The Effects of Anchor Length, Test Difficulty, Population Ability Differences, Mixture of Populations and Sample Size on the Psychometric Properties of Levine Observed Score Linear Equating Method for Different Assumptions

BY

Jorge E. Carvajal-Espinoza

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

____
Chairperson

Committee members* ______________________*

____

____

____

Date defended: __________

The Dissertation Committee for Jorge Carvajal-Espinoza certifies
that this is the approved version of the following dissertation:

The Effects of Anchor Length, Test Difficulty, Population Ability Differences, Mixture of Populations and Sample Size on the Psychometric Properties of Levine Observed Score Linear Equating Method for Different Assumptions

____
Chairperson

Date approved:_________________________

Abstract

The Non-Equivalent groups with Anchor Test equating (NEAT) design is a widely used equating design in large scale testing that involves two groups that do not have to be of equal ability. One group P gets form X and a group of items A and the other group Q gets form Y and the same group of items A. One of the most commonly used equating methods in the NEAT design is the Levine Observed Score method for linear equating. The purpose of this study was to compare two different assumptions for the Levine Observed Score method of linear equating and to establish how accurately these two assumptions recover the true equating function.

These two assumptions were compared using simulated data at synthetic population level and at sample level by manipulating anchor length, differences in ability distribution for populations P and Q, differences in test difficulty, mixture of populations and sample size. The traditional assumption outperformed the alternative assumption in conditions with larger difference in standard deviation for the ability distribution and shorter anchor length.

Acknowledgements

My sincere thanks to the members of my dissertation committee: Dr. Bruce Frey, Dr. Yaozhong Hu, Dr. Neal Kingston, Dr. William Skorupski and Dr. Greg Welch for their willingness to serve in the committee and for their suggestions. Special thanks to my advisor Dr. Skorupski for his unconditional support and encouragement and for the fun of working with him.

Thanks to Dr. Michael Walker, from Educational Testing Service, for his patience with my questions.

To my wife Kemly, for the sacrifice that pursing my degree has meant for her.

To my kids José Ignacio, María Paz, Marta Elena and Mónica, who are my inspiration.

To my dad, who is now being tested by sickness, for his example and support.

To my sister, for her continuous support and prayers.

To my mom, who is no longer physically with us, for her continued love.

TABLE OF CONTENTS

List of Tables

Table Page

Table Page

List of Figures

Figure Page

Figure Page

List of Equations

Chapter One – Introduction

Statement of the Problem

Currently many testing programs administer several versions of a given test. Given the importance of the use of results from standardized tests, it is desirable that their scores provide a fair and equitable measure of the students' abilities. The use of multiple forms ensures that no student has advantage for having previous knowledge about the questions in the test (Cook & Eignor, 1991). Test security reasons also prompt the use of different versions of a standardized test (Suh et al. 2009). Although these versions are built based on the same content and statistical specifications, differences in test difficulty are still likely to occur (Tanguma, 2000).

To address these differences among test forms, statistical techniques have been developed. These techniques, known as equating, aim to make scores from different test versions comparable. As stated by Dorans and Holland (2000, p. 281) "Test equating techniques are those statistical and psychometric methods used to adjust scores obtained on different tests measuring the same construct so that they are comparable." According to Lord (1980), two test forms, X and Y, are considered equated if it is a matter of indifference to test takers of every level of ability which form they take.

Equating techniques can be grouped in Classical Test Theory (CTT) and Item Response Theory (IRT) methods. This dissertation focuses on one of the techniques in CTT known as linear equating and therefore IRT equating methods are not discussed here.

Before an equating technique is applied, it is necessary to have a design for collecting the data. A widely used design in large scale testing involves two groups that do not have to be of equal ability. One group P gets form X and a group of items A and the other group Q gets form Y and the same group of items A. In other words, A is a test of common items. The group of common items measures differences in group ability (Kolen and Brennan, 2004). Anchor test is another name for this group of common items and for this reason von Davier et al. (2004) called this design Non-Equivalent groups with Anchor Test equating (NEAT) design.

The equating under NEAT involves the estimation of two unknown pieces of information: the performance of group P on test Y and the performance of group Q on test X. "The different anchor equating methods are different because they make different assumptions as to which aspects of the statistical relationship will generalize to the target population" (Livingston, 2004, p. 45).

One of the most widely used methods for such estimation is the Levine Observed Score method for linear equating. In the Levine Observed Score method, three assumptions are traditionally made so that unknown information from Test Y for group P and Test X for group Q can be estimated. An alternative assumption to the third of these assumptions has been recently proposed by Holland (2004; see also Holland & Walker, 2006). Using this alternative assumption the formulas for the computation of unknown pieces of information under the NEAT design are different.

Only one study (Carvajal et al. 2008) has compared the performance of this third assumption versus the traditional Levine method. They found that although the estimation from the traditional assumption appeared to be closer to the actual values than the estimation from the alternative assumption, the differences were not of practical significance given that these differences are not expected to result in a change in actual reported scores. However, the only factor considered in that study was the mixture of populations (Braun & Holland, 1982). The effects of anchor length, population ability differences for populations P and Q, test difficulty differences and sample size of the groups taking X and Y were not addressed in that study and replications of the simulation were not conducted.

Therefore, the current study was warranted in order to determine the degree of concordance or discrepancy between the estimation from the traditional assumptions in the Levine method and the estimation based on the alternative assumption. Because this was a simulation study, the true equating function was known, so this study compared this function to the one obtained under the alternative assumption and the one obtained under the traditional assumption (in conjunction with other two traditional assumptions in each case).

Purpose

The purpose of this study was to compare two different sets of assumptions in the Levine Observed Score method of linear equating and to establish how accurately these two sets of assumptions recover the true equating function. The two sets of assumptions share two assumptions:

L1) $X = \tau_X + e_X$, $Y = \tau_Y + e_Y$, and $A = \tau_A + e_A$, along with the assumption that the error terms, $e_{\boldsymbol{X}}$, $e_{\boldsymbol{Y}}$, and $e_{\boldsymbol{A}}$, are uncorrelated with their corresponding true scores, $\tau_{\boldsymbol{X}}$, $\tau_{\boldsymbol{Y}}$, and $\tau_{\boldsymbol{A}}$, where X and Y are the tests to be equated and A is the anchor. Specifically, X denotes the new form, i.e. the X scores will be equated to Y scores.

L2) $\tau_X = a + b\tau_{\mathbf{A}}$, and $\tau_{\boldsymbol{Y}} = c + d\tau_{\boldsymbol{A}}$. i.e. the true scores of X and A and Y and A are linearly related. This is the congenericity assumption.

However, the two sets differ on the third assumption that they make. For the first set, the third assumption is

L3) The error variances $\sigma^2_{e_X/S}$, $\sigma^2_{e_Y/S}$ and $\sigma^2_{e_A/S}$ are the same for any S of the synthetic form $S = wP + (1 - w)Q$, where $0 \leq w \leq 1$ (Braun & Holland, 1982).

On the other hand, the third assumption for the second set is

L3*) The ratios $\frac{\rho_{X/S}}{\rho_{A/S}}$ and $\frac{\rho_{Y/S}}{\rho_{A/S}}$ are constant as functions of *S* of the synthetic form (Holland & Walker, 2006); that is, the ratios of the square roots of the reliabilities $\frac{\rho_{X/S}}{\rho_{A/S}}$ and $\frac{\rho_{Y/S}}{\rho_{A/S}}$ are population invariant.

For performing such a comparison, a simulation was designed for which a population P and four populations Q were combined and crossed with four test conditions: two conditions for anchor length and two conditions for difficulty of the test Y as well as five mixtures of populations. The result was 80 synthetic populations for which the three equating functions were computed and compared: a true equating function, an equation form the traditional assumption L3 and an equation from the alternative assumption L3*. A sampling design from populations P and Q was also implemented. This sampling design will be described in Chapter Three. The variables for this study are indicated next.

Variables

There were several variables in this study. The dependant variable was the equating function of X to Y. The equating function can be computed in three ways: the true equated score and the equated scores obtained under the first and the second set of assumptions.

The independent variables included anchor length, differences in ability distribution for populations P and Q, differences in test difficulty, mixture of populations and sample size.

Research Questions

The research questions addressed in this study were the following:

1) What is the relative performance of L3 and L3* in recovering the true equated score at the synthetic population level?

2) What is the relative performance of L3 and L3* in recovering the true equated score at sample level?

Hypotheses

It was hypothesized that the alternative assumption of the constant square root of reliability ratio recovers the true equated score in a more accurate way than the traditional assumption of constant error variance in the Levine Observed Method of linear equating both at synthetic population level as well as at sample level.

Chapter Two - Literature Review

Currently many testing programs administer several versions of a given test. Given the importance of the use of results from standardized tests, it is desirable that their scores provide a fair and equitable measure of the students' abilities. The use of multiple forms ensures that no student has advantage for having previous knowledge about the questions in the test (Cook & Eignor, 1991). Test security reasons also prompt the use of different versions of a standardized test (Suh et al. 2009). Although these versions are built based on the same content and statistical specifications, differences in test difficulty are still likely to occur (Tanguma, 2000).

What is Equating?

To address these differences among test forms, statistical techniques have been developed. These techniques, known as equating, aim to make scores from different test versions comparable. As stated by Dorans and Holland (2000, p. 281) "Test equating techniques are those statistical and psychometric methods used to adjust scores obtained on different tests measuring the same construct so that they are comparable." According to Lord (1980), two test forms, X and Y, are considered equated if it is a matter of indifference to test takers of every level of ability which form they take.

It is important to emphasize that equating adjusts for differences in difficulty, not for differences in content. A related process to equating is linking or vertical scaling, which involves a comparison of scores from tests for different grade levels; however, because the content from these tests is different, these scores cannot be used interchangeably as in the case of equated scores (Kolen and Brennan, 2004).

Dorans (1990) explains four Lord (1980) conditions that a equating of test X to test Y must meet as follows. The first condition is that X and Y must measure the same construct. For achieving this condition, also known as equatability, the two tests need to be as parallel as possible from a content point of view; "…that is they should contain the same context mix of items." (Dorans, 1990, p. 5). This is precisely what distinguishes equating from scaling. The second condition is equity, which states that it must be a matter of indifference to the test taker whether he or she takes X or Y. Although equatability is a prerequisite for equity, it does not imply equity because two tests that measure the same construct can differ in difficulty. The third property is symmetry, which requires that the one to one relationship between scores from X and Y be the same regardless of whether X is equated to Y or Y is equated to X. This is why a regression does not work as an equating method; the regression X to Y and Y to X does not, in general, produce two functions that are inverse of each other. Finally the fourth property is population invariance, which requires that equating transformations should be unique and identical across subpopulations from the population.

Dorans and Holland (2000) in addition, explain the equal reliability requirement for equating. They say that this concept is important because in an extreme case, a test could be equated to an item, which they consider incorrect. However, they indicate that in practice this requirement can be violated and still result on satisfactory equating. That is why they think this is a secondary consideration but not a fundamental requirement for test equating. They indicate that violations to this requirement can lead to violations of the population invariance requirement and that the concern about equal reliability should be complemented with the question of amount of reliability in the sense that more reliability is better for equating.

Equating techniques can be grouped in Classical Test Theory (CTT) and Item Response Theory (IRT) methods. This dissertation focuses on one of the techniques in CTT known as linear equating and therefore IRT equating methods are not discussed here. In addition, before an equating technique is applied, it is necessary to have a design for collecting the data. So a mention of different data collections is in order before referring to the equating technique used in this study.

Equating Designs

Kolen and Brennan (2004) identify and describe three types of equating designs: random groups design, single group, and common-item nonequivalent groups. The technique used in this study corresponds to the third category, therefore the first two will only be briefly described, and then a more detailed description of the last category is presented.

In the *random group design* the forms of the test are randomly assigned to test takers, which leads to comparable randomly equivalents groups that take form X and form Y and the difference between of performance between groups is attributed to the difference in difficulty between the two versions of the test (Kolen and Brennan, 2004).

In the *single group design* both forms X and Y are given to the same group of examinees. Obviously fatigue and order effect could be disadvantages of this design so it is not used frequently. A variation of this design is the *single group with counterbalancing* in which half of the tests administered have form X first and then form Y and the other half have form Y and then form X. The tests are then assigned alternatively to the examinees so the result is that the group taking X form first is comparable to the group taking Y form first. In this design, the relationship between the forms (X to Y for example) can be used when they are taken first. In addition, if the effect of taking X after taking Y is the same as taking Y after X, the equating relationship will be the same between the forms taken first as between the forms taken second. If this does not happen then there is a differential order effect and the data for the form taken second might not be used, which can produce instability in the equating and waste time from the examinees. An obvious disadvantage of the single group design is that the testing time doubles (Kolen and Brennan, 2004).

The common-item nonequivalent group is a widely used design in large scale testing. It involves two groups that do not have to be of equal ability. One group P gets form X and a group of items A and the other group Q gets form Y and the same group of items A. In other words, A is a test of common items. The group of common items measures differences in group ability (Kolen and Brennan, 2004). Anchor test is another name for this group of common items and for this reason von Davier et al. (2004) called this design Non-Equivalent groups with Anchor Test equating (NEAT) design and this is the term that will be used throughout this study. The anchor test is called *internal* when the scores from test A contribute to the total scores for both tests X and Y and *external* if they don't contribute to those total scores. Table 1 illustrates the NEAT design.

Table 1
Illustration of the NEAT design

	Group P	Group Q
Form X	✓	
Form Y		✓
Anchor A	✓	✓

In the NEAT design, the test A is a shorter and less reliable test that measures the same constructs that X and Y measure and its purpose is quantifying the difference between the two groups that affect their performance in tests X and Y (Holland and Dorans, 2006).

The object of equating techniques in the context of the NEAT design is to estimate how some total group S, which is a weighted combination of groups P and Q, would perform on both tests X and Y. The notion of a *synthetic population* S was proposed by Braun and Holland

(1982), where $S = wP + (1 - w)Q$, and w is greater or equal than 0 and less or equal than 1. The estimation under NEAT involves the estimation of two unknown pieces of information: the performance of group P on test Y and the performance of group Q on test X. Once these estimates are obtained, the performance on tests X and Y can be directly compared in S and the two tests can be thus equated to each other (Kolen and Brennan, 2004).

A Classification of Equating Methods

Depending on the equating design for data collection, different equating techniques can be used. Following, I will focus on the CTT techniques for the NEAT design.

There is a wide variety of CTT equating methods used under the NEAT design. Livingston (2004) presents a classification of NEAT design methods in a table similar to the following.

Table 2
A Classification of Equating Methods

	Chained equating	Conditioning on the anchor
Linear	Chained linear	Tucker Levine
Equipercentile	Chained equipercentile	Frequency estimation equipercentile method

Livingston (2004) explains that there are two ways to use information from the anchor test, the first being chained equating and the second being conditioning on the anchor, as the table shows. In chained equating (CE) the scores on form X are equated to the anchor scores and then the anchor scores are equated to scores on form Y. So by using this chain the scores on X are equated to scores on Y.

A second approach is conditioning on the anchor, which corresponds to the second column of the table. Livingston (2004) states that this approach is also known as poststratification (PSE). The anchor in this case is used as if it were a predictor variable: for each score on the anchor, the distribution (or possibly just the mean and the standard deviation) of scores on X and on Y is estimated in the target population. Such estimates are used for the equating, as if they were observed in the target population.

The two rows in Table 2 present two commonly used types of equating: linear and equipercentile. First a brief description of equipercentile methods is presented and then the focus is set on linear methods, specifically the Levine Observed Score method which is the topic of this study.

The basic idea in equipercentile methods is that a score on form X and a score on form Y are equivalent in a given group if they have the same percentile rank in the group (Livingston, 2004). In this sense, a score on form X is transformed to a score on form Y that has the same percentile rank in that group. Equipercentile equating will make the distribution of the adjusted X scores very similar to the distribution of scores on Y, and because of this the mean and the standard deviation on both distributions will be nearly the same (Livingston, 2004).

The basic concept in linear equating is that a score on X and a score on Y are equivalent in a group of examinees if the two scores are the same number of standard deviations above or below the mean for that group. This implies that in order to equate a score on X to a score on Y, one needs to transform the score on X to a score on Y that is the same number of standard deviations above or below the mean of the group (Livingston, 2004). This concept can be represented by the following formula:

$$\frac{Y - mean(Y)}{SD(Y)} = \frac{X - mean(X)}{SD(X)} \,, \qquad (1)$$

where X and Y are two equivalent scores and the group and the mean and standard deviations are computed for the group of examinees.

The previous equation can be solved for Y to obtain:

$$Y = \left(\frac{SD(Y)}{SD(X)} \right) X + \left[mean(Y) - \left(\frac{SD(Y)}{SD(X)} \right) mean(X) \right] = adjusted\, X \qquad (2)$$

In this way, it is possible to adjust any X score (Livingston, 2004).

If this linear equating basic idea is used in a random group design, single-group design or a counterbalanced design it is assumed that the observed relationship will generalize to the target population. By using the observed means and standard deviations to establish an equating relationship the assumption is not that they are a good estimate of those means and standard deviations in the population but that the equating relationship observed in the samples is a good estimate of the equating relationship in the population (Livingston, 2004).

However, in the design with an anchor, the case is more complex because the information from the anchor is needed to adjust for the differences between the group of examinees taking X and Y. Any method in an anchor design assumes that “something about the statistical relationship between scores on the new form [X] and the anchor in the group that actually took the new form will generalize to the target population –and similarly for the reference form [Y]. The different anchor equating methods are different because they make different assumptions as to which aspects of the statistical relationship will generalize to the target population” (Livingston, 2004, p. 45).

Evaluating the Accuracy of Equating Methods

Holland and Dorans (2006) present some measures of statistical accuracy for equating functions. The standard error of equating (SEE) is defined as the standard deviation of the distribution of the estimated linking function at a particular score. It provides a measure of how accurately the equating function is estimated. The difference that matters (DTM) addresses the fact of whether or not the difference between two equating functions has important consequences for reported scores. This is dependant of the test and its use. For example on the SAT the DTM is 5 reported-score points because SAT scores are reported and rounded in steps of 10 points.

They also present measures for checking the sensitivity of equating functions to the populations where they are estimated. Those constitute population invariance checks. For example, the root mean squared difference (RMSD) between the equating functions for each subpopulation and the function for the overall target population.

Mekahael (2009) presents specific formulas for these measures when they are used in the context of a simulation study.

The conditional standard error of equating (CSEE) is computed by

$$CSEE_j = \sqrt{\frac{1}{I}\sum_i \left(\hat{e}_y(x_{ij}) - \frac{\sum_i \hat{e}_y(x_{ij})}{I} \right)^2} \tag{3}$$

where I is the number of replications of the simulation study, and $\hat{e}_y(x_{ij})$ is the X to Y equated score at score x_j estimated for a given replication i. The CSEE across all score points can be used to compute an overall SEE:

$$Avg\ SEE = \sqrt{\sum_j p_j CSEE_j^2} \tag{4}$$

where p_j is the raw proportion of examinees at score x_j.

For the conditional bias the formula is:

$$CBias_j = \frac{1}{I}\sum_i [\hat{e}_y(x_{ij}) - e_y(x_j)] \tag{5}$$

where $e_y(x_j)$ is the criterion equating at score x_j estimated with population data. An estimate of the overall bias can be obtained by summing all score points:

$$Bias = \sum_j p_j CBias \qquad (6)$$

By squaring both sides this equation an estimate of the average bias squared is obtained. The average bias squared is useful when calculating the average root mean squared error (RMSE). It provides an estimate based on combining information from systematic error (bias) and random error (SEE):

$$RMSE = \sqrt{\sum_j p_j (CBias^2 + SEE^2)} \qquad (7)$$

Comparing Methods of Equating

Before the assumptions and details of the Levine Observed Score method - the focus of this study- are discussed it is important to present a brief description of previous results on comparing different CTT methods under the NEAT design.

When comparing CE and frequency estimation methods, Harris and Kolen (1990) found that these two methods produce different results when there are differences in ability in the samples for X and Y forms and they suggested the use of PSE methods because of their better theoretical base. However, Livingston et al. (1990) recommended the use of CE when there is a large ability difference in the groups taking the two forms of the test.

Kolen and Brenan (2004) indicate that Levine methods are more appropriate than the Tucker method when groups are rather dissimilar, but at the same time they indicate that if the populations are too dissimilar, any equating is suspect. They add that it is impossible to provide a strict guide as to what “too dissimilar” means but an example of too much dissimilarity are forms that do not share common content and statistical specifications. However they indicate that in

NEAT, differences between the two groups of approximately .1 or less standard deviation unit on the anchor appear to cause few problems for any of the methods. On the other hand, mean group differences of about .3 or more standard deviation unit can produce substantial differences among methods. Finally, ratios of group standard deviation on the anchor of less than .8 or greater than 1.2 are associated with substantial differences among methods.

Holland (2004) compared CE, Tucker and Levine and concluded that if the equating is performed from a more able population to a less able one, the three equating functions will lie in a fixed relationship to each other. Tucker will lie below chain, which will lie below Levine. If the equating is performed from a less able population to a more able one, the order will be reversed. He also concluded that the more different the populations P y Q are, regarding the mean performance on A, the anchor test, the more the three equating functions will differ. Finally, he showed that for a less reliable test, two groups will have smaller standardized mean differences than they will have for a more reliable test, which suggests that the less reliable the anchor test A is, the less different the three linear equating will be.

Mekhael et al. (2009) compared Tucker and Levine examining the effect of two factors: standardized mean ability difference and anchor-total correlation. They found that the average bias for Tucker was greater than the average bias for Levine. Additionally, the standard error of equating (SES) was almost identical across ability differences.

Puhan (2010) compared CE, Tucker and Levine and found that Tucker performed the worst in terms of bias and RMSE. The Levine method had the lowest bias and RMSE was similar for CE and Levine. Puhan (2010) concluded that CE, Tucker or Levine may be used when the difference in X and Y is small and the correlation between the anchor and total test is at least moderately high. On the other hand, CE appears to be preferable when the groups taking X and Y differ in ability since CE produce the lowest RMSE.

