer, 2004).

In addition to calculating the conditional SEE's at each score point, KE also provides the SEED statistics for calculating the standard error of equating difference between two KE functions at each score point. Von Davier, Holland, and Thayer (2004) used SEED to decide whether the equating results of two KE methods are significantly different from each other.

2.4 Equating Error

Equating error reflects the difference between the equated scores estimated from the sample and the equated scores from the population. It consists of two sources of error – random equating error and systematic equating error. Random equating error is the error simply due to sampling. Systematic equating error arises if 1) the equating design is inappropriately executed; 2) the statistical assumptions of an equating method are violated; 3) equating procedure is inappropriately implemented, for example, applying an IRT equating to a multidimensional test. The definition of random error and systematic error determines that the magnitude of the random equating error closely depends on the sample size, while the systematic equating error does not depend on the number of examinees in the equating (Kolen & Brennan, 2004).

2.5 Evaluating the Results of Equating

After equating is conducted, the results of equating can be evaluated with several criteria. According to Harris and Crouse (1993) and other evaluation studies of KE, the evaluation criteria for equating results include:

1) Standard error of equating conditional on scores;
2) Root Mean Squared Deviation (RMSD) index and "average equating error" index (Klein & Jarjoura, 1985; Livingston, Dorans, & Wright, 1990) for evaluating overall equating accuracy;
3) Conditional equating bias and "average equating bias" (Livingston, 1993);
4) Root Mean Square Error (RMSE) for overall adequacy of equating (Mao, von Davier, & Rupp, 2005);
5) Standard Error of Equating Difference calculated under the KE framework

(von Davier, Holland, & Thayer, 2004).

2.5.1 Standard Error of Equating

The Standard Error of Equating (SEE) is useful in indicating the amount of random error in equating which is due to sampling of examinees. There are two ways of calculating SEE's: analytic methods, and computational methods such as a bootstrap resampling method or other empirical methods. The delta method is an analytic method replying on asymptotic statistical assumptions. It uses normal distribution to approximate the probability distribution of a statistical estimator. The assumption of asymptotic normality holds only when sample size is relatively large. When sample size is small, the delta method will not be accurate unless strong normality assumption holds for the population.

Using a real data with a common item nonequivalent group design, Hanson, Zeng, and Kolen (1993) compared the delta method standard errors of equating with the bootstrap standard errors of equating for Levine observed score and true score linear equating. The sample size is over 700. The results of their study indicate that compared to the bootstrap SEE, the random equating errors for scores at the higher end were overestimated by the delta method with a normality assumption while the random equating errors for scores at the lower end were underestimated. Lu and Kolen (1994) used the delta method and the bootstrap method to estimate SEE's of Tucker linear equating for a common item nonequivalent group design. They compared the differences between standard errors derived from the delta method and the bootstrap method given different sample sizes and different number of bootstrap replications. They also found that the difference between standard errors calculated by the delta method and the

bootstrap method become larger as sample size decreases and as the number of bootstrap replications decreases.

Bootstrap method refers to the resampling procedure of selecting random samples with replacement from a given sample with size N repeatedly. The theoretical framework for the bootstrap method and the applications of the bootstrap method were decribed in Efron (1982), Efron and Tibshirani (1993) and Kolen and Brennan (2004). Suppose in a random equivalent group design, two groups of examinees of size n_1 and n_2 took test forms X and Y respectively, Form Y is equated to Form X using equating method B, Then a typical bootstrap method has the following steps: 1) Draw a sample of size n_1 with replacement from the group of examinees taking test form X (size = n_1); 2) Draw a random bootstrap sample of size n_2 with replacement from the group of examinees taking test form Y (size = n_2); 3) Conduct equating on the random bootstrap samples and obtain an equating function; 4) Repeat step 1 through step 3 for a large number of times and equate Y to X every time; 5) All the equating results at each score point form a distribution. Calculate standard deviation of the equating results at each score point. The result is called the estimated bootstrap standard error of equating conditional on every score point. Then the bootstrap standard error of this equating procedure conditional on each score level will be:

$$SEE = \sqrt{\frac{1}{n-1}\sum_{1}^{n}(\hat{e}_X(y_k) - \bar{\hat{e}}_X(y_k))^2} \qquad (15)$$

where n is the total number of replications; y_k represents the k^{th} score on Form Y; $e_X(y_k)$ is the equated score on Form X corresponding to score y_k; $\bar{\hat{e}}_X(y_k)$ is the mean

of equated scores at score y_k over the n replications. Parshall, Houghton, and Kromrey (1995) used bootstrap standard error of equating and statistical bias in equating to study the adequacy of equating. Their results incidate that as sample size decreased, equating bias remains stable but the bootstrap SEE increased substantially. Therefore, they argued for using the bootstrap method instead of the delta method to calculate SEE for samall samples (Tsai, 1995).

Livingston (1993a) compared the standard errors of kernel equating methods with traditional equipercentile methods using a common item nonequivalent group design. He calculated random standard error of equating using an empirical method different from the typical bootstrap method. He selected 50 small random samples of size n without replacement from a big population dataset of size N. He then obtained equating results for each of the 50 small samples. Standard deviation of the 50 equated scores from the population criterion equating result at each raw score point is regarded as the conditional standard error of equating at each score point. Instead of using the mean of the 50 equated scores for each raw score point ($\bar{\hat{e}}_X(y_k)$ in formula 15), he used the equated score on the population criterion.

The simulation study in this dissertation follows the same procedure as described in Livingston (1993) to calculate empirical standard error of equating. The bootstrap method was applied on the real datasets to calculate standard error of equating.

2.5.2 Root Mean Squared Deviation (RMSD)

The root mean squared deviation (RMSD), is a measure of the overall equating accuracy (Livingston, Dorans, & Wright, 1990; Livington, 1993; Schmitt, Cook, Dorans, & Eignor, 1990). It can be calculated by:

$$RMSD = \sqrt{\frac{\sum n_{y_k} (\hat{x}_{y_k} - x_{y_k})^2}{\sum n_{y_k}}} \tag{16}$$

where x_{y_k} is the equated score on Form X corresponding to score y using the criterion equating method; $\hat{x}_{y_k}$ is the equated score on test form X corresponding to score y using other equating methods; n_{y_k} is the number of observations at each score level of test Y. The RMSD is basically an average of the conditional random equating errors. An alternative summary statistics is the average equating error, which is simply the average of the conditional standard error of equatings over all the score points on test Form Y (Klein & Jarjoura, 1985).

2.5.3 Equating Bias

Equating bias is useful in indicating systematic error in equating. In equating practice, equating bias is often estimated when comparing equating results with an arbitrarily selected sound criterion. Generally, results from equipercentile equating are a good candidate for such a criterion. Yen (1985) suggested using the results from equipercentile equating as a criterion because it is as accurate as the IRT-based equating results. Livingston (1993a and 1993b) used the equipercentile equtaing results for a very large sample as a baseline criterion. Alternatively, the true equating relationship can be found from simulated data. In simulation studies, the population equating relationship is known and can be reckoned as a comparison criterion for calculating equating bias, but the degree to which the simulated data can represent real data is questionable.

Use the same notation defined above, the equating bias conditional on each score

level can be caculated by:

$$\hat{x}_{y_k} - x_{y_k} \tag{17}$$

The overall bias of equating can be calculated by:

$$\sum n_{y_k} (\hat{x}_{y_k} - x_{y_k}) / \sum n_{y_k} \tag{18}$$

2.5.4 Root Mean Square Error

As described above, SEE and RMSD reflects random equating error and systematic equating error respectively. Tsai (1995) and Mao, von Davier, and Rupp (2005) adopted the Root Mean Square Error (RMSE) index. Tsai (1995) explained why this statistics takes into account the random equating error and systematic equating error simultaneously.

$$RMSE = \sqrt{(\bar{d})^2 + (sd_d)^2} \tag{19}$$

Where $\bar{d}$ is the mean of the equating differences at each score level, and sd_d is the standard deviation of the equating differences between two methods. It reflects how biased and how accurate the equating results are comparing to an equating criterion.

2.5.5 Standard Error of Equating Difference

SEED calculated in KE can be used to determine whether the equating difference between two KE methods is significant or not. Von Davier, Holland, and Thayer (2004) used SEED to decide if equating bias in a CB design is significantly big. When equating using a CB design, the equating function of the 2SG approach with weights of (1, 1) is unbiased since the data from tests taken first is not affected by order effects. If a 2SG

method with certain weights is compared with the unbiased 2SG(1, 1) method, and their equating difference falls within the range of ± 2SEED, then the equating bias of this 2SG method is small enough to be neglected. The standard error of equating will become the only statistics to compare when selecting an equating function.

TABLE 5. *KE methods and corresponding traditional equating methods*

2SG(.5, .5) KE linear	Traditional SG linear equating
2SG(1, 1) KE linear	Traditional EG linear equating
2SG(.5, .5) KE equipercentile	Traditional SG equipercentile equating
2SG(1, 1) KE equipercentile	Traditional EG equipercentile equating
2SG with other weights	*Not available*

TABLE 6. *All equating methods compared in this study for simulated data*

	Equating	Explanation
2SG Design Linear	2SG(.5,.5)	Log-linear smoothing; Treat data as two independent groups; Using weights of (.5,.5) for X and Y
	2SG(.5,.75)	Log-linear smoothing; Treat data as two independent groups; Using weights of (.5,.75) for X and Y
	2SG(.6,.5)	Log-linear smoothing; Treat data as two independent groups; Using weights of (.6,.5) for X and Y
	2SG(.6,.6)	Log-linear smoothing; Treat data as two independent groups; Using weights of (.6,.6) for X and Y
	2SG(.75,.5)	Log-linear smoothing; Treat data as two independent groups; Using weights of (.75,.5) for X and Y
	2SG(.75,.75)	Log-linear smoothing; Treat data as two independent groups; Using weights of (.75,.75) for X and Y
	2SG(.9,.5)	Log-linear smoothing; Treat data as two independent groups; Using weights of (.9,.5) for X and Y
	2SG(.9,.9)	Log-linear smoothing; Treat data as two independent groups; Using weights of (.9,.9) for X and Y
	2SG(1,1)	Log-linear smoothing; Treat data as two independent groups; Using weights of (1,1) for X and Y
2SG Design Equi-percentile	2SG(.5,.5)	Log-linear smoothing; Treat data as two independent groups; Using weights of (.5,.5) for X and Y
	2SG(.5,.75)	Log-linear smoothing; Treat data as two independent groups; Using weights of (.5,.75) for X and Y
	2SG(.6,.5)	Log-linear smoothing; Treat data as two independent groups; Using weights of (.6,.5) for X and Y
	2SG(.6,.6)	Log-linear smoothing; Treat data as two independent groups; Using weights of (.6,.6) for X and Y
	2SG(.75,.5)	Log-linear smoothing; Treat data as two independent groups; Using weights of (.75,.5) for X and Y
	2SG(.75,.75)	Log-linear smoothing; Treat data as two independent groups; Using weights of (.75,.75) for X and Y
	2SG(.9,.5)	Log-linear smoothing; Treat data as two independent groups; Using weights of (.9,.5) for X and Y
	2SG(.9,.9)	Log-linear smoothing; Treat data as two independent groups; Using weights of (.9,.9) for X and Y
	2SG(1,1)	Log-linear smoothing; Treat data as two independent groups; Using weights of (1,1) for X and Y
SG design	SG_Lin	Linear-interpolation; Traditional linear equating
	SG_Equi	Linear-interpolation; Traditional equipercentile equating
EG design	EG Linear	Linear-interpolation for continuization; Traditional linear equating
	EG Equi	Linear-interpolation for continuization; Traditional equipercentile equating

Among these methods, the EG linear, EG equipercentile, SG linear and SG equipercentile equating methods are the corresponding traditional equating methods for

the 2SG(1, 1) linear, 2SG(1, 1) KE, SG KE linear and SG KE methods.

Chapter III: Methods

3.1 Quantification of Differential Order Effect

This study draws on DOE as $(\bar{X}_1 - \bar{Y}_1) - (\bar{X}_2 - \bar{Y}_2)$ (Kolen and Brennan, 2004) to further introduce Hypothesis Testing and effective size and estimate order effects in a CB design.

The following is a derivation for a hypothesis testing of the statistical significance of DOE:

$$DOE = (\hat{\mu}_{X_1} - \hat{\mu}_{Y_1}) - (\hat{\mu}_{X_2} - \hat{\mu}_{Y_2}) = (\hat{\mu}_{X_1} + \hat{\mu}_{Y_2}) - (\hat{\mu}_{X_2} + \hat{\mu}_{Y_1})$$

$$= \left(\frac{\sum X_1}{N_1} + \frac{\sum Y_2}{N_1}\right) - \left(\frac{\sum X_2}{N_2} + \frac{\sum Y_1}{N_2}\right)$$

$$= \frac{\sum (X_1 + Y_2)}{N_1} - \frac{\sum (X_2 + Y_1)}{N_2} = \hat{\mu}_{(X_1+Y_2)} - \hat{\mu}_{(X_2+Y_1)} \qquad (20)$$

where $\hat{\mu}_{(X_1+Y_2)}$ is the average sum scores of X_1 and Y_2 for sample 1, $\hat{\mu}_{(X_2+Y_1)}$ is the average sum scores of X_2 and Y_1 for sample 2; N_1 is the number of examinees in sample 1, and N_2 is the total number of examinees in sample 2.

Therefore, the hypothesis testing for the significance of DOE is actually equivalent to a two independent sample t-test for the mean difference of *Sum12* and *Sum21*. The null hypothesis for DOE becomes: $H_0 : \mu_{(X_1+Y_2)} - \mu_{(X_2+Y_1)} = 0$;

and the t test is: $t = \dfrac{DOE}{\sqrt{s_p^2\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$ (21)

where s_p is the square root of the pooled variance of the two sum scores,

$$s_p = \sqrt{\frac{(n_1 - 1)s^2_{(X_1+Y_2)} + (n_2 - 1)s^2_{(Y_1+X_2)}}{n_1 + n_2 - 2}} \tag{22}$$

The statistical significance of DOE, however, relies heavily on sample sizes. To avoid the influence of sample size on the quantification of differential order effects, the effect sizes of DOE can be calculated:

$$\text{Effect size } \hat{d} = \frac{Mean_{(X_1+Y_2)} - Mean_{(Y_1+X_2)}}{s_p} \tag{23}$$

3.2 Data

This study uses 2 real datasets and 6 simulated datasets with CB designs. The six simulated datasets are generated in a systematic way with different sizes of DOE.

3.2.1 Real Data

Real data 1: Von Davier, Holland, and Thayer (2004) provided a real dataset from a small field study of an international testing program. In their dataset, both test forms X and Y are number-right scored. They have 75 items and 76 items respectively and their correlation is $r_{(X_1,Y_2)} = r_{(X_2,Y_1)} = 0.88$.

TABLE 7. *Summary statistics for real data 1*

	X_1	Y_2	X_2	Y_1	X	Y	*Sum12*	*Sum21*
N	143	143	140	140	283	283	143	140
Mean	52.65	51.42	50.64	51.39	51.66	51.41	**104.07**	**102.04**
SD	12.41	11.03	13.83	12.18	13.15	11.59	22.72	25.23
Skew	-0.52	-0.37	-0.54	-0.58	-0.55	-0.49	-0.45	-0.57
Kurt	-0.15	-0.64	-0.82	-0.52	-0.50	-0.55	-0.40	-0.67
Min	16	27	19	18	16	18	45	45
Max	74	71	72	71	74	71	142	142

**X* and *Y* are scores for combined groups; *Sum12* is the sum of scores on test X_1 and Y_2 for the first group; *Sum21* is the sum of scores on test X_2 and Y_1 for the second group.

The differential order effect in this dataset is DOE $= (\bar{X}_1 - \bar{Y}_1) - (\bar{X}_2 - \bar{Y}_2) =$ 2.03, which has an effect size of 0.08 approximately. T-test is not significant.

Real data 2: The second real data was collected using a CB design for an algebra test. Each of the equating forms has 25 multiple-choice items. Group one has 399 students, who took Form X first and Form Y second, and Group two has 362 students, who took Form Y first and Form X second. Both test forms X and Y are number-right scored and their total score correlations are $r_{(X_1,Y_2)} = 0.64$ and $r_{(X_2,Y_1)} = 0.74$ respectively.

TABLE 8. *Summary statistics for real data 2*

	X_1	Y_2	Y_1	X_2	X	Y	*Sum12*	*Sum21*
N	399	399	362	362	761	761	399	362
Mean	13.04	13.00	12.14	11.84	12.47	12.59	**26.04**	**23.98**
SD	3.94	4.35	4.15	4.66	4.33	4.27	7.50	8.22
Skew	-0.22	-0.25	0.25	0.22	-0.03	-0.01	-0.07	0.37
Kurt	0.21	0.40	-0.34	-0.15	-0.06	-0.02	0.19	-0.28
Min	0	0	2	0	0	0	0	4
Max	23	25	23	25	23	25	48	48

**X* and *Y* are scores for combined groups; *Sum12* is the sum of scores on test X_1 and Y_2 for the first group; *Sum21* is the sum of scores on test X_2 and Y_1 for the second group.

The differential order effect in this dataset is 2.06, which has an effect size of 0.26 approximately.

3.2.2 Simulated Data

In compliance with Davey, Nering, and Thompson's (1997) purpose of simulating realistic item response data, this study made an effort to generate data as close as possible to the first real data described earlier. The reason for selecting real data 1 as a target is that the two test forms in this dataset have equal test-retest reliabilities, which is an important assumption for linear and equipercentile equating. There are 75 items on each simulated test form.

Six population datasets were simulated with different sizes of order effects using a 3 parameter logistic Item Response Theory model (3PL IRT model). In Lord (1980), a 3PL IRT model takes the form as below:

$$P_{(\theta)} = c + \frac{1-c}{1+e^{-1.7a(\theta-b)}} \qquad (24)$$

where θ is the underlying ability to be measured, a is the item discrimination parameter, b is item difficulty, and c is the item guessing parameter indicating the probability that a person completely lacking in ability will answer the item correctly.

Each of the six simulated datasets has two samples, each with size of 100,000. Each sample takes two tests X and Y but in different order. A 75 by 100,000 item-person response matrix with 0 and 1 scores was generated for each sample using the 3PL IRT model. The scores on each item were then totaled to get an observed test score for each examinee. After the simulation of data for two independent group taking two test forms in different order, data from the two independent samples were simply combined together to form the dataset with a pooled SG design. Please see the design below:

For a CB design: $sample1:(X_1, Y_2)$
$sample2:(X_2, Y_1)$

For a SG design: pooled sample $= \begin{pmatrix} X_1 & Y_2 \\ X_2 & Y_1 \end{pmatrix}$

However, one drawback of using real data 1 is its lack of item response data. Without the item response block, it is more difficult to estimate the item parameters of the real test items and use the estimated parameters for simulation. In this simulation, the parameter distributions were decided based on empirical experience.

To ensure that the generated item discriminant parameter *a* and item guessing level *c* are positive, parameter *a's* were randomly selected from a log-normal distribution, and parameter *c's* were randomly selected from a beta distribution. Furthermore, in order to make the simulated data more realistic, means and variances of the distributions of parameter *a*, *b*, and *c* were adjusted to be certain values to best emulate the first real data set used in this study. Specifically, the mean and variance for the log-normal distribution of parameter *a* was fixed as 1 and 0.12; the mean and variance for the normal distribution of parameter *b* was fixed as -0.3 and 0.8 and the mean and variance for the beta distribution of parameter *c* was fixed as 0.25 and 0.008.

Order effects were considered as a second dimension of examinee's underlying abilities when taking the second test and the size of order effects varies across examinees. Assume that the changes in examinees' performances reflect the changes in their underlying abilities, then,

$$\theta_{12k} = \theta_{11k} + o_{1k} \quad \text{(sample 1);} \tag{25}$$

$$\theta_{22k} = \theta_{21k} + o_{2k} \quad \text{(sample 2);} \tag{26}$$

where k is the number of examinees;

θ_{11k} denotes the underlying abilities of examinees in sample 1 taking the first test (X_1);

θ_{12k} denotes the abilities of examinees in sample 1 taking the second test (Y_2);

o_{1k} denotes the order effects of examinees in sample 1 taking test X first and Y second;

θ_{21k} denotes the underlying abilities of examinees in sample 2 taking the first test (Y_1);

θ_{22k} denotes the abilities of examinees in sample 2 taking the second test (X_2);

o_{2k} denotes the order effects of examinees in sample 2 taking test Y first and X second;

It was assumed that θ_{11k} and θ_{12k} (or θ_{21k} and θ_{22k}) follows a bivariate normal distribution with the same standard deviations. The correlation between θ_{11k} and θ_{12k} (or θ_{21k} and θ_{22k}) may not be perfect since order effects are not constant across examinees. It was set to be 0.94 in this study in order to achieve a correlation of observed score at 0.88. o_{1k} and o_{2k} both have variances of $(1\text{-}0.94)^2$. When all the parameters a, b, c, and θ were randomly selected, calculate the probability of each examinee with certain θ level answering each item correctly from the 3PL IRT model. If the probability of a correct response is greater than a random number from a uniform distribution, the item response for a person on a specific item will be 1, otherwise it will be 0.

In this study, the effect sizes of differential order effects were controlled to be

changing from 0 to 0.2 in the simulated datasets. In order to meet this restriction and make simulated data as real as possible, different means for the distributions of θ_{11} and θ_{12} (or θ_{21} and θ_{22}) were tried and DOE's were calculated afterwards until order effects are within the range and the simulated test scores share similar descriptive statistics as test scores in the first real dataset. The distributions and descriptive statistics of the six simulated datasets are provided below. As shown in table 9 to table 14, the simulated data has similar distribution moments as the first real dataset.