The Levine Observed Score Method in NEAT

According to Kolen and Brennan (2004) the Levine Observed Score method relates observed scores on X to observed scores on Y but its assumptions pertain to true scores, which are assumed to be related to observed scores according to the classical test theory, whereby the observed score is the sum of true and error scores, the expectation of the error scores is zero and error scores are uncorrelated with true scores.

In the Levine Observed Score method, three assumptions are traditionally made so that unknown information from test Y for group P and Test X for group Q can be estimated. Following Holland and Walker (2006) these three assumptions will be denoted L1, L2 and L3 and are presented below.

L1) $X = \tau_X + e_X$, $Y = \tau_Y + e_Y$, and $A = \tau_A + e_A$, along with the assumption that the error terms $e_{\boldsymbol{X}}$, $e_{\boldsymbol{Y}}$, and $e_{\boldsymbol{A}}$ are uncorrelated with their corresponding true scores $\tau_{\boldsymbol{X}}$, $\tau_{\boldsymbol{Y}}$, and $\tau_{\boldsymbol{A}}$.

L2) $\tau_X = a + b\tau_{\mathbf{A}}$, and $\tau_{\boldsymbol{Y}} = c + d\tau_{\mathbf{A}}$. i.e. the true scores of X and A and Y and A are linearly related. This is the congenericity assumption.

The Levine method rests on the assumption that L1 and L2 hold for any population S of the synthetic form

$$S = wP + (1 - w)Q, \tag{8}$$

where $0 \leq w \leq 1$ (Braun & Holland, 1982).

L3) The error variances $\sigma^2_{e_X/S}$, $\sigma^2_{e_Y/S}$ and $\sigma^2_{e_A/S}$ are the same for any S of the synthetic form. This assumption, along with an assumption of proportional error variances for X, Y, and A, is used in the computation of Angoff's (1971) reliability estimates. L3 will be referred to as Angoff's assumption.

An alternative assumption to L3 has been recently proposed by Holland (2004; see also Holland & Walker, 2006) and it is denoted here by L3*.

L3*) The ratios $\frac{\rho_{X/S}}{\rho_{A/S}}$ and $\frac{\rho_{Y/S}}{\rho_{A/S}}$ are constant as functions of S of the synthetic form (Holland & Walker, 2006); that is, the ratios of the square roots of the reliabilities $\frac{\rho_{X/S}}{\rho_{A/S}}$ and $\frac{\rho_{Y/S}}{\rho_{A/S}}$ are population invariant.

The idea behind this L3* assumption is that from L1 and L2 it can be shown that

$\sigma_{X/S} = \sigma_{X/P}\,\frac{\sigma_{A/S}}{\sigma_{A/P}}\left[\frac{\rho_{X/P}}{\rho_{A/P}}\bigg/\frac{\rho_{X/S}}{\rho_{A/S}}\right]$ and $\sigma_{Y/S} = \sigma_{Y/Q}\,\frac{\sigma_{A/S}}{\sigma_{A/Q}}\left[\frac{\rho_{Y/Q}}{\rho_{A/Q}}\bigg/\frac{\rho_{Y/S}}{\rho_{A/S}}\right]$ (Holland & Walker, 2006). Under L3* the previous two formulas reduce to

$\sigma_{X/S} = \sigma_{X/P}\,\frac{\sigma_{A/S}}{\sigma_{A/P}}$ and $\sigma_{Y/S} = \sigma_{Y/Q}\frac{\sigma_{A/S}}{\sigma_{A/Q}}$ because the value in brackets will be equal to 1.0.

The next section in this dissertation presents the derivation of such formulas and other formulas related to the Levine method.

Derivation of Formulas

The following derivation of formulas is based on Holland & Walker (2006).

Results from L1 and L2

Several simplifying consequences can be derived from assumptions L1 and L2.

From L1 for any *S*, the mean of X and of τ_X over *S* are the same, i.e.,

$$\mu_{X/S} = \mathrm{E}(X|\,S) = \mathrm{E}(\tau_X|\,S) = \mu_{\tau_X/S} \qquad (9)$$

Similar results hold for Y and A as well.

By taking expectations over *T* of the linear equations in L2, and by then letting *w* = 1 so that *S* = *P*, it follows that

$$a = \mu_{X/S} - b\mu_{A/S} = \mu_{X/P} - b\mu_{A/P},$$

by the rules of expectations of functions, and implies the following basic formula for μ_{XT} in terms of quantities that can be estimated directly in the NEAT design plus the unknown value of *b*,

$$\mu_{X/S} = \mu_{X/P} + b(\mu_{A/S} - \mu_{A/P}). \qquad (10)$$

By an analogous argument a formula for $\mu_{Y/S}$ is obtained:

$$\mu_{Y/S} = \mu_{Y/Q} + d(\mu_{A/S} - \mu_{A/Q}). \qquad (11)$$

In addition, taking variances over S of the linear equations in L2, and then letting *w* = 1 so that *S* = *P*, results in

$$\sigma^2_{\tau_X/S} = b^2\sigma^2_{\tau_A/S} \text{ and } \sigma^2_{\tau_X/P} = b^2\sigma^2_{\tau_A/P}. \qquad (12)$$

This follows directly from the definition of the variance of a function. Equation (12) implies the following formula for *b* and shows the sense in which it is the “effective length” of X relative to A,

$$b = \frac{\sigma_{\tau_X/S}}{\sigma_{\tau_A/S}} = \frac{\sigma_{\tau_X/P}}{\sigma_{\tau_A/P}}. \qquad (13)$$

The notion of *effective test length* is expressed as the ratio of the true score standard deviations. By an analogous argument the corresponding formula for *d*, can be derived:

$$d = \frac{\sigma_{\tau_Y/S}}{\sigma_{\tau_A/S}} = \frac{\sigma_{\tau_Y/Q}}{\sigma_{\tau_A/Q}}. \qquad (14)$$

Note that L2 gets its strength as an assumption from the requirement that it holds for any T, and is therefore *population invariant*.

Formulas for the Variances of X and Y over S

Two ways to obtain expressions for the variances of X and Y over S are presented. The first assumes L1 and L2 and makes a population invariance assumption concerning the ratio of the reliabilities of X and A and of Y and A. In the second approach, the traditional one, also L1 and L2 are assumed but a different population invariance assumption concerning the error variances is made.

For the first approach, the usual formulas for test reliability are used to first express the relationship in (13) in slightly different terms. Defining the *reliabilities* of X and A in S, as usual, as

$$\rho^2_{X/S} = \frac{\sigma^2_{\tau_X/S}}{\sigma^2_{X/S}} \text{ and } \rho^2_{A/S} = \frac{\sigma^2_{\tau_A/S}}{\sigma^2_{A/S}}, \tag{15}$$

results

$$\sigma_{\tau_X/S} = \rho_{X/S}\sigma_{X/S} \text{ and } \sigma_{\tau_A/S} = \rho_{A/S}\sigma_{A/S}$$

and therefore from (6) it follows that

$$b = \frac{\rho_{X/S}\sigma_{X/S}}{\rho_{A/S}\sigma_{A/S}} = \frac{\rho_{X/P}\sigma_{X/P}}{\rho_{A/P}\sigma_{A/P}}. \tag{16}$$

From (16) it results that

$$\sigma_{X/S} = \sigma_{X/P}\,\frac{\sigma_{A/S}}{\sigma_{A/P}}\left[\frac{\rho_{X/P}}{\rho_{A/P}}\middle/\frac{\rho_{X/S}}{\rho_{A/S}}\right]. \tag{17}$$

Analogously, from L1 and L2 it can be shown that

$$\sigma_{Y/S} = \sigma_{Y/Q}\,\frac{\sigma_{A/S}}{\sigma_{A/Q}}\left[\frac{\rho_{Y/Q}}{\rho_{A/Q}}\middle/\frac{\rho_{Y/S}}{\rho_{A/S}}\right]. \tag{18}$$

At this point the idea behind the first approach to estimating $\sigma_{X/S}$ and $\sigma_{Y/S}$ is to assume that the expressions in brackets in (17) and (18) have the value 1.0; that is, to assume that the *ratios* of the square roots of the reliabilities, $\frac{\rho_{X/S}}{\rho_{A/S}}$ and $\frac{\rho_{Y/S}}{\rho_{A/S}}$ are *population invariant*. This is assumption L3*. Under this assumption, the standard deviations of X and Y over S are given by

$$\sigma_{X/S} = \sigma_{X/P}\,\frac{\sigma_{A/S}}{\sigma_{A/P}} \text{ and } \sigma_{Y/S} = \sigma_{Y/Q}\frac{\sigma_{A/S}}{\sigma_{A/Q}}. \tag{19}$$

The second approach uses the well known decomposition of observed test score variance into true score variance and error variance; that is,

$$\sigma^2_{X/S} = \sigma^2_{\tau_X/S} + \sigma^2_{e_X/S}.$$

The population invariance assumption is on the error variances, such that the variances $\sigma^2_{e_X/S}$, $\sigma^2_{e_Y/S}$ and $\sigma^2_{e_A/S}$ are constant for any S of the synthetic form. This is assumption L3. From L3 it results that

$$\sigma^2_{X/S} - \sigma^2_{\tau_X/S} = \sigma^2_{X/P} - \sigma^2_{\tau_X/P}, \tag{20}$$

or

$$\sigma^2_{X/S} = \sigma^2_{X/P} + (\sigma^2_{\tau_X/S} - \sigma^2_{\tau_X/P}). \tag{21}$$

Using the equations in (5) it follows that (14) may be expressed as

$$\sigma^2_{X/S} = \sigma^2_{X/P} + b^2(\sigma^2_{\tau_A/S} - \sigma^2_{\tau_A/P}), \tag{22}$$

Moreover, L3 also implies that

$$\sigma^2_{\tau_A/S} - \sigma^2_{\tau_A/P} = \sigma^2_{A/S} - \sigma^2_{A/P},$$

so that (15) can be written as

$$\sigma^2_{X/S} = \sigma^2_{X/P} + b^2(\sigma^2_{A/S} - \sigma^2_{A/P}). \tag{23}$$

An analogous result follows for $\sigma^2_{Y/S}$, i.e.,

$$\sigma^2_{Y/S} = \sigma^2_{Y/Q} + d^2 (\sigma^2_{A/S} - \sigma^2_{A/Q}). \tag{24}$$

Note that equations (23) and (24) are similar to (10) and (11) in that they depend on the variances that can be directly estimated in the NEAT design as well as on the unknown values, *b* and *d*.

Therefore in order to use (10), (11), (23) and (24), the values of *b* and *d* need to be estimated. From (16) *b* can be estimated by

$$b = \frac{\rho_{X/P}\sigma_{X/P}}{\rho_{A/P}\sigma_{A/P}}. \tag{25}$$

For *d*, the corresponding formula is

$$d = \frac{\rho_{Y/Q}\sigma_{Y/Q}}{\rho_{A/Q}\sigma_{A/Q}}. \tag{26}$$

*Comparing L3 and L3**

Carvajal et al. (2008) conducted a study to determine which of two assumptions - Angoff's constant error variance assumption (L3) or Holland's constant reliability ratio assumption (L3*) - is more viable across a range of populations S. However, that study, although based on simulated data, did not include replications of the simulation and did not take into consideration factors such as anchor length, differences in test difficulty, differences in ability distribution for populations P and Q and sample size of the groups taking X and Y. No other studies have been found that address a comparison between L3 and L3*. In consequence this dissertation study compared L3 and L3* by manipulating those four factors. This dissertation study also included the mixture of populations, which was included as a factor in the Carvajal et al. (2008) study.

Carvajal et al. (2008) used an anchor test with 50 items and tests X and Y with 100 items each. For the anchor test A in population P, true scores were generated under N(25, 64). In population Q true scores for A were generated under N(27, 81), where 64 and 81 represent variances, so that population Q was more able and more variable than population P. Observed scores on test A were generated under L1 by using a binomial error model to generate error terms. Under the binomial error model the (squared) conditional standard error of measurement (CSEM) is determined by

$$\sigma^2_{A|\tau} = {\tau_A (n - \tau_A)} \big/ {(n-1)} \tag{27}$$

where n is the number of items in the test (Lord & Novick, 1968).

Once the error scores for A were generated under $N\,(0, \sigma^2_{A|\tau}\,)$, the observed A scores were obtained by adding the true scores and the corresponding error scores.

True X scores and true Y scores were generated under L2 ($\tau_{\boldsymbol{X}} = a + b\,\tau_{\mathbf{A}}$ and $\tau_{\boldsymbol{Y}} = c + d\,\tau_{\mathbf{A}}$). The choice of *a, b, c* and *d* was made to replicate reasonable values in practical settings, given the desired maximum score of 100 for both tests: *a*=2, *b*=2.1, *c*=4, and *d*=1.9. The result was a test X that was somewhat easier across the majority of the score range. Observed X and Y scores were generated under the binomial error model in a similar way as observed scores for A were generated.

In that study, the populations P and Q were combined to produce the synthetic form *S* = *wP* + (1 – *w*)*Q*, (Braun & Holland, 1982) where *w* varies from 0 to 1. Eleven different weights for *w* where used, ranging from 0 to 1 in increments of .1. For example when *w* = 0, S= Q and when w = 1, S = P. When *w* = 0.1, S is the combination of a random sample of 10% from P and 90% from Q.

In the Carvajal et al. (2008) study, populations P and Q were created with 100,000 subjects each. Every population S had 100,000 cases as well. The actual variances and reliabilities for X and Y were computed directly in each synthetic population S, because full information on X and Y was available for every case. Then the estimates of variances and reliabilities under the L3 and L3* assumptions were computed, using only information on X in P and information on Y in Q. These estimates were then compared with the actual values.

They found that although the estimation from the traditional assumption L3 appeared to be closer to the estimation from the alternative L3*, the differences were not of practical significance given that these differences were not expected to result in a change in actual reported scores. However, the only factor considered in that study was the mixture of populations (Braun & Holland, 1982). The effects of anchor length, sample size, population ability differences for populations P and Q, and test difficulty differences were not addressed in that study and replications of the simulation were not conducted.

No other studies have been found that address such factors. Therefore, the current study is warranted to determine the degree of concordance or discrepancy between L3 and L3* under various conditions related to those factors.

The educational implication of this dissertation resides in the evaluation of the usually untestable invariance assumptions inherent in NEAT equating. This study compared the assumption of the constant square root of reliability ratio (L3*) to the traditional assumption of constant error variance (L3) in the Levine Observed Score method of linear equating.

Chapter Three - Methods

General Description

In the interest of achieving greater degree of fairness in standardized testing, equating methods need to be used as to adjust for unwanted differences among forms. The greater the knowledge about the performance of a specific equating method and its relative performance with other methods, the better decisions can be taken regarding its application. The methods described here attempt to examine the relative performance of the Levine Observed Score method of linear equating under two different assumptions, namely L3 and L3*.

Because this is a simulation study, the true equating function is known, so this study compared this function to the one obtained under L3 and the one obtained under L3* (in conjunction with assumptions L1 and L2 in each case). Therefore the dependent variable is the equating function, which was evaluated using bias and RMSE.

The independent variables are anchor length, differences in test difficulty, differences in ability distribution for populations P and Q, mixture of populations, and sample size of groups taking X and Y. The conditions for these independent variables or factors are described later.

Following the data generation in Carvajal et al. (2008), this study generated true scores for the anchor test A in population P and in population Q. Then observed scores for test A were generated under L1 by using a binomial error model to generate error terms. Under the binomial error model the (squared) conditional standard error of measurement (CSEM) is determined by $\sigma^2_{A|\tau} = {}^{\tau_A(n-\tau_A)}\!\diagup_{(n-1)}$ where *n* is the number of items in the test (Lord & Novick, 1968).

Once the error scores for A were generated under $N\,(0, \sigma^2_{A|\tau}\,)$, the observed A scores were obtained by adding the true scores and the corresponding error scores.

True X scores and true Y scores were generated under L2 ($\tau_X = a + b\,\tau_A$ and $\tau_Y = c + d\,\tau_A$). Different choices of a, *b, c* and *d* were made as described later. This resulted in different level of difficulty between X and Y. Once true X scores and true Y scores were generated, observed X and Y scores were generated by using again the binomial error model in a similar way as observed scores for *A* were generated. Because the way X and Y scores were generated, whereby errors for the observed scores are all independent of each other, the test A is considered an external anchor. For an internal anchor model, the error for observed score A would be contained in the error for observed scores X and Y. Note however that the derivation of formulas derived in Chapter Two did not make use of internal or external anchor properties; therefore those formulas should be applicable to either case.

Factors for the Study

This study manipulated five factors: ability distribution, anchor length, difference in test difficulty, mixture of populations and sample size of groups taking X and Y.

Following, a description of the factors or independent variables for this study is presented.

Ability Distribution

Population P is the baseline so there is only one condition for it and population Q was varied to have more conditions.

The condition for population P is denoted as N(50%, 15%) where the percentages are relative to anchor test length. For example, for an anchor test of 32 items, N(50%, 15%) means N(16, 4.8) where 16 is the mean and 4.8 the standard deviation of the normal distribution of true scores for anchor test A in population P.

The same notation is used for population Q for which four conditions were defined:

N(50%, 15%) (no difference)
N(55%, 15%) (difference only in mean ability)

N(50%, 18%) (difference only in variability)
N(55%, 18%) (difference in mean ability and variability)

For example, for an anchor test of 32 items, N(55%, 18%) means N(17.6, 5.76) for the true scores of anchor A in Q.

Population P and the four populations Q were generated to have 100,000 subjects each.

The data were originally generated for the conditions N(50%, 10%) for P and N(50%, 10%), N(55%, 10%), N(50%, 15%) and N(55, 15%) for Q but this produced such low reliabilities that the results would not be applicable to real data situations. Therefore a decision was made to increase the percentage corresponding to the standard deviations to 15% and 18%. This is reported in more detail in Chapter Four.

Anchor Length

The second factor manipulated in this dissertation study was anchor length. With a fixed 80 item test length for X and Y, two anchor length conditions were defined: 40% and 20% of test length. This resulted in two anchor lengths of 32 and 16.

The corresponding distributions for P and Q under the conditions of ability distribution and anchor length are presented in detail in Chapter Four.

Test Difficulty

The third factor manipulated in this study was the difference in difficulty between tests X and Y. This was achieved by controlling the coefficients a, b, c, d in the equations $\tau_{\boldsymbol{X}} = a + b\tau_{\mathbf{A}}$, and $\tau_{\boldsymbol{Y}} = c + d\tau_{\mathbf{A}}$ of assumption L2 in the following manner:

- For test X the coefficient *a* equals 0.
- For test Y the coefficient *d* equals the ratio number of test items to the number of anchor items whereas the coefficient *c* equals either +4% *n* or -4% *n*, where *n* corresponds to number of items of test X and Y. Adding 4% *n* makes test Y

easier, subtracting 4% n makes test Y more difficult. Given that n is 80 across this study, *c* is either 3.2 or -3.2.

- Coefficient *b* for test X was set to be equal to coefficient *d* for test Y.

The corresponding distributions for P and Q under the conditions of ability distribution and anchor length are presented in detail in Chapter Four.

Mixture of populations

P and Q were combined to produce the synthetic form $S = wP + (1 – w)Q$ (Braun & Holland, 1982) where *w* varies from 0 to 1. According to Kolen and Brennan (2004), although the NEAT design involves two populations, an equating function is typically viewed as being defined for a single population; therefore populations P and Q must be combined to obtain an equating relationship. According to these authors in the vast majority of real equating contexts the choice of *w* makes little practical difference; however, they indicate that many equations are simplified considerably by choosing $w = 1$ and that furthermore, setting $w= 1$ means that the synthetic group is the new population, which is often the only population that will take the new form X. They indicate that *w*= N1/(N1+N2), where N1 and N2 are the sizes of groups P and Q, is a common choice but that, ultimately, the choice of *w* is a judgment that should be based on an investigator’s conceptualization of the synthetic population.

Therefore the values $w = 0.25$, $w = 0.50$, $w = 0.75$ and $w = 1$ were selected to cover the possible choices described by Kolen and Brennan (2004) and w = 0 was added to complete the range of possible values for *w*.

Sample Size

The combination of the factors ability distribution (four conditions), anchor length (2 conditions) and test difficulty (2 conditions) produced 16 conditions. For each of the 16 conditions 100 samples of size 500, 1000 and 2000 from population P and 100 samples of size 500, 1000 and 2000 from population Q were randomly extracted in a bootstrap with replacement fashion whereby after a subject that is selected for the sample is returned to the population so that he could be selected again. This produced 300 pairs of samples P and Q for each of 16 conditions, i.e. in total 4800 pairs of samples were generated.

For each of these pairs of samples five equating functions were computed under the L3 assumption and five equating functions were computed under the L3* assumption. Each of these five equating functions in each case was produced using a different *w* weight, where *w:* 0, 0.25, 0.50, 0.75, and 1. The steps for the equating at sample level are described in a later section.

Summary of Conditions

The settings for the five factors (ability distribution, anchor length, test difficulty, mixture of populations, and sample size) manipulated in this study produced 4x2x2x5x3 = 240 conditions. Table 3 summarizes the conditions for the study.

Table 3
Summary of Completely Crossed Conditions for the Study

Ability Distribution (4)	Anchor Length (2)	Test Difficulty (2)	Mixture of Populations (5)	Sample Size (3)
1. P N(50%, 15%)	1. 40%	1. + 4%	1. 0	1. 500
	2. 20%	2. - 4%	2. 0.25	2. 1000
1. Q N(50%, 15%)			3. 0.5	3. 2000
2. Q N(55%, 15%)			4. 0.75	
3. Q N(50%, 18%)			5. 1	
4. Q N(55%, 18%)				

Data Generation and Equating at Synthetic Population Level

The following steps describe in detail how the data were generated for each condition in the study and how the means, variances and equating functions were computed at synthetic population level. The equating at sample level will be explained later in this chapter.