Simulated data 1 with insignificant order effects (DOE = -0.04)

- Sample 1 (N=100000):

$$\theta \sim \left(\left(\mu_{\theta_{11}} = 0 \quad \mu_{\theta_{12}} = 0.01 \right), \begin{pmatrix} \sigma^2_{\theta_{11}} = 1 & \sigma_{\theta_{11}\theta_{12}} = .94 \\ \sigma_{\theta_{11}\theta_{12}} = .94 & \sigma^2_{\theta_{12}} = 1 \end{pmatrix} \right)$$

- Sample 2 (N=100000):

$$\theta \sim \left(\left(\mu_{\theta_{21}} = 0 \quad \mu_{\theta_{22}} = 0.01 \right), \begin{pmatrix} \sigma^2_{\theta_{21}} = 1 & \sigma_{\theta_{21}\theta_{22}} = .94 \\ \sigma_{\theta_{21}\theta_{22}} = .94 & \sigma^2_{\theta_{22}} = 1 \end{pmatrix} \right)$$

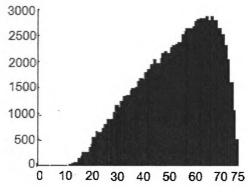
$$a \sim \left(\mu_a = 1, \sigma_b^2 = 0.12 \right); \; b \sim \left(\mu_b = -0.3, \sigma_b^2 = 0.8 \right);$$

$$c \sim \left(\mu_c = 0.25, \sigma_b^2 = 0.008 \right)$$

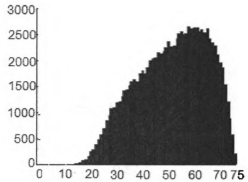
TABLE 9. *Descriptive statistics for simulated data 1*

Test	Min.	Max.	Mean	Std	Skewness	Kurtosis
X_1	10	75	52.52	13.78	-0.46	-0.67
Y_2	9	75	50.50	13.57	-0.32	-0.81
X_2	10	75	50.51	13.59	-0.31	-0.81
Y_1	8	75	52.55	13.80	-0.45	-0.68

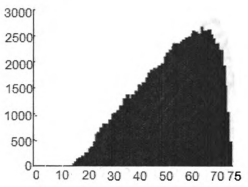
$r_{(X1, Y2)} = r_{(Y1, X2)} \approx 0.88$



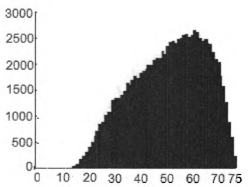
X_1 (skewness=-0.46)



Y_2 (skewness=-0.32)



Y_1 (skewness=-0.45)



X_2 (skewness=-0.31)

Simulated data2 with significant order effects (DOE= -0.58, effect size of DOE = 0.025)

- Sample 1 (N=100000):

$$\theta \sim \left(\left(\mu_{\theta_{11}} = 0 \quad \mu_{\theta_{12}} = -0.025 \right), \begin{pmatrix} \sigma^2_{\theta_{11}} = 1 & \sigma_{\theta_{11}\theta_{12}} = .94 \\ \sigma_{\theta_{11}\theta_{12}} = .94 & \sigma^2_{\theta_{12}} = 1 \end{pmatrix} \right)$$

- Sample 2 (N=100000):

$$\theta \sim \left(\left(\mu_{\theta_{21}} = 0 \quad \mu_{\theta_{22}} = 0.025 \right), \begin{pmatrix} \sigma^2_{\theta_{21}} = 1 & \sigma_{\theta_{21}\theta_{22}} = .94 \\ \sigma_{\theta_{21}\theta_{22}} = .94 & \sigma^2_{\theta_{22}} = 1 \end{pmatrix} \right)$$

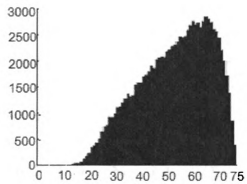
$a \sim \left(\mu_a = 1, \sigma_b^2 = 0.12\right)$; $b \sim \left(\mu_b = -0.3, \sigma_b^2 = 0.8\right)$;

$c \sim \left(\mu_c = 0.25, \sigma_b^2 = 0.008\right)$

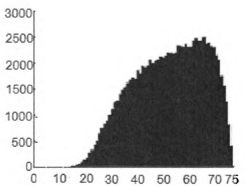
TABLE 10. *Descriptive statistics for simulated data 2*

Test	Min.	Max.	Mean	Std	Skewness	Kurtosis
X_1	9	75	52.01	13.71	-0.43	-0.71
Y_2	10	75	50.54	14.01	-0.27	-0.89
X_2	11	75	51.15	13.90	-0.30	-0.87
Y_1	10	75	51.98	13.66	-0.41	-0.73

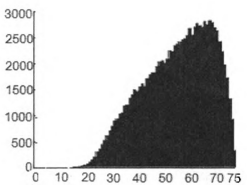
$r_{(X1,\ Y2)} = r_{(Y1,\ X2)} \approx 0.88$



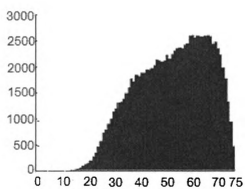
X_1 (skewness=-0.43)



Y_2 (skewness=-0.27)



Y_1 (skewness=-0.41)



X_2 (skewness=-0.3)

Simulated data3 with significant order effects (DOE= 1.41, effect size of DOE = 0.05)

- Sample 1 (N=100000):

$$\theta \sim \left(\begin{pmatrix} \mu_{\theta_{11}} = 0 & \mu_{\theta_{12}} = 0.05 \end{pmatrix}, \begin{pmatrix} \sigma^2_{\theta_{11}} = 1 & \sigma_{\theta_{11}\theta_{12}} = .94 \\ \sigma_{\theta_{11}\theta_{12}} = .94 & \sigma^2_{\theta_{12}} = 1 \end{pmatrix} \right)$$

- Sample 2 (N=100000):

$$\theta \sim \left(\begin{pmatrix} \mu_{\theta_{21}} = 0 & \mu_{\theta_{22}} = -0.05 \end{pmatrix}, \begin{pmatrix} \sigma^2_{\theta_{21}} = 1 & \sigma_{\theta_{21}\theta_{22}} = .94 \\ \sigma_{\theta_{21}\theta_{22}} = .94 & \sigma^2_{\theta_{22}} = 1 \end{pmatrix} \right)$$

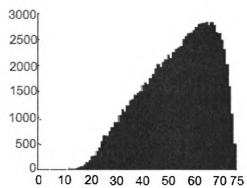
$$a \sim \left(\mu_a = 1, \sigma_b^2 = 0.12 \right); \; b \sim \left(\mu_b = -0.3, \sigma_b^2 = 0.8 \right);$$

$$c \sim \left(\mu_c = 0.25, \sigma_b^2 = 0.008 \right)$$

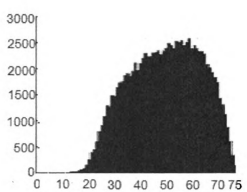
TABLE 11. *Descriptive statistics for simulated data 3*

Test	Min.	Max.	Mean	Std	Skewness	Kurtosis
X_1	9	75	52.01	13.71	-0.43	-0.71
Y_2	10	75	51.54	13.90	-0.34	-0.84
X_2	11	75	50.15	14.00	-0.24	-0.92
Y_1	10	75	51.98	13.66	-0.41	-0.73

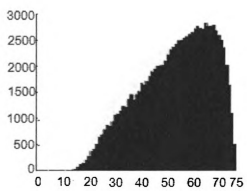
$r_{(X1, Y2)} = r_{(Y1, X2)} \approx 0.88$



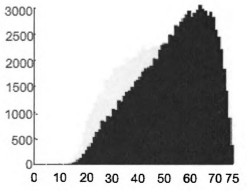
X_1 (skewness=-0.43)



Y_2 (skewness=-0.34)



Y_1 (skewness=-0.41)



X_2 (skewness=-0.24)

Simulated data4 with significant order effects (DOE= -2.75, effect size of DOE = 0.1)

- Sample 1 (N=100000):

$$\theta \sim \left(\left(\mu_{\theta_{11}} = 0 \quad \mu_{\theta_{12}} = -0.1 \right), \begin{pmatrix} \sigma^2_{\theta_{11}} = 1 & \sigma_{\theta_{11}\theta_{12}} = .94 \\ \sigma_{\theta_{11}\theta_{12}} = .94 & \sigma^2_{\theta_{12}} = 1 \end{pmatrix} \right)$$

- Sample 2 (N=100000):

$$\theta \sim \left(\left(\mu_{\theta_{21}} = 0 \quad \mu_{\theta_{22}} = 0.1 \right), \begin{pmatrix} \sigma^2_{\theta_{21}} = 1 & \sigma_{\theta_{21}\theta_{22}} = .94 \\ \sigma_{\theta_{21}\theta_{22}} = .94 & \sigma^2_{\theta_{22}} = 1 \end{pmatrix} \right)$$

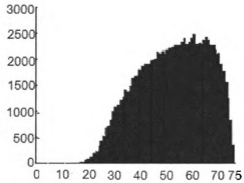
$$a \sim \left(\mu_a = 1, \sigma_b^2 = 0.12 \right); \; b \sim \left(\mu_b = -0.3, \sigma_b^2 = 0.8 \right);$$

$$c \sim \left(\mu_c = 0.25, \sigma_b^2 = 0.008 \right)$$

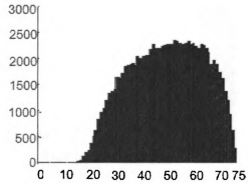
TABLE 12. *Descriptive statistics for simulated data 4*

Test	Min.	Max.	Mean	Std	Skewness	Kurtosis
X_1	10	75	50.34	13.50	-0.31	-0.80
Y_2	10	75	48.64	13.57	-0.29	-0.84
X_2	11	75	51.35	13.27	-0.45	-0.67
Y_1	9	75	50.39	13.56	-0.30	-0.81

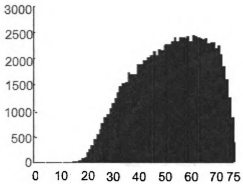
$r_{(X1, Y2)} = r_{(Y1, X2)} \approx 0.88$



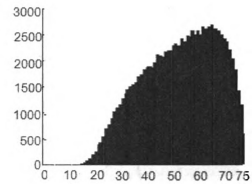
X_1 (skewness=-0.31)



Y_2 (skewness=-0.29)



Y_1 (skewness=-0.30)



X_2 (skewness=-0.45)

Simulated data5 with significant order effects (DOE= -3.76, effect size of DOE = 0.15)

- Sample 1 (N=100000):

$$\theta \sim \left(\begin{pmatrix} \mu_{\theta_{11}} = 0 & \mu_{\theta_{12}} = -0.1 \end{pmatrix}, \begin{pmatrix} \sigma^2_{\theta_{11}} = 1 & \sigma_{\theta_{11}\theta_{12}} = .94 \\ \sigma_{\theta_{11}\theta_{12}} = .94 & \sigma^2_{\theta_{12}} = 1 \end{pmatrix} \right)$$

- Sample 2 (N=100000):

$$\theta \sim \left(\begin{pmatrix} \mu_{\theta_{21}} = 0 & \mu_{\theta_{22}} = 0.2 \end{pmatrix}, \begin{pmatrix} \sigma^2_{\theta_{21}} = 1 & \sigma_{\theta_{21}\theta_{22}} = .94 \\ \sigma_{\theta_{21}\theta_{22}} = .94 & \sigma^2_{\theta_{22}} = 1 \end{pmatrix} \right)$$

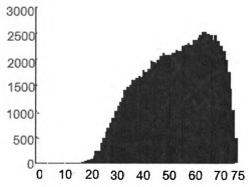
$$a \sim \left(\mu_a = 1, \sigma_b^2 = 0.12\right);\ b \sim \left(\mu_b = -0.3, \sigma_b^2 = 0.8\right);$$

$$c \sim \left(\mu_c = 0.25, \sigma_b^2 = 0.008\right)$$

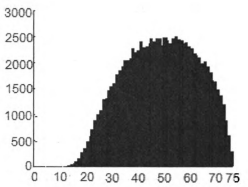
TABLE 13. *Descriptive statistics for simulated data 5*

Test	Min.	Max.	Mean	Std	Skewness	Kurtosis
X_1	10	75	50.99	14.07	-0.29	-0.89
Y_2	9	75	48.50	13.65	-0.11	-0.88
X_2	11	75	52.33	13.36	-0.34	-0.75
Y_1	9	75	50.92	14.10	-0.29	-0.88

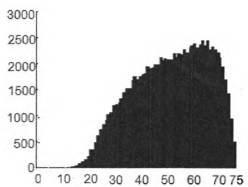
$r_{(X1,\ Y2)} = r_{(Y1,\ X2)} \approx 0.88$



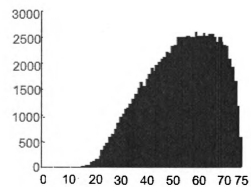
X_1 (skewness=-0.29)



Y_2 (skewness=-0.11)



Y_1 (skewness=-0.29)



X_2 (skewness=-0.34)

Simulated data 6 with significant order effects (DOE= -5.22, effect size of DOE = 0.2)

- Sample 1 (N=100000):

$$\theta \sim \left(\left(\mu_{\theta_{11}} = 0 \quad \mu_{\theta_{12}} = -0.2 \right), \begin{pmatrix} \sigma^2_{\theta_{11}} = 1 & \sigma_{\theta_{11}\theta_{12}} = .94 \\ \sigma_{\theta_{11}\theta_{12}} = .94 & \sigma^2_{\theta_{12}} = 1 \end{pmatrix} \right)$$

- Sample 2 (N=100000):

$$\theta \sim \left(\left(\mu_{\theta_{21}} = 0 \quad \mu_{\theta_{22}} = 0.2 \right), \begin{pmatrix} \sigma^2_{\theta_{21}} = 1 & \sigma_{\theta_{21}\theta_{22}} = .94 \\ \sigma_{\theta_{21}\theta_{22}} = .94 & \sigma^2_{\theta_{22}} = 1 \end{pmatrix} \right)$$

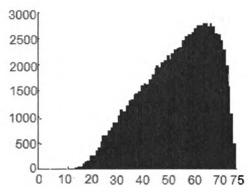
$$a \sim \left(\mu_a = 1, \sigma^2_b = 0.12 \right); \; b \sim \left(\mu_b = -0.3, \sigma^2_b = 0.8 \right);$$

$$c \sim \left(\mu_c = 0.25, \sigma^2_b = 0.008 \right)$$

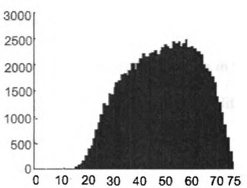
TABLE 14. *Descriptive statistics for simulated data 6*

Test	Min.	Max.	Mean	Std	Skewness	Kurtosis
X_1	9	75	52.52	13.78	-0.26	-0.88
Y_2	9	75	47.75	13.79	-0.05	-0.96
X_2	11	75	52.93	13.24	-0.37	-0.79
Y_1	11	75	52.55	13.80	-0.25	-0.89

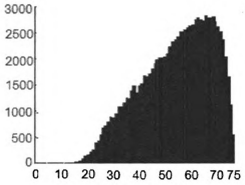
$r_{(X1,\ Y2)} = r_{(Y1,\ X2)} \approx 0.88$



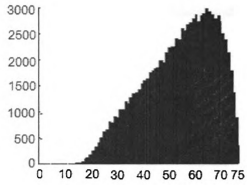
X_1 (skewness=-0.26)



Y_2 (skewness=-0.05)



Y_1 (skewness=-0.25)



X_2 (skewness=-0.37)

3.3 Analysis

The analysis of real data and simulated data in this study differs slightly. For the two real datasets, the bootstrap method was employed to calculate standard error of 14 out of the total 22 equatings (as listed in Table 15 and Table 16). The equating results were evaluated by SEE and RMSE. For the simulated datasets, empirical standard errors of equating were calculated for 22 equating methods as displayed in Table 6. The equating functions were evaluated by SEE, equating bias relative to the large sample standard, RMSE and SEED. Computer software SAS, MATLAB, Compaq Visual Fortran, and MATLAB were used to simulate data and conduct equating procedures.

3.3.1 Equating Methods Applied for Simulated Data

Table 6 lists the names of all the equatings conducted for simulated data in this study and provides detailed explanations for each equating. The results of the traditional equipercentile equating (EG_Equi) on each population dataset were considered as criterion equating results. All the other equating results were compared to this criterion equating for each population. In this study, all the equatings are from test Form Y to test

Form X, i.e., the equating function takes the form of $e_x(y)$, which is a function of score y.

3.3.2 Procedure for Estimating Empirical SEE for Simulated Data

Once the population datasets were generated, 500 random samples were selected from each of the four populations without replacement. The estimation of empirical SEE for the simulated datasets followed procedures as below:

1. Randomly select one sample (n=50) from each of the two independent samples from population 1 without replacement. Selected sample 1 has scores for Form X, which is taken first and Form Y, which is taken second. Selected sample 2 will have scores for Form X, which is taken second, and Form Y, which is taken first. Data from the two independent samples were simply combined to form a data with the pooled single group design.
2. Apply the 22 equatings to the samples selected from the population. When the sample size is greater than 100, two log-linear models were fit to the data for all the KE equating methods. The first log-linear model (model (2, 2, 1)) preserves the first bivariate moment (the correlation of scores on Form X and Form Y) and the first two univariate moments of each variable (mean and standard deviation). The second log-linear model (model (4, 4, 1)) preserves the first bivariate moment and the first four univariate moments of each variable.
3. Replace the test-takers into the corresponding population and repeat

sampling for 500 times. Then the 500 replications build up a conditional distribution of equating results at each score point. The mean of this conditional distribution is the equating results at each score point and the standard deviation of this conditional distribution is the empirical conditional SEE at each score point.

4. Repeat step 1 to 3, change the selected sample size from 50, to 100, 300, 500 and 1000.

5. Repeat the above procedures for simulated data 2 to data 6.

The bandwidth for KE linear equating was set at 200. The weighting parameter w_x or w_y took values from 0.5 to 1.

3.3.3 Evaluating Equating Results from Simulated Data

For the simulated data, traditional equipercentile equating results with the EG design were considered as the criterion. All the other equating methods were compared to this criterion and were evaluated in terms of Standard Error of Equating, equating bias relative to the large sample standard, Root Mean Square Error and Standard Error of Equating Difference. For the two real datasets, only bootstrap SEE and RMSE were reported.

Equating Bias Relative to the Large Sample Standard

To calculate equating bias at each score point, for each of the 22 equatings under each of the six population conditions, the mean of the 500 replications' equating results were subtracted from the criterion equating results (EG_Equi) at each score level (as in formula 17). Conditional equating bias was not reported for simulated data. Instead, the average of all the conditional biases at each score level was calculated and reported in

chapter IV.

Root Mean Square Error (RMSE)

The Root Mean Square Error of each equating compared to the criterion equating is equal to the square root of the sum of squared average bias and variance of bias over possible score points: $RMSE = \sqrt{(\bar{d})^2 + (sd_d)^2}$, where $\bar{d}$ is the mean of the equating differences and sd_d is the standard deviation of the differences between the equating results of one method and the criterion equating results. It reflects how biased and how accurate the equating results are compared to the population criterion.

Standard Error of Equating (SEE)

The empirical conditional standard error of equating was considered as the standard deviation of the conditional distribution formed by the equating results for 500 replications. It can be calculated using the following formula. In chapter IV, only the average of these conditional SEE's over different score points was reported for each equating method.

$$SEE = \sqrt{\frac{1}{499}\sum_{j=1}^{500}(\hat{e}_X(y_k) - e_X(y_k))^2} \qquad (27)$$

where j = 1 to 500 is the number of selected samples; k = 1 to K is the possible score points on Form Y; $\hat{e}_X(y_k)$ is the equated score from Form Y to Form X for the j^{th} replication; $e_X(y_k)$ is the equated score of $\boldsymbol{X}$ corresponding to score y_k from the population dataset.

In this study, SEED was calculated directly by the KE software.

Chapter IV: Results

4.1 Real Data 1

Real data 1 has a DOE of 2.03, which is not statistically significant (t=.713, se=2.85, p=.476, df=281), i.e., the order effect can be almost cancelled out by pooling together the two groups of data in this specific example. The effect size of DOE is 0.08. Levene's test of homogeneity of variance (Levene, 1960) is not significant (F=1.67, p=.197). The best fit model for the KE methods is model (2, 2, 1): $T_X = T_Y = 2$ and I = L = 1. The following figures show the observed score distributions for X_1, Y_2, X_2, and Y_1 and their fitted data distributions.

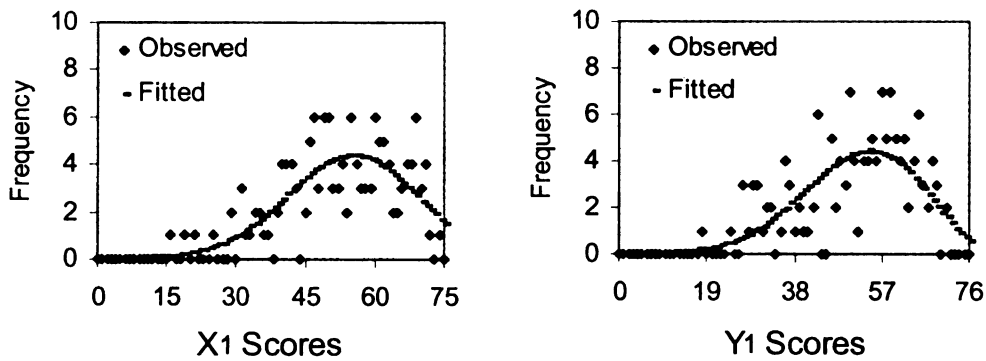


FIGURE 1. *Observed score distributions for X_1 and Y_1 in real data 1.*

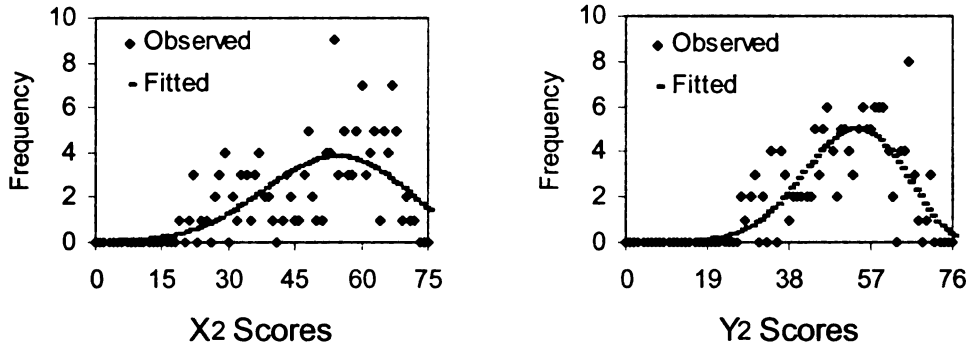


FIGURE 2. *Observed score distributions for* X_2 and Y_2 *in real data 1.*

4.1.1 Selecting the Best Equating Function Using RMSE

All equating methods were compared to the *traditional equipercentile equating with an EG design* (EG Equi.). It shows that, when DOE is insignificant, 2SG(.5,.5) and SG_KE has similar equating results with almost the smallest SEE's over the whole score point scale, but they have bigger RMSE compared to the EG design. Not much difference was found between the equating results of traditional equating and Kernel Equating. No large difference was found between linear and equipercentile equating methods except for traditional EG linear and traditional EG equipercentile. This is because the sample size for EG design is only about 70 for each sample in this dataset, which is too small for equipercentile equating. Equating results of 2SG (.75, .75) have relatively small SEE and RMSE. It is the only method that best represents the criterion equating results.

TABLE 15. *Evaluation of equating results from real data 1*

	2SG KE					SG	EG
	(.5,.5)	(.5,.75)	(.75,.5)	(.75,.75)	(1,1)	traditional	traditional
Linear							
Mean SEE	0.663	0.839	0.884	1.252	2.381	0.663	2.384
SD SEE	0.313	0.371	0.42	0.565	1.113	0.313	1.118
Min. SEE	0.32	0.44	0.425	0.646	1.164	0.32	1.164
Max. SEE	1.334	1.634	1.776	2.433	4.674	1.334	4.648
Mean Diff	2.066	1.769	1.229	0.908	-0.403	2.066	-0.418
RMSE	2.92	2.5	2.1	1.76	1.68	2.92	1.69
Equipercentile							
Mean SEE	0.692	0.833	0.846	1.147	**2.196**	2.133	**3.1**
SD SEE	0.343	0.346	0.343	0.408	**0.928**	2.241	**1.926**
Min. SEE	0.332	0.385	0.419	0.429	**0.491**	0	**0**
Max. SEE	1.384	1.485	1.43	1.714	**3.557**	6.778	**6.821**
Mean Diff	2.29	2.04	1.518	1.262	**-0.062**	1.369	**0**
RMSE	3.09	2.72	2.31	1.98	**1.42**	2.26	**0**

*Criterion equating = traditional EG equipercentile equating

The 2SG approach with weights of (1, 1) has the smallest RMSE when taking the EG traditional equipercentile equating function as a baseline. Therefore, the 2SG (1, 1) equipercentile method is the best equating function when using RMSE as an index.

4.1.2 Selecting the Best Equating Function Using SEED

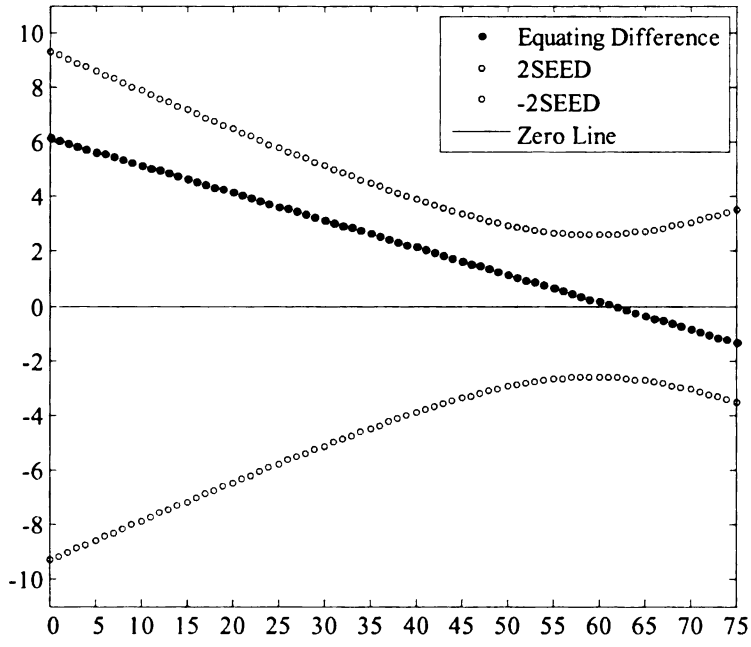


FIGURE 3. *Equating difference between 2SG(1, 1) linear and 2SG(.5, .5) linear and the ± 2SEED confidence interval band around zero line, real data 1.*

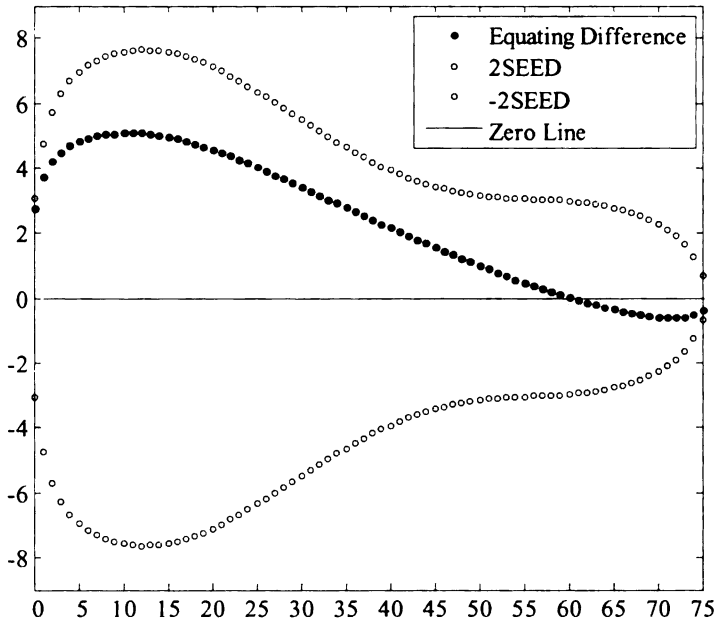


FIGURE 4. *Equating difference between 2SG(1, 1) equipercentile and 2SG(.5, .5) equipercentile and the ± 2SEED confidence interval band around zero line, real data 1.*

Figure 3 and Figure 4 indicate that the differences between the two KE linear and the two KE equipercentile methods using weights of (1, 1) and weights of (0.5, 0.5) are small in comparison with the ± 2SEED band. According to von Davier, Holland, and Thayer (2004), this indicates that the equating bias introduced by order effects is small enough to be ignored. Thus, the best equating function can be selected solely based on the random equating error, i.e., the standard error of equating. In this case, the 2SG linear or equipercentile equating with weights of (.5, .5) will be considered as the best ones. Their equating difference can be tested against SEED again to decide which one to choose.

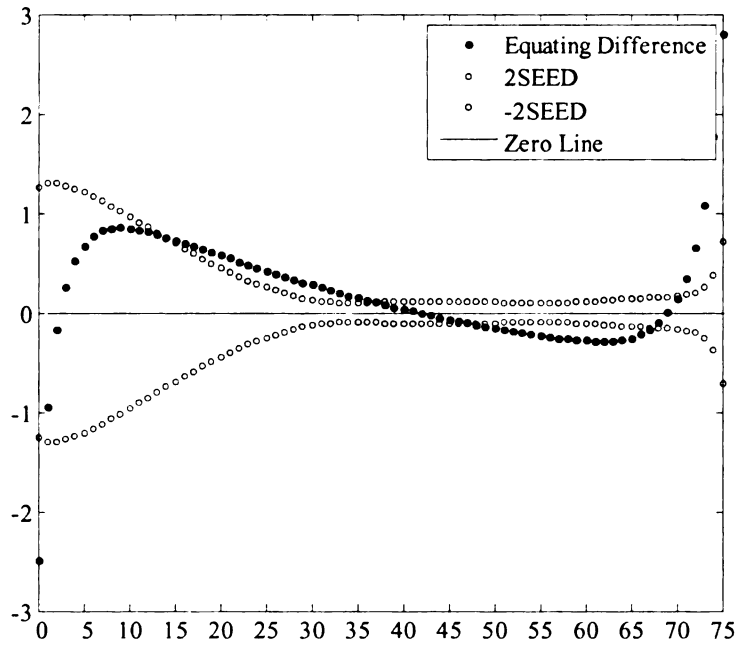


FIGURE 5. *Equating difference between 2SG(.5, .5) linear and 2SG(.5, .5) equipercentile and the ± 2SEED confidence interval band around zero line, real data 1.*

As shown in Figure 5, the difference between the KE linear and the KE equipercentile equating functions falls beyond the 95% confidence intervals along the whole score scale except the lower end. The equating function deviates from a linear function. Therefore, the 2SG equipercentile equating function with weights of (.5, .5) is preferable to the 2SG linear equating function with weights of (.5, .5) (von Davier, Holland, & Thayer, 2004).

4.2 Real Data 2

The second real data has a DOE of 2.06. This is significant as the order effect can not be cancelled out by pooling together the two groups of data in this example. The

effect size of DOE is 0.26. The best fit model for the KE methods is model (2, 2, 1) ($T_X = T_Y = 2, I = L = 1$) for group 1 and model (4, 4, 1) ($T_X = T_Y = 4, I = L = 1$) for group 2. The following figures show the observed score distributions for X_1, Y_2, X_2, and Y_1 and their best-fit log-linear models.

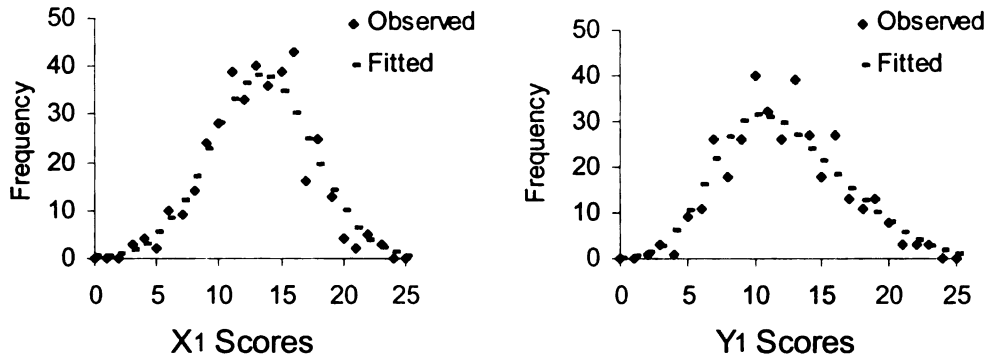


FIGURE 6. *Observed score distributions for X_1, and Y_1 in real data 2.*

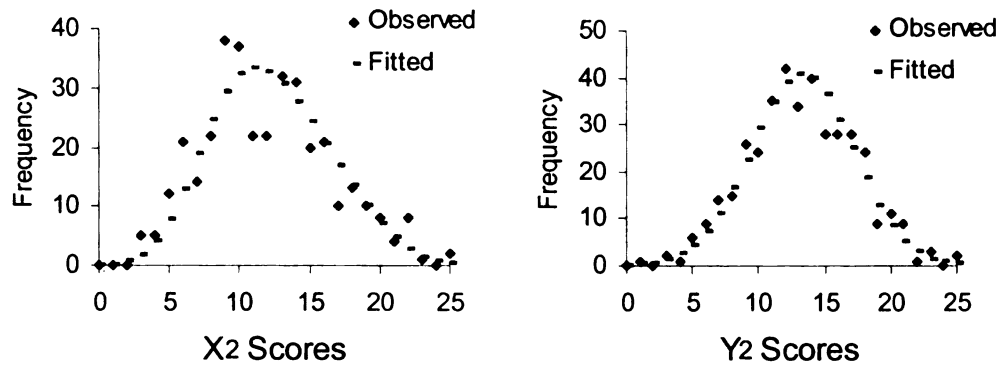


FIGURE 7. *Observed score distributions for X_2, and Y_2 in real data 2.*

4.2.1 Selecting the Best Equating Function Using RMSE

TABLE 16. *Evaluation of equating results from real data 2*

	2SG KE					SG	EG
	(.5,.5)	(.5,.75)	(.75,.5)	(.75,.75)	(1,1)	traditional	traditional
Linear							
Mean SEE	0.205	0.223	0.243	0.296	0.51	0.205	0.51
SD SEE	0.07	0.068	0.079	0.083	0.143	0.07	0.143
Min SEE	0.117	0.138	0.145	0.193	0.333	0.117	0.333
Max SEE	0.341	0.354	0.403	0.454	0.767	0.341	0.767
Mean Diff	0.749	0.498	0.448	0.198	-0.387	0.774	-0.29
RMSE	1.007	0.832	0.753	0.638	0.876	1.161	0.673
Equipercentile							
Mean SEE	0.254	0.284	0.25	0.304	**0.49**	0.382	0.54
SD SEE	0.114	0.124	0.075	0.077	**0.113**	0.241	0.274
Min SEE	0.134	0.166	0.165	0.224	**0.343**	0	0
Max SEE	0.484	0.548	0.399	0.463	**0.72**	0.845	0.96
Mean Diff	0.671	0.526	0.505	0.362	**0.033**	0.965	0
RMSE	0.97	0.857	0.774	0.681	**0.624**	1.317	0

*Criterion equating = traditional EG equipercentile equating.

In Table 16, the 2SG equipercentile equating with weights of (1, 1) has the smallest RMSE when taking the EG traditional equipercentile equating function as a baseline. Therefore, the 2SG (1, 1) equipercentile method is the best equating function when using RMSE as an index.

4.2.2 Selecting the Best Equating Function Using SEED

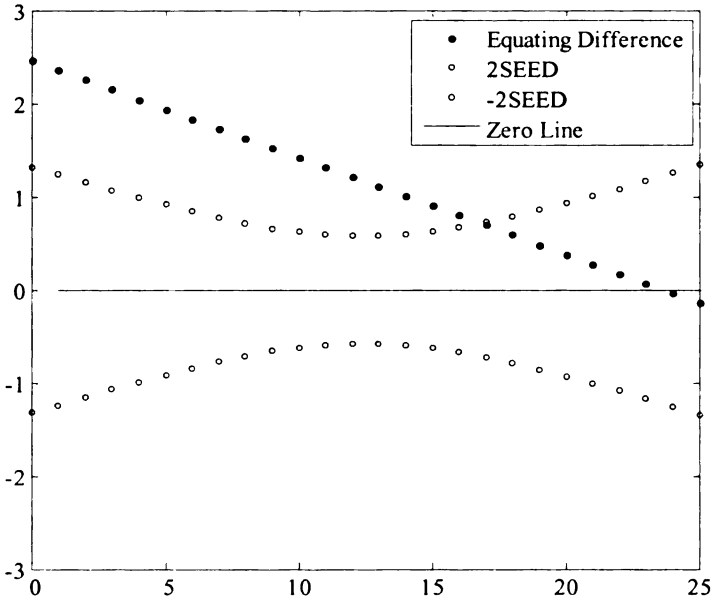


FIGURE 8. *Equating difference between 2SG(1, 1) linear and 2SG(.5, .5) linear and the ± 2SEED confidence interval band around zero line, real data 2.*

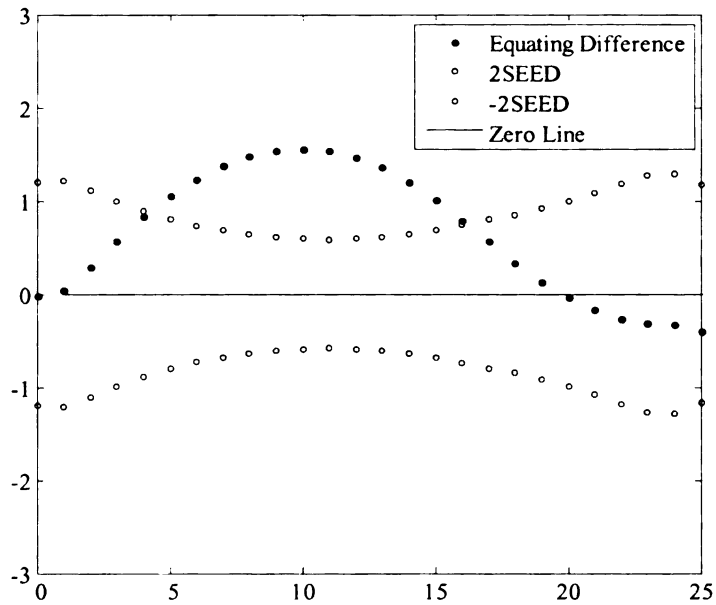


FIGURE 9. *Equating difference between 2SG(1, 1) equipercentile and 2SG(.5, .5) equipercentile and the ± 2SEED confidence interval band around zero line, real data 2.*

Figure 8 and Figure 9 indicate that the differences between the two KE linear and the two KE equipercentile methods using weights of (1, 1) and weights of (0.5, 0.5) are beyond the ± 2SEED band in the middle part of the score scale, where most of the scores distributed. For von Davier, Holland, and Thayer (2004), this indicates that the equating bias introduced by the use of the data from form X_2 and Y_2 cannot be ignored. The best solution would be to discard data from tests taken second, that is, to treat the data collected by a CB design as an EG design. After the weights are decided, the SEED plots can be used again to decide which equating function to choose, the 2SG linear equating with weights of (1, 1) or the 2SG equipercentile equating with weights of (1, 1).

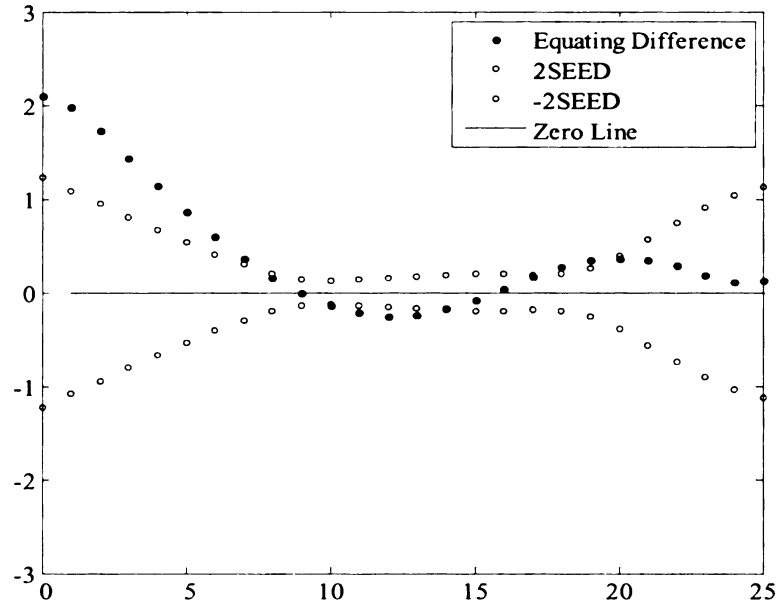


FIGURE 10. *Equating difference between 2SG(1, 1) linear and 2SG(1, 1) equipercentile, and the ± 2SEED confidence interval band around zero line, real data 2.*

As shown in Figure 10, the difference between the 2SG (1, 1) linear and the 2SG (1, 1) equipercentile equating functions falls beyond the 95% confidence intervals at the lower and the middle score scale end. This indicates the equating function deviates from a linear function. Therefore the 2SG(1, 1) equipercentile equating function is preferable to the 2SG (1, 1) linear equating function (von Davier, Holland, & Thayer, 2004).

4.3 Simulated Data

All the simulated data can be fitted by a log-linear model of (2, 2, 1) with adequate model fit. Fitting a model with more parameters did not reduce the likelihood ratio chi-square statistics significantly. In addition, the Freeman -Tukey residual plots are within the range of (-3, +3) for all the simulated data when fitted with a model of log-linear model of (2, 2, 1) like in Figure 13.

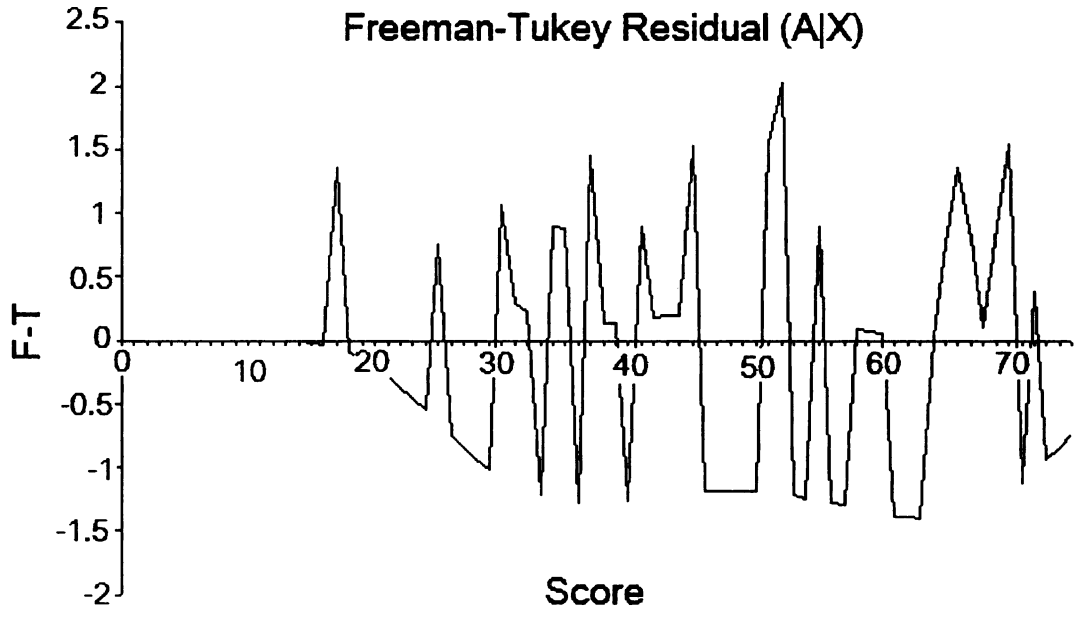


FIGURE 11. *One example of Freeman-Tukey residual plot for POP3.*

4.3.1 Model Fit

Various log-linear models were fitted to the simulated sample datasets. The results indicate that, when sample size is 50, model (2, 2, 1) is the best fit model. When sample size is 100, 300, 500 or 1000, both model (2, 2, 1) and model (4, 4, 1) have

fairly good model fit. In this study, only the equating results of fitting model (2, 2, 1) are reported since the equating results of fitting model (4, 4, 1) are very similar to the equating results of fitting model (2, 2, 1).

4.3.2 Evaluating the Equating Results by RMSE

As shown in Table 17 and Table 18, the pooled SG and 2SG(.5,.5) approaches under the KE framework have the lowest SEE and RMSE when DOE is almost zero. This indicates that when order effect can be cancelled out, the pooled SG method or 2SG(.5,.5) method can both provide optimal equating results.

Table 19 and Table 20 show the equating results for population data 2 where DOE has an effect size of 0.025. The 2SG linear and equipercentile equating methods with weights of (.5, .75) for X and Y have the smallest RMSE. When the differential order effect gets larger, as in data 3 where the effect size of DOE is 0.05, the 2SG linear equating methods with weights of (.9, .9) have the smallest RMSE (Table 21 and Table 22). When the effect size of DOE approaches to 0.1, the pooled SG approach and the 2SG(.5, .5) approach are apparently not the best (Table 23 and Table 24). Instead, the 2SG linear equating method with weights of (1, 1) (i.e., EG KE linear method) or the EG traditional linear method has the smallest RMSE. Furthermore, in population data 5 and data 6 when the effect size of DOE is around 0.15 and 0.2, the benefit of using weights of (1, 1) in the 2SG approach becomes outstandingly bigger. As shown in Table 25 to Table 28, the EG KE linear or EG traditional linear methods have much smaller RMSE than those methods which treat data as a single group design.

TABLE 17. *Summary statistics for POP1 linear equating methods*

	2SG									SG	EG
	Linear									Traditional	Traditional
POP1	2SG (.5,.5)	2SG (.5,.75)	2SG (.6,.5)	2SG (.6,.6)	2SG (.75,.5)	2SG (.75,.75)	2SG (.9,.5)	2SG (.9,.9)	2SG (1,1)	Linear	Linear
n=50											
RMSE	**0.16**	0.456	0.36	0.262	0.588	0.192	1.282	0.822	0.338	**0.16**	0.338
SEE	**1.265**	1.645	1.331	1.545	1.637	2.385	2.307	3.97	4.466	**1.265**	4.466
Bias	**-0.058**	0.429	-0.328	-0.208	-0.568	-0.097	-1.247	-0.743	-0.231	**-0.058**	-0.231
n=100											
RMSE	**0.149**	0.491	0.31	0.195	0.511	0.147	0.994	0.441	0.16	**0.149**	0.16
SEE	**0.907**	1.172	0.95	1.107	1.172	1.667	1.545	2.627	3.007	**0.907**	3.007
Bias	**-0.031**	0.467	-0.275	-0.123	-0.487	0.003	-0.98	-0.353	0	**-0.031**	0
n=300											
RMSE	**0.147**	0.494	0.255	0.147	0.498	0.146	0.849	0.188	0.15	**0.147**	0.15
SEE	**0.529**	0.676	0.591	0.654	0.678	0.965	0.888	1.461	1.743	**0.529**	1.743
Bias	**-0.018**	0.469	-0.207	-0.031	-0.471	0.011	-0.835	-0.114	0.03	**-0.018**	0.03
n=500											
RMSE	**0.147**	0.495	0.224	0.152	0.488	0.147	0.763	0.145	0.148	**0.147**	0.148
SEE	**0.412**	0.505	0.434	0.473	0.521	0.725	0.668	1.071	1.31	**0.412**	1.31
Bias	**-0.004**	0.471	-0.154	0.025	-0.458	0.014	-0.74	-0.005	0.026	**-0.004**	0.026
n=1000											
RMSE	**0.148**	0.482	0.244	0.147	0.501	0.146	0.818	0.176	0.147	**0.148**	0.147
SEE	**0.294**	0.36	0.307	0.341	0.368	0.515	0.483	0.825	0.933	**0.294**	0.933
Bias	**-0.018**	0.457	-0.187	-0.016	-0.472	0.001	-0.799	-0.1	0.017	**-0.018**	0.017

TABLE 18. *Summary statistics for POP1 equipercentile equating methods*

	2SG									SG	EG
	Equipercentile									Traditional	Traditional
POP1	2SG (.5,.5)	2SG (.5,.75)	2SG (.6,.5)	2SG (.6,.6)	2SG (.75,.5)	2SG (.75,.75)	2SG (.9,.5)	2SG (.9,.9)	2SG (1,1)	Equi.	Equi.
n=50											
RMSE	**0.176**	0.436	0.332	0.242	0.535	0.202	1.097	0.598	0.379	0.318	0.394
SEE	**1.218**	1.526	1.252	1.411	1.505	2.123	1.972	3.3	3.955	2.145	3.925
Bias	**-0.041**	0.383	-0.277	-0.167	-0.473	-0.05	-1.046	-0.475	-0.069	-0.027	0.005
n=100											
RMSE	**0.158**	0.461	0.29	0.191	0.476	0.172	0.888	0.373	0.25	0.377	0.356
SEE	**0.862**	1.095	0.879	1.02	1.089	1.526	1.397	2.366	2.752	1.772	3.026
Bias	**-0.018**	0.411	-0.231	-0.1	-0.411	0.017	-0.843	-0.266	0.048	0.004	0.099
n=300											
RMSE	**0.149**	0.458	0.244	0.151	0.465	0.153	0.769	0.188	0.173	0.33	0.31
SEE	**0.51**	0.637	0.552	0.613	0.637	0.887	0.828	1.362	1.587	1.228	1.996
Bias	**-0.014**	0.4	-0.179	-0.026	-0.406	0.008	-0.713	-0.081	0.028	-0.032	0.071
n=500											
RMSE	**0.148**	0.456	0.219	0.155	0.459	0.15	0.7	0.156	0.16	0.363	0.316
SEE	**0.398**	0.481	0.41	0.451	0.494	0.675	0.625	0.989	1.203	1.112	1.684
Bias	**-0.003**	0.401	-0.128	0.029	-0.394	0.01	-0.628	0.001	0.021	0.01	0.061
n=1000											
RMSE	**0.148**	0.445	0.236	0.148	0.47	0.148	0.747	0.173	0.152	0.328	0.305
SEE	**0.288**	0.347	0.3	0.324	0.353	0.483	0.456	0.759	0.861	0.897	1.344
Bias	**-0.016**	0.387	-0.163	-0.017	-0.407	-0.004	-0.68	-0.079	0.009	-0.039	-0.006