1) For population P, generate true score for A according to the corresponding distribution.
2) By using the corresponding *a* and *b* parameters and L2 assumption determine the true score for X. Do the same for the true score for Y by using *c* and *d* parameters and L2.
3) Use the binomial error model (Lord & Novick, 1968) to determine the error variances at each true score level for tests A, X and Y. Generate random errors for each score in the population for tests A, X and Y.
4) Compute the observed A, X, and Y scores as true score plus error (using L1).
5) Compute observed score mean and variance, true score mean and variance, and error mean and variance, for X, Y and A in P. Since all the data are known such means and variances can be computed directly by using the mean and variance formulas.
6) Repeat steps 1-5 for population Q.
7) Create synthetic populations $S = wP + (1 - w)Q$, for chosen w weights ranging from 0 to 1.
8) Compute analogous means and variances to those indicated in point 5 for X, Y and A in *S* using full information. This is accomplished in similar way as indicated in point 5.
9) Now, discard information on X from Q and information on Y from P. Estimate the observed score mean and variance of X and Y in S using assumptions L3 and L3*. Note that information on A does not need to be estimated, because information on A is available in both P and Q.

To accomplish point 9 is necessary to note that, according to the development of formulas presented in Chapter Two, the formulas for the means are common for L3 and L3* whereas the formulas for variances are different under L3 than under L3*.

For the means the formulas presented in Chapter Two are

$$\mu_{X/S} = \mu_{XP} + b(\mu_{A/S} - \mu_{A/P}) \text{ and} \qquad (10)$$

$$\mu_{Y/S} = \mu_{Y/Q} + d(\mu_{A/S} - \mu_{A/Q}), \qquad (11)$$

where *b* and *d* can be estimated by

$$b = \frac{\rho_{X/P}\sigma_{X/P}}{\rho_{A/P}\sigma_{A/P}} \qquad (25)$$

and

$$d = \frac{\rho_{Y/Q}\sigma_{Y/Q}}{\rho_{A/Q}\sigma_{A/Q}} . \qquad (26)$$

The four reliabilities in formulas (25) and (26) for *b* and *d* can be computed using the ratio of true score variance to observed score variance.

For the variance under L3 the formulas are

$$\sigma^2_{X/S} = \sigma^2_{X/P} + b^2(\sigma^2_{A/S} - \sigma^2_{A/P}) \text{ and} \qquad (23)$$

$$\sigma^2_{Y/S} = \sigma^2_{Y/Q} + d^2(\sigma^2_{A/S} - \sigma^2_{A/Q}), \qquad (24)$$

where *b* and *d* can be estimated in the same way as indicated for the means.

The variances under L3* can be computed by squaring the standard deviations, which formulas are

$$\sigma_{X/S} = \sigma_{X/P}\frac{\sigma_{A/S}}{\sigma_{A/P}} \text{ and } \sigma_{Y/S} = \sigma_{Y/Q}\frac{\sigma_{A/S}}{\sigma_{A/Q}} . \qquad (19)$$

10) All the previous steps make information available in order to compute three different equating functions: the criterion equating function or true equating function, and the two equating functions corresponding to L3 and L3*.

Specifically if $Y = \alpha X + \beta$ denotes the equating function, α and β are computed each in three different ways: one for the criterion, one for L3 and one for L3* as follows.

α is computed by the formula $\alpha = \dfrac{\sigma_{Y/S}}{\sigma_{X/S}}$. (28)

For the criterion the standard deviations are computed from the full information in the synthetic population. For L3 and L3* the standard deviations are estimated using the corresponding formulas indicated in point 9.

The β coefficient is computed with the formula $\beta = \mu_{Y/S} - \alpha\mu_{X/S}$. (29)

For the criterion these means are computed using the full information while for L3 and L3* these means are estimated using the formulas indicated in point 9.

Equating at Sample Level

The following steps describe in detail how the equating at sample level was conducted.

1) Each of the 4800 pairs of samples is constituted by a sample P taking X and a sample Q taking Y.
2) Specify a *w* weight, namely 0, 0.25, 0.50, 0.75, or 1.
3) Once the *w* weight is specified, the mean and variance of the anchor in the corresponding hypothetical synthetic population can be estimated using the following two properties of a mixture distribution (Kolen and Brennan, 2004)

$\mu_{A/S} = w\mu_{A/P} + (1-w)\mu_{A/S}$ and (30)

$\sigma^2_{A/S} = w\sigma^2_{A/P} + (1-w)\sigma^2_{A/Q} + w(1-w)\left(\mu_{A/P} - \mu_{A/Q}\right)^2$, (31) where $S = wP + (1 - w)Q$ and w varies from 0 to 1.

Note that although these equations are formulated in terms of parameters, in practice these parameters are substituted by the sample estimates.

4) Once the mean and variance of the anchor in S are estimated it is possible to apply the formulas for L3 and L3* (formulas (10), (11), (25), (26), (23), (24), (19), (28), and (29) presented in that order in the previous section) and the equating procedure described for the population level. Note that those formulas are written with population parameters but they are substituted from estimates from the samples.

5) For the samples in this study the four reliabilities in formulas (25) and (26) were computed using the ratio of true score variance to observed variance, which reflects sampling variability.

6) Each one of the 300 pairs of samples for each of the 16 described conditions at population level can be associated with a synthetic population corresponding condition. For each pair of samples the two equating functions under L3 and L3* can be computed and then compared to the true equating function in the mixture of synthetic population corresponding to that w weight. In this manner bias and RMSE across the 100 samples can be computed.

7) This produced 300 pairs of samples P and Q for each of 16 conditions i.e. in total 4800 pairs of samples were generated.

8) For each of these pairs of samples five equating functions were computed under the L3 assumption and five equating functions were computed under the L3* assumption. Each of these five equating functions in each case was produced using a different w weight, where w: 0, 0.25, 0.50, 0.75, and 1.

9) As can be noted 4800 x 5= 24,000 equating functions were computed for L3 and 24,000 for L3*. Each of these 24,000 cases can be related to a true equating function in the corresponding synthetic population.

10) Therefore for each of 24,000 cases it is possible to compute the bias for each of the X observed scores included in the sample P of the pair and then obtain a bias average over the sample size of that particular sample. Similarly, for each X score in P, a squared bias can be computed and then used to compute an RMSE for that sample over its sample size. These 24000 average biases and 24,000 RMSE for L3 and L3* were used to perform two repeated measures ANOVA whereby the within subjects factor is the type of assumption (this has two levels, the bias under L3 and L3* in one case, and the RMSE under L3 and L3* in the other case). In each case the between subjects factors are sample size (SS), weight (WGT), length of anchor (ANC), whether test Y is easier or more difficult than X (DIF), and ability distribution of the population from where the sample comes from (ABIL). SS has three levels: 500, 1000 and 2000, WGT has five levels: 0, 0.25, 0.50, 0.75, and 1. ANC has two levels: e (easier) and d (more difficult). ABIL has four levels: 05, 08, 55, and 58 which indicate respectively N(50%, 15%), N(55%, 15%), N(50%, 18%), and N(55%, 18%).

An Illustrative Example with Real Data

To illustrate the use of L3 and L3* equating with a real data example these methods were applied to two 36-item forms X and Y. The data sets were obtained from the software CIPE referred by Kolen and Brennan (2004). For these forms the anchor A is formed by every third item (items 3, 6, 9, …, 36). Since scores on A are contained in X and Y, A is an internal anchor.

However, according to what was indicated in Chapter Two, the formulas provided by L3 and L3* can be applied to either external or internal anchors.

The application of such formulas is the same as explained for the equating at sample level. On the other hand, for this example there is not a true equating function available, as it is the case with any real data application. Therefore instead of bias and RMSE, the difference between the equating function produced by L3 and L3* was computed and as well as the RMSD.

Software

SPSS 17.0 was used to compute skewness in populations P and Q and to conduct the ANOVA analysis at sample size level. FORTRAN 95 was used for the data generation, data management, sampling, conducting the equating and computing bias and RMSE. For this purpose 10 FORTRAN 95 programs were written by the author of this study. Figures were built in Excel.

Chapter Four – Results

The purpose of this study was to compare the equating functions resulting from the assumption of the constant squared root of reliability ratio (L3*) to that of the traditional assumption of constant error variance (L3) in the Levine Observed Score method of linear equating. Because this is a simulation study, it was possible to compute the true equating function and to compare this function to the one obtained under L3 and the one obtained under L3*.

This study was conducted in the context of the NEAT design whereby population P takes test X and population Q takes test Y. However for the sake of computing the true equating function in this simulation study both X and Y were administered to both P and Q. Then different assumptions were used to estimate the equating function using only information from X in P, Y in Q and from the anchor A in both P and Q.

To accomplish this purpose, data for one population P and four populations Q were generated. Anchor test true scores were generated under different conditions as explained in Chapter Three.

Descriptive Statistics of the Generated Data

Table 4 shows the designed mean and standard deviation for the anchor test true score based on anchor length, test difficulty and ability distribution. The first four data rows of the table refer to population P. Note that there is only one condition for the ability distribution for population P, which is N(50%, 15%). Two anchor lengths and two conditions for test difficulty make up the four conditions shown for P in this table. For example, the second data row of the table shows that for an anchor length of 40% of test length, a more difficult test Y, and an ability distribution of N(50%, 15%), the anchor length is 32, the ratio test length to anchor length (coefficient *b*) is 2.5, the difference in difficulty is -3.2 (coefficient *c*), and the designed mean and standard deviation for the anchor true score are 16 and 4.8 respectively.

The remaining 16 rows of Table 4 show the conditions for the four populations Q. For example the second to last row of the table shows that for an anchor length of 20% of test length, an easier test Y, and an ability distribution of N(55%, 18%), the anchor length is 16, the ratio test length to anchor length (coefficient *b*) is 5, the difference in difficulty is 3.2 (coefficient *c*), and the designed mean and standard deviation for the anchor true score are 8.8 and 2.88 respectively.

Table 4
Designed Mean and Standard Deviation of Anchor Test True Scores

Population	Anchor Length %	Test Difficulty	Ability Distribution	Anchor Length	Ratio Test/Anchor Length (b)	±4% Test Length (c)	True A Mean	True A SD
P	40	e	05	32	2.5	3.2	16	4.8
	40	d	05	32	2.5	-3.2	16	4.8
	20	e	05	16	5	3.2	8	2.4
	20	d	05	16	5	-3.2	8	2.4
Q	40	e	05	32	2.5	3.2	16	4.8
	40	d	05	32	2.5	-3.2	16	4.8
	20	e	05	16	5	3.2	8	2.4
	20	d	05	16	5	-3.2	8	2.4
	40	e	55	32	2.5	3.2	17.6	4.8
	40	d	55	32	2.5	-3.2	17.6	4.8
	20	e	55	16	5	3.2	8.8	2.4
	20	d	55	16	5	-3.2	8.8	2.4
	40	e	08	32	2.5	3.2	16	5.76
	40	d	08	32	2.5	-3.2	16	5.76
	20	e	08	16	5	3.2	8	2.88
	20	d	08	16	5	-3.2	8	2.88
	40	e	58	32	2.5	3.2	17.6	5.76
	40	d	58	32	2.5	-3.2	17.6	5.76
	20	e	58	16	5	3.2	8.8	2.88
	20	d	58	16	5	-3.2	8.8	2.88

[a]Ability Distribution 05: N(50%, 15%) 55: N(55%, 15%) 08: N(50%, 18%) 58:N(55%, 18%)
[b]Test Y Difficulty e: Test Y is easier d: Test Y is more difficult

To give an idea of the correspondence between the designed mean and standard deviations for anchor test true scores and the obtained mean and standard deviations for the anchor test true scores, Table 5 shows the obtained mean and standard deviation for the anchor test true scores based on anchor length, test difficulty and ability distribution. For the first of the two examples just provided for Table 4 the designed mean and variances were 16 and 4.8 and Table 5 shows that the corresponding obtained mean and standard deviation were 15.995 and 4.802. For the second example the designed mean and standard deviation were 8.8 and 2.88 and Table 5 shows that the corresponding obtained mean and standard deviation were 8.798 and 2.853.

Table 5
Actual Mean and Standard Deviation of Anchor Test True Scores

Population	Anchor Length %	Test Difficulty	Ability Distribution	Actual A True Mean	Actual A True SD
	40	e	05	16.003	4.797
P	40	d	05	15.995	4.802
	20	e	05	7.989	2.397
	20	d	05	8.010	2.402
	40	e	05	16.031	4.808
	40	d	05	16.029	4.803
	20	e	05	8.004	2.398
	20	d	05	8.000	2.397
	40	e	55	17.587	4.788
	40	d	55	17.583	4.795
	20	e	55	8.798	2.389
Q	20	d	55	8.802	2.395
	40	e	08	16.008	5.739
	40	d	08	15.996	5.732
	20	e	08	8.005	2.870
	20	d	08	7.997	2.861
	40	e	58	17.578	5.740
	40	d	58	17.586	5.734
	20	e	58	8.798	2.853
	20	d	58	8.793	2.868

[a]Ability Distribution 05: N(50%, 15%) 55: N(55%, 15%) 08: N(50%, 18%) 58:N(55%, 18%)
[b]Test Y Difficulty e: Test Y is easier d: Test Y is more difficult

Observed scores for A, X and Y were obtained as described in Chapter Three. Table 6 shows the mean and SD for these scores.

Table 6
Actual Mean and Standard Deviation of Observed Scores of Anchor A and Test X and Y

Population	Anchor	Test	Ability	A	A	X	X	Y	Y
	Length	Difficulty	Distribution	Obs.	Obs.	Obs.	Obs.	Obs.	Obs.
	%			Mean	SD	Mean	SD	Mean	SD
P	40	e	05	16.01	5.52	40.02	12.73	43.20	12.72
	40	d	05	16.00	5.54	39.99	12.74	36.81	12.74
	20	e	05	7.98	3.12	39.95	12.73	43.15	12.73
	20	d	05	8.01	3.11	40.05	12.76	36.83	12.75
Q	40	e	05	16.02	5.54	40.10	12.74	43.28	12.74
	40	d	05	16.03	5.55	40.09	12.73	36.88	12.76
	20	e	05	8.01	3.12	40.03	12.73	43.20	12.71
	20	d	05	8.00	3.11	40.00	12.71	36.80	12.73
	40	e	55	17.59	5.51	43.98	12.69	47.17	12.67
	40	d	55	17.59	5.52	43.96	12.73	40.76	12.72
	20	e	55	8.79	3.09	43.97	12.69	47.19	12.65
	20	d	55	8.80	3.10	44.00	12.71	40.81	12.72
	40	e	08	16.01	6.34	40.04	14.95	43.22	14.92
	40	d	08	16.00	6.32	40.01	14.94	36.78	14.90
	20	e	08	8.00	3.47	40.03	14.95	43.21	14.92
	20	d	08	7.99	3.45	39.98	14.91	36.78	14.89
	40	e	58	17.56	6.32	43.96	14.93	47.12	14.87
	40	d	58	17.59	6.31	43.96	14.93	40.79	14.95
	20	e	58	8.79	3.42	44.00	14.87	47.17	14.78
	20	d	58	8.79	3.45	43.96	14.93	40.78	14.95

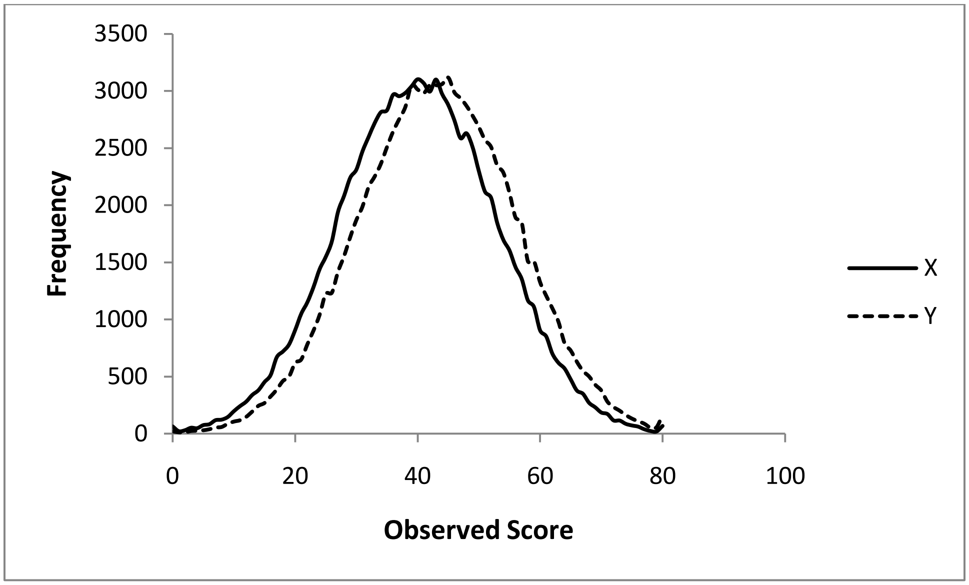
[a]Ability Distribution 05: N(50%, 15%) 55: N(55%, 15%) 08: N(50%, 18%) 58:N(55%, 18%)
[b]Test Y Difficulty e: Test Y is easier d: Test Y is more difficult

As a check for accuracy it can be noted that the means of anchor A observed scores in Table 6 are very similar to the means of anchor A true scores in Table 5. For example, for the second data row for population P in Table 6, the mean A observed score is 16.00 and the corresponding mean for A true score in Table 5 is 15.995. This is due to the fact that error scores should have a mean of 0 and observed scores equal true scores plus error scores.

Another check for accuracy can be done in Table 6 by noting that when test Y is more difficult than test X, the mean of Y observed score should be about 3.2 lower than the mean of X observed score. For example, for the second data row in Table 6, the mean X observed is 39.99 and the mean Y observed is 36.81 which is 3.18 lower than 39.99. When test Y is easier than test X, this situation should be reversed. For example, for the second to last row in Table 6, the mean X observed is 44.00 and the mean Y observed is 47.17 which is 3.17 higher than 44.00.

It is important to indicate that to make the data as close to reality as possible, A, X and Y scores were rounded to the closest integer and that observed scores that were out of range were set equal to the maximum or minimum possible score. For example a score of 81.6 in X or Y was rounded to 80.

As was noted in Chapter Three the data were originally generated for the conditions N (50%, 10%) for P and N (50%, 10%), N (55%, 10%), N (50%, 15%) and N (55, 15%) for Q, but this resulted in such low reliabilities that the results would not be applicable to real data situations. Therefore a decision was made to increase the standard deviations to 15% and 18% instead of 10% and 15%. On the other hand, these new conditions with 15% and 18% SD tended to produce a larger amount of out range scores for certain conditions. This situation as well as the normality of the data and the approximate 3.20 unit shifting between X and Y is illustrated in figures 1 to 10 which contain a frequency graph for the four P conditions and the 16 Q conditions.



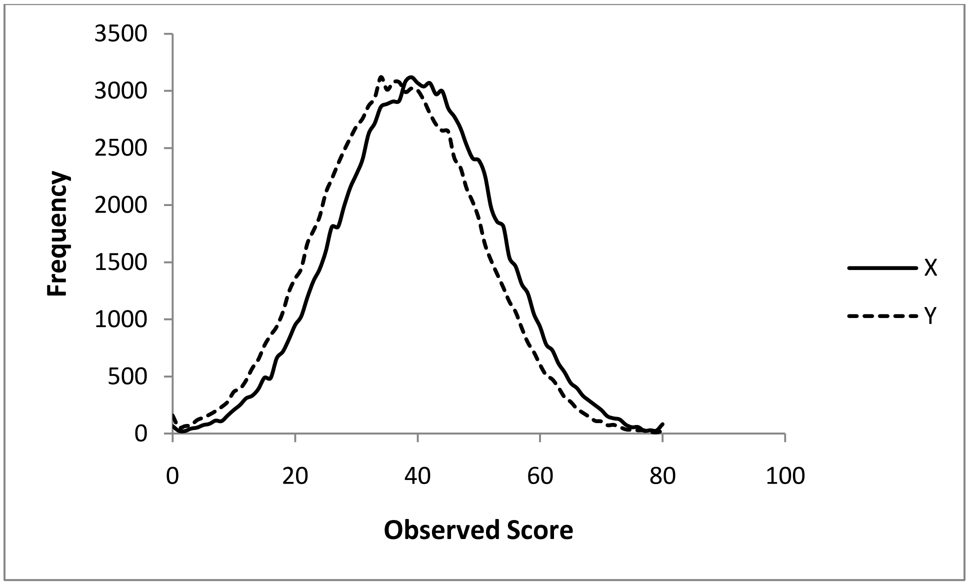
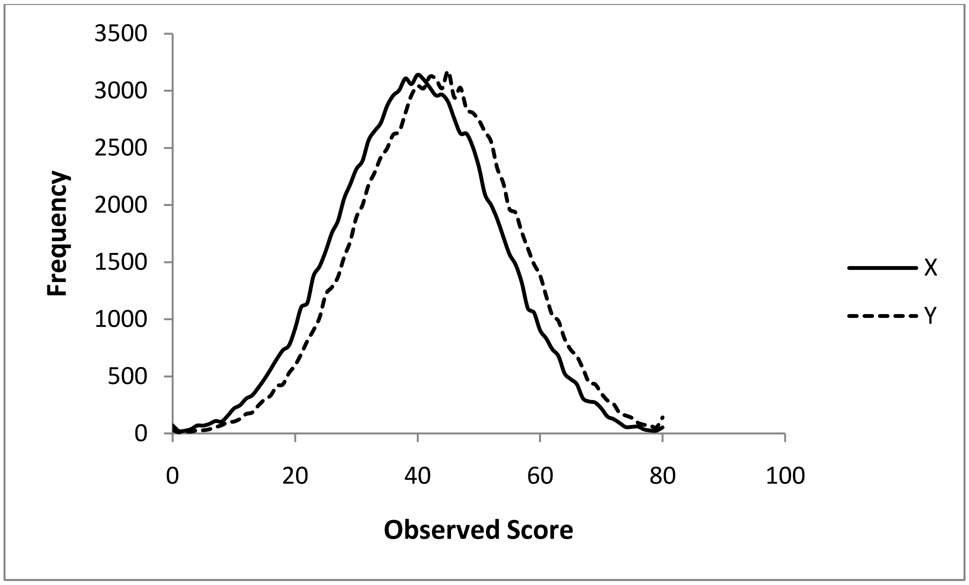


Figure 1. Frequency for X and Y observed scores for population P, anchor 40%, easier test Y and ability N(50%, 15%) (top) and population P, anchor length 40%, more difficult test Y and ability N(50%, 15%) (bottom).