TABLE 19. *Summary statistics for POP2 linear equating methods*

POP2	2SG									SG	EG
	Linear									Traditional	Traditional
	2SG (.5,.5)	2SG (.5,.75)	2SG (.6,.5)	2SG (.6,.6)	2SG (.75,.5)	2SG (.75,.75)	2SG (.9,.5)	2SG (.9,.9)	2SG (1,1)	Linear	Linear
n=50											
RMSE	0.332	**0.164**	0.469	0.321	0.712	0.322	1.003	0.349	0.386	0.29	0.452
SEE	1.193	**1.513**	1.221	1.36	1.473	2.156	1.909	3.213	3.992	1.192	3.996
Bias	-0.302	**0.089**	-0.442	-0.293	-0.676	-0.294	-0.945	-0.319	-0.352	0.114	-0.074
n=100											
RMSE	0.435	**0.128**	0.548	0.372	0.735	0.288	0.948	0.22	0.185	0.301	0.426
SEE	0.873	**1.089**	0.907	1	1.102	1.526	1.418	2.222	2.733	0.866	2.688
Bias	-0.399	**0.042**	-0.51	-0.338	-0.692	-0.255	-0.891	-0.182	-0.141	0.221	-0.154
n=300											
RMSE	0.414	**0.194**	0.515	0.312	0.679	0.173	0.854	0.126	0.189	0.26	0.54
SEE	0.555	**0.703**	0.589	0.652	0.724	1.016	0.925	1.478	1.811	0.551	1.837
Bias	-0.363	**0.154**	-0.465	-0.261	-0.623	-0.108	-0.79	0.045	0.145	0.194	-0.376
n=500											
RMSE	0.452	**0.162**	0.552	0.353	0.711	0.214	0.878	0.124	0.153	0.267	0.452
SEE	0.429	**0.543**	0.462	0.512	0.572	0.781	0.729	1.118	1.36	0.426	1.379
Bias	-0.389	**0.109**	-0.49	-0.293	-0.646	-0.147	-0.807	0	0.099	0.222	-0.3
n=1000											
RMSE	0.511	**0.142**	0.607	0.407	0.757	0.254	0.913	0.134	0.139	0.305	0.403
SEE	0.297	**0.38**	0.319	0.355	0.391	0.539	0.492	0.768	0.932	0.293	0.934
Bias	-0.441	**0.056**	-0.539	-0.341	-0.687	-0.188	-0.838	-0.032	0.073	0.276	-0.254

TABLE 20. *Summary statistics for POP2 equipercentile equating methods*

POP2	2SG									SG	EG
	Equipercentile									Traditional	Traditional
	2SG (.5,.5)	2SG (.5,.75)	2SG (.6,.5)	2SG (.6,.6)	2SG (.75,.5)	2SG (.75,.75)	2SG (.9,.5)	2SG (.9,.9)	2SG (1,1)	Equi.	Equi.
n=50											
RMSE	0.295	**0.168**	0.413	0.283	0.626	0.269	0.891	0.288	0.373	0.333	0.387
SEE	1.19	**1.42**	1.197	1.297	1.39	1.951	1.769	2.897	3.665	1.989	3.714
Bias	-0.261	**0.099**	-0.385	-0.251	-0.595	-0.233	-0.846	-0.211	-0.192	-0.254	-0.249
n=100											
RMSE	0.385	**0.12**	0.486	0.329	0.664	0.253	0.862	0.183	0.182	0.443	0.252
SEE	0.864	**1.024**	0.882	0.951	1.037	1.39	1.303	2.014	2.509	1.62	2.881
Bias	-0.355	**0.051**	-0.456	-0.3	-0.629	-0.223	-0.817	-0.131	-0.078	-0.367	-0.068
n=300											
RMSE	0.386	**0.178**	0.481	0.303	0.637	0.188	0.815	0.141	0.18	0.387	0.283
SEE	0.553	**0.656**	0.567	0.61	0.669	0.922	0.842	1.359	1.694	1.274	2.103
Bias	-0.338	**0.132**	-0.434	-0.251	-0.587	-0.117	-0.754	0.018	0.112	-0.272	0.044
n=500											
RMSE	0.423	**0.158**	0.517	0.347	0.674	0.24	0.834	0.156	0.156	0.443	0.331
SEE	0.421	**0.513**	0.441	0.477	0.532	0.721	0.672	1.045	1.287	1.217	1.883
Bias	-0.365	**0.087**	-0.461	-0.285	-0.613	-0.161	-0.767	-0.028	0.062	-0.246	0.004
n=1000											
RMSE	0.485	**0.154**	0.58	0.406	0.723	0.283	0.873	0.183	0.16	0.451	0.215
SEE	0.286	**0.357**	0.304	0.336	0.363	0.498	0.455	0.716	0.878	0.973	1.406
Bias	-0.419	**0.034**	-0.514	-0.336	-0.655	-0.203	-0.801	-0.066	0.028	-0.323	0.078

TABLE 21. *Summary statistics for POP3 linear equating methods*

POP3	2SG									SG	EG
	Linear									Traditional	Traditional
	2SG (.5,.5)	2SG (.5,.75)	2SG (.6,.5)	2SG (.6,.6)	2SG (.75,.5)	2SG (.75,.75)	2SG (.9,.5)	2SG (.9,.9)	2SG (1,1)	Linear	Linear
n=50											
RMSE	0.887	1.007	0.636	0.675	0.241	0.336	0.315	**0.14**	0.386	1.178	0.452
SEE	1.182	1.515	1.208	1.353	1.46	2.155	1.908	**3.225**	4.02	1.218	4.024
Bias	0.864	0.967	0.616	0.65	0.213	0.306	-0.235	**-0.073**	-0.349	-1.064	-0.072
n=100											
RMSE	0.773	0.942	0.552	0.614	0.221	0.364	0.267	**0.133**	0.185	1.037	0.426
SEE	0.865	1.096	0.901	1	1.099	1.53	1.425	**2.232**	2.752	0.886	2.708
Bias	0.758	0.911	0.537	0.596	0.188	0.338	-0.187	**0.06**	-0.139	-0.942	-0.152
n=300											
RMSE	0.81	1.051	0.599	0.691	0.288	0.506	0.204	**0.313**	0.188	1.052	0.54
SEE	0.544	0.697	0.579	0.645	0.718	1.014	0.928	**1.483**	1.824	0.558	1.85
Bias	0.799	1.026	0.588	0.678	0.261	0.488	-0.081	**0.288**	0.147	-0.973	-0.374
n=500											
RMSE	0.776	0.998	0.567	0.652	0.271	0.461	0.234	**0.268**	0.153	1.003	0.452
SEE	0.422	0.537	0.456	0.506	0.569	0.779	0.732	**1.122**	1.37	0.433	1.389
Bias	0.768	0.979	0.557	0.642	0.233	0.446	-0.104	**0.242**	0.101	-0.937	-0.298
n=1000											
RMSE	0.718	0.937	0.517	0.599	0.244	0.417	0.265	**0.237**	0.139	0.934	0.403
SEE	0.297	0.381	0.321	0.356	0.393	0.54	0.496	**0.772**	0.939	0.303	0.94
Bias	0.71	0.923	0.504	0.589	0.187	0.402	-0.139	**0.208**	0.075	-0.876	-0.252

TABLE 22. *Summary statistics for POP3 equipercentile equating methods*

POP3	2SG									SG	EG
	KE Equipercentile									Traditional	Traditional
	2SG (.5,.5)	2SG (.5,.75)	2SG (.6,.5)	2SG (.6,.6)	2SG (.75,.5)	2SG (.75,.75)	2SG (.9,.5)	2SG (.9,.9)	2SG (1,1)	Equi.	Equi.
n=50											
RMSE	0.808	0.963	0.599	0.649	0.262	0.435	**0.215**	0.357	0.476	0.712	0.387
SEE	1.17	1.407	1.176	1.278	1.365	1.919	**1.743**	2.842	3.598	2.023	3.626
Bias	0.773	0.912	0.571	0.617	0.236	0.378	**-0.147**	0.131	-0.034	0.559	-0.245
n=100											
RMSE	0.713	0.903	0.524	0.591	0.228	0.421	**0.175**	0.316	0.322	0.67	0.252
SEE	0.851	1.021	0.867	0.94	1.019	1.367	**1.292**	1.977	2.465	1.724	2.813
Bias	0.683	0.861	0.501	0.565	0.205	0.386	**-0.116**	0.204	0.074	0.453	-0.069
n=300											
RMSE	0.74	0.987	0.555	0.645	0.278	0.51	**0.158**	0.389	0.331	0.749	0.283
SEE	0.54	0.643	0.557	0.599	0.656	0.903	**0.836**	1.334	1.661	1.294	2.054
Bias	0.707	0.947	0.527	0.617	0.246	0.488	**-0.05**	0.354	0.263	0.395	0.039
n=500											
RMSE	0.711	0.937	0.529	0.613	0.263	0.466	**0.181**	0.33	0.265	0.744	0.331
SEE	0.412	0.505	0.434	0.476	0.527	0.711	**0.669**	1.026	1.263	1.177	1.839
Bias	0.676	0.902	0.497	0.587	0.22	0.448	**-0.069**	0.305	0.21	0.408	0.001
n=1000											
RMSE	0.659	0.877	0.482	0.56	0.241	0.414	**0.221**	0.285	0.219	0.666	0.215
SEE	0.29	0.358	0.309	0.34	0.367	0.496	**0.458**	0.71	0.863	0.925	1.372
Bias	0.619	0.845	0.441	0.532	0.169	0.397	**-0.108**	0.266	0.175	0.475	0.075

TABLE 23. *Summary statistics for POP4 linear equating methods*

POP4	2SG									SG	EG
	Linear									Traditional	Traditional
	2SG (.5,.5)	2SG (.5,.75)	2SG (.6,.5)	2SG (.6,.6)	2SG (.75,.5)	2SG (.75,.75)	2SG (.9,.5)	2SG (.9,.9)	2SG (1,1)	Linear	Linear
n=50											
RMSE	1.246	1.516	1.132	1.296	0.522	0.796	0.591	1.124	**0.534**	1.247	**0.534**
SEE	1.366	1.85	1.433	1.611	1.835	2.596	2.292	3.55	**4.4**	1.366	**4.4**
Bias	-1.24	-1.511	-1.107	-1.268	-0.489	-0.774	-0.402	-1.011	**-0.392**	-1.24	**-0.392**
n=100											
RMSE	1.255	1.556	0.982	1.104	0.47	0.771	0.263	0.626	**0.464**	1.255	**0.464**
SEE	0.945	1.249	1.002	1.118	1.226	1.729	1.528	2.355	**2.937**	0.946	**2.937**
Bias	-1.248	-1.549	-0.974	-1.095	-0.444	-0.749	-0.094	-0.553	**-0.308**	-1.248	**-0.308**
n=300											
RMSE	1.232	1.514	0.951	1.082	0.408	0.688	0.149	0.556	**0.239**	1.233	**0.239**
SEE	0.559	0.739	0.623	0.694	0.724	1.015	0.891	1.408	**1.694**	0.559	**1.694**
Bias	-1.227	-1.51	-0.943	-1.074	-0.395	-0.679	-0.017	-0.529	**-0.18**	-1.228	**-0.18**
n=500											
RMSE	1.219	1.476	0.906	1.014	0.367	0.62	0.144	0.364	**0.123**	1.219	**0.123**
SEE	0.418	0.573	0.453	0.527	0.557	0.792	0.712	1.176	**1.321**	0.418	**1.321**
Bias	-1.214	-1.472	-0.899	-1.008	-0.353	-0.612	0.079	-0.346	**-0.062**	-1.215	**-0.062**
n=1000											
RMSE	1.228	1.474	0.885	0.987	0.362	0.605	0.161	0.305	**0.103**	1.228	**0.103**
SEE	0.301	0.414	0.335	0.385	0.404	0.571	0.497	0.787	**0.944**	0.301	**0.944**
Bias	-1.223	-1.47	-0.878	-0.981	-0.348	-0.597	0.122	-0.286	**-0.022**	-1.223	**-0.022**

TABLE 24. *Summary statistics for POP4 equipercentile equating methods*

POP4	2SG									SG	EG
	Equipercentile									Traditional	Traditional
	2SG (.5,.5)	2SG (.5,.75)	2SG (.6,.5)	2SG (.6,.6)	2SG (.75,.5)	2SG (.75,.75)	2SG (.9,.5)	2SG (.9,.9)	2SG (1,1)	Equi.	Equi.
n=50											
RMSE	1.156	1.413	1.031	1.179	0.56	0.802	0.591	1.015	0.557	1.074	0.367
SEE	1.348	1.746	1.364	1.517	1.741	2.363	2.089	3.159	3.926	2.181	3.821
Bias	-1.135	-1.384	-1.001	-1.144	-0.517	-0.766	-0.43	-0.922	-0.395	-0.769	-0.096
n=100											
RMSE	1.176	1.474	0.933	1.056	0.53	0.821	0.36	0.728	0.599	1.072	0.379
SEE	0.913	1.186	0.977	1.081	1.176	1.6	1.436	2.134	2.624	1.807	2.867
Bias	-1.16	-1.453	-0.914	-1.033	-0.492	-0.785	-0.187	-0.627	-0.417	-0.803	-0.065
n=300											
RMSE	1.154	1.449	0.897	1.03	0.453	0.748	0.23	0.653	0.454	1.055	0.296
SEE	0.548	0.713	0.615	0.678	0.707	0.958	0.857	1.302	1.537	1.271	1.908
Bias	-1.139	-1.433	-0.881	-1.012	-0.436	-0.73	-0.1	-0.6	-0.329	-0.741	-0.034
n=500											
RMSE	1.141	1.416	0.859	0.973	0.41	0.684	0.191	0.505	0.36	1.04	0.271
SEE	0.413	0.555	0.445	0.509	0.544	0.749	0.672	1.057	1.2	1.133	1.656
Bias	-1.126	-1.4	-0.843	-0.956	-0.395	-0.669	-0.015	-0.454	-0.229	-0.796	0.007
n=1000											
RMSE	1.149	1.419	0.836	0.945	0.405	0.675	0.164	0.445	0.345	1.044	0.299
SEE	0.299	0.399	0.322	0.365	0.393	0.537	0.465	0.709	0.855	0.955	1.345
Bias	-1.135	-1.404	-0.821	-0.929	-0.391	-0.661	0.027	-0.401	-0.201	-0.808	0.052

TABLE 25. *Summary statistics for POP5 linear equating methods*

POP5	2SG									SG	EG
	Linear									Traditional	Traditional
	2SG (.5,.5)	2SG (.5,.75)	2SG (.6,.5)	2SG (.6,.6)	2SG (.75,.5)	2SG (.75,.75)	2SG (.9,.5)	2SG (.9,.9)	2SG (1,1)	Linear	Linear
n=50											
RMSE	2.049	1.51	1.872	1.725	1.517	0.972	1.468	0.839	**0.12**	2.049	**0.12**
SEE	1.232	1.634	1.303	1.528	1.629	2.354	2.118	3.581	**4.173**	1.232	**4.173**
Bias	-2.032	-1.464	-1.863	-1.703	-1.517	-0.96	-1.465	-0.786	**0.064**	-2.033	**0.064**
n=100											
RMSE	2.094	1.562	1.93	1.759	1.546	0.994	1.438	0.694	**0.149**	2.094	**0.149**
SEE	0.85	1.113	0.945	1.072	1.132	1.621	1.495	2.382	**2.858**	0.85	**2.858**
Bias	-2.076	-1.514	-1.919	-1.738	-1.546	-0.988	-1.431	-0.677	**0.082**	-2.076	**0.082**
n=300											
RMSE	2.05	1.56	1.829	1.64	1.508	0.988	1.282	0.46	**0.145**	2.051	**0.144**
SEE	0.493	0.66	0.594	0.667	0.661	0.944	0.89	1.395	**1.635**	0.493	**1.635**
Bias	-2.037	-1.513	-1.824	-1.63	-1.507	-0.983	-1.253	-0.445	**0.073**	-2.037	**0.073**
n=500											
RMSE	2.034	1.549	1.801	1.608	1.496	0.98	1.235	0.395	**0.149**	2.034	**0.148**
SEE	0.381	0.503	0.456	0.497	0.498	0.721	0.644	1.046	**1.265**	0.381	**1.265**
Bias	-2.021	-1.504	-1.796	-1.597	-1.495	-0.975	-1.207	-0.38	**0.074**	-2.022	**0.074**
n=1000											
RMSE	2.045	1.574	1.823	1.632	1.513	1.009	1.237	0.403	**0.107**	2.045	**0.107**
SEE	0.286	0.378	0.313	0.351	0.376	0.533	0.467	0.778	**0.916**	0.286	**0.916**
Bias	-2.032	-1.525	-1.815	-1.617	-1.512	-1.002	-1.216	-0.392	**0.035**	-2.032	**0.035**

TABLE 26. *Summary statistics for POP5 equipercentile equating methods*

POP5	2SG									SG	EG
	Equipercentile									Traditional	Traditional
	2SG (.5,.5)	2SG (.5,.75)	2SG (.6,.5)	2SG (.6,.6)	2SG (.75,.5)	2SG (.75,.75)	2SG (.9,.5)	2SG (.9,.9)	2SG (1,1)	Equi.	Equi.
n=50											
RMSE	1.792	1.327	1.655	1.523	1.395	0.915	1.362	0.787	0.268	1.667	0.533
SEE	1.163	1.528	1.258	1.444	1.526	2.14	1.941	3.155	3.727	2.108	3.836
Bias	-1.714	-1.262	-1.607	-1.479	-1.332	-0.877	-1.321	-0.725	-0.015	-1.362	0.216
n=100											
RMSE	1.839	1.385	1.71	1.561	1.422	0.942	1.346	0.698	0.21	1.684	0.417
SEE	0.802	1.051	0.905	1.017	1.064	1.492	1.38	2.169	2.586	1.719	2.993
Bias	-1.764	-1.32	-1.666	-1.522	-1.358	-0.909	-1.299	-0.669	-0.047	-1.385	0.122
n=300											
RMSE	1.814	1.404	1.622	1.457	1.393	0.949	1.206	0.496	0.251	1.645	0.321
SEE	0.471	0.628	0.569	0.635	0.631	0.88	0.829	1.282	1.488	1.242	2.062
Bias	-1.742	-1.34	-1.573	-1.419	-1.327	-0.919	-1.122	-0.478	-0.098	-1.304	0.07
n=500											
RMSE	1.802	1.397	1.6	1.431	1.384	0.943	1.165	0.463	0.257	1.633	0.285
SEE	0.365	0.481	0.434	0.475	0.478	0.676	0.608	0.964	1.152	1.031	1.698
Bias	-1.731	-1.334	-1.553	-1.396	-1.317	-0.914	-1.091	-0.443	-0.102	-1.339	0.05
n=1000											
RMSE	1.815	1.424	1.616	1.452	1.398	0.973	1.164	0.496	0.31	1.644	0.314
SEE	0.279	0.364	0.301	0.339	0.362	0.5	0.439	0.709	0.831	0.931	1.379
Bias	-1.745	-1.359	-1.573	-1.418	-1.335	-0.943	-1.101	-0.467	-0.149	-1.326	0.067

TABLE 27. *Summary statistics for POP6 linear equating methods*

POP6	2SG									SG	EG
	Linear									Traditional	Traditional
	2SG (.5,.5)	2SG (.5,.75)	2SG (.6,.5)	2SG (.6,.6)	2SG (.75,.5)	2SG (.75,.75)	2SG (.9,.5)	2SG (.9,.9)	2SG (1,1)	Linear	Linear
n=50											
RMSE	3.041	2.145	3.018	2.738	2.537	1.66	2.598	1.455	**0.394**	3.042	**0.394**
SEE	1.317	1.635	1.401	1.578	1.645	2.288	2.21	3.632	**4.041**	1.317	**4.041**
Bias	-3.006	-2.11	-2.951	-2.662	-2.497	-1.615	-2.483	-1.284	**-0.252**	-3.007	**-0.252**
n=100											
RMSE	3.046	2.137	2.957	2.63	2.493	1.595	2.378	1.02	**0.186**	3.046	**0.186**
SEE	0.914	1.172	0.981	1.11	1.164	1.659	1.546	2.537	**2.932**	0.914	**2.932**
Bias	-3.007	-2.102	-2.908	-2.579	-2.46	-1.563	-2.324	-0.965	**-0.122**	-3.008	**-0.122**
n=300											
RMSE	3.043	2.14	2.819	2.466	2.453	1.555	2.134	0.678	**0.153**	3.044	**0.153**
SEE	0.52	0.657	0.547	0.604	0.655	0.925	0.833	1.387	**1.618**	0.52	**1.618**
Bias	-3.003	-2.098	-2.781	-2.425	-2.419	-1.518	-2.098	-0.626	**-0.018**	-3.004	**-0.018**
n=500											
RMSE	3.054	2.154	2.834	2.471	2.466	1.57	2.136	0.634	**0.143**	3.054	**0.144**
SEE	0.398	0.501	0.418	0.441	0.49	0.686	0.562	0.87	**1.187**	0.399	**1.187**
Bias	-3.014	-2.113	-2.796	-2.433	-2.433	-1.535	-2.104	-0.605	**-0.038**	-3.015	**-0.039**
n=1000											
RMSE	3.047	2.147	2.83	2.474	2.453	1.556	2.134	0.659	**0.104**	3.048	**0.105**
SEE	0.288	0.361	0.293	0.325	0.356	0.496	0.437	0.703	**0.857**	0.288	**0.857**
Bias	-3.01	-2.109	-2.796	-2.439	-2.424	-1.526	-2.108	-0.639	**-0.025**	-3.01	**-0.025**

TABLE 28. *Summary statistics for POP6 equipercentile equating methods*

	2SG									SG	EG
	Equipercentile									Traditional	Traditional
POP6	2SG (.5,.5)	2SG (.5,.75)	2SG (.6,.5)	2SG (.6,.6)	2SG (.75,.5)	2SG (.75,.75)	2SG (.9,.5)	2SG (.9,.9)	2SG (1,1)	Equi.	Equi.
n=50											
RMSE	2.727	1.999	2.691	2.465	2.35	1.644	2.403	1.459	0.73	2.395	0.252
SEE	1.201	1.47	1.272	1.418	1.473	2.01	1.95	3.102	3.562	2.12	3.706
Bias	-2.625	-1.923	-2.609	-2.383	-2.261	-1.56	-2.288	-1.293	-0.459	-2.05	-0.038
n=100											
RMSE	2.744	2.007	2.657	2.394	2.327	1.608	2.238	1.165	0.646	2.39	0.21
SEE	0.835	1.058	0.888	0.997	1.048	1.472	1.372	2.244	2.625	1.679	2.911
Bias	-2.643	-1.931	-2.589	-2.329	-2.243	-1.532	-2.168	-1.06	-0.385	-2.067	-0.056
n=300											
RMSE	2.747	2.02	2.545	2.261	2.297	1.592	2.034	0.99	0.728	2.37	0.261
SEE	0.478	0.599	0.499	0.55	0.592	0.831	0.744	1.24	1.461	1.235	1.956
Bias	-2.648	-1.941	-2.484	-2.205	-2.215	-1.508	-1.977	-0.822	-0.348	-2.002	0.03
n=500											
RMSE	2.758	2.034	2.559	2.268	2.311	1.606	2.037	0.947	0.735	2.374	0.25
SEE	0.368	0.461	0.382	0.405	0.447	0.622	0.511	0.812	1.079	1.092	1.606
Bias	-2.659	-1.955	-2.499	-2.213	-2.228	-1.523	-1.984	-0.803	-0.37	-1.993	0.015
n=1000											
RMSE	2.755	2.029	2.56	2.274	2.301	1.593	2.035	0.96	0.703	2.379	0.243
SEE	0.271	0.338	0.275	0.302	0.334	0.458	0.4	0.65	0.786	0.918	1.309
Bias	-2.657	-1.953	-2.503	-2.221	-2.222	-1.516	-1.987	-0.831	-0.358	-2.014	0.029

The results indicate the KE methods can approximate their corresponding traditional equating methods. No large differences were found between the KE equating methods and their corresponding traditional equating methods (e.g., KE linear and traditional linear, KE equipercentile and traditional equipercentile). This is consistent with the results of evaluation studies for KE, such as Mao, von Davier, and Rupp (2005), von Davier, Holland, Livingston, and others (2005).

Compared to the standard error of equating, the equating bias index is more sample size independent. Given the same equating method, the equating bias does not change a great deal as sample size increases. However, the standard error of equating decreases conspicuously as sample size increases. The more data we have, the more information we can use to estimate the equating relationship; the less equating error there will be. This feature of SEE is inherited from its calculation formula.

When using RMSE as a means of evaluating equating functions, it was found that: a) When DOE is almost zero, pooling the two samples together or using the 2SG approach with weights of (.5, .5) are the optimal equating methods with small standard error of equating and small bias; b) As DOE increases, the 2SG methods under the KE framework with different weights can provide optimal equating results with smallest RMSE. The weights for the 2SG approach gets larger as DOE increases; c) When the size of DOE approaches to a certain point, treating data collected in a CB design as an EG design will be the best equating solution. The weights of the 2SG approach will become 1. The equating method could be either 2SG (1, 1) or traditional linear or equipercentile method.

4.3.3 Evaluating the Equating Results by SEED

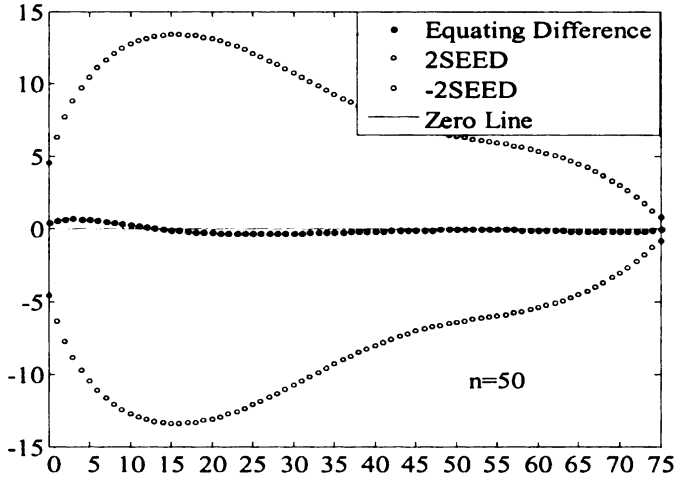
Equating differences were compared against their 95% confidence intervals for all the sample size conditions under each population. The last graph in Figure 12 plots the equating differences between 2SG(.5, .5) linear and 2SG(.5, .5) equipercentile methods for simulated data 1 when sample size is 1000. The straight horizontal line in the middle is the zero line. The equating differences represented by solid dots are around the zero line within the range of the ± 2SEED band. The other five graphs present the equating differences between the 2SG equipercentile equating with weights of (.5, .5) and (1, 1) for different sample sizes drawn from simulated data 1.

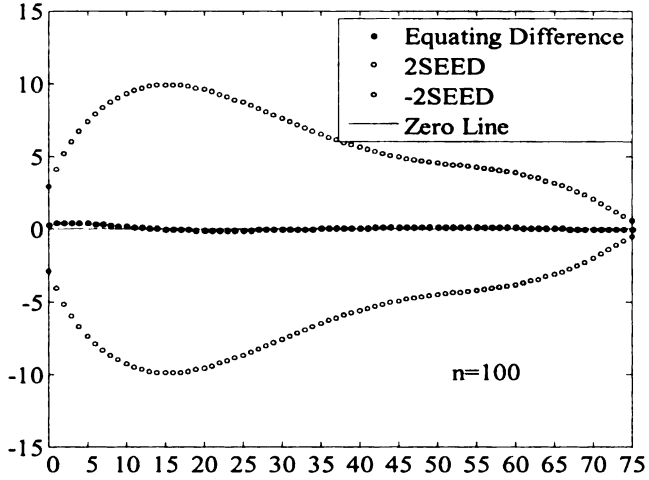
It can be seen from these plots that SEED gets larger when the equating methods are different from each other and when sample size decreases. Among the graphs in Figure 12, the last graph exhibits the smallest SEED, showing that the 2SG methods with the same weighting parameters provide more similar equating results than the 2SG methods with different weighting parameters. Furthermore, the plots in Figure 12 indicate that under a certain order effect situation, the equating difference stays relatively unchanged, but SEED decreases as sample size increases. Therefore, the significance of the equating difference mostly depends on the sample size. If the equating differences between two methods fall beyond the ± 2SEED band when sample size is 500, they must also be out of the band when sample size is 1000. Reversely, if the equating difference is not significant when sample size is 1000, then it must not be significant when sample size is 500.

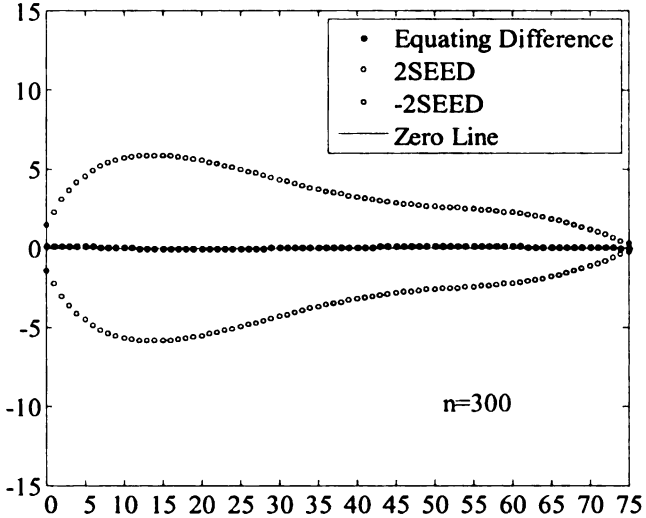
More SEED plots are provided in the appendix. Most of the SEED plots are for the differences between the 2SG(.5, .5) method and the 2SG(1, 1) method. The rational of

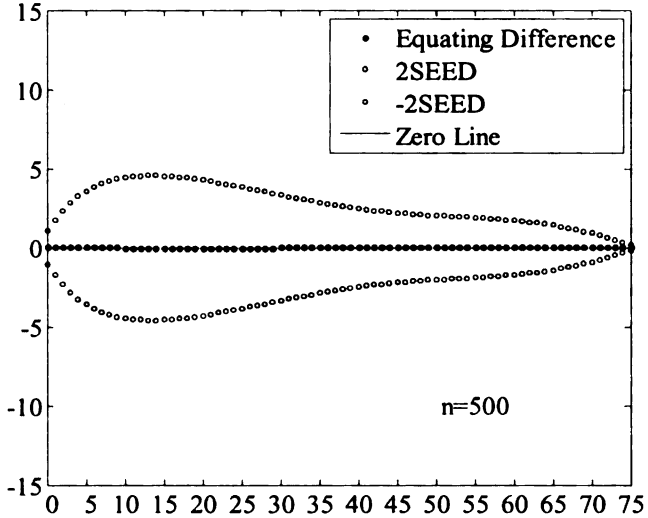
not comparing the equating difference between the 2SG(1, 1) method with the 2SG method with any weights between 0.5 and 1 is provided here: In equating for a CB design with differential order effect, the 2SG approach with weights of (1, 1) has no equating bias. The 2SG approach with weights of (.5, .5) will have the biggest equating bias. If the equating difference between 2SG(.5, .5) and 2SG(1, 1) is not significant, then the equating difference between 2SG(1, 1) and a 2SG approach with any weights between 0.5 and 1 will not be significant.

All the SEED plots for all the simulated datasets indicate that none of the equating differences between methods 2SG(.5, .5) and 2SG(1, 1) under different sample size conditions of population data 1, data 2 and data 3 are significant. Therefore the bias introduced by using data from tests taken second can be ignored. Thus the 2SG approach with weights of (.5, .5) can be selected as the best equating line for simulated data 1, data 2 and data 3 when the effect size of DOE is relatively small.


Equating Difference
2SEED
-2SEED
Zero Line
n=50
15
10
5
0
-5
-10
-15
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75


Equating Difference
2SEED
-2SEED
Zero Line
n=100
15
10
5
0
-5
-10
-15
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75


Equating Difference
2SEED
-2SEED
Zero Line
n=300
15
10
5
0
-5
-10
-15
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75


Equating Difference
2SEED
-2SEED
Zero Line
n=500
15
10
5
0
-5
-10
-15
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75

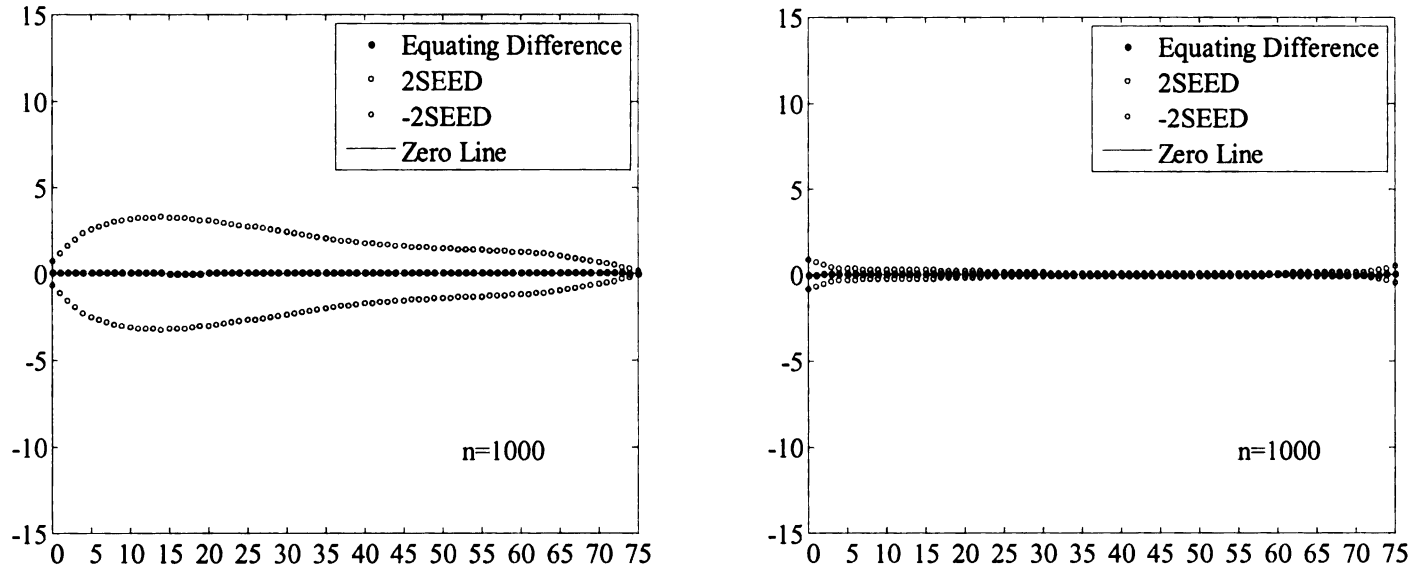


FIGURE 12. *Equating differences and the ± 2SEED band for simulated data 1.*

As DOE increases in simulated data 4, the equating difference between methods 2SG(.5, .5) and 2SG(1, 1) falls beyond the 95% confidence interval when sample size is 1000. In this case, the 2SG approach with weights of (1, 1) is preferred to avoid the equating bias introduced by including data from X_2 and Y_2. This is also the case for data 5 when sample size is 500 and 1000 and for data 6 when sample size is 300, 500 and 1000.

Table 29 summarizes the equating functions selected by using SEED plots for different samples under different order effect situations. It reflects that the EG design (the 2SG approach with weights of (1, 1)) is more appropriate at the lower right corner when DOE gets larger and when sample size gets bigger.

TABLE 29. *Selected equating function based on SEED*

DOE	n=50	n=100	n=300	n=500	n=1000
d=0	**2SG (.5, .5)**	**2SG (.5, .5)**	**2SG (.5, .5)**	**2SG (.5, .5)**	**2SG (.5, .5)**
d=0.025	2SG (.5, .5)	2SG (.5, .5)	2SG (.5, .5)	2SG (.5, .5)	2SG (.5, .5)
d=0.05	2SG (.5, .5)	2SG (.5, .5)	2SG (.5, .5)	2SG (.5, .5)	2SG (.5, .5)
d=0.1	2SG (.5, .5)	2SG (.5, .5)	2SG (.5, .5)	2SG (.5, .5)	**2SG (1, 1)**
d=0.15	2SG (.5, .5)	2SG (.5, .5)	2SG (.5, .5)	**2SG (1, 1)**	**2SG (1, 1)**
d=0.2	2SG (.5, .5)	2SG (.5, .5)	**2SG (1, 1)**	**2SG (1, 1)**	**2SG (1, 1)**

TABLE 30. *Selected equating function based on RMSE*

DOE	n=50	n=100	n=300	n=500	n=1000
d=0	**2SG (.5, .5)**	**2SG (.5, .5)**	**2SG (.5, .5)**	**2SG (.5, .5)**	**2SG (.5, .5)**
d=0.025	2SG (.5, .75)	2SG (.5, .75)	2SG (.5, .75)	2SG (.5, .75)	2SG (.5, .75)
d=0.05	2SG (.9, .9)	2SG (.9, .9)	2SG (.9, .9)	2SG (.9, .9)	2SG (.9, .9)
d=0.1	2SG (1, 1)	2SG (1, 1)	2SG (1, 1)	2SG (1, 1)	**2SG (1, 1)**
d=0.15	2SG (1, 1)	2SG (1, 1)	2SG (1, 1)	**2SG (1, 1)**	**2SG (1, 1)**
d=0.2	2SG (1, 1)	2SG (1, 1)	**2SG (1, 1)**	**2SG (1, 1)**	**2SG (1, 1)**

Comparing Table 30 with Table 29, it can be found that the RMSE and SEED statistical indices produce same results when DOE is almost zero and when DOE is large

(effect size > 0.2 in this case). When the effect size of DOE is within a certain small range, the RMSE can provide more fine-grained equating solution. This is when the weighting method comes into place.

Chapter V: Discussion

5.1 Performance of the KE Methods

The results of this study are consistent with previous studies that compared the KE methods with the traditional equating methods. In general, the KE methods produce results very similar to their corresponding traditional equating methods. These similarities in equating results support KE method as a promising unified approach to test equating based on a flexible family of equipercentile-like equating functions. The entire classic observed score equating methods can be incorporated into its framework. The summary statistics in Table 17 to Table 28 indicate that the 2SG(.5, .5) linear method and the SG linear method produce very similar equating results in terms of SEE, equating bias and RMSE. Similarly, the 2SG(1, 1) linear and traditional EG linear equating methods provide equating results very close to each other; so are the 2SG(.5, .5) equipercentile, SG KE equipercentile and traditional SG equipercentile equating methods. The equating differences between 2SG(1, 1) equipercentile method and the traditional EG equipercentile method are small as well. Although the summary statistics in Table 17 to Table 28 indicate their equating difference is relatively larger compared to the equating differences between the other previously-discussed approximation pairs. The actual differences of their equating functions are smaller than 1 raw score point for any score point above chance score, which are not large differences. Figure A28 to Figure A34 plot the equating differences between the 2SG(1, 1) equipercentile method and the

traditional EG equipercentile method for selected cases. The equating differences between these two methods are the biggest in simulated data 6.

KE provides the SEED statistics for examining the equating difference between two KE methods. The usefulness of this statistics is discussed below.

5.2 Effects of the Weighting Method

The overall equating accuracy consists of two parts: random equating error (SEE) and systematic error (equating bias). When a CB design is used to collect data for an equating, the 2SG approach under KE framework attempts to provide an optimal equating solution with the least overall equating error, which is indicated by the magnitude of RMSE in this study.

In the rest of this section, the effect of the weighting method in enhancing overall equating accuracy is discussed in terms of both equating bias and the overall equating error.

The study results based on both real and simulated data indicate that the weighting mechanism is effective in some extent. As DOE gets larger, the weights with smallest RMSE also increase (as indicated in Table 30 for simulated data 2 and data 3). Because random equating error increases as weights increase, the reduction in RMSE must be due to the reduction of equating bias. Therefore, the results of this study demonstrate that the 2SG approach can reduce systematic equating error by adjusting the weights placed on the data from tests taken first. However, the reduction in equating bias is not significant as indicated by the SEED plots (as indicated in Table 29 for simulated data 2 and data 3). The reduction of equating bias is only significant when sample size is large enough and when DOE is big enough. When this happens, the weights in the 2SG

approach will be (1, 1), which indicates an EG design.

The reason for the small amount of improvement in terms of RMSE is because, as DOE gets larger, examinee's performance on the second test will be more affected by order effects and will be less accurate. Thus the 2SG approach assigns more weights on the tests taken first to reduce bias introduced by order effects. The bigger the order effects, the more weights will be put on the tests taken first to reduce bias. However, the more weights on the first tests, the bigger the random equating errors are. Because of this trade-off between random equating error and system equating error, when both random and systematic equating errors are considered together, the equating error in terms of RMSE does not seem to be reduced much.

The findings of this study support the 2SG approach as a sensitive approach with the flexibility of using optimal data information as the size of order effects changes. The RMSE index provides more detailed information and can help decide which weights to use. However, the way of trying every possible weight between 0.5 and 1 to decide the fine-grained weights using the criterion of RMSE involves lengthy calculations.

Other possible ways of determining how to treat the data collected by a CB design could be the hypothesis testing of DOE introduced in the method section and the SEED method applied in this study. If the hypothesis test of DOE is not significant, the data collected by a CB design shall be pooled together as a SG design. Otherwise, the data shall be treated as an EG design. The SEED plot method tests the significance of the equating difference between 2SG(.5, .5) and 2SG(1, 1). If the equating difference is not significant, the 2SG(.5, .5) method will be used, i.e., data from the two samples will be pooled together and will be treated as a SG design. Otherwise, if the equating difference

is significant, the 2SG(1, 1) method will be used, i.e., the data in a CB design will be treated as an EG design. These two methods may not be as accurate as the RMSE method, but they are simpler to be carried out in practice. Further study can investigate how consistent the decisions are when using these three methods to select the best equating design.

Finally, the results of this study suggest that the advantage of collecting data using a CB design over an EG design appears only when the magnitude of DOE is small. When DOE is within a small range, data from the two groups can be pooled together using different weights to reduce the overall equating error. However, when DOE is large, information from tests taken second will make no contribution to improve the overall equating accuracy. On the other hand, this study alerts us to the importance of implementing random sampling and random assignment in a CB design.

5.3 Limitations of This Study

One concern about real data 2 is that test X and test Y has different test-retest reliabilities, e.g., $r_{(X_1,Y_2)}$ =0.64, $r_{(X_2,Y_1)}$=0.74. Effort was made to enhance the reliability of test X and to make it equal to the reliability of test Y. One way was to remove items on test X that had low correlation with test score of Y_2. This purpose has not been achieved successfully. It turned out that the reliability of test Y increased by a similar amount as the reliability of test X increased. As a result, the equatings were conducted to real data 2 disregarding the issue of unequal reliabilities.

The average equating bias reported in this study also has its disadvantages. That is, when averaging all the conditional equating differences, the negative bias at individual

score levels will cancel out the positive bias at each individual raw-score level.

5.3.1 Arbitrary Nature of the Equating Criterion

In this study, the equating criterion for each population was selected to be the results of traditional equipercentile equating. It might be interesting to regard the results of an IRT-based equating method as the equating criterion for each population. However, this will not make too much change to the patterns of the equating differences between different methods from the author's point of view since Lord and Wingersky (1984) found the IRT true score equating and equipercentile observed score equating yields almost indistinguishable results using a sample of size around 3000.

5.3.2 Problem with Simulated Data

Besides the 3PL IRT model, the one parameter IRT model and two parameter IRT model were also applied to simulate data in this study. Comparing to the 1PL or 2PL model, the distributions of data simulated by using the 3PL model better represent the distributions of real data 1 in terms of the minimum observed score level, the mean scores, the skewness and the kurtosis statistics. Although efforts were made to make the simulated data as close as possible to a real dataset, like many simulation studies, it is unsure to what extent that the simulated data represents real order effects in a real CB design.

5.4 Future Study

The 95% confidence interval in the current SEED plot is two times of the conditional standard error of equating difference at each raw score level, which indicates

that the current SEED plot conduct independent t-test at each score level to examine the significance of equating difference. One drawback of the current SEED plot is that it does not control the family-wise error rate. Since the error rate at each score level is 0.05, the overall error rate across the whole score scale must be larger than 0.05. When the attention is on the equating difference at a particular cut score or within a small score range, it is fine to apply the ± 2SEED confidence interval at each score level. Nevertheless, when it is needed to make a statement on the overall equating differences across the whole score scale, a multivariate global test will need to take into account the dependency among each score point and to control for the family-wise error rate. Future study can explore how to develop such an overall test for the significance of global equating difference between two equating methods.