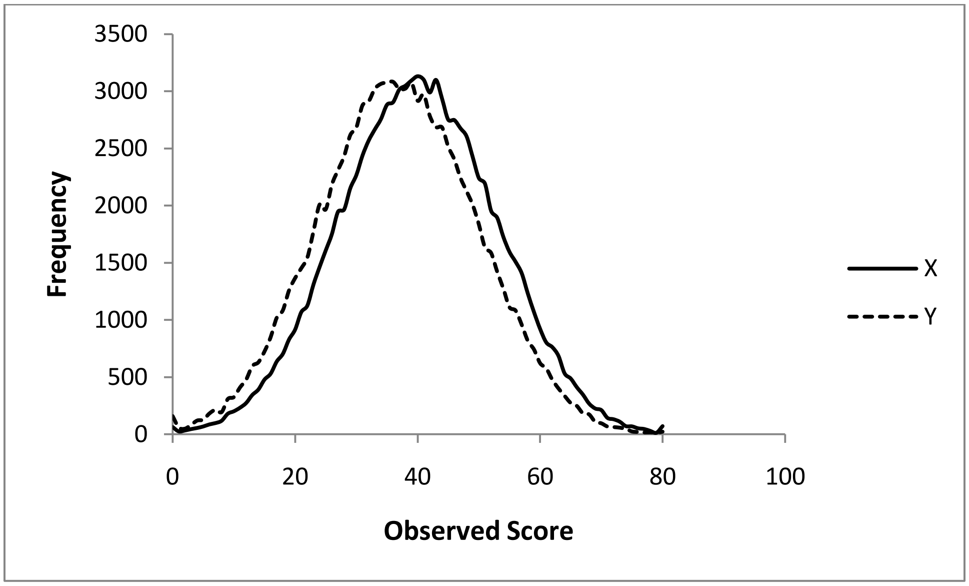
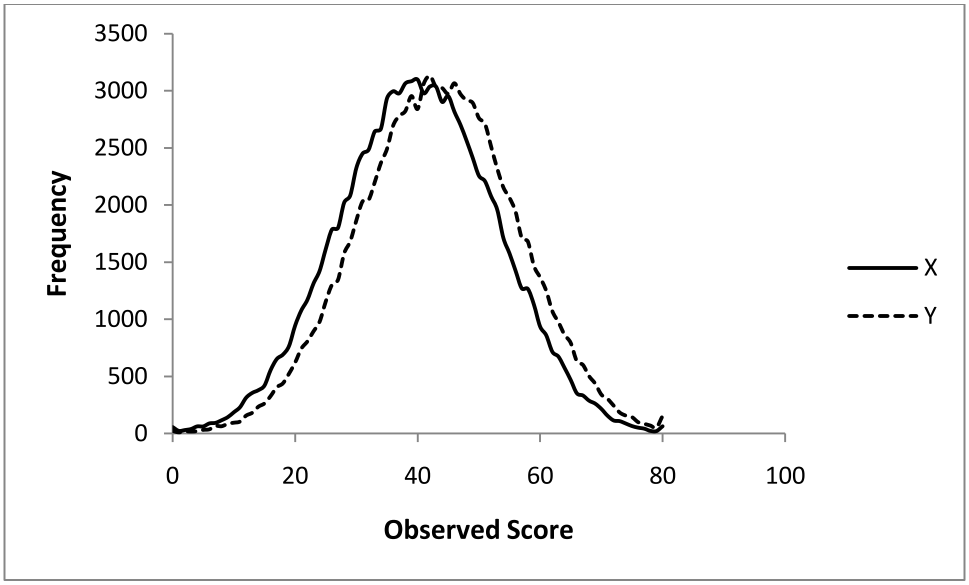


Figure 2. Frequency for X and Y observed scores for population P, anchor 20%, easier test Y and ability N(50%, 15%) (top) and population P, anchor length 20%, more difficult test Y and ability N(50%, 15%) (bottom).



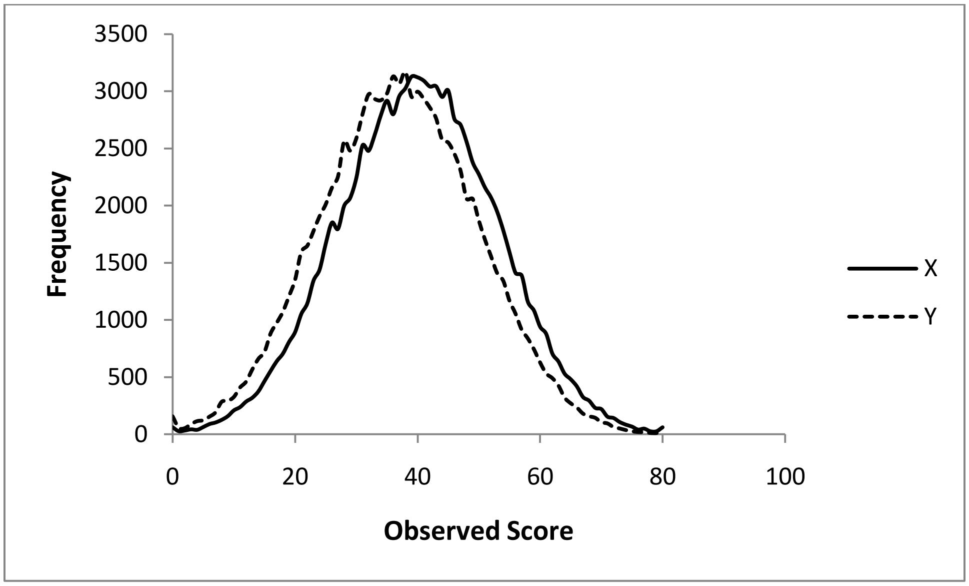


Figure 3. Frequency for X and Y observed scores for population Q, anchor 40%, easier test Y and ability N(50%, 15%) (top) and population Q, anchor length 40%, more difficult test Y and ability N(50%, 15%) (bottom).

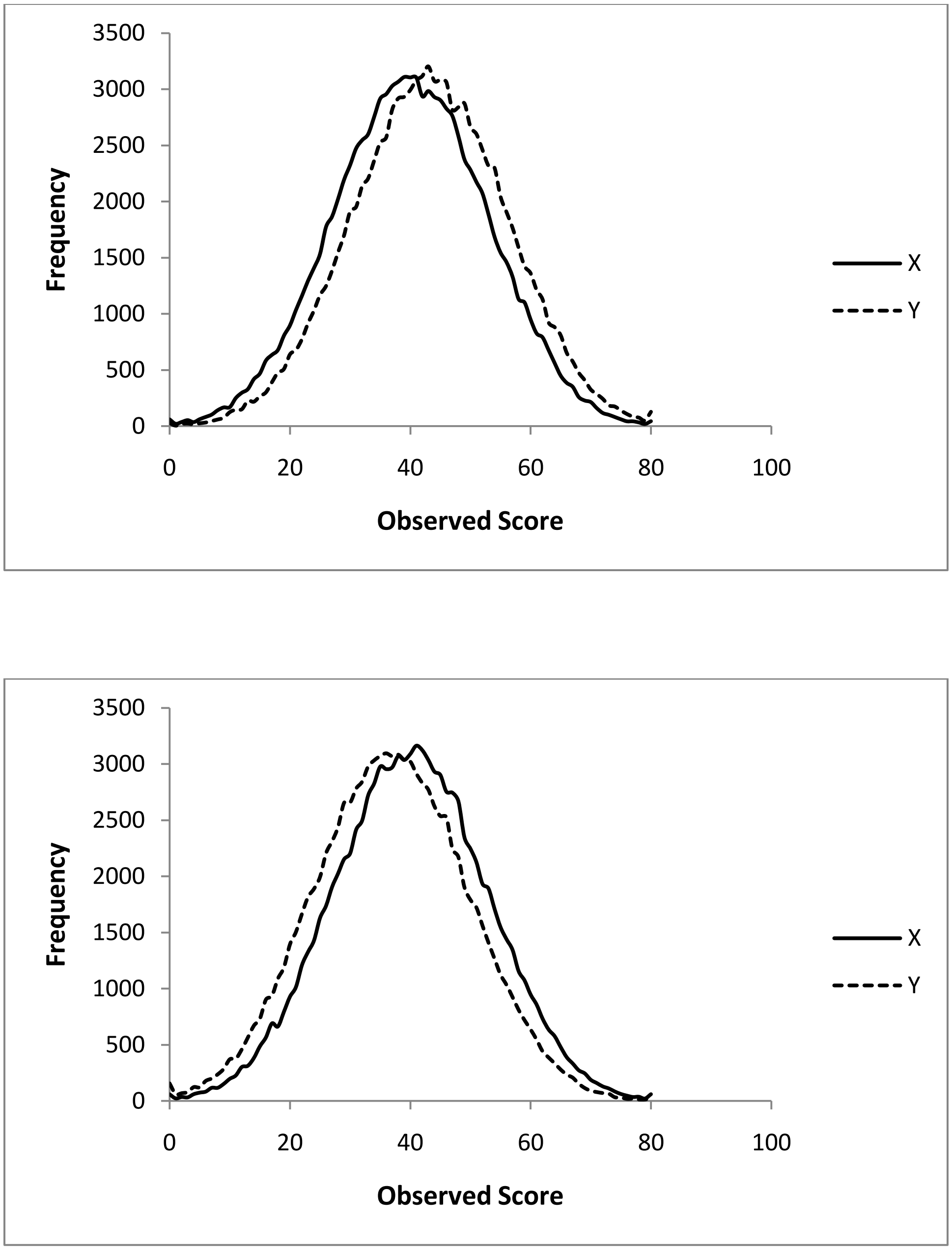
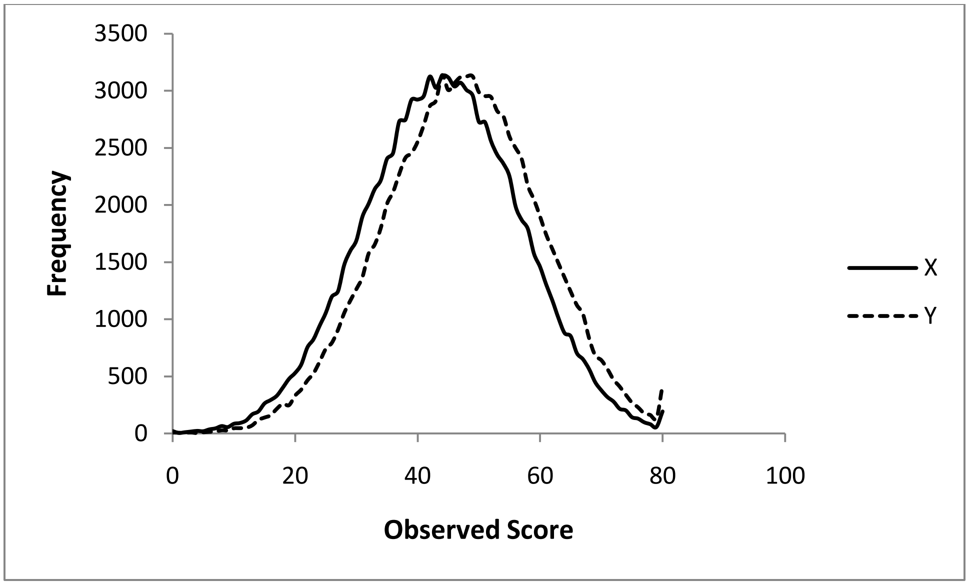


Figure 4. Frequency for X and Y observed scores for population Q, anchor 20%, easier test Y and ability N(50%, 15%) (top) and population Q, anchor length 20%, more difficult test Y and ability N(50%, 15%) (bottom).



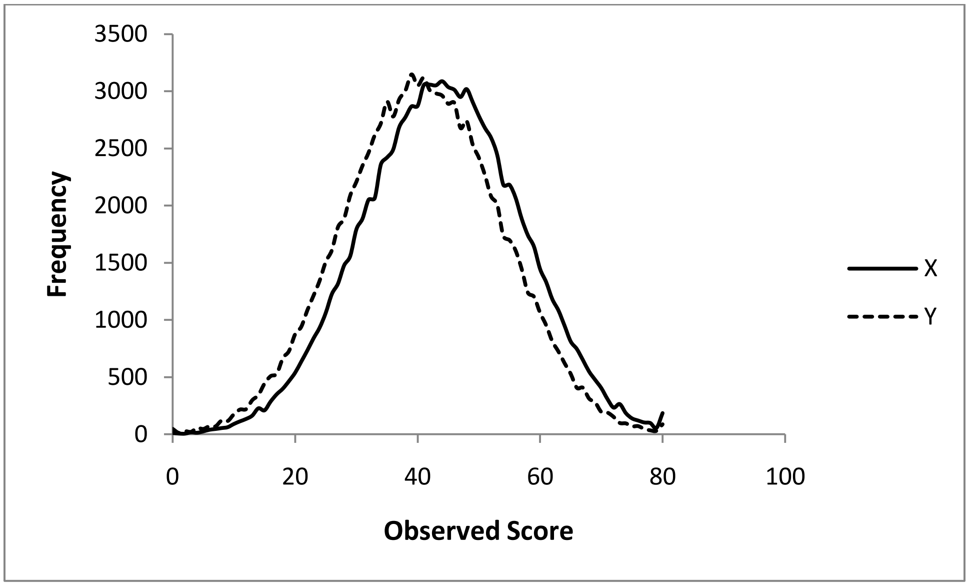
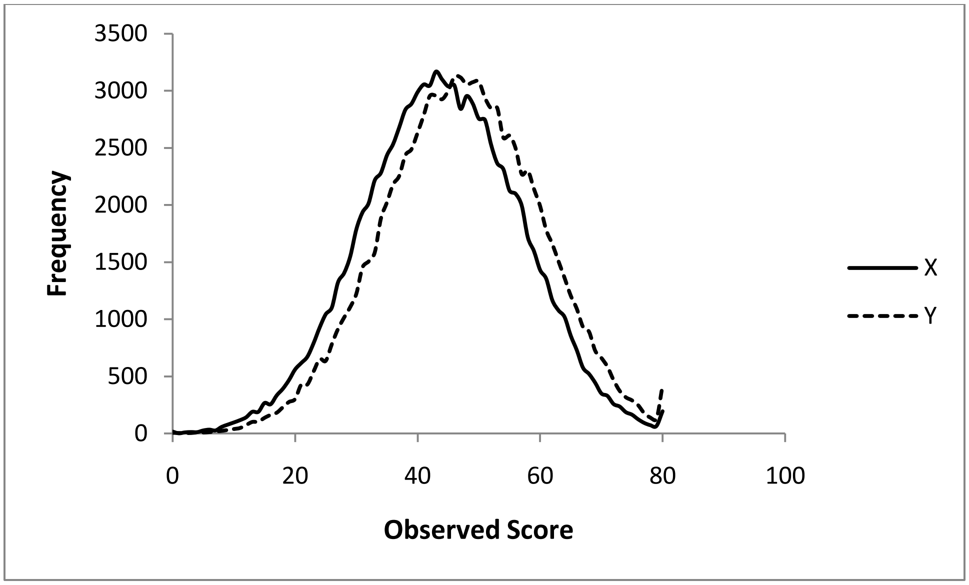


Figure 5. Frequency for X and Y observed scores for population Q, anchor 40%, easier test Y and ability N(55%, 15%) (top) and population Q, anchor length 40%, more difficult test Y and ability N(55%, 15%) (bottom).



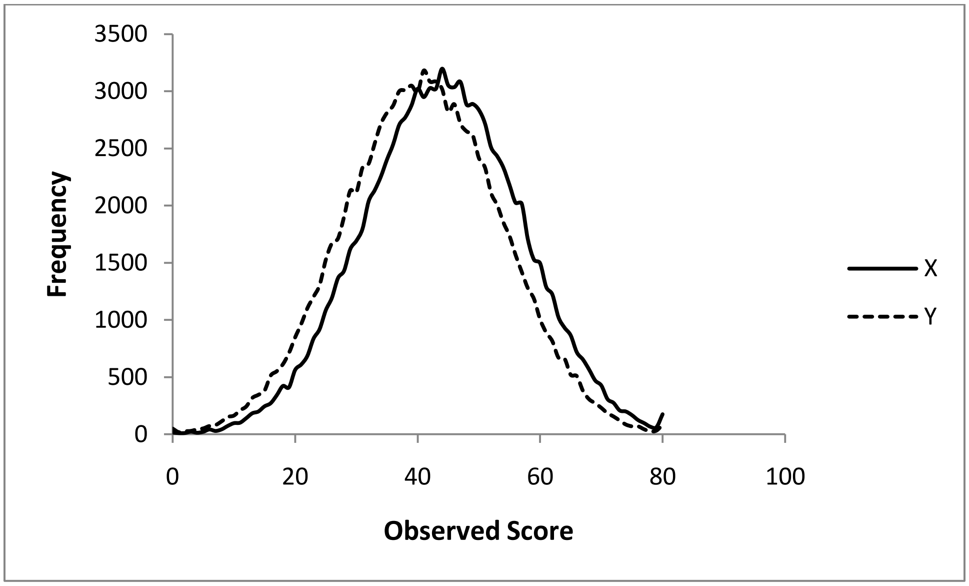
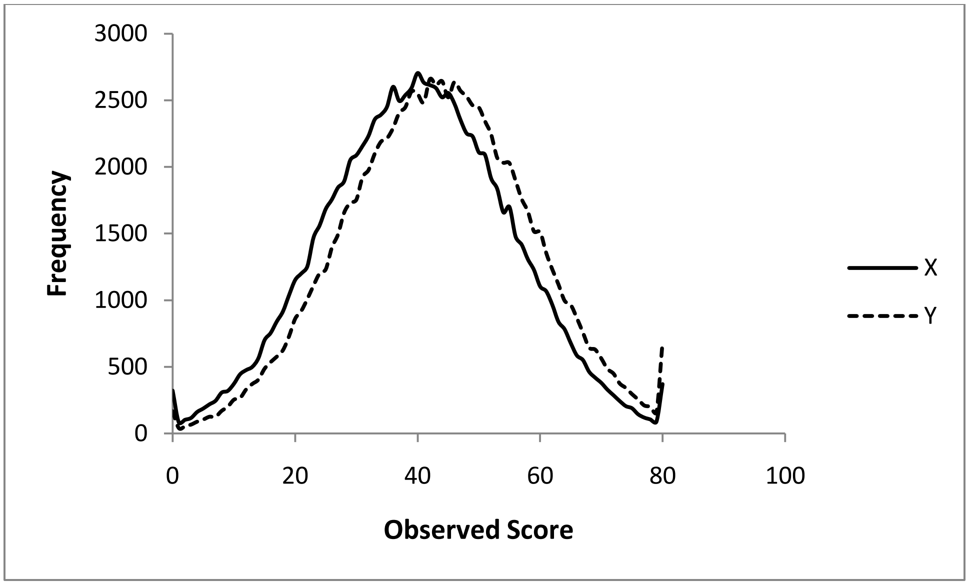


Figure 6. Frequency for X and Y observed scores for population Q, anchor 20%, easier test Y and ability N(55%, 15%) (top) and population Q, anchor length 20%, more difficult test Y and ability N(55%, 15%) (bottom).



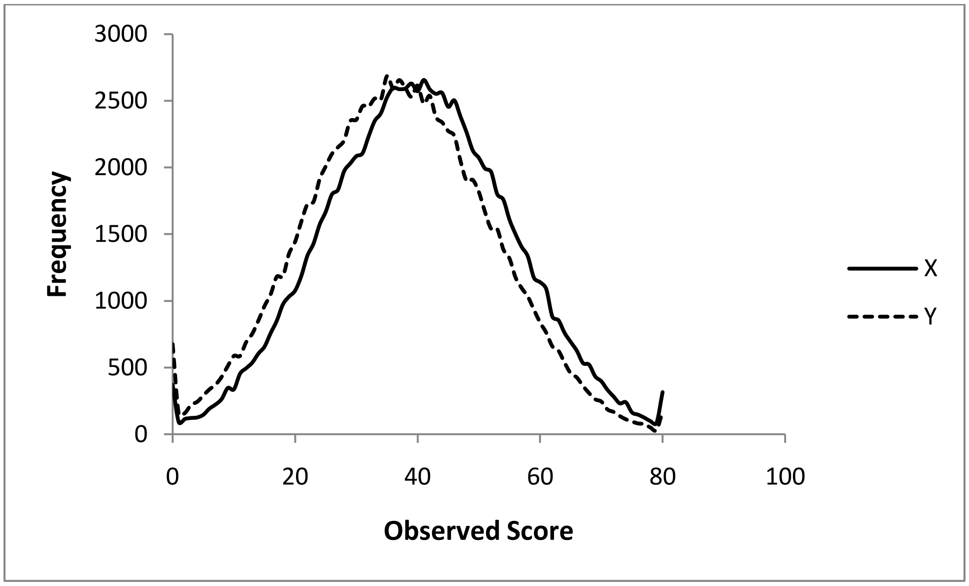
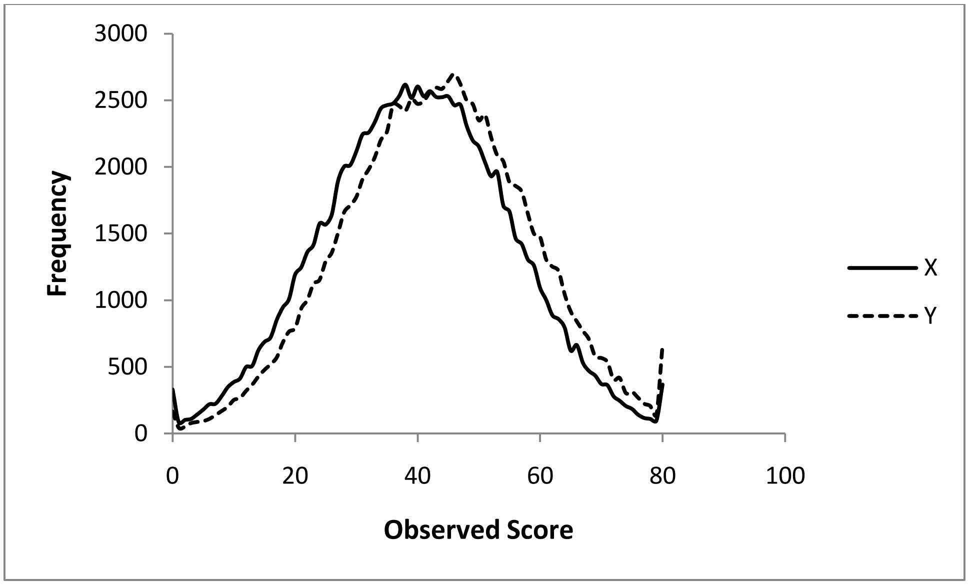


Figure 7. Frequency for X and Y observed scores for population Q, anchor 40%, easier test Y and ability N(50%, 18%) (top) and population Q, anchor length 40%, more difficult test Y and ability N(50%, 18%) (bottom).



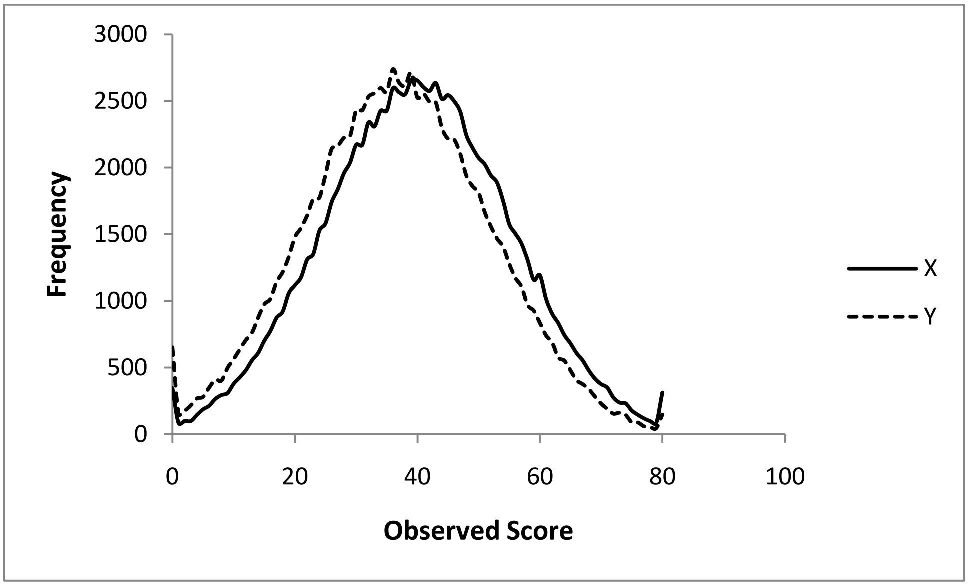
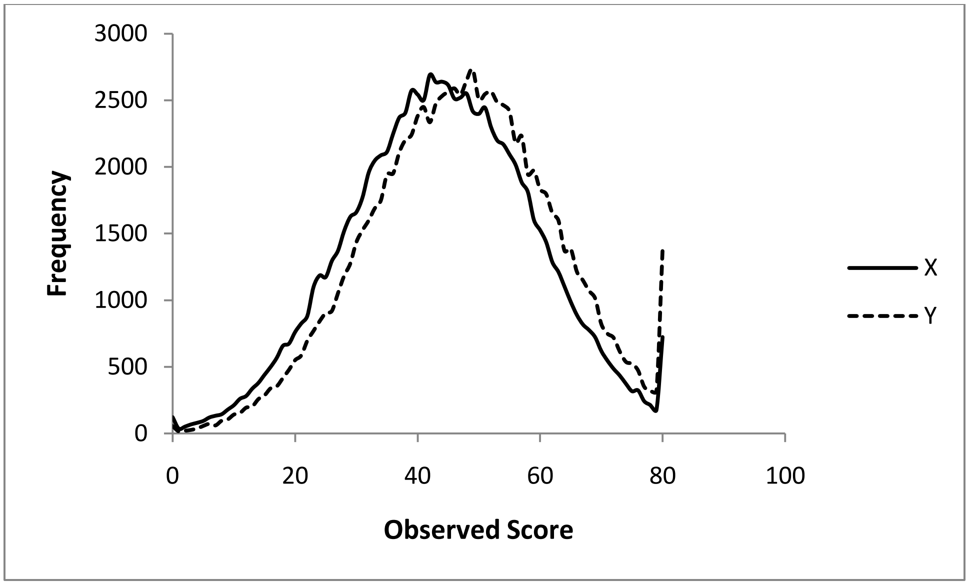


Figure 8. Frequency for X and Y observed scores for population Q, anchor 20%, easier test Y and ability N(50%, 18%) (top) and population Q, anchor length 20%, more difficult test Y and ability N(50%, 18%) (bottom).



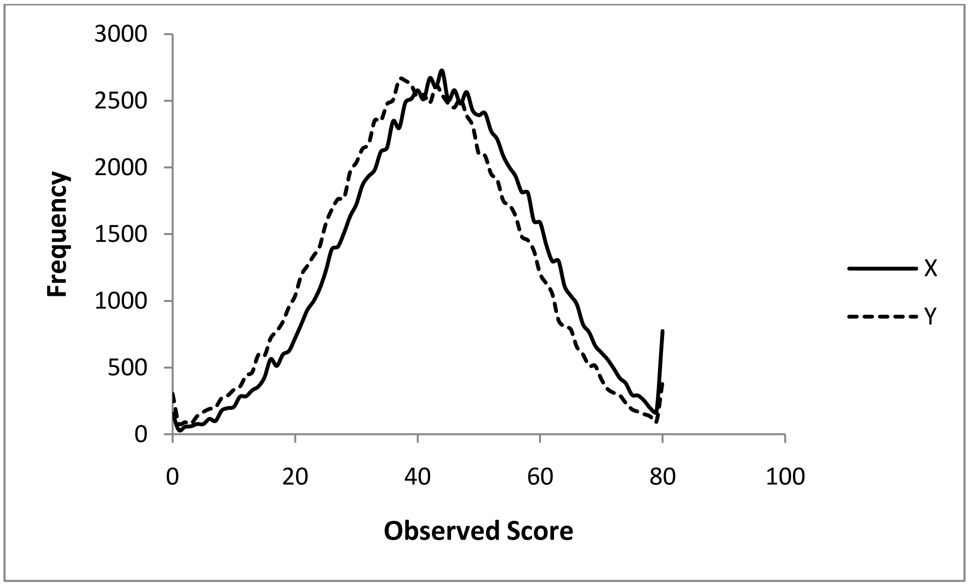
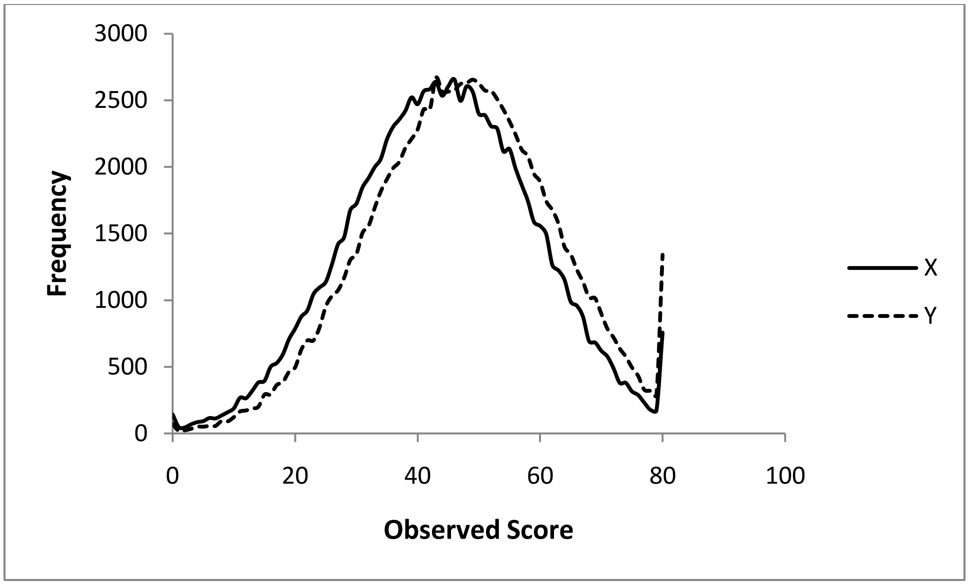


Figure 9. Frequency for X and Y observed scores for population Q, anchor 40%, easier test Y and ability N(55%, 18%) (top) and population Q, anchor length 40%, more difficult test Y and ability N(55%, 18%) (bottom).



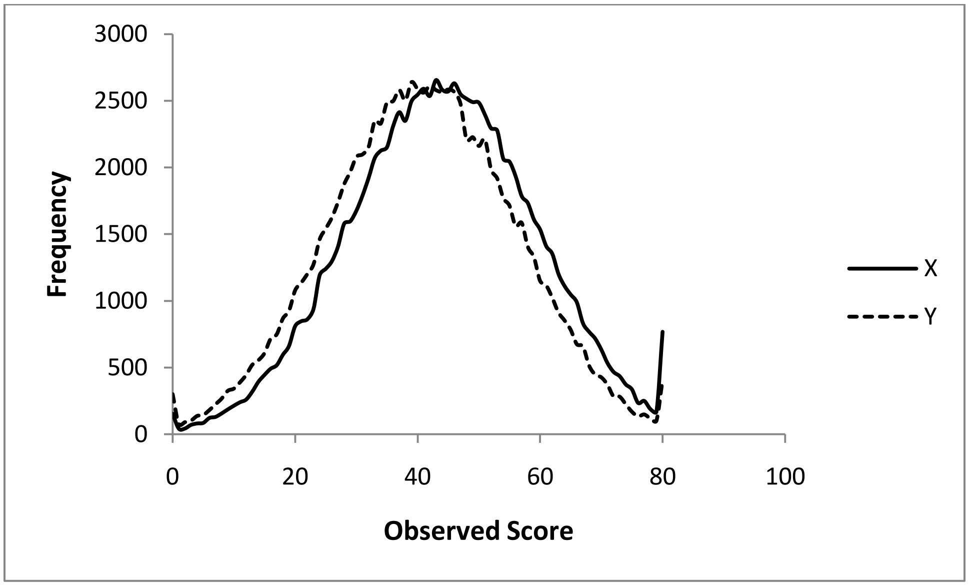


Figure 10. Frequency for X and Y observed scores for population Q, anchor 20%, easier test Y and ability N(55%, 18%) (top) and population Q, anchor length 20%, more difficult test Y and ability N(55%, 18%) (bottom).

As can be noted in the previous figures, the accumulation of out of range occurs to a greater extent in the conditions that have an 18% SD and it is accentuated for conditions in which the test Y is easier. On the other hand, the accumulation of out of range scores on the right side of the distribution at some extent models a ceiling effect in which very able subjects are not allowed to score higher than the maximum score in the test.

To further explore the impact of this out of range scores in the shape of the distribution of scores, the skewness was computed for X and Y observed scores across the 20 conditions and are presented in Table 7.

Table 7

Skewness for Tests X and Y Observed Scores in Populations P and Q

Population	Anchor Length %	Test Difficulty	Ability Distribution	Skewness Test X	Skewness Test Y
P	40	e	05	0.0005	-0.0190
	40	d	05	-0.0009	0.0218
	20	e	05	-0.0075	-0.0331
	20	d	05	-0.0041	0.0225
Q	40	e	05	-0.0046	-0.0271
	40	d	05	-0.0048	0.0140
	20	e	05	-0.0104	-0.0242
	20	d	05	0.0003	0.0172
	40	e	55	-0.0331	-0.0569
	40	d	55	-0.0242	0.0025
	20	e	55	-0.0270	-0.0550
	20	d	55	-0.0332	-0.0103
	40	e	08	0.0077	-0.0323
	40	d	08	0.0025	0.0346
	20	e	08	0.0055	-0.0286
	20	d	08	-0.0034	0.0329
	40	e	58	-0.0501	-0.0899
	40	d	58	-0.0458	-0.0078
	20	e	58	-0.0543	-0.0930
	20	d	58	-0.0516	-0.0188

[a]Ability Distribution 05: N(50%, 15%) 55: N(55%, 15%) 08: N(50%, 18%) 58:N(55%, 18%)

[b]Test Y Difficulty e: Test Y is easier d: Test Y is more difficult

The data in Table 7 confirms that the largest skewness occur within the conditions that include N(55%, 18%), however, the magnitude of the skewness is small. Given that the data generation with SD 15% and 18% produced much higher reliabilities than the data generation with SD 10% and 15%, that the former models to some extent the ceiling effect whereby very able students are not allowed to score higher than the maximum score in the test and that the skewness produced by the former can be considered minor, the SD 15% and 18% were kept for this study.

Table 8

Reliability for Generated Data for Anchor A, and Tests X and Y in Populations P and Q

Population	Anchor Length %	Test Difficulty	Ability Distribution	Reliability A	Reliability X	Reliability Y
	40	e	05	0.76	0.89	0.89
P	40	d	05	0.75	0.89	0.89
	20	e	05	0.59	0.89	0.89
	20	d	05	0.60	0.89	0.89
	40	e	05	0.75	0.89	0.89
	40	d	05	0.75	0.89	0.89
	20	e	05	0.59	0.89	0.89
	20	d	05	0.59	0.89	0.89
	40	e	55	0.76	0.89	0.89
	40	d	55	0.75	0.89	0.89
	20	e	55	0.60	0.89	0.89
Q	20	d	55	0.60	0.89	0.89
	40	e	08	0.82	0.92	0.92
	40	d	08	0.82	0.92	0.92
	20	e	08	0.69	0.92	0.92
	20	d	08	0.69	0.92	0.92
	40	e	58	0.82	0.92	0.92
	40	d	58	0.82	0.92	0.92
	20	e	58	0.69	0.92	0.92
	20	d	58	0.69	0.92	0.92

[a]Ability Distribution 05: N(50%, 15%) 55: N(55%, 15%) 08: N(50%, 18%) 58:N(55%, 18%)
[b]Test Y Difficulty e: Test Y is easier d: Test Y is more difficult

As can be noted in Table 8, the highest reliabilities occur within populations with greater SD as is the case with conditions with 18% SD. This makes sense, as the data generated more variable true scores within these conditions and greater true score variance means higher reliability. On the other hand, the original data generation whereby the ability distributions for anchor true scores were N(50%, 10%) for P and N(50%, 10%), N(55%, 10%), N(50%, 15%) and N(55, 15%) for Q, produced reliabilities as low as 0.35 for the anchor A and therefore that original data generation was discarded.

Equating in the Synthetic Populations

Each one of the 16 conditions for Q was matched with the corresponding condition for P. For example, the last condition in Table 7, which is anchor length 20%, test Y more difficult than test X and ability distribution N(55%, 18%) was matched with the last condition for P which is anchor length 20%, test Y more difficult than test X and ability distribution N(50%, 15%). Since for population P the ability distribution is always N(50%, 15%), each condition for P got matched with four conditions for Q. This matching gave origin to 16 cases for each of which five different mixtures of synthetic populations were created by mixing P and Q into populations S of the synthetic form $S = wP + (1 - w)Q$, where $0 \leq w \leq 1$ (Braun & Holland, 1982). This was accomplished by using 5 values for *w:* 0, 0.25, 0.50, 0.75, and 1. For example, when w=0.25, a random 25% of P and a random 75% of Q were combined to create a population S for that condition.

This produced 80 different synthetic populations. For each of these 80 synthetic populations, a true equating function was computed as described in Chapter Three. In addition, for each synthetic population an equating function using the L3 assumption and another equating function using L3* were computed as also described in Chapter Three.

These triplets of equating functions were compared by using the following procedure:

a) Recall that each Population P and Q had 100,000 subjects. Therefore each synthetic population *S* has 100,000 subjects as well. For each of the 100,000 subjects in S the corresponding X score were plugged into each of the equating functions (true, L3 and L3*).

b) The amount of bias for each subject was computed for both L3 and L3* equating functions using the true Y equated score as the criterion.

c) An average of the amount of bias was computed across the 100,000 subjects in S.

d) For each subject the bias was squared and then an average across the synthetic population was computed and the square root obtained to get the RMSE (root mean squared error).

The resulting bias and RMSE are shown in Tables 9 and 10.

Table 9

*Bias for the 80 Synthetic Populations for L3 and L3**

Anchor	Test		w									
length	Y		L3					L3*				
%	Diff		0.00	0.25	0.50	0.75	1.00	0.00	0.25	0.50	0.75	1.00
40	e	05	0.04	0.06	0.04	0.05	0.05	0.04	0.06	0.04	0.05	0.05
40	d	05	0.03	0.01	0.00	0.01	0.01	0.03	0.01	0.00	0.01	0.01
20	e	05	-0.07	-0.07	-0.06	-0.08	-0.09	-0.07	-0.07	-0.06	-0.08	-0.09
20	d	05	0.00	-0.01	-0.01	0.00	0.01	0.00	-0.01	-0.01	0.00	0.01
40	e	55	-0.01	0.00	-0.01	0.00	0.01	-0.01	0.00	-0.01	0.00	0.01
40	d	55	0.00	-0.02	-0.04	-0.03	-0.01	0.00	-0.02	-0.04	-0.03	-0.01
20	e	55	0.00	0.02	0.03	0.02	0.03	0.00	0.02	0.03	0.02	0.03
20	d	55	-0.03	-0.03	-0.02	0.00	-0.01	-0.03	-0.03	-0.02	0.00	-0.01
40	e	08	0.00	0.00	-0.01	-0.01	0.01	0.00	0.00	-0.01	-0.01	0.01
40	d	08	0.03	0.01	0.00	-0.01	-0.02	0.03	0.01	0.00	-0.01	-0.02
20	e	08	-0.03	-0.02	-0.04	-0.05	-0.05	-0.03	-0.02	-0.04	-0.05	-0.05
20	d	08	0.03	0.04	0.05	0.04	0.05	0.03	0.04	0.05	0.04	0.05
40	e	58	0.05	0.06	0.05	0.06	0.05	0.05	0.06	0.05	0.06	0.05
40	d	58	0.00	0.01	-0.01	0.00	0.01	0.00	0.01	-0.01	0.00	0.01
20	e	58	-0.01	0.00	0.01	-0.01	-0.02	-0.01	0.00	0.01	-0.01	-0.02
20	d	58	-0.03	-0.02	-0.01	-0.01	0.00	-0.03	-0.02	-0.01	-0.01	0.00

[a] Ability Distribution 05: N(50%, 15%) 55: N(55%, 15%) 08: N(50%, 18%) 58:N(55%, 18%)

[b] Test Y Diff. e: Test Y is easier d: Test Y is more difficult.

Table 10

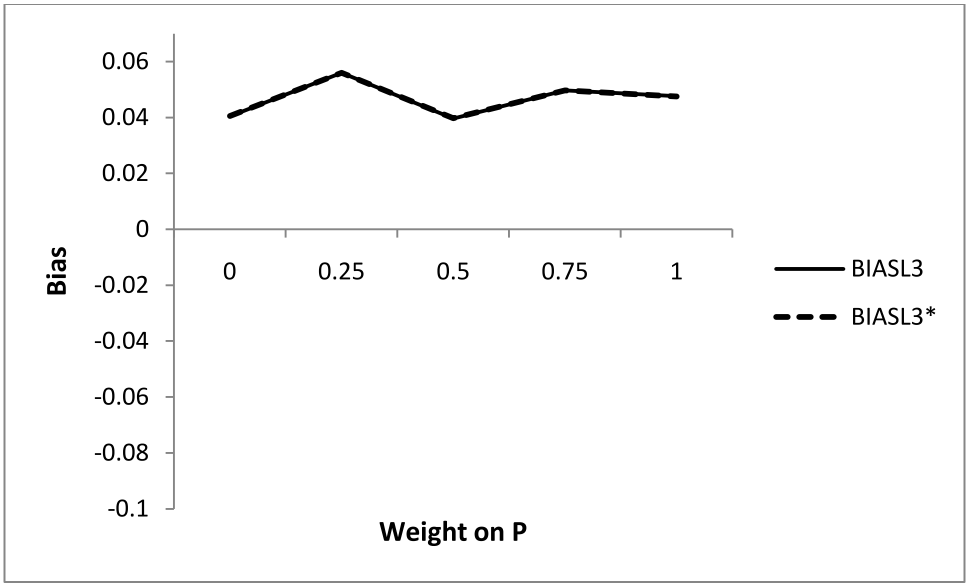
*RMSE for the 80 Synthetic Populations for L3 and L3**

Anchor	Test		w									
Length	Y		L3					L3*				
%	Diff		0.00	0.25	0.50	0.75	1.00	0.00	0.25	0.50	0.75	1.00
40	e	05	0.06	0.08	0.07	0.08	0.07	0.05	0.07	0.06	0.07	0.06
40	d	05	0.04	0.01	0.00	0.01	0.01	0.04	0.01	0.00	0.01	0.01
20	e	05	0.07	0.07	0.06	0.08	0.09	0.07	0.07	0.06	0.08	0.09
20	d	05	0.05	0.04	0.03	0.02	0.02	0.05	0.04	0.03	0.02	0.02
40	e	55	0.02	0.01	0.03	0.04	0.04	0.02	0.02	0.04	0.04	0.04
40	d	55	0.04	0.04	0.05	0.04	0.03	0.03	0.03	0.05	0.04	0.02
20	e	55	0.16	0.15	0.14	0.12	0.11	0.10	0.09	0.08	0.06	0.06
20	d	55	0.03	0.03	0.02	0.03	0.02	0.03	0.03	0.02	0.01	0.01
40	e	08	0.02	0.02	0.02	0.01	0.01	0.33	0.32	0.30	0.28	0.27
40	d	08	0.10	0.11	0.10	0.09	0.08	0.41	0.41	0.37	0.35	0.32
20	e	08	0.17	0.16	0.15	0.16	0.17	0.87	0.82	0.77	0.75	0.72
20	d	08	0.10	0.11	0.10	0.09	0.10	0.77	0.75	0.70	0.67	0.64
40	e	58	0.07	0.07	0.06	0.06	0.05	0.36	0.34	0.32	0.28	0.26
40	d	58	0.11	0.12	0.13	0.12	0.14	0.42	0.42	0.41	0.38	0.38
20	e	58	0.34	0.32	0.32	0.31	0.29	0.97	0.92	0.88	0.82	0.75
20	d	58	0.16	0.17	0.19	0.20	0.20	0.82	0.81	0.80	0.77	0.73

[a] Ability Distribution 05: N(50%, 15%) 55: N(55%, 15%) 08: N(50%, 18%) 58:N(55%, 18%)

[b] Test Y Diff. e: Test Y is easier d: Test Y is more difficult

Figures 11 to 26 illustrate the values of bias and RMSE for the 80 synthetic populations.