APPENDICES

TABLE A1. *Standard error of linear equating for real data 1*

Data1	2SG KE					SG	EG
x	2SG (.5,.5)	2SG (.5,.75)	2SG (.75,.5)	2SG (.75,.75)	2SG (1, 1)	Traditional Linear	Traditional Linear
0	1.334	1.634	1.776	2.433	4.674	1.334	4.648
1	1.311	1.607	1.745	2.393	4.533	1.311	4.575
2	1.288	1.579	1.715	2.354	4.405	1.288	4.503
3	1.265	1.552	1.685	2.315	4.355	1.265	4.431
4	1.242	1.525	1.655	2.276	4.326	1.242	4.359
5	1.219	1.497	1.624	2.236	4.287	1.218	4.287
6	1.196	1.47	1.594	2.197	4.215	1.195	4.215
7	1.173	1.443	1.564	2.158	4.144	1.172	4.143
8	1.15	1.415	1.534	2.12	4.072	1.149	4.071
9	1.127	1.388	1.504	2.081	4	1.127	3.999
10	1.104	1.361	1.474	2.042	3.928	1.104	3.928
11	1.081	1.334	1.444	2.003	3.857	1.081	3.856
12	1.058	1.307	1.414	1.965	3.786	1.058	3.785
13	1.035	1.28	1.385	1.926	3.715	1.035	3.714
14	1.013	1.253	1.355	1.888	3.643	1.013	3.643
15	0.99	1.227	1.325	1.849	3.573	0.99	3.572
16	0.967	1.2	1.296	1.811	3.502	0.967	3.501
17	0.945	1.174	1.266	1.773	3.431	0.945	3.431
18	0.923	1.147	1.237	1.735	3.361	0.922	3.361
19	0.9	1.121	1.208	1.697	3.291	0.9	3.291
20	0.878	1.095	1.178	1.66	3.221	0.878	3.221
21	0.856	1.068	1.149	1.622	3.151	0.856	3.151
22	0.834	1.042	1.12	1.585	3.082	0.834	3.082
23	0.812	1.017	1.092	1.548	3.013	0.812	3.012
24	0.79	0.991	1.063	1.511	2.944	0.79	2.943
25	0.768	0.965	1.034	1.474	2.875	0.768	2.875
26	0.746	0.94	1.006	1.437	2.807	0.746	2.807
27	0.725	0.915	0.978	1.401	2.739	0.725	2.739
28	0.703	0.89	0.95	1.365	2.671	0.703	2.671
29	0.682	0.865	0.922	1.329	2.604	0.682	2.604
30	0.661	0.841	0.894	1.293	2.537	0.661	2.537
31	0.64	0.816	0.867	1.258	2.471	0.64	2.471
32	0.62	0.793	0.84	1.223	2.405	0.62	2.405
33	0.6	0.769	0.813	1.189	2.34	0.6	2.34
34	0.58	0.746	0.786	1.155	2.275	0.58	2.275
35	0.56	0.723	0.76	1.121	2.211	0.56	2.211
36	0.54	0.7	0.735	1.088	2.148	0.54	2.148

TABLE A1. *Continued*

Data1	2SG KE					SG	EG
x	2SG (.5,.5)	2SG (.5,.75)	2SG (.75,.5)	2SG (.75,.75)	2SG (1, 1)	Traditional Linear	Traditional Linear
37	0.521	0.678	0.709	1.056	2.086	0.521	2.086
38	0.503	0.657	0.685	1.024	2.024	0.503	2.024
39	0.485	0.636	0.661	0.992	1.963	0.485	1.963
40	0.467	0.616	0.637	0.962	1.903	0.467	1.903
41	0.45	0.596	0.614	0.932	1.845	0.45	1.845
42	0.433	0.577	0.592	0.903	1.787	0.433	1.787
43	0.418	0.559	0.571	0.875	1.731	0.418	1.731
44	0.403	0.542	0.55	0.848	1.677	0.403	1.677
45	0.389	0.526	0.531	0.822	1.623	0.389	1.623
46	0.376	0.511	0.513	0.797	1.572	0.376	1.572
47	0.364	0.497	0.496	0.774	1.523	0.364	1.523
48	0.353	0.484	0.481	0.752	1.476	0.353	1.476
49	0.344	0.473	0.467	0.732	1.431	0.344	1.431
50	0.336	0.463	0.455	0.714	1.389	0.336	1.389
51	0.33	0.455	0.445	0.697	1.349	0.33	1.349
52	0.325	0.449	0.437	0.683	1.313	0.325	1.313
53	0.322	0.444	0.431	0.671	1.28	0.322	1.28
54	0.32	0.441	0.427	0.661	1.251	0.32	1.251
55	0.321	0.44	0.425	0.653	1.225	0.321	1.225
56	0.323	0.441	0.426	0.648	1.204	0.323	1.204
57	0.327	0.444	0.429	0.646	1.187	0.327	1.187
58	0.333	0.448	0.434	0.646	1.175	0.333	1.175
59	0.34	0.455	0.442	0.649	1.167	0.34	1.167
60	0.349	0.463	0.451	0.654	1.164	0.349	1.164
61	0.359	0.472	0.463	0.662	1.166	0.359	1.166
62	0.37	0.483	0.476	0.672	1.172	0.37	1.172
63	0.383	0.496	0.491	0.684	1.183	0.383	1.183
64	0.397	0.509	0.507	0.699	1.199	0.396	1.199
65	0.411	0.524	0.525	0.715	1.22	0.411	1.219
66	0.426	0.54	0.543	0.734	1.244	0.426	1.244
67	0.443	0.557	0.563	0.754	1.272	0.442	1.272
68	0.459	0.575	0.584	0.776	1.304	0.459	1.304
69	0.477	0.594	0.606	0.799	1.34	0.477	1.34
70	0.495	0.614	0.629	0.824	1.378	0.495	1.379
71	0.513	0.634	0.652	0.85	1.42	0.513	1.42
72	0.532	0.655	0.676	0.877	1.464	0.532	1.464
73	0.551	0.676	0.701	0.905	1.511	0.551	1.511
74	0.571	0.698	0.726	0.934	1.559	0.571	1.56
75	0.591	0.72	0.751	0.964	1.61	0.591	1.61

TABLE A2. *Standard error of equipercentile equating for real data 1*

Data1	2SG KE					SG	EG
x	2SG (.5,.5)	2SG (.5,.75)	2SG (.75,.5)	2SG (.75,.75)	2SG (1, 1)	Traditional Equipercentile	Traditional Equipercentile
0	1.218	1.269	1.169	1.272	1.778	0	0
1	1.354	1.432	1.349	1.511	2.328	1.159	1.025
2	1.384	1.477	1.409	1.609	2.664	2.318	2.049
3	1.383	1.485	1.428	1.657	2.896	3.478	3.073
4	1.369	1.478	1.43	1.683	3.068	4.512	3.981
5	1.348	1.464	1.424	1.698	3.2	5.325	4.751
6	1.324	1.445	1.414	1.707	3.303	5.928	5.341
7	1.298	1.424	1.4	1.712	3.383	6.335	5.791
8	1.27	1.401	1.385	1.714	3.444	6.65	6.231
9	1.241	1.378	1.369	1.714	3.49	6.724	6.366
10	1.212	1.353	1.352	1.712	3.523	6.774	6.485
11	1.182	1.328	1.335	1.708	3.545	6.777	6.551
12	1.152	1.303	1.317	1.703	3.556	6.762	6.582
13	1.122	1.278	1.298	1.697	3.557	6.755	6.622
14	1.092	1.252	1.28	1.689	3.551	6.747	6.673
15	1.063	1.227	1.26	1.679	3.536	6.758	6.744
16	1.033	1.201	1.241	1.668	3.515	6.778	6.821
17	1.004	1.176	1.221	1.655	3.486	5.972	6.441
18	0.975	1.15	1.2	1.64	3.451	5.548	6.292
19	0.947	1.125	1.179	1.624	3.41	5.347	6.214
20	0.919	1.1	1.158	1.606	3.363	3.977	5.787
21	0.892	1.075	1.136	1.586	3.311	3.277	5.531
22	0.865	1.05	1.113	1.564	3.254	1.957	5.042
23	0.839	1.026	1.09	1.541	3.193	1.516	4.665
24	0.813	1.001	1.067	1.516	3.127	1.284	4.371
25	0.788	0.977	1.043	1.49	3.058	1.018	4.131
26	0.763	0.953	1.018	1.462	2.987	0.866	3.697
27	0.739	0.929	0.993	1.433	2.912	0.801	3.397
28	0.715	0.905	0.968	1.403	2.837	0.786	3.201
29	0.692	0.881	0.942	1.371	2.759	1.121	3.049
30	0.669	0.858	0.916	1.339	2.681	1.197	2.39
31	0.646	0.834	0.889	1.305	2.603	1.267	2.266
32	0.624	0.811	0.862	1.271	2.524	1.201	2.194
33	0.603	0.788	0.836	1.237	2.446	1.027	2.283
34	0.582	0.765	0.809	1.202	2.368	0.762	2.493
35	0.561	0.742	0.782	1.167	2.292	0.835	2.671
36	0.541	0.72	0.755	1.131	2.217	1.085	2.757
37	0.521	0.697	0.728	1.096	2.143	1.344	2.797
38	0.502	0.675	0.702	1.062	2.071	1.305	2.848
39	0.483	0.654	0.676	1.028	2.001	1.279	2.884

TABLE A2. *Continued*

Data1	2SG KE					SG	EG
x	2SG (.5,.5)	2SG (.5,.75)	2SG (.75,.5)	2SG (.75,.75)	2SG (1, 1)	Traditional Equipercentile	Traditional Equipercentile
40	0.465	0.633	0.651	0.995	1.934	1.071	3.091
41	0.448	0.613	0.626	0.962	1.869	0.849	3.152
42	0.431	0.594	0.603	0.931	1.807	0.875	2.997
43	0.415	0.575	0.58	0.902	1.748	1.047	2.818
44	0.401	0.558	0.559	0.873	1.693	1.008	2.715
45	0.387	0.542	0.539	0.847	1.641	0.929	2.726
46	0.374	0.527	0.52	0.823	1.593	0.772	2.568
47	0.363	0.513	0.504	0.801	1.548	0.831	2.431
48	0.353	0.502	0.489	0.781	1.508	0.896	2.241
49	0.345	0.492	0.477	0.764	1.472	0.764	2.024
50	0.338	0.484	0.467	0.749	1.441	0.687	1.94
51	0.334	0.478	0.46	0.737	1.413	0.75	1.983
52	0.332	0.474	0.455	0.727	1.389	0.934	1.989
53	0.332	0.473	0.454	0.721	1.369	0.988	1.907
54	0.334	0.473	0.454	0.717	1.353	0.745	1.832
55	0.338	0.476	0.458	0.715	1.339	0.619	1.626
56	0.344	0.48	0.464	0.715	1.328	0.59	1.442
57	0.352	0.487	0.472	0.718	1.319	0.574	1.353
58	0.362	0.495	0.482	0.722	1.312	0.539	1.322
59	0.374	0.504	0.494	0.727	1.306	0.541	1.276
60	0.387	0.515	0.508	0.734	1.299	0.5	1.242
61	0.401	0.526	0.522	0.741	1.293	0.534	1.308
62	0.416	0.538	0.538	0.748	1.285	0.623	1.442
63	0.432	0.551	0.554	0.755	1.276	0.86	1.501
64	0.448	0.563	0.569	0.761	1.265	0.928	1.51
65	0.464	0.575	0.585	0.767	1.251	0.984	1.471
66	0.479	0.586	0.599	0.771	1.234	0.682	1.4
67	0.494	0.596	0.613	0.772	1.212	0.486	1.276
68	0.508	0.604	0.624	0.771	1.185	0.613	1.177
69	0.519	0.608	0.632	0.766	1.152	0.745	1.23
70	0.527	0.609	0.635	0.755	1.109	0.945	1.032
71	0.53	0.603	0.632	0.736	1.054	0.859	1.051
72	0.524	0.585	0.618	0.704	0.98	0.58	1.142
73	0.502	0.549	0.585	0.648	0.876	0.88	1.353
74	0.452	0.479	0.516	0.553	0.717	1.326	1.576
75	0.38	0.385	0.419	0.429	0.491	0	0

TABLE A3. *Standard error of linear equating for real data 2*

Data1	2SG KE					SG	EG
x	2SG (.5,.5)	2SG (.5,.75)	2SG (.75,.5)	2SG (.75,.75)	2SG (1, 1)	Traditional Linear	Traditional Linear
0	0.341	0.354	0.403	0.454	0.767	0.341	0.767
1	0.318	0.331	0.377	0.425	0.718	0.318	0.718
2	0.296	0.309	0.352	0.397	0.67	0.296	0.67
3	0.274	0.287	0.327	0.37	0.623	0.274	0.623
4	0.252	0.265	0.302	0.343	0.577	0.252	0.577
5	0.231	0.244	0.278	0.317	0.533	0.231	0.533
6	0.211	0.224	0.255	0.293	0.491	0.211	0.491
7	0.191	0.205	0.233	0.27	0.452	0.191	0.453
8	0.173	0.187	0.212	0.249	0.417	0.173	0.417
9	0.156	0.172	0.193	0.23	0.386	0.156	0.387
10	0.141	0.158	0.176	0.215	0.362	0.141	0.362
11	0.129	0.148	0.162	0.203	0.344	0.129	0.344
12	0.121	0.141	0.151	0.195	0.334	0.121	0.334
13	0.117	0.138	0.146	0.193	0.333	0.117	0.333
14	0.118	0.14	0.145	0.196	0.341	0.118	0.341
15	0.124	0.147	0.15	0.204	0.357	0.124	0.357
16	0.134	0.157	0.16	0.216	0.381	0.134	0.381
17	0.147	0.17	0.173	0.232	0.411	0.147	0.411
18	0.163	0.185	0.189	0.251	0.445	0.163	0.445
19	0.18	0.203	0.208	0.272	0.483	0.18	0.483
20	0.199	0.222	0.229	0.295	0.525	0.199	0.525
21	0.219	0.242	0.251	0.32	0.568	0.219	0.568
22	0.24	0.262	0.274	0.346	0.613	0.24	0.613
23	0.261	0.284	0.298	0.373	0.66	0.261	0.66
24	0.283	0.306	0.323	0.4	0.708	0.283	0.708
25	0.305	0.328	0.347	0.428	0.757	0.305	0.757

TABLE A4. *Standard error of equipercentile equating for real data 2*

Data1	2SG KE					SG	EG
x	2SG (.5,.5)	2SG (.5,.75)	2SG (.75,.5)	2SG (.75,.75)	2SG (1, 1)	Traditional Equipercentile	Traditional Equipercentile
0	0.484	0.548	0.399	0.456	0.72	0	0
1	0.469	0.532	0.393	0.463	0.67	0.711	0.785
2	0.448	0.498	0.393	0.451	0.624	0.827	0.916
3	0.393	0.433	0.362	0.411	0.575	0.845	0.96
4	0.328	0.36	0.318	0.361	0.525	0.448	0.832
5	0.272	0.295	0.277	0.315	0.479	0.353	0.696
6	0.229	0.247	0.244	0.279	0.439	0.239	0.535
7	0.202	0.216	0.222	0.255	0.404	0.286	0.423
8	0.185	0.2	0.207	0.241	0.378	0.268	0.348
9	0.172	0.191	0.198	0.234	0.358	0.201	0.513
10	0.16	0.184	0.189	0.229	0.347	0.196	0.425
11	0.149	0.177	0.18	0.226	0.343	0.198	0.334
12	0.139	0.171	0.172	0.224	0.346	0.192	0.422
13	0.134	0.167	0.166	0.224	0.358	0.179	0.454
14	0.134	0.166	0.165	0.227	0.377	0.216	0.374
15	0.139	0.17	0.169	0.235	0.402	0.213	0.532
16	0.149	0.177	0.178	0.245	0.431	0.231	0.5
17	0.161	0.186	0.19	0.257	0.459	0.24	0.662
18	0.176	0.197	0.204	0.269	0.486	0.305	0.598
19	0.195	0.213	0.219	0.282	0.511	0.312	0.74
20	0.222	0.239	0.235	0.297	0.538	0.345	0.883
21	0.259	0.276	0.254	0.316	0.566	0.406	0.952
22	0.307	0.326	0.277	0.34	0.592	0.439	0.769
23	0.354	0.379	0.297	0.362	0.61	0.839	0.262
24	0.375	0.413	0.298	0.365	0.609	0.779	0.131
25	0.374	0.418	0.304	0.351	0.605	0.671	0

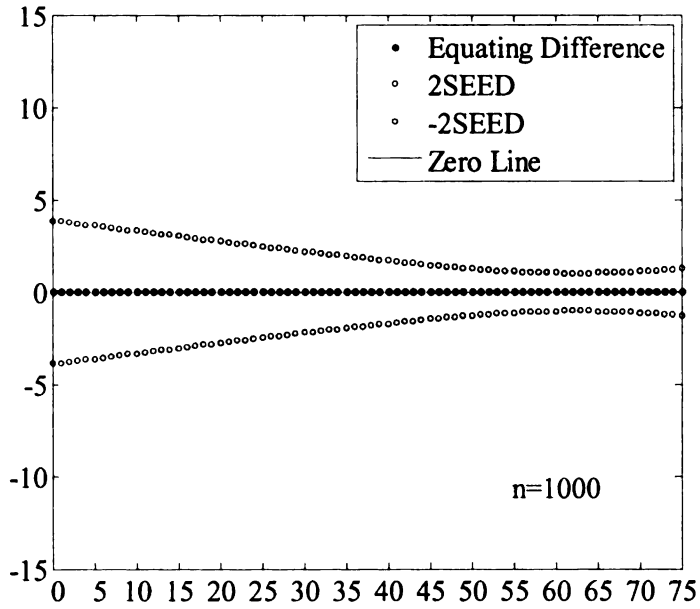


FIGURE A1. *Equating difference between 2SG(1,1) linear and 2SG(.5,.5) linear, POP1, n=1000.*

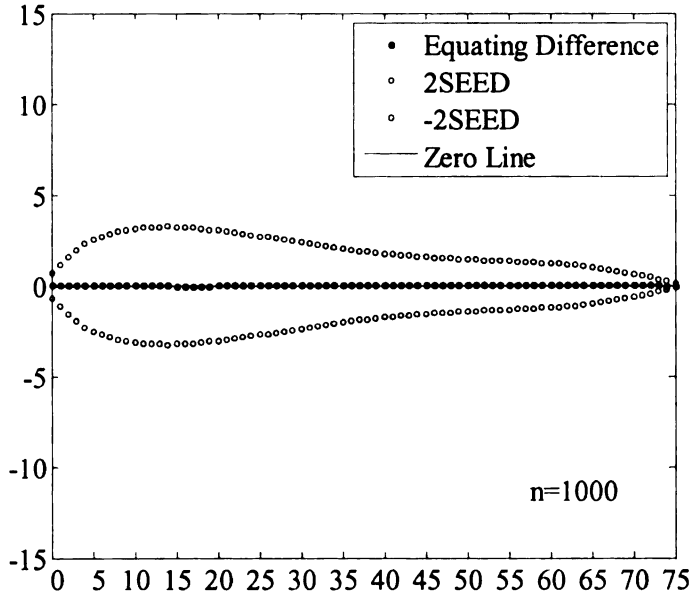


FIGURE A2. *Equating difference between 2SG(1, 1) equipercentile and 2SG(.5, .5) equipercentile, POP1, n=1000.*

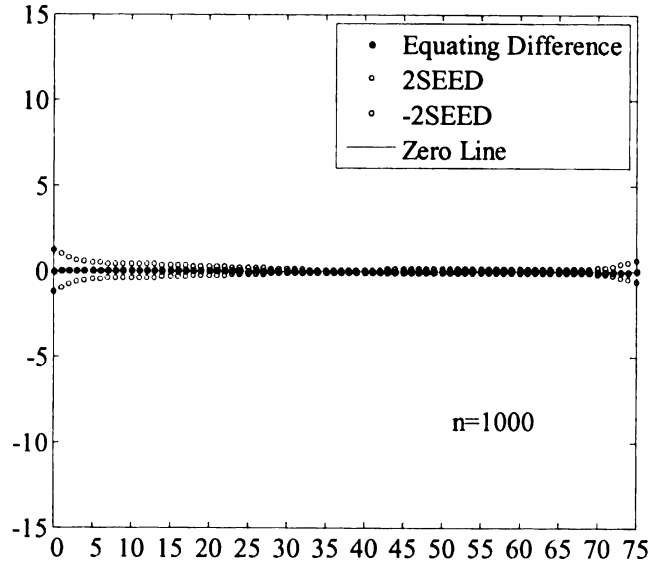


FIGURE A3. *Equating difference between 2SG(.5, .5) linear and 2SG(.5, .5) equipercentile, POP1, n=1000.*

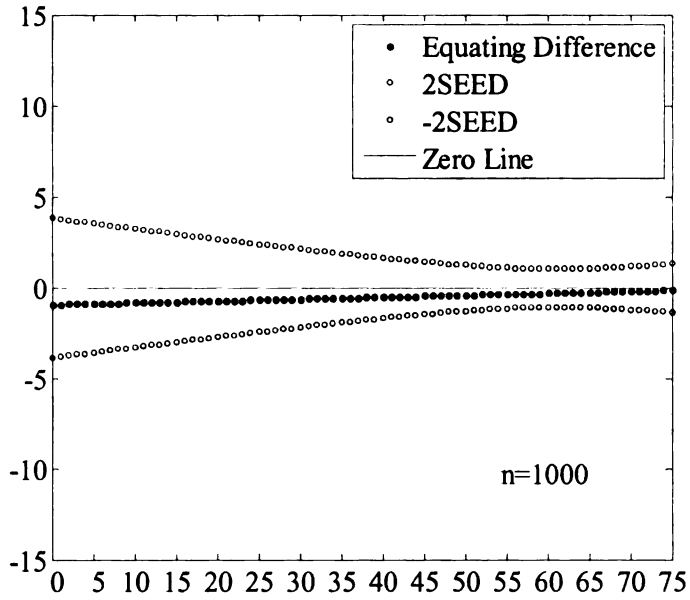


FIGURE A4. Equating difference between 2SG(1,1) linear and 2SG(.5,.5) linear, POP2, n=1000.

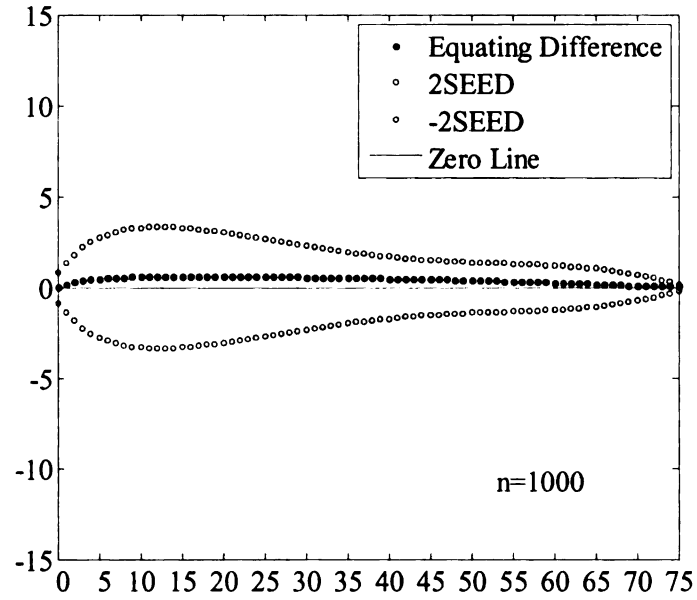


FIGURE A5. Equating difference between 2SG(1, 1) equipercentile and 2SG(.5, .5) equipercentile, POP2, n=1000.

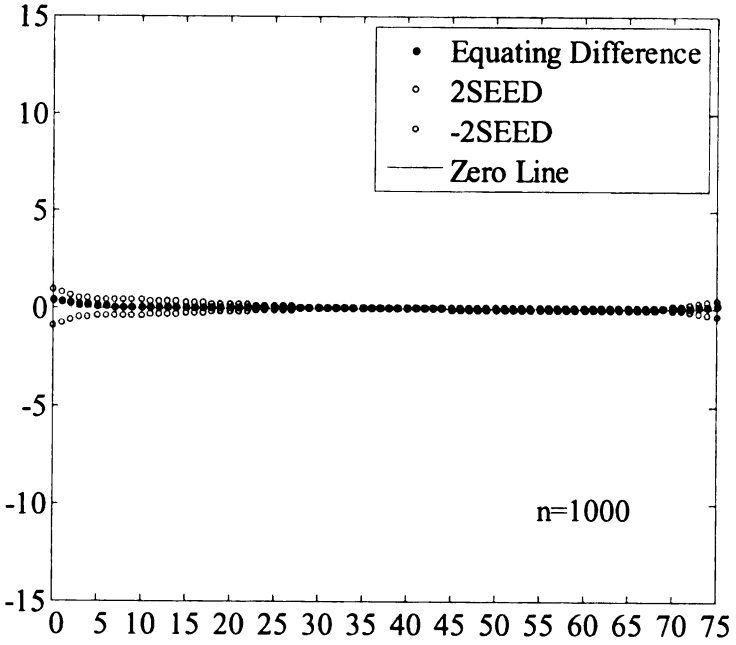


FIGURE A6. *Equating difference between 2SG(.5, .5) linear and 2SG(.5, .5) equipercentile, POP2, n=1000.*

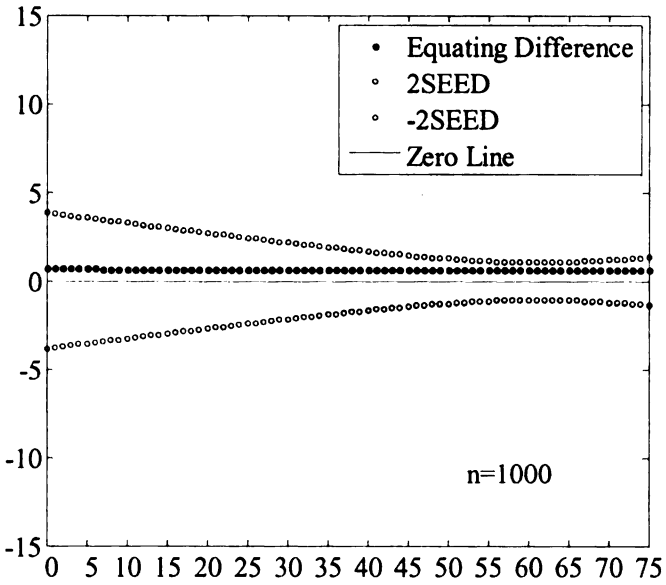


FIGURE A7. Equating difference between 2SG(1,1) linear and 2SG(.5,.5) linear, POP3, n=1000.

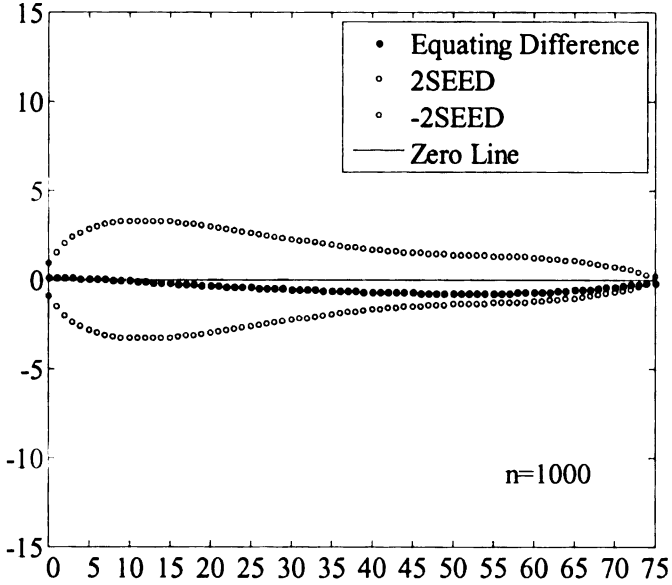


FIGURE A8. Equating difference between 2SG(1, 1) equipercentile and 2SG(.5, .5) equipercentile, POP3, n=1000.

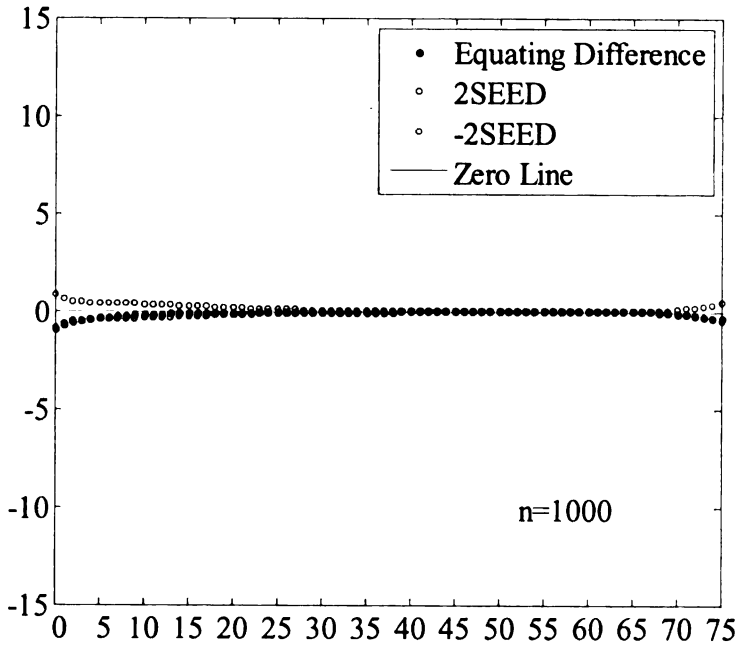


FIGURE A9. *Equating difference between 2SG(.5, .5) linear and 2SG(.5, .5) equipercentile, POP3, n=1000.*

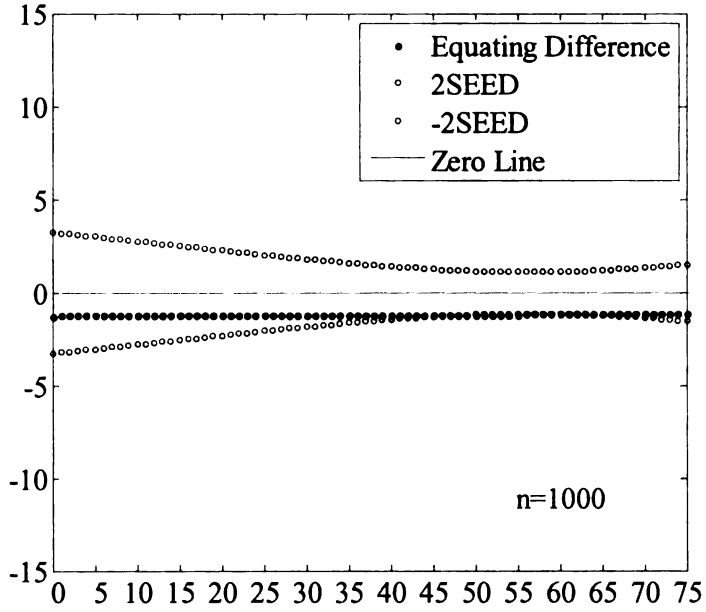


FIGURE A10. Equating difference between 2SG(1,1) linear and 2SG(.5,.5) linear, POP4, n=1000.

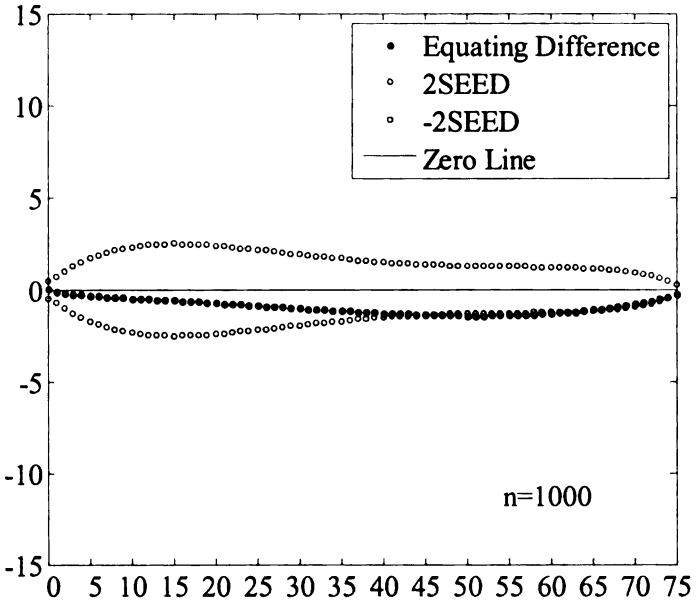


FIGURE A11. Equating difference between 2SG(1, 1) equipercentile and 2SG(.5, .5) equipercentile, POP4, n=1000.

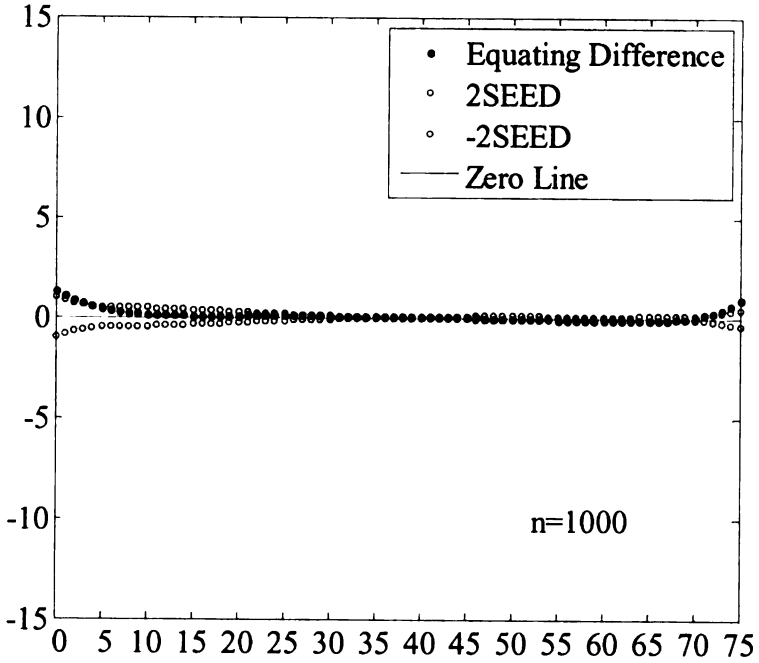


FIGURE A12. *Equating difference between 2SG(.5, .5) linear and 2SG(.5, .5) equipercentile, POP4, n=1000.*

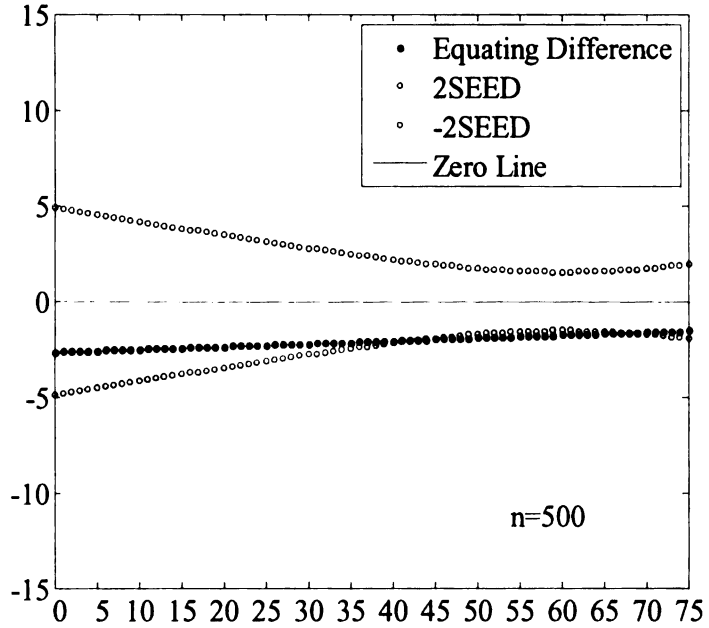


FIGURE A13. Equating difference between 2SG(1,1) linear and 2SG(.5,.5) linear, POP5, n=500.

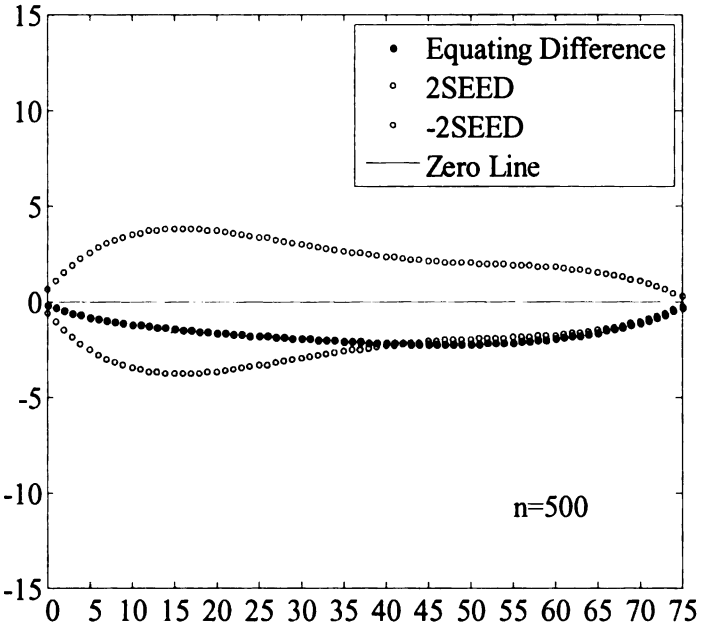


FIGURE A14. Equating difference between 2SG(1, 1) equipercentile and 2SG(.5, .5) equipercentile, POP5, n=500.

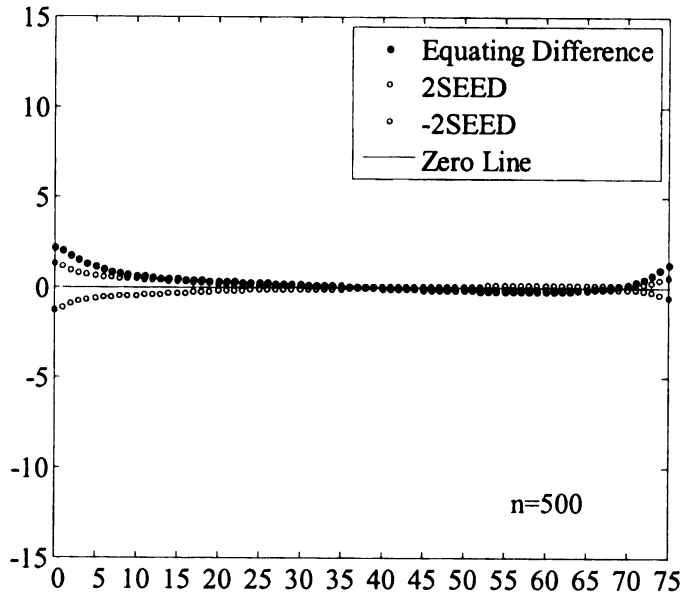


FIGURE A15. *Equating difference between 2SG(.5, .5) linear and 2SG(.5, .5) equipercentile, POP5, n=500.*

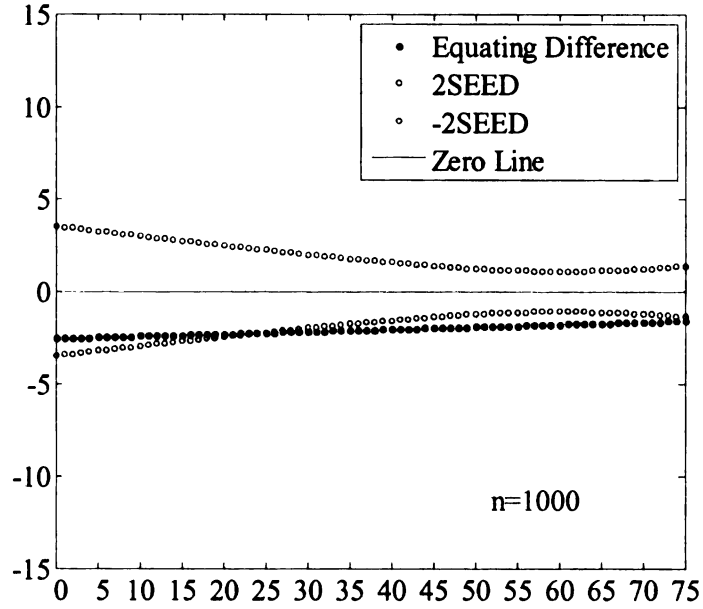


FIGURE A16. Equating difference between 2SG(1,1) linear and 2SG(.5,.5) linear, POP5, n=1000.

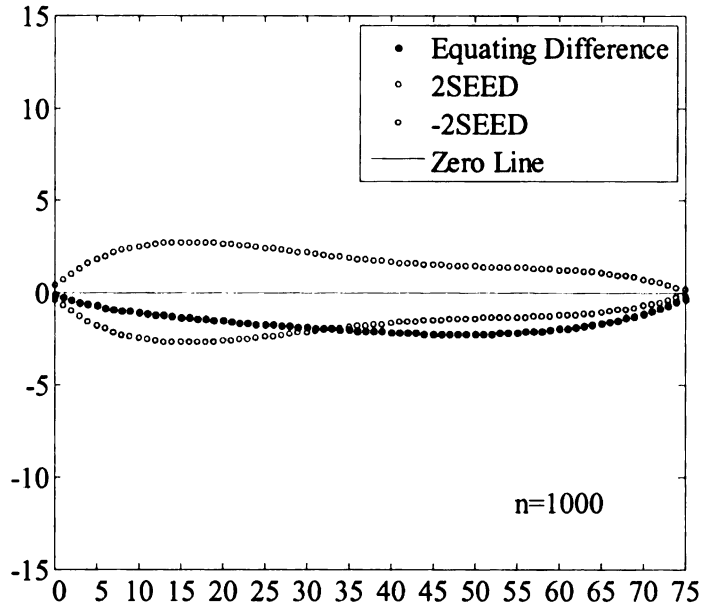


FIGURE A17. Equating difference between 2SG(1, 1) equipercentile and 2SG(.5, .5) equipercentile, POP5, n=1000.

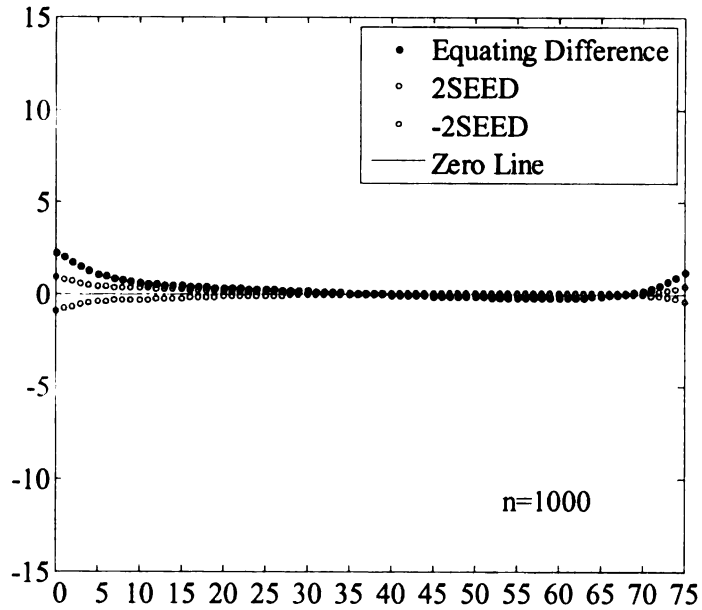


FIGURE A18. *Equating difference between 2SG(.5, .5) linear and 2SG(.5, .5) equipercentile, POP5, n=1000.*

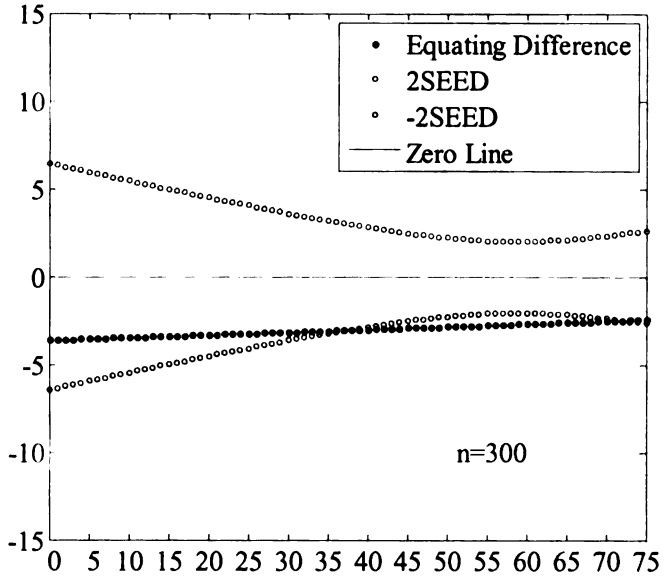


FIGURE A19. Equating difference between 2SG(1, 1) linear and 2SG(.5, .5) linear, POP6, n=300.

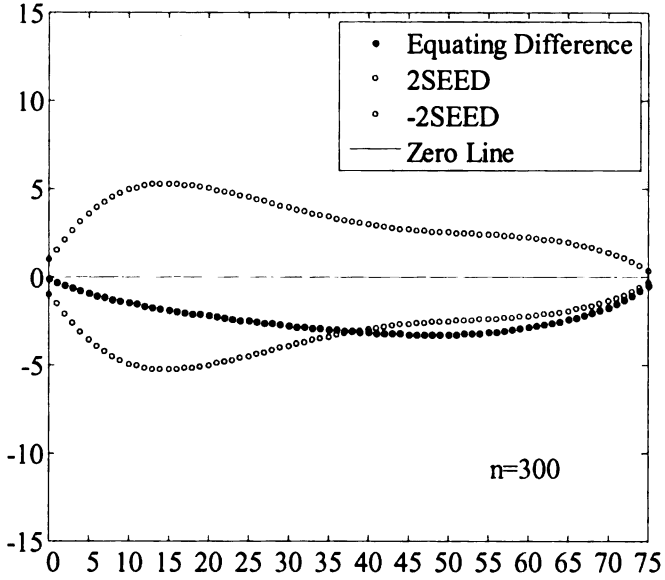


FIGURE A20. Equating difference between 2SG(1, 1) equipercentile and 2SG(.5, .5) equipercentile, POP6, n=300.

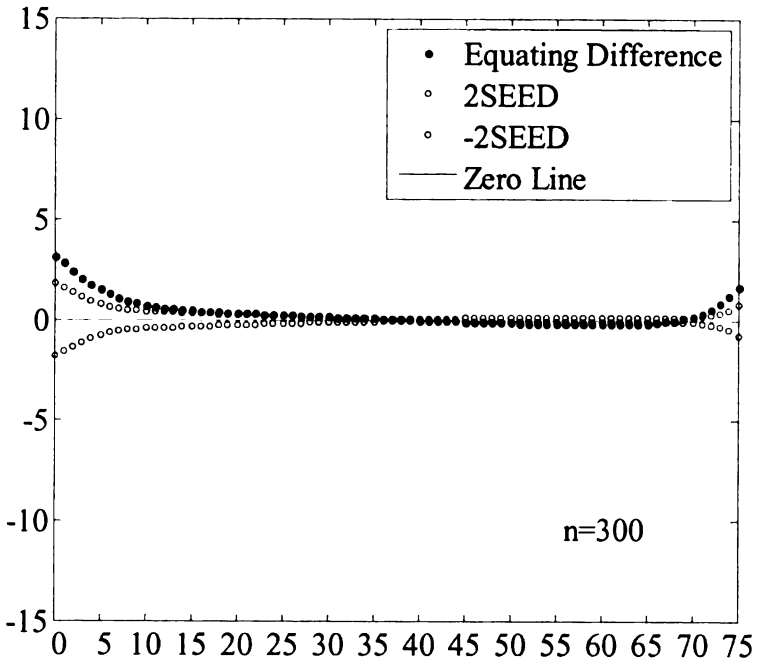


FIGURE A21. *Equating difference between 2SG(.5, .5) linear and 2SG(.5, .5) equipercentile, POP6, n=300.*

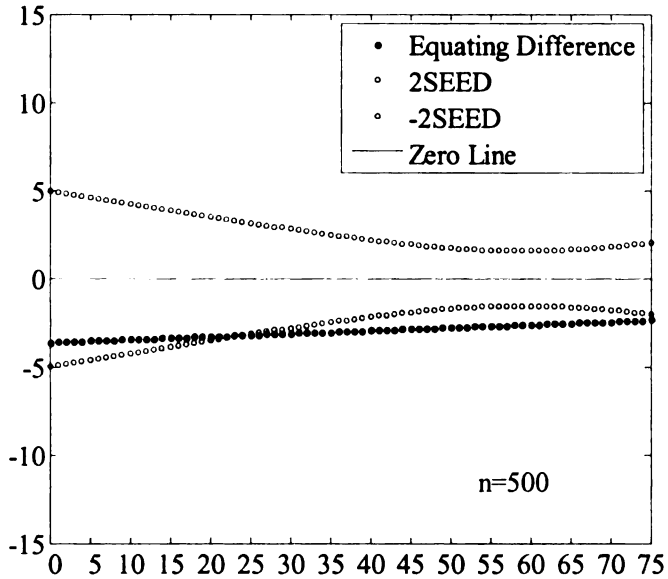


FIGURE A22. Equating difference between 2SG(1, 1) linear and 2SG(.5, .5) linear, POP6, n=500.