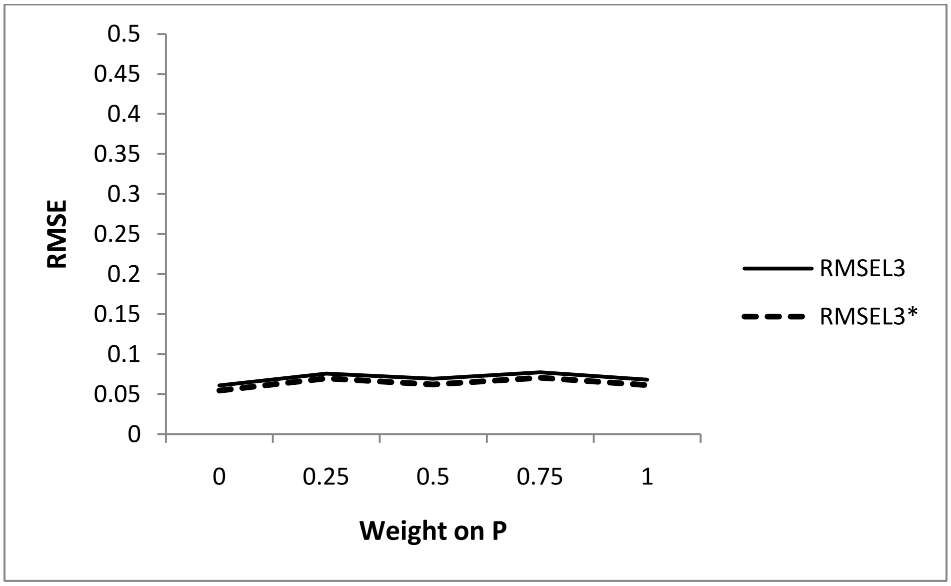
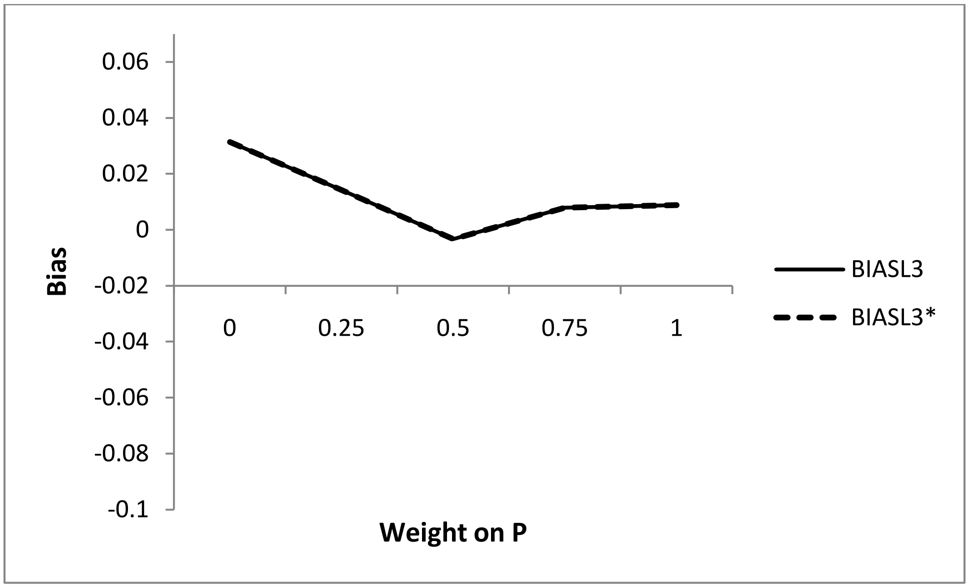


Figure 11. Bias (top) and RMSE (bottom) for five synthetic populations, anchor 40%, easier test Y and ability N(50%, 15%).



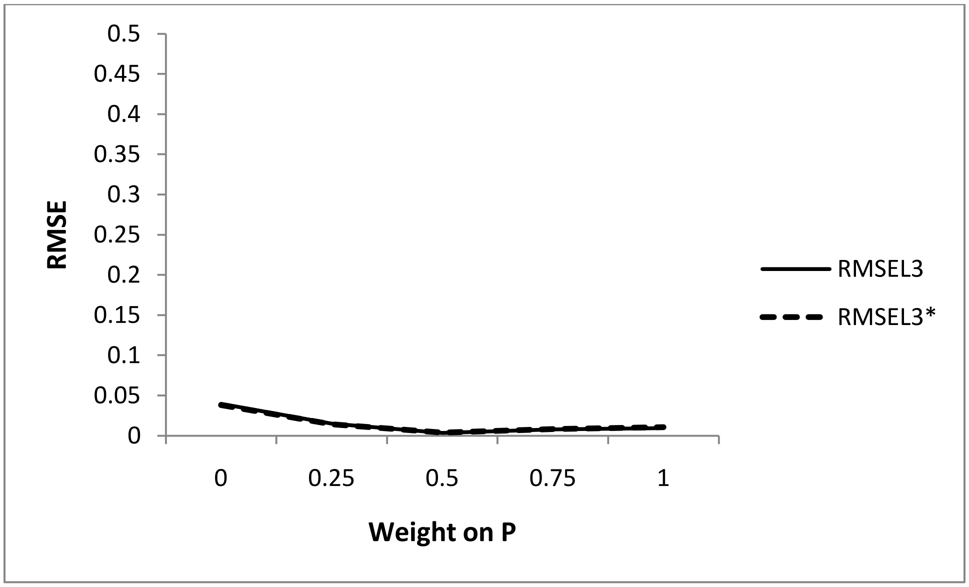
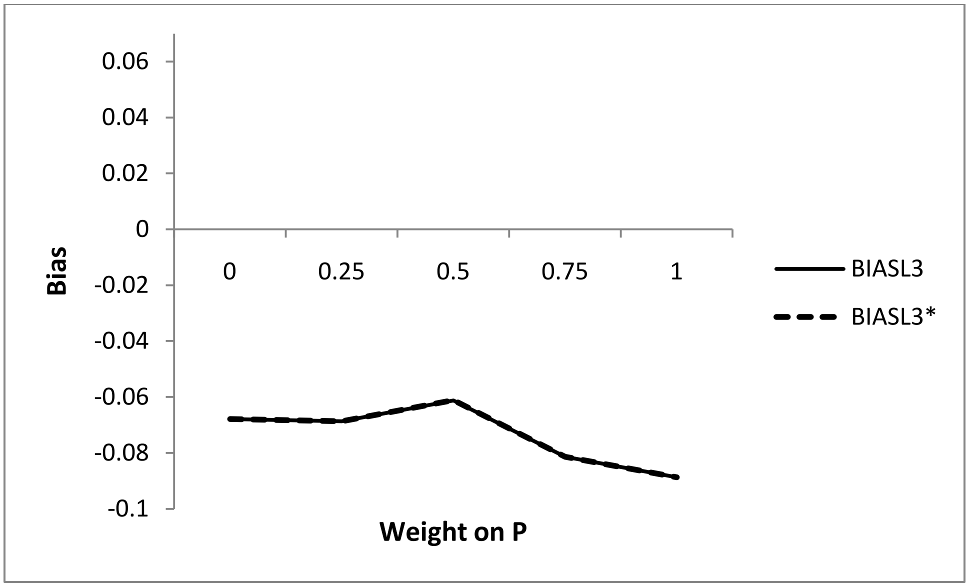


Figure 12. Bias (top) and RMSE (bottom) for five synthetic populations, anchor 40%, more difficult test Y and ability N(50%, 15%).



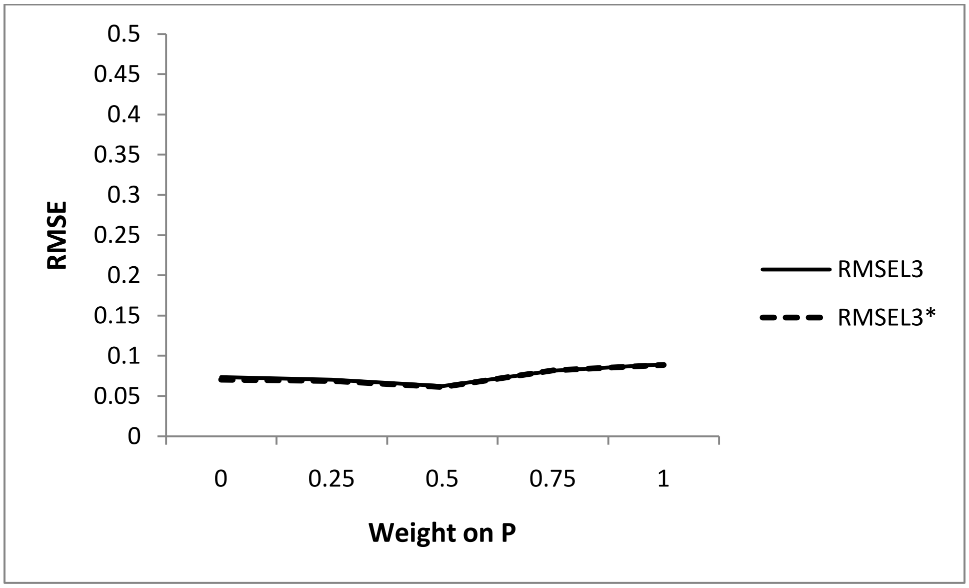


Figure 13. Bias (top) and RMSE (bottom) for five synthetic populations, anchor 20%, easier test Y and ability N(50%, 15%).

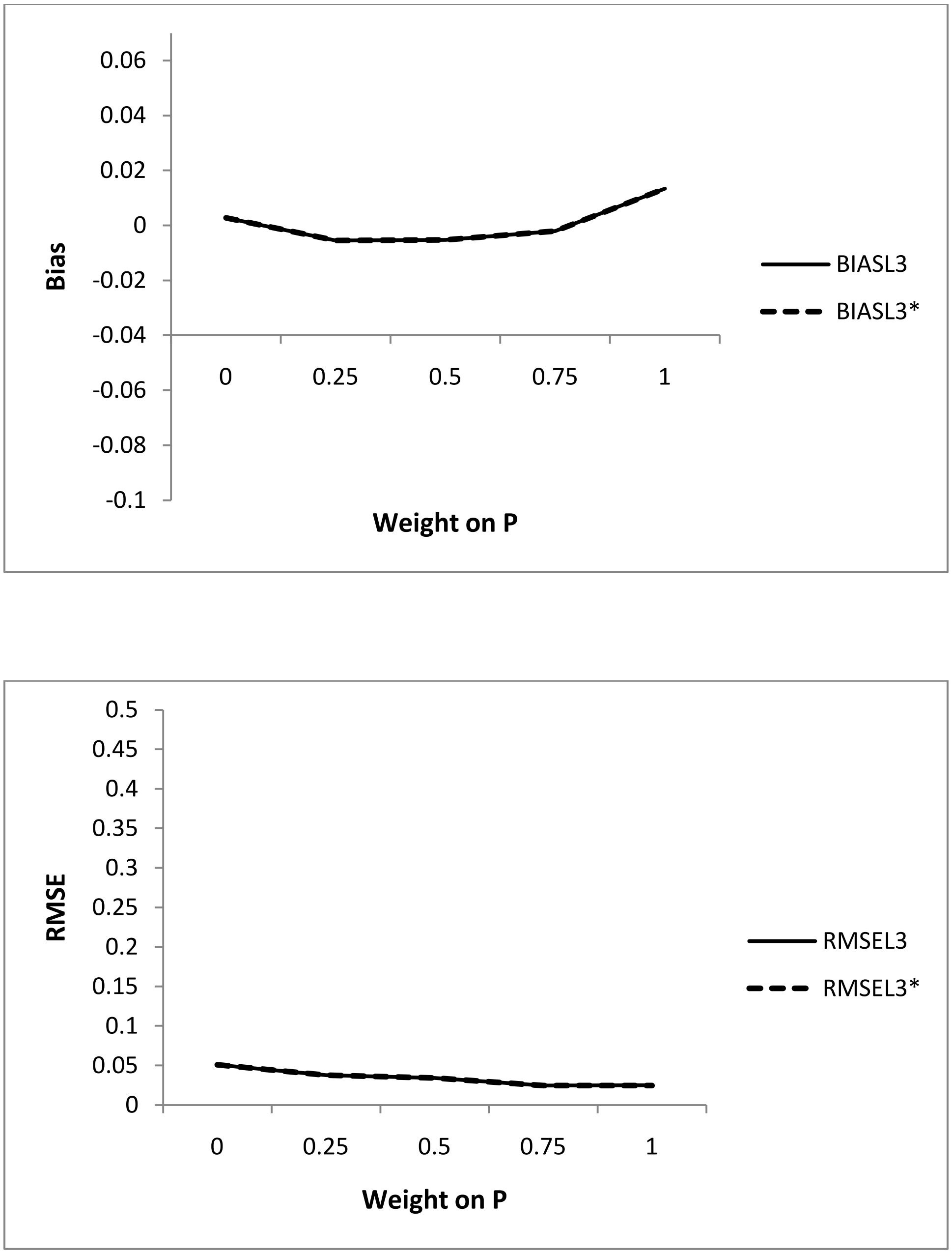


Figure 14. Bias (top) and RMSE (bottom) for five synthetic populations, anchor 20%, more difficult test Y and ability N(50%, 15%).

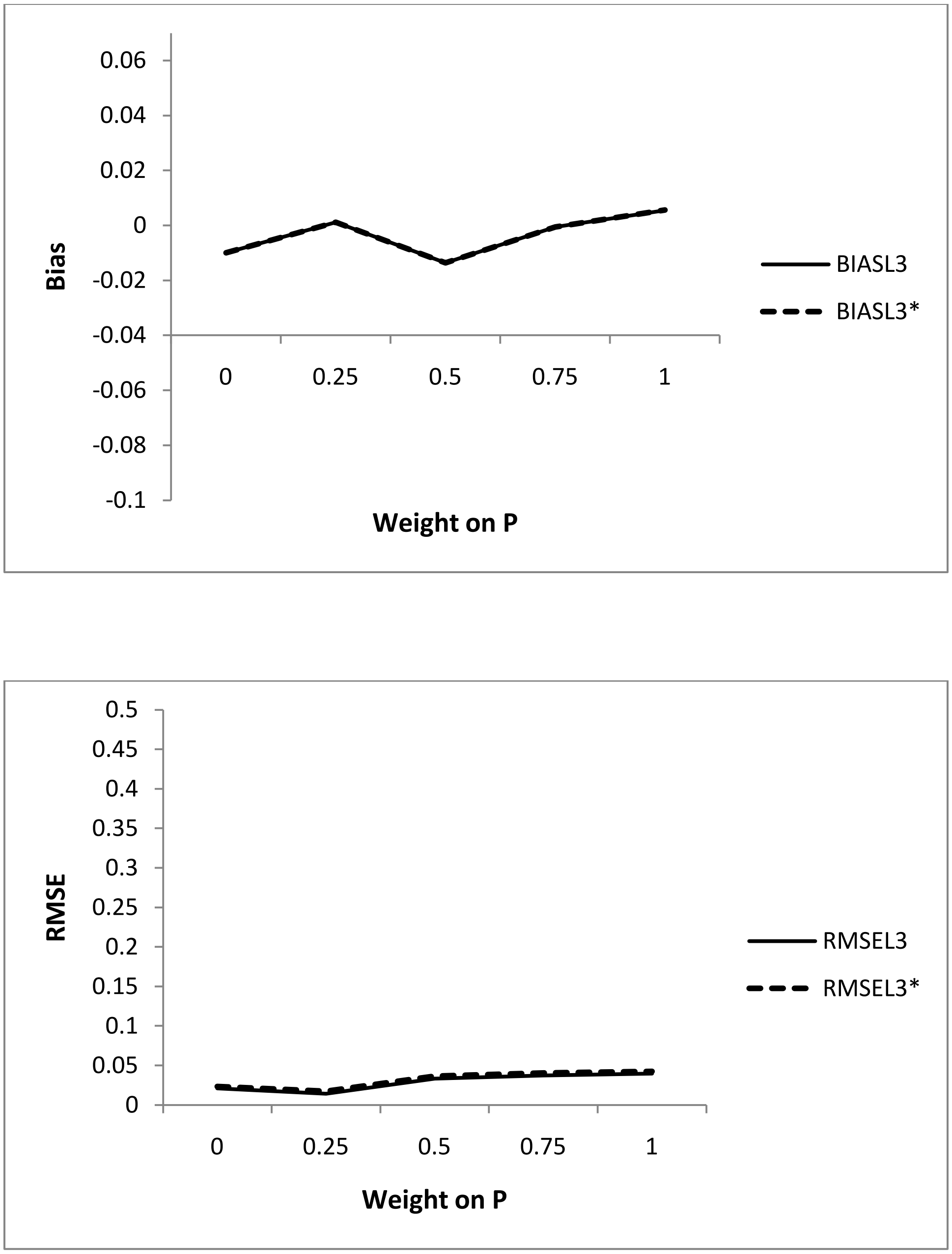
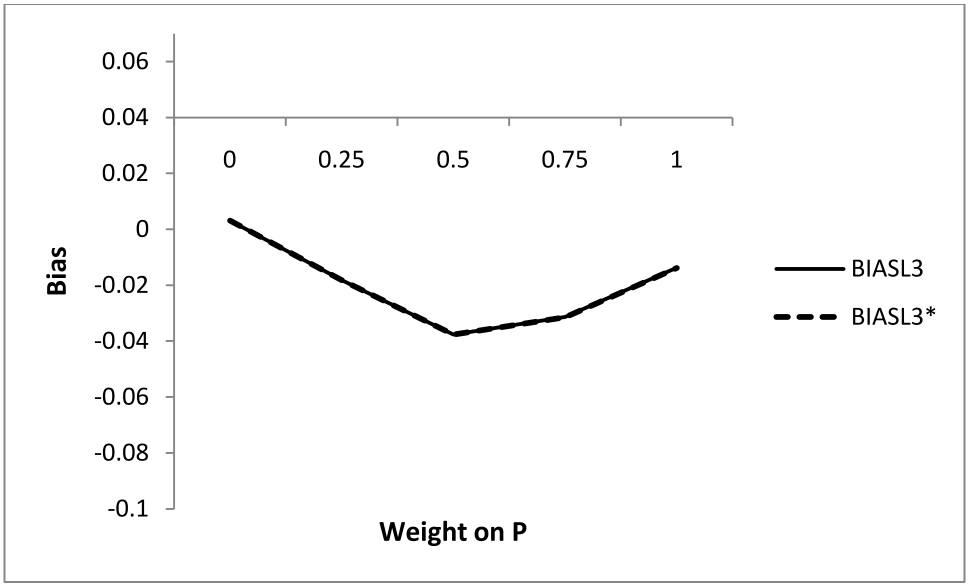


Figure 15. Bias (top) and RMSE (bottom) for five synthetic populations, anchor 40%, easier test Y and ability N(55%, 15%).



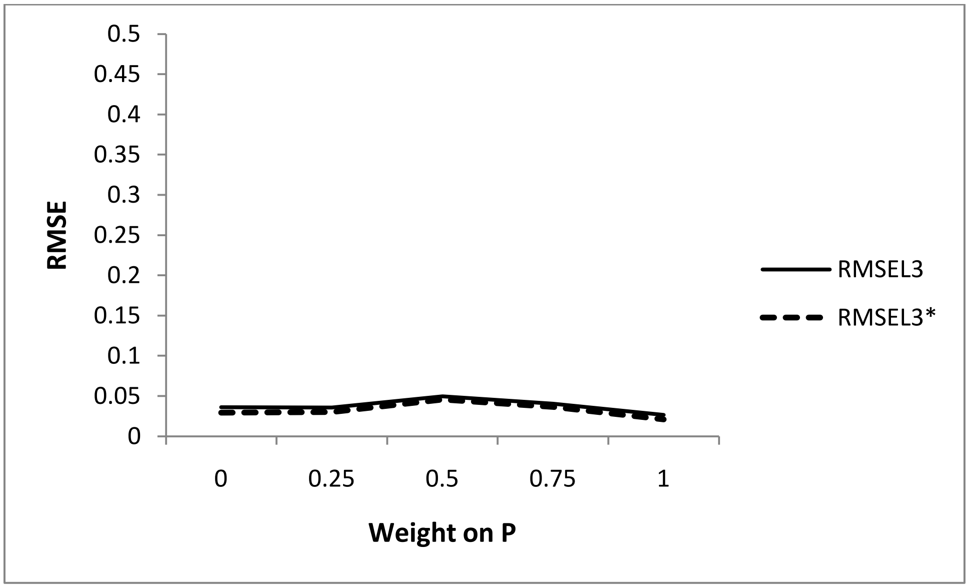
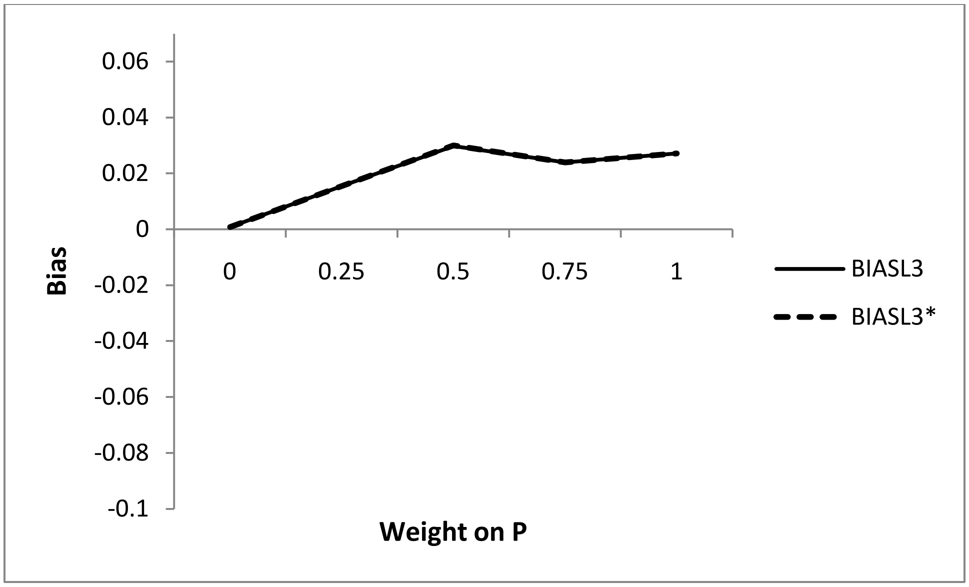


Figure 16. Bias (top) and RMSE (bottom) for five synthetic populations, anchor 40%, more difficult test Y and ability N(55%, 15%).



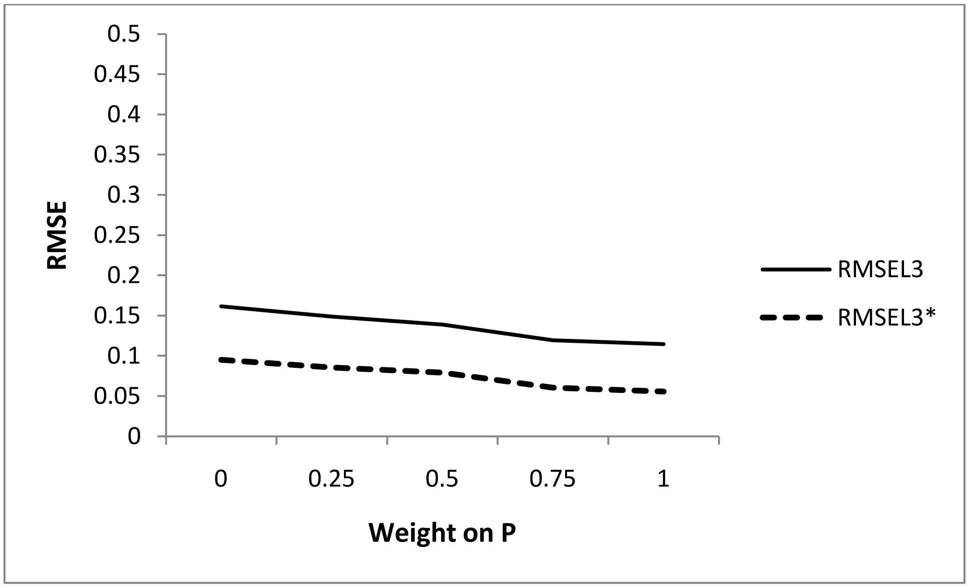


Figure 17. Bias (top) and RMSE (bottom) for five synthetic populations, anchor 20%, easier test Y and ability N(55%, 15%).

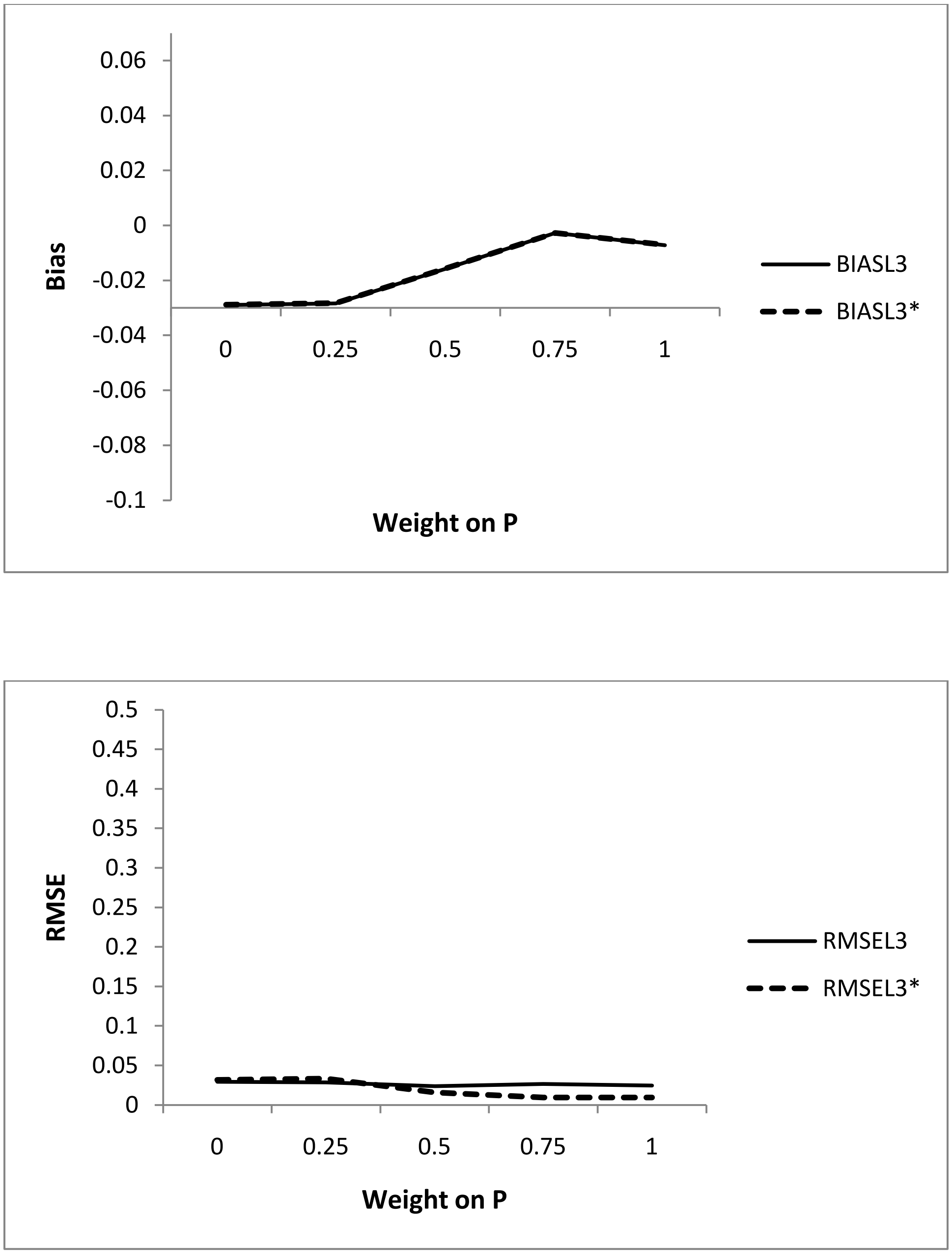
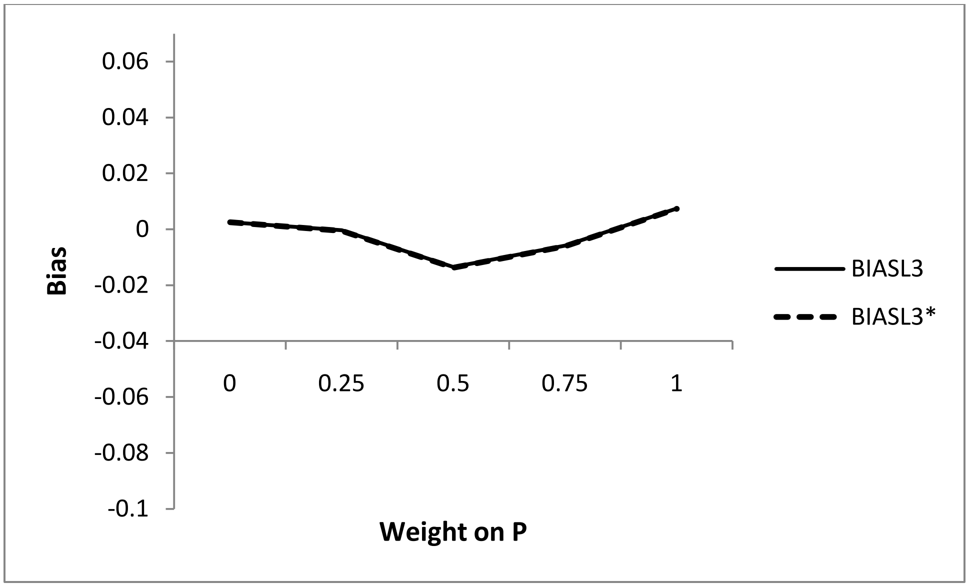


Figure 18. Bias (top) and RMSE (bottom) for five synthetic populations, anchor 20%, more difficult test Y and ability N(55%, 15%).



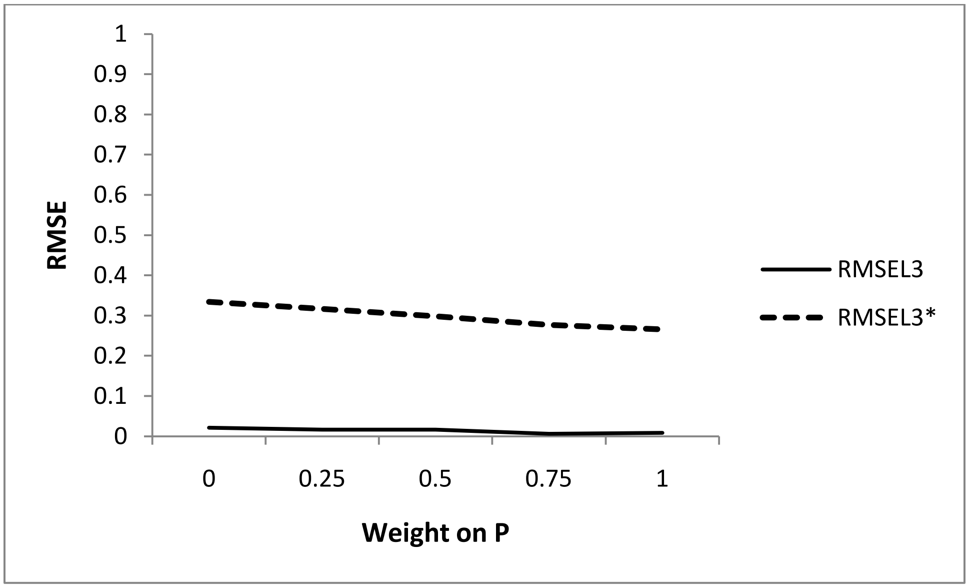


Figure 19. Bias (top) and RMSE (bottom) for five synthetic populations, anchor 40%, easier test Y and ability N(50%, 18%).

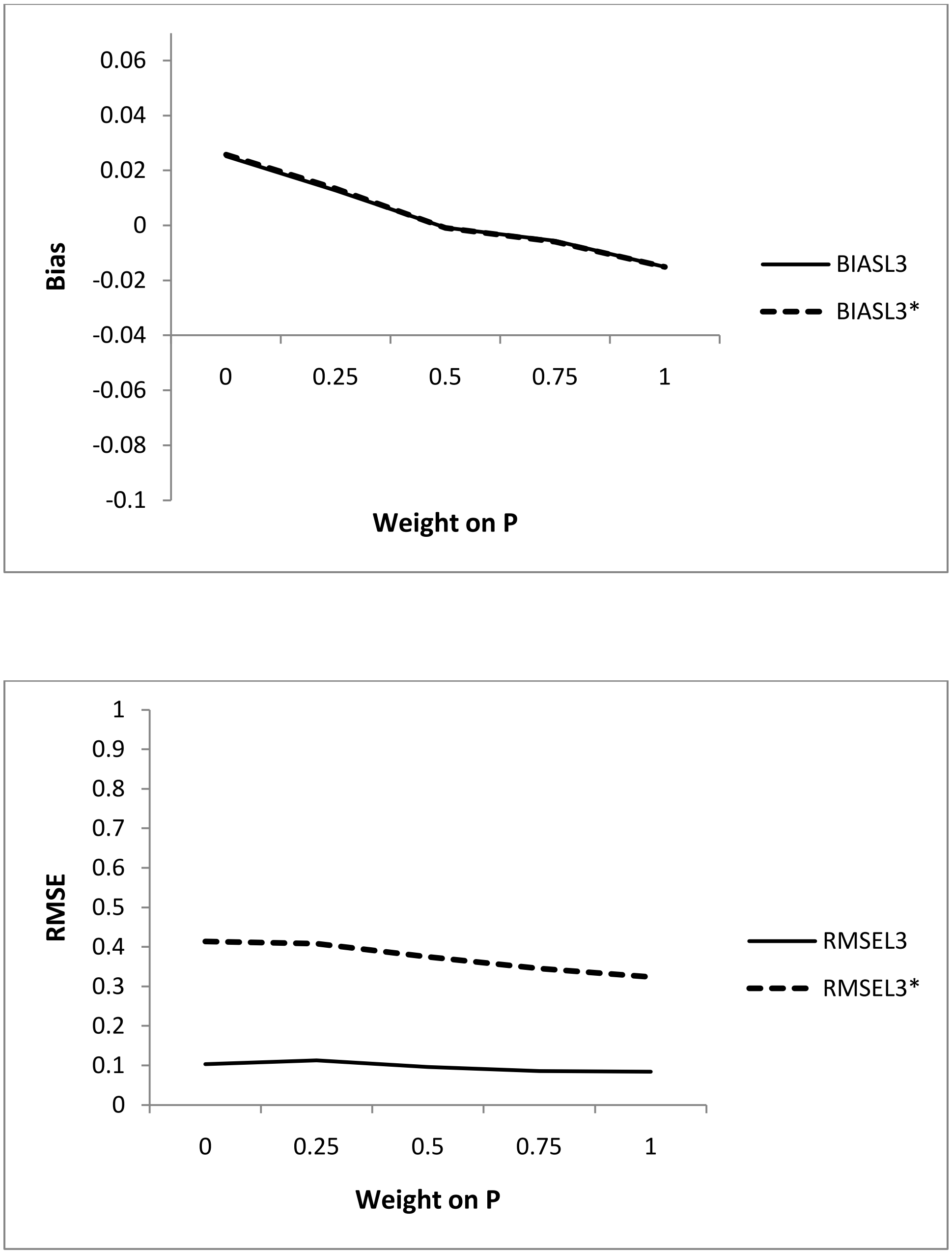
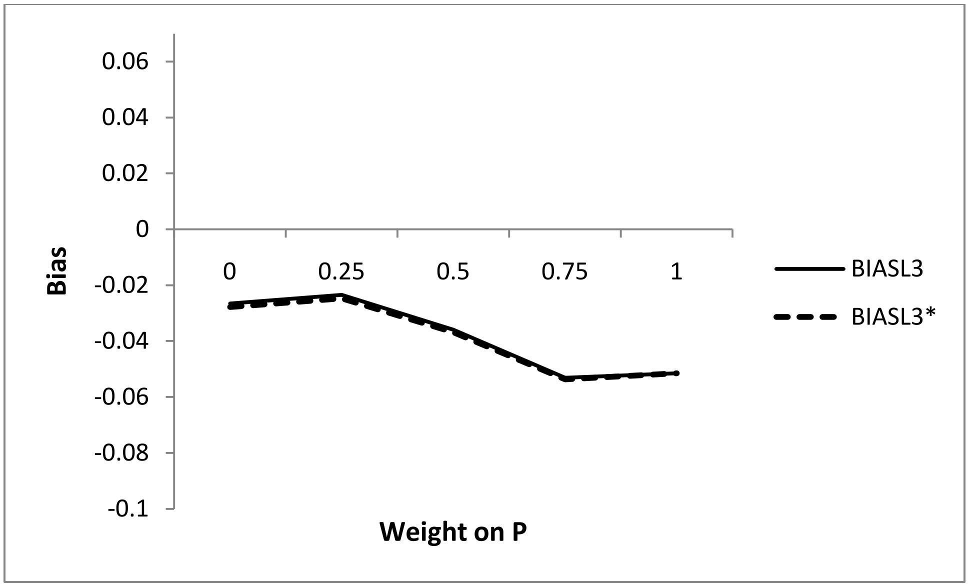


Figure 20. Bias (top) and RMSE (bottom) for five synthetic populations, anchor 40%, more difficult test Y and ability N(50%, 18%).



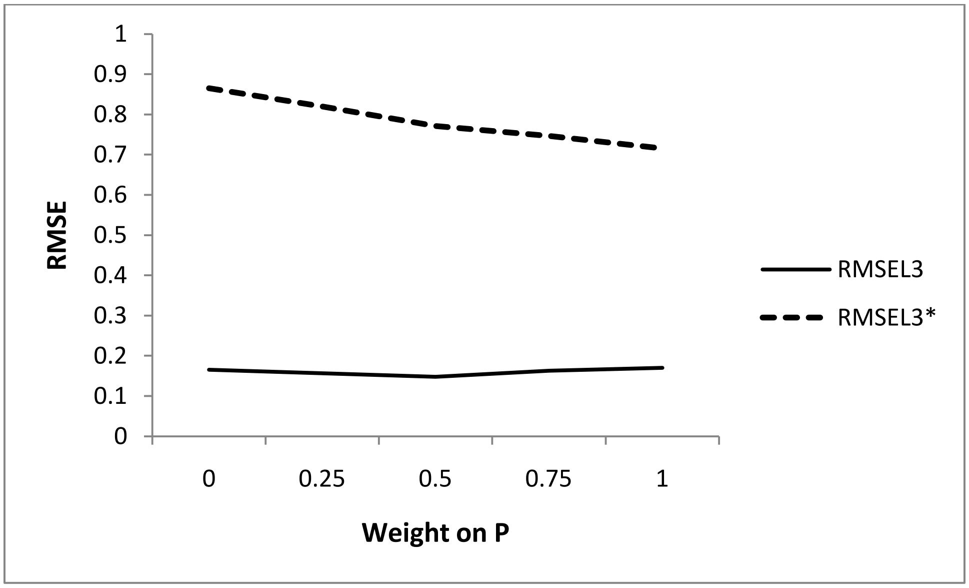


Figure 21. Bias (top) and RMSE (bottom) for five synthetic populations, anchor 20%, easier test Y and ability N(50%, 18%).

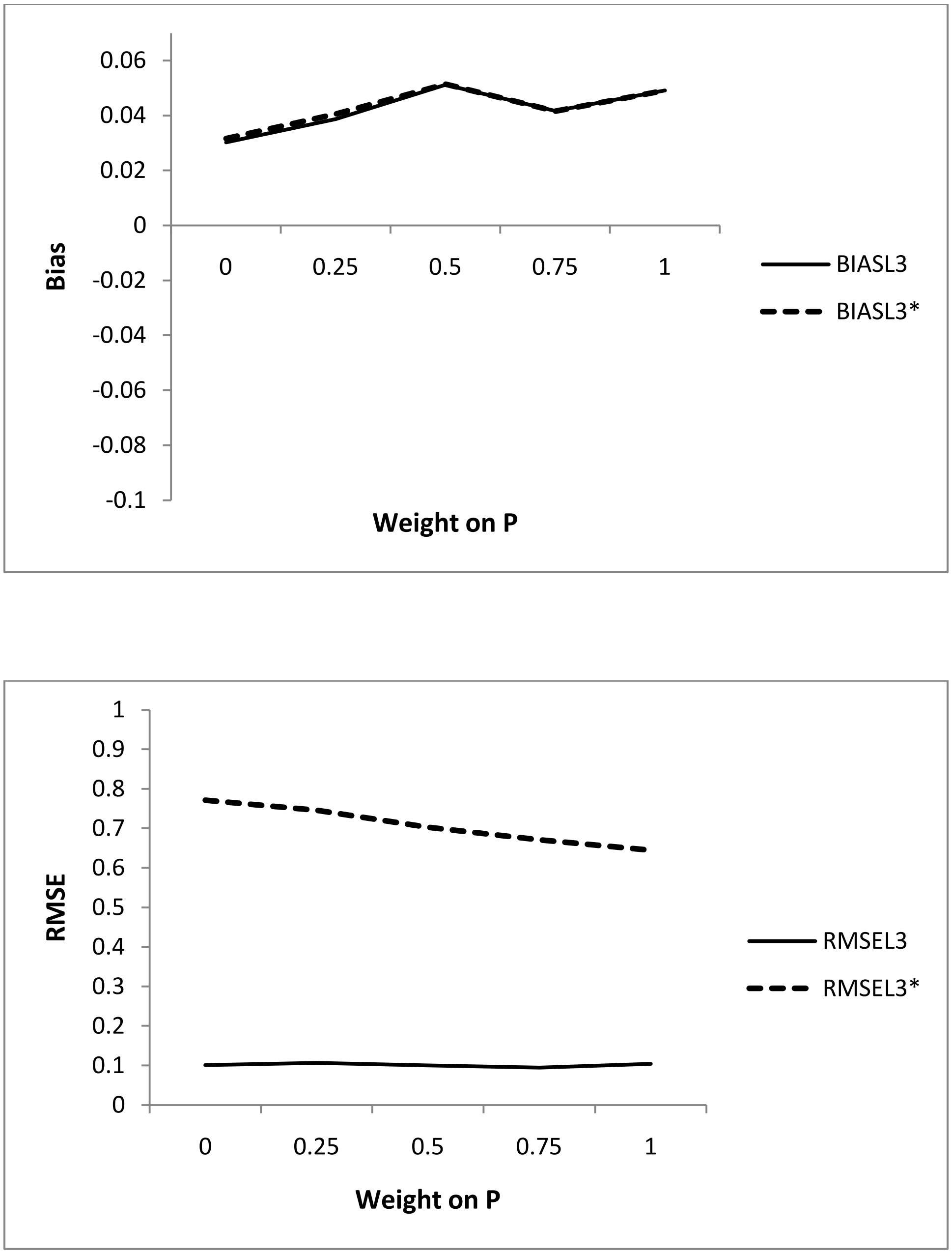
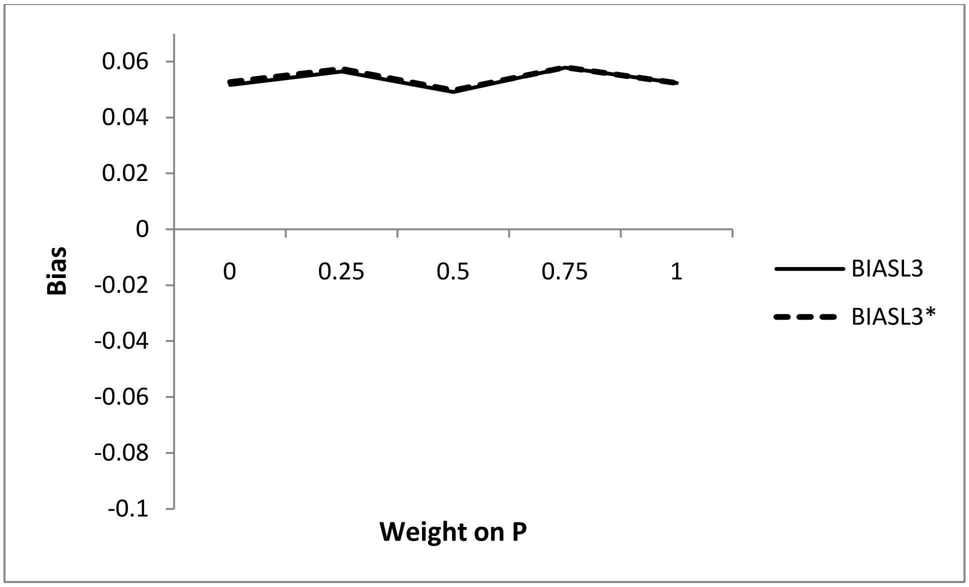


Figure 22. Bias (top) and RMSE (bottom) for five synthetic populations, anchor 20%, more difficult test Y and ability N(50%, 18%).



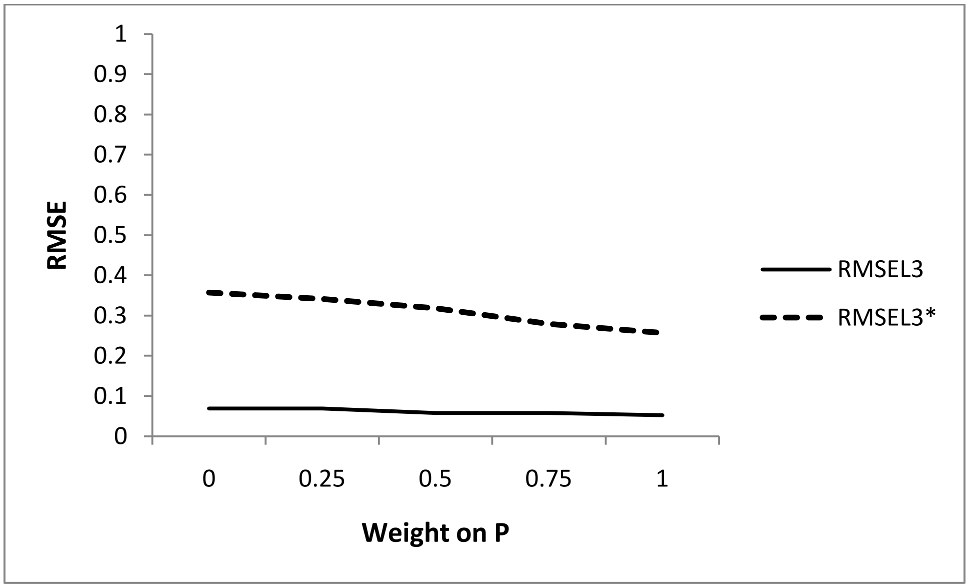


Figure 23. Bias (top) and RMSE (bottom) for five synthetic populations, anchor 40%, easier test Y and ability N(55%, 18%).

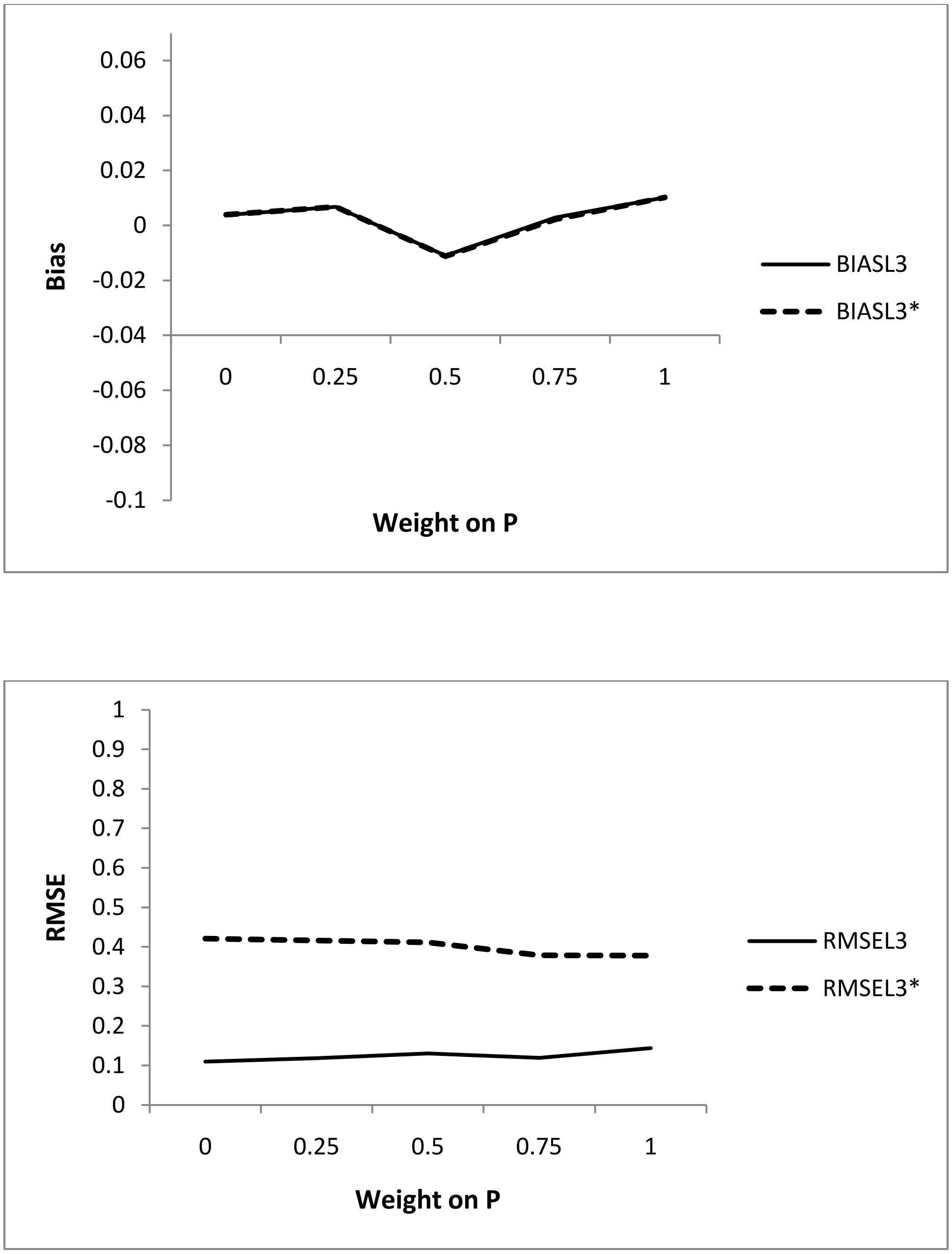


Figure 24. Bias (top) and RMSE (bottom) for five synthetic populations, anchor 40%, more difficult test Y and ability N(55%, 18%).

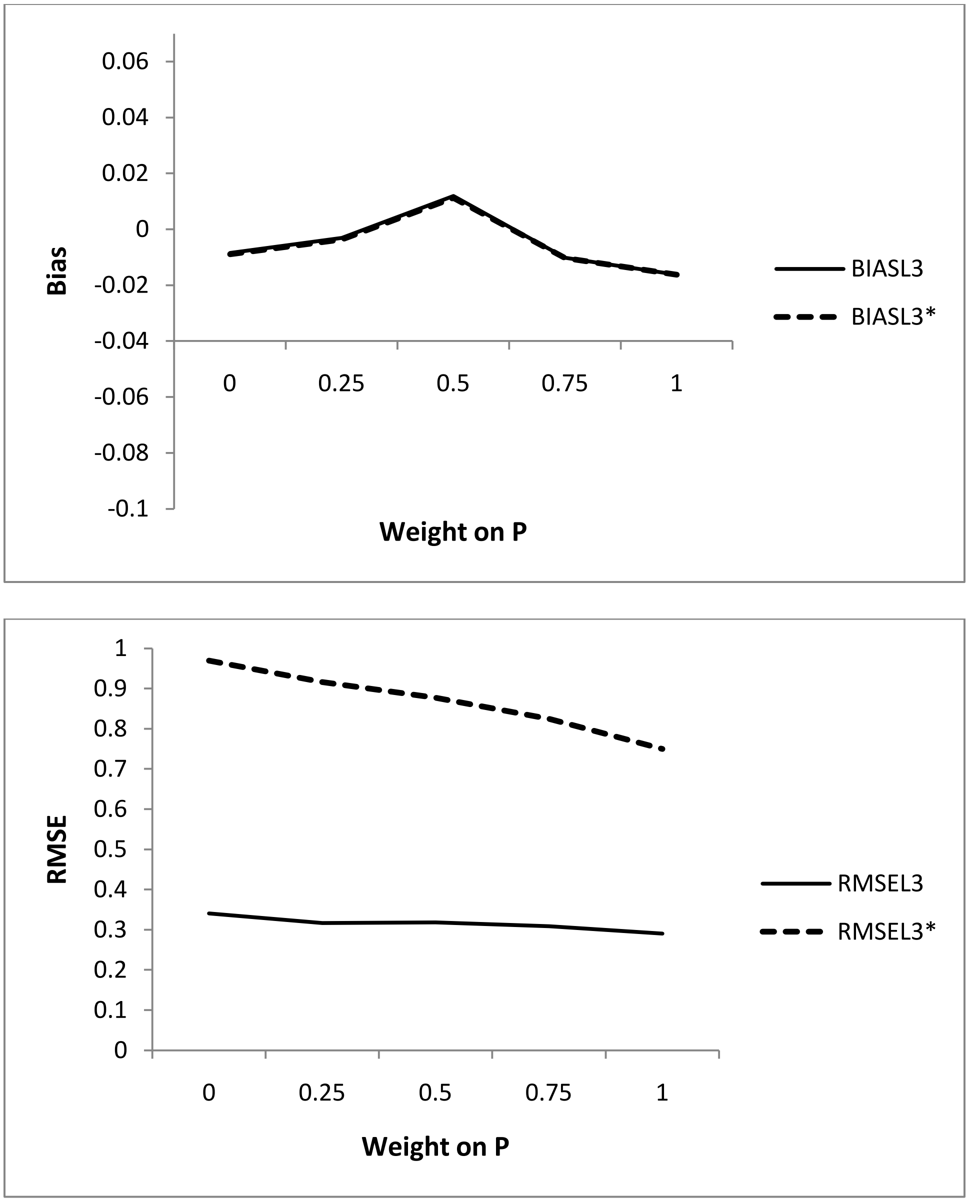
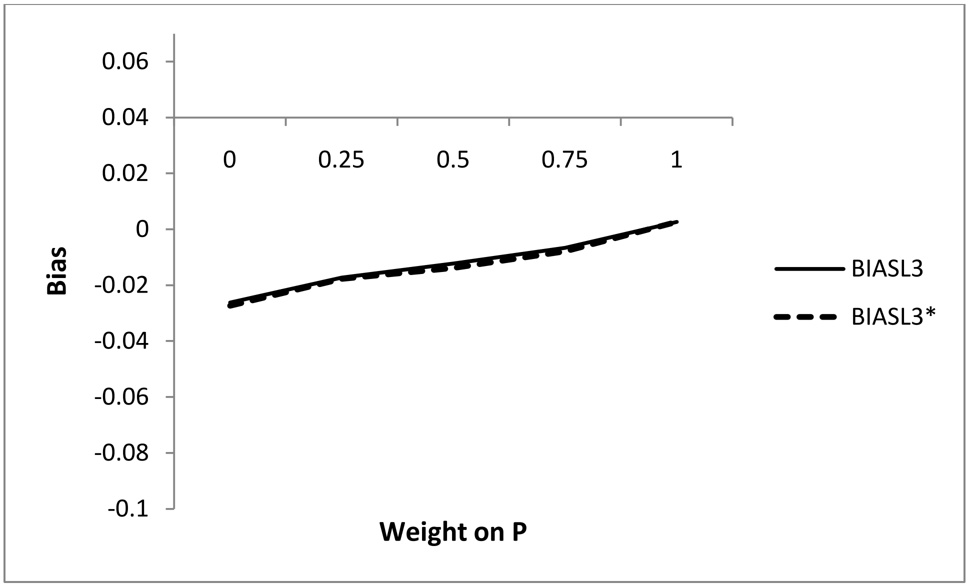


Figure 25. Bias (top) and RMSE (bottom) for five synthetic populations, anchor 20%, easier test Y and ability N(55%, 18%).



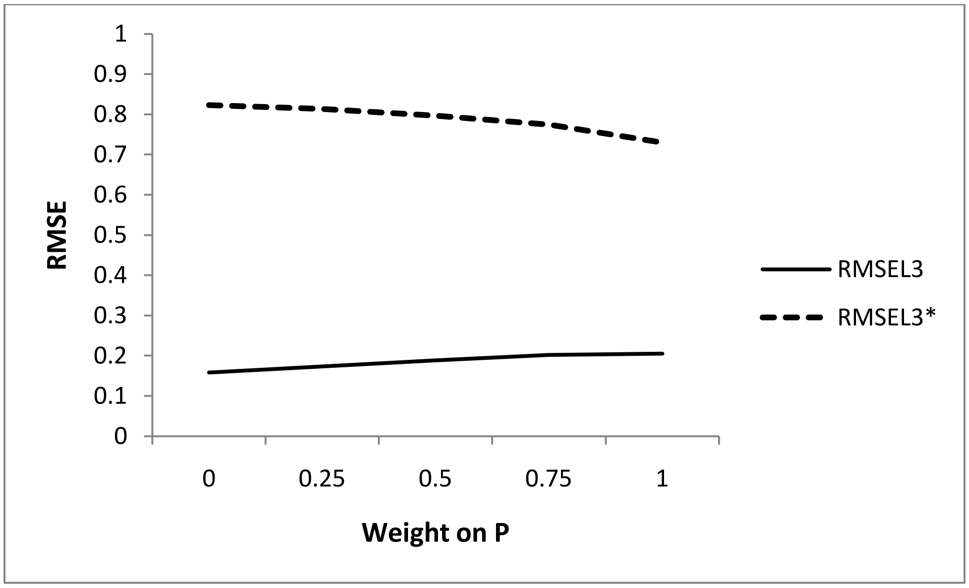


Figure 26. Bias (top) and RMSE (bottom) for five synthetic populations, anchor 20%, more difficult test Y and ability N(55%, 18%).

From Table 9 and figures 11 to 26 graphs it is apparent that the bias for the 16 cases of synthetic populations is the same.

From Table 10 and figures 11 to 14 it is apparent that the RMSE is the same for all the synthetic populations associated to equal ability distribution.

From Table 10 and figures 15 to 18 it can be observed that RMSE is the same for all the synthetic populations associated with N(50%, 15%) for P and N(55%, 15%) for Q except for the case in which the anchor is anchor 20% and test Y is easier, in which case L3* shows a lower RMSE than L3.

Figures 19 to 26 illustrate the cases in which the population Q related to the synthetic population has ability distribution with a standard deviation of 18%. In all these cases L3* shows a much larger RMSE than L3 and the RMSE is even larger for L3* for shorter anchor length (20%). The largest RMSE for L3* was observed for the case in which anchor length was 20%, the ability distribution for Q was N(55%, 18%), test Y was easier, and w=0. This value was 0.97. As w increased in this case, RMSE decreased to 0.75 for w=1.

Equating in Samples

As explained in Chapter Three, the combination of the factors ability distribution (four conditions), anchor length (2 conditions) and test difficulty (2 conditions) produced 16 conditions. For each of the 16 conditions 100 samples of size 500, 1000 and 2000 from population P and 100 samples of size 500, 1000 and 2000 form population Q were randomly extracted in a bootstrap with replacement fashion whereby after a subject is selected for the sample, he is returned to the population so that he could be selected again. This produced 300 pairs of samples P and Q for each of 16 conditions i.e. in total 4800 pairs of samples were generated.

For each of these pairs of samples five equating functions were computed under the L3 assumption and five equating functions were computed under the L3* assumption. Each of these five equating functions in each case was produced using a different *w* weight, where *w*: 0, 0.25, 0.50, 0.75, and 1.

For each of these pairs of samples five equating functions were computed under the L3 assumption and five equating functions were computed under the L3* assumption. Each of these five equating functions in each case was produced using a different w weight, where *w:* 0, 0.25, 0.50, 0.75, and 1.

As can be noted 4800 x 5= 24000 equating functions were computed for L3 and 24,000 for L3*. Each of these 24,000 cases can be related to a true equating function in the corresponding synthetic population.

Therefore for each of 24,000 cases it is possible to compute the bias for each of the X observed scores included in the sample P of the pair and then obtain a bias average over the sample size of that particular sample. Similarly, for each X score in P, a squared bias can be computed and then used to compute an RMSE for that sample over its sample size. These 24,000 average biases and 24,000 RMSE for L3 and L3* were used to perform two repeated measures ANOVA whereby the within subjects factor is the type of assumption (this has two levels, the bias under L3 and L3* in one case, and the RMSE under L3 and L3* in the other case). In each case the between subjects factors are sample size (SS), weight (WGT), length of anchor (ANC), whether test Y is easier or more difficult than X (DIF), and ability distribution of the population from where the sample comes (ABIL). SS has three levels: 500, 1000 and 2000, W has five levels: 0, 0.25, 0.50, 0.75, and 1. ANC has two levels: e (easier test Y) and d (more difficult test Y). ABIL has four levels: 05, 08, 55, and 58 which indicate respectively N(50%, 15%), N(55%, 15%), N(50%, 18%), and N(55%, 18%).

In combination with a statistical significance at $\alpha = 0.05$, Partial Eta Squared (PES) was used as a measure of effect size. A cut-off of PES = 0.1 was set for considering a result of practical significance. It is important to use a practical significance criterion because due to large sample size, it is expected that most interactions and main effects are significant.

With these criteria, it was found that sample size did not have any practical significance. None of the interactions in which sample size was present had practical significance. Therefore sample size was dropped as a factor and two new ANOVAs were run, this time with four between subjects factors (ABIL, DIF, WGT, and ANC) and with only the 8000 cases for sample size 2000.

Table 11 and 12 present the results for these two new ANOVAs. Table 11 presents the results for bias and Table 12 presents the results of the ANOVA for RMSE. Both tables are organized with the significant (statistically and practically) main effects or interactions first and then the non-significant are presented.

Table 11
Significance and Partial Eta Square of the Repeated Measures ANOVA for Bias for Samples

Factor	F	df	Sig.	Partial Eta Squared
Assumption	7953.471	(1,7920)	.000	.501
Assumption*ABIL	9446.569	(3, 7920)	.000	.782
Assumption*WGT	1012.303	(4, 7920)	.000	.338
Assumption*ANC	961.022	(1, 7920)	.000	.108
Assumption*WGT*ABIL	1231.819	(12, 7920)	.000	.651
Assumption*ANC*ABIL	1187.581	(3, 7920)	.000	.310
Assumption*WGT*ANC*ABIL	155.130	(12,7920)	.000	.190
Assumption *DIF	1.698	(1, 7920)	.193	.000
Assumption*WGT*ANC	122.021	(4, 7920)	.000	.058
Assumption *DIF*ABIL	15.606	(3, 7920)	.000	.006
Assumption*ANC*DIF	2.224	(1, 7920)	.136	.000
Assumption*WGT*DIF	0.221	(4, 7920)	.927	.000
Assumption * ANC *DIF*ABIL	14.835	(3, 7920)	.000	.006
Assumption * WGT *DIF*ABIL	2.214	(12, 7920)	.009	.003
Assumption * WGT * ANC * DIF	.184	(4, 7920)	.947	.000
Assumption * WGT * ANC * DIF* ABIL	2.257	(12, 7920)	.000	.006

[a] df=degrees of freedom

Table 12

Significance and Partial Eta Square of the Repeated Measures ANOVA for RMSE for Samples

Factor	F	df	Sig.	Partial Eta Squared
Assumption	5367.979	(1,7920)	.000	.404
Assumption*ABIL	3115.795	(3, 7920)	.000	.541
Assumption*ANC	945.805	(1, 7920)	.000	.107
Assumption*ANC*ABIL	659.966	(3, 7920)	.000	.200
Assumption*WGT	16.975	(4, 7920)	.000	.009
Assumption *DIF	0.176	(1, 7920)	.675	.000
Assumption*ANC*DIF	37.607	(1, 7920)	.000	.005
Assumption*WGT*ABIL	7.233	(12, 7920)	.000	.011
Assumption *DIF*ABIL	3.155	(3, 7920)	.024	.001
Assumption*WGT*ANC	0.582	(4, 7920)	.676	.000
Assumption*WGT*DIF	0.240	(4, 7920)	.916	.000
Assumption * ANC *DIF*ABIL