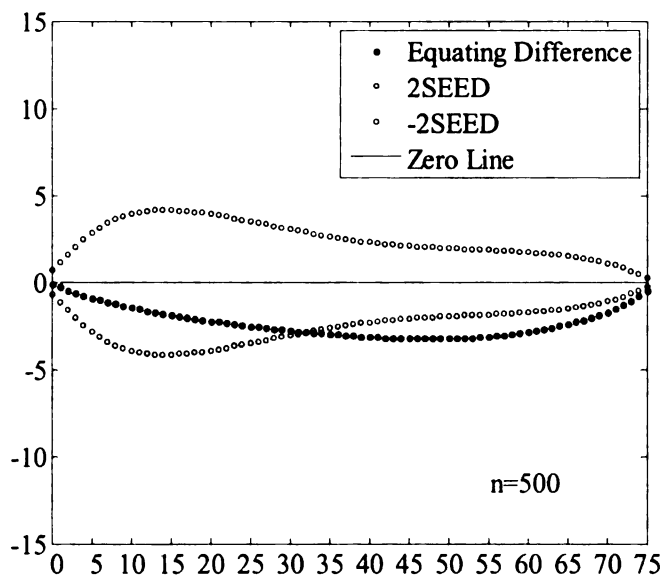


FIGURE A23. Equating difference between 2SG(1, 1) equipercentile and 2SG(.5, .5) equipercentile, POP6, n=500.

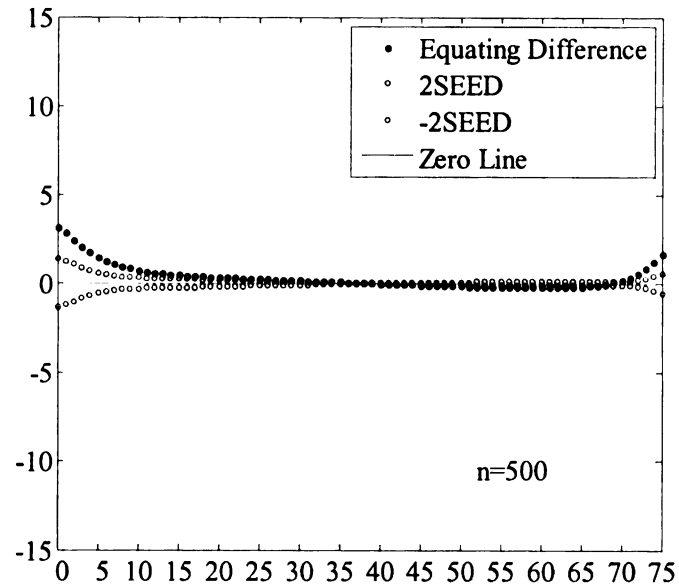


FIGURE A24. *Equating difference between 2SG(.5, .5) linear and 2SG(.5, .5) equipercentile, POP6, n=500.*

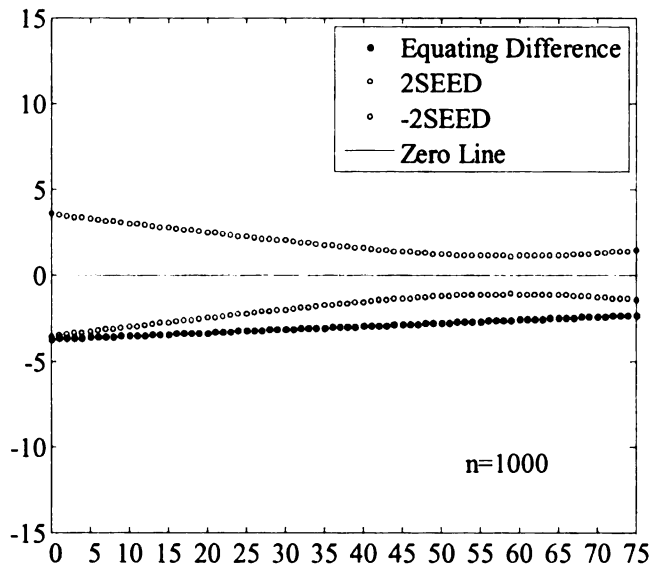


FIGURE A25. Equating difference between 2SG(1, 1) linear and 2SG(.5, .5) linear, POP6, n=1000.

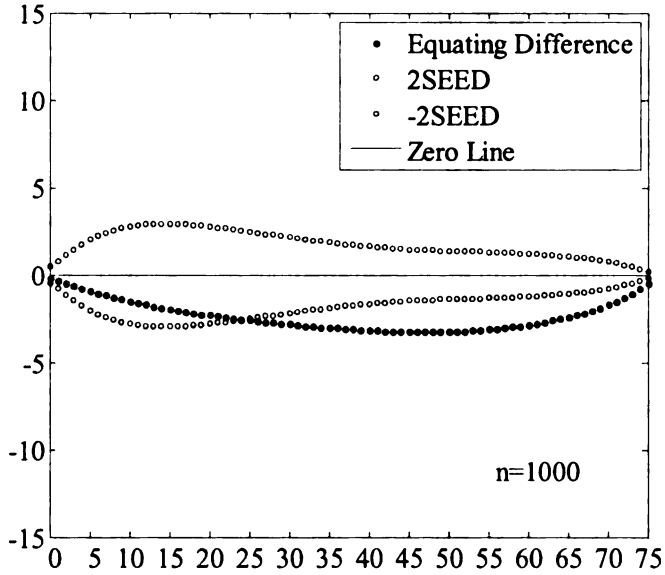


FIGURE A26. Equating difference between 2SG(1, 1) equipercentile and 2SG(.5, .5) equipercentile, POP6, n=1000.

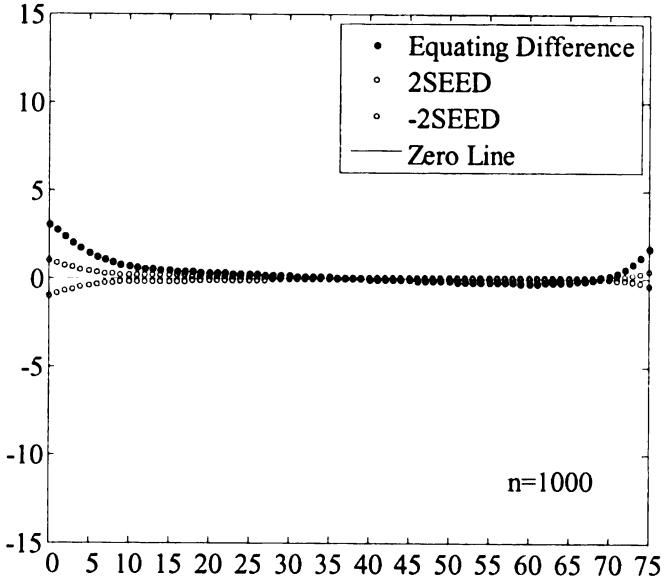


FIGURE A27. *Equating difference between 2SG(.5, .5) linear and 2SG(.5, .5) equipercentile, POP6, n=1000.*

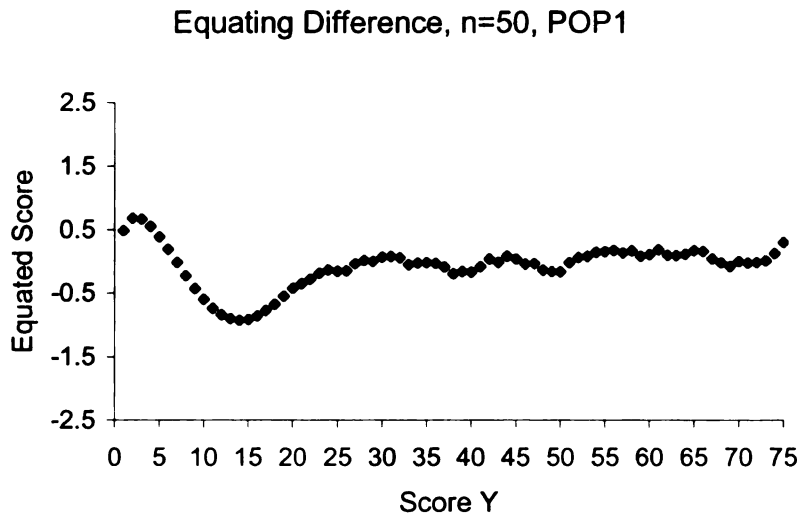


FIGURE A28. *Equating difference between 2SG(1, 1) equipercentile and EG equipercentile, POP1, n=50.*

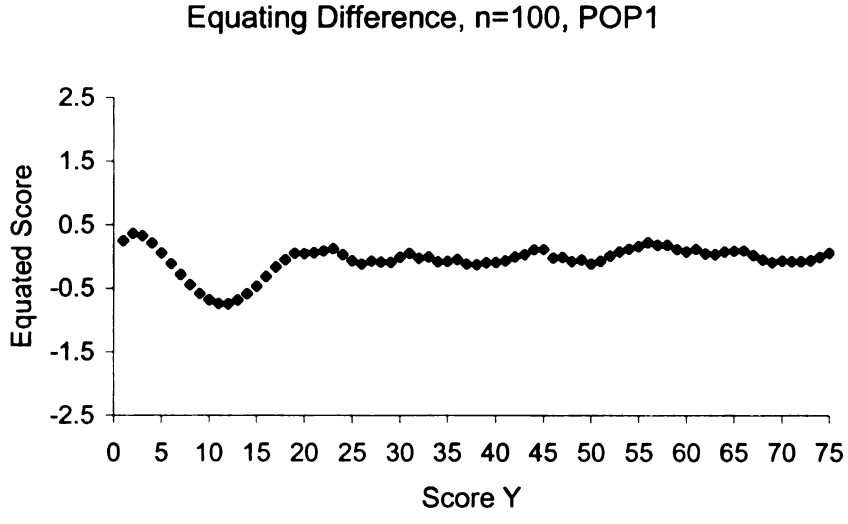


FIGURE A29. *Equating difference between 2SG(1, 1) equipercentile and EG equipercentile, POP1, n=100.*

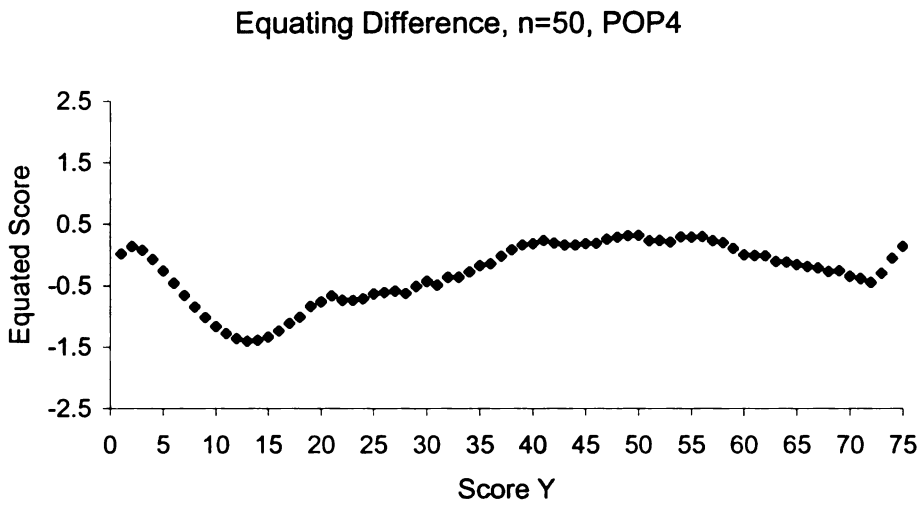


FIGURE A30. *Equating difference between 2SG(1, 1) equipercentile and EG equipercentile, POP4, n=50.*

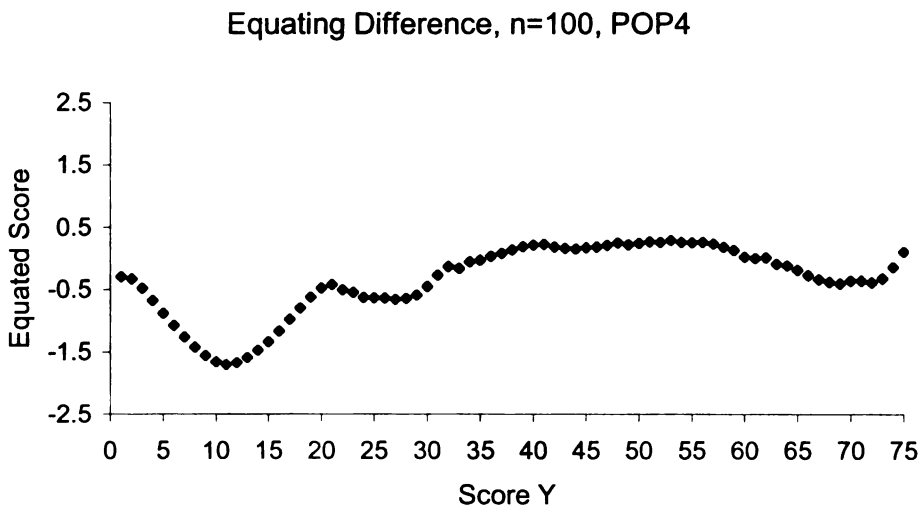


FIGURE A31. *Equating difference between 2SG(1, 1) equipercentile and EG equipercentile, POP4, n=100.*

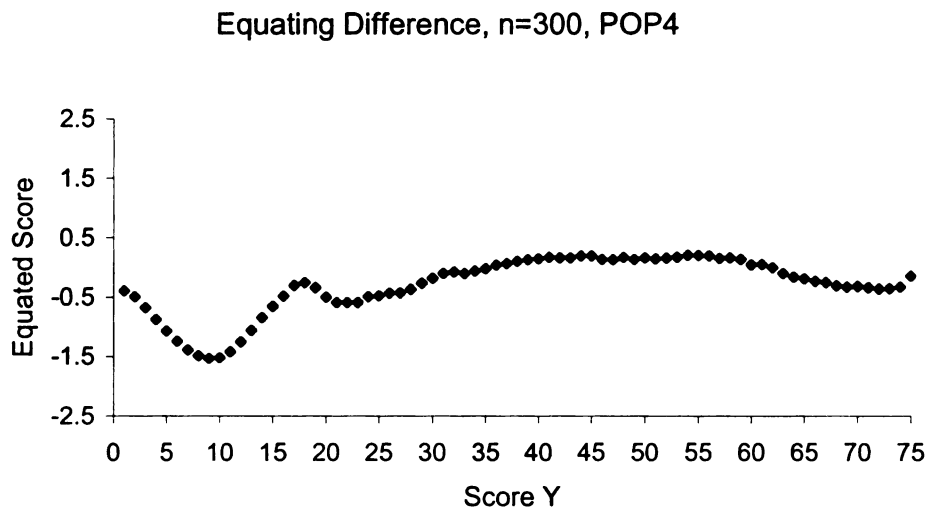


FIGURE A32. *Equating difference between 2SG(1, 1) equipercentile and EG equipercentile, POP4, n=300.*

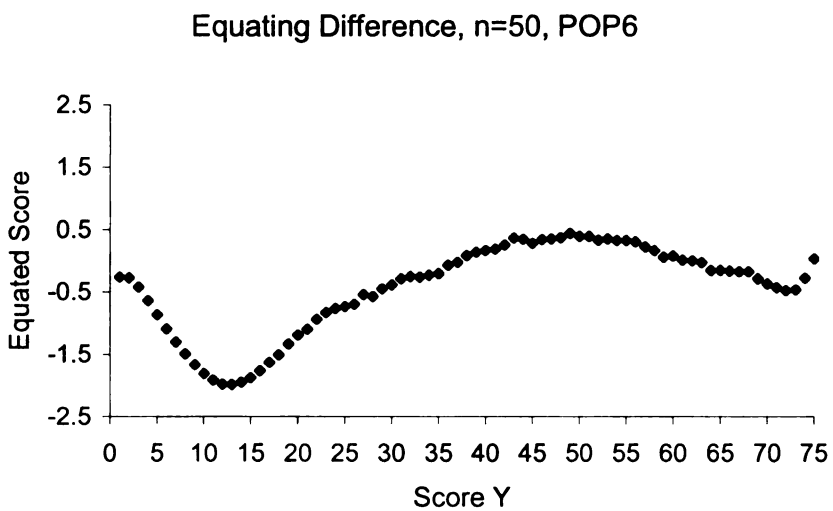


FIGURE A33. *Equating difference between 2SG(1, 1) equipercentile and EG equipercentile, POP6, n=50.*

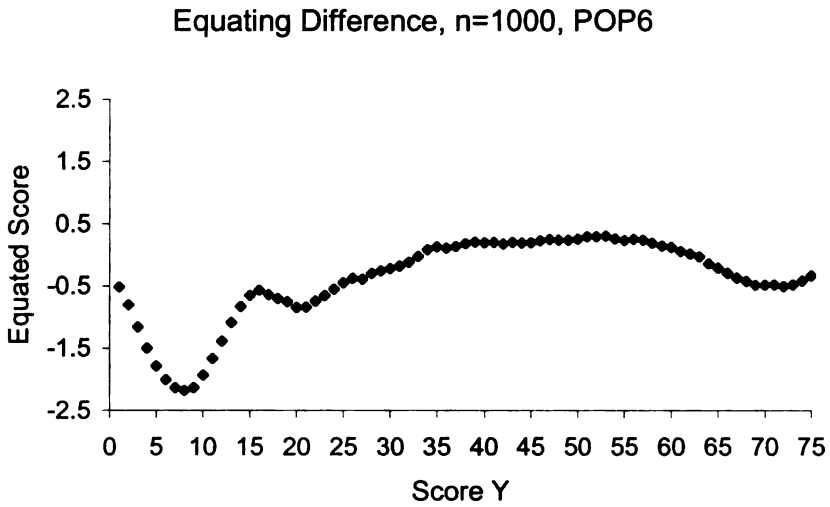


FIGURE A34. *Equating difference between 2SG(1, 1) equipercentile and EG equipercentile, POP6, n=1000.*

REFERENCES

Angoff, W.H. (1971). Scales, norms, and equivalent scores. In R.L. Thorndike (Ed)., *Educational Measurement* (2nd ed., pp. 508-600). Washington, DC: American Council on Education. (Reprinted as W.H. Angoff, *Scales, Norms, and Equivalent Scores*. Princeton, NJ: Educational Testing Service, 1984).

Cochran, W.G., & Cox G.M. (1957). *Experimental Designs* (2nd Ed.), New York: Wiley.

Compaq Visual Fortan 6.5. (2000). Compaq Computer Corporation.

Cook, L.L., & Eignor, D.R. (1991). IRT Equating Methods. *Educational Measurement: Issues and Practice, 10(3),* 37-45.

Davey, T., Nering, M.L. & Thompson, T. (1997). *Realistic simulation of item response data.* (ACT Research Report Series 97-4). Iowa City, IA: American College Testing.

von Davier, A.A., Holland, P.W., & Thayer, D.T. (2004). *The kernel method of test equating.* New York: Springer Verlag.

von Davier, A.A., Holland, P.W., Livingston, S.A., Casablanca, J., Grant, M.C., & Martin, K. (2005). *An evaluation of the kernel equating method in a non-equivalent groups design with an external anchor-- a special study with pseudo-tests from real test data.* Paper presented at the National Council of Measurement in Education, Montréal, Canada.

von Davier, A.A. & Kong, N. (2005). A unified approach to linear equating for the nonequivalent groups design. *Journal of Educational and Behavioral Statistics, 30*(3), 313-342.

Efron, B. (1982). The jackknife, the bootstrap, and other resampling plans. Philadelphia, PA: Society for Industrial and Applied Mathematics.

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

Han, N., Li, S., & Hambleton, R. K. (2005). *Comparing kernel and IRT equating methods.* Paper presented at the National Council of Measurement in Education, Montréal, Canada.

Hanson, B.A., Zeng, L., & Kolen, M.J. (1993). Standard errors of Levine linear equating. *Applied Psychological Measurement, 17*, 225-237.

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

Holland, P.W., & Thayer, D.T. (1989). *The kernel method of equating score distributions. Program statistics research technical report no. 89-84. Access ERIC: Fulltext* (142 Reports--Evaluative No. ETS-RR-89-7). New Jersey: Educational Testing Service, Princeton, NJ.

Holland, P.W., & Thayer, D.T. (2000). Univariate and bivariate log linear models for discrete test score distributions. *Journal of Educational and Behavioral Statistics, 25*, 133-183.

Holland, P.W., Liu, M., & Thayer, D.T. (2005). *Exploring the population sensitivity of linking functions to differences in test constructs and reliability using the Dorans-Holland measures, kernel equating and data from the last.* Paper presented at the National Council of Measurement in Education, Montréal, Canada.

KE Software (2004). Computer Program. Princeton, NJ: Educational Testing Service.

Klein, L.W., & Jarjoura, D. (1985). The importance of content representation for common-item equating with nonrandom 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*, 1-11.

Kolen, M.J. (1984). Effectiveness of analytic smoothing in equipercentile equating. *Journal of Educational Statistics, 9*, 25-44.

Kolen, M.J., (1988). Traditional equating methodology. *Educational Measurement: Issues and Practice, 7*(4), 29-36.

Kolen, M.J., & Brennan, R.J. (2004). *Test Equating: Methods and Practices* (2nd ed.). New York: Springer.

Liou, M., & Cheng, P.E. (1995). Asymptotic standard error of equipercentile equating. *Journal of Educational and Behavioral Statistics, 20*, 259-286.

Liou, M., Cheng, P.E., & Johnson, E.G. (1997). Standard errors of the kernel equating methods under the common-item design. *Applied Psychological Measurement, 21*(4), 349-369.

Liu, J.H., Allspach, J.R., Feigenbaum, M., Oh, H.J., & Burton, N. (2004). A study of fatigue effects from the new SAT. (Research Report 2004-5 & RR-04-46). New York: College Entrance Examination Board, & Princeton, NJ: Educational Testing Service.

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

Livingston, S.A. (1993a). *An empirical tryout of kernel equating* (142 Reports--Evaluative No. ETS-RR-93-33). New Jersey: Educational Testing Service, Princeton, NJ.

Livingston, S.A. (1993b). Small sample equating with log-linear smoothing. *Journal of Educational Measurement, 30*(1), 23-39.

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

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

Lord, F.M. (1982a). The standard error of equipercentile equating. *Journal of Educational Statistics, 7*, 165-174.

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

Lord, F.M., & Wingersky, M.S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings". *Applied Psychological Measurement, 8*, 452-461.

Lu, S., & Kolen, M.J. (1994). *Bootstrap standard errors and confidence intervals in linear equating.* Paper presents at the annual meeting of the American Educational Research Assosciation, New Orleans.

Mao, X., von Davier, A.A., & Rupp, S. (2005). *Comparisons of the kernel equating method with the traditional equating methods on praxis data.* Paper presented at the National Council of Measurement in Education, Montréal, Canada.

MATLAB version 7.1, (1984-2005). The MathWorks, Inc.

Montogomery D.C. (2000). *Design and analysis of experiments (5^{th} edition).* New York : Wiley.

Moses, T., Yang, W., & Wilson, C. (2005). *Using kernel equating to check the statistical equivalence of nearly identical test editions.* Paper presented at the National Council of Measurement in Education, Montréal, Canada.

Moses, T.P., & von Davier, A.A. (2005). *A SAS macro for log linear smoothing: Applications and implications.* Paper presented at the American Educational Research Association, Montréal, Canada.

Parr, W.C. (1983). A note on the jackknife, the bootstrap and the delta method estimators of bias and variance. *Biometrika, 70, 3,* 719-22.

Parshall, C.G., Houghton, Du Bose P., & Kromrey J.D. (1995). Equating error and statistical bias in small sample linear equating. *Journal of Educational Measurement, 32,* 37-54.

Qu, Y. & Von Davier, A. (2006). *Comparison of two approaches for Counter-Balanced design in a Kernel Equating framework.* Paper presented at the National Council of Measurement in Education, San Francesco, USA.

Rice, J.A. (1988). *Mathematical statistics and data analysis.* Monterey, Calif. : Brooks/Cole.

SAS version 9, (2002). SAS Institute Inc., Cary, NC, USA.

Tsai, T.H. (1998). *A comparison of bootstrap standard errors of IRT equating methods for the common item nonequivalent groups design.* Unpublished Dissertation. Iowa City: University of Iowa.

Yu, L., Anderson, D.O., & Zeller, K. (2003). *Report of the counterbalanced equating study for the Algebra End-Of-Course assessment* (Research report SR – 2003 – 56). Princeton, NJ: Educational Testing Service.

Zeng, L., & Cope, R. (1995). Standard error of linear equating for the counterbalanced design. *Journal of Educational and Behavioral Statistics, 20*(4), 337-348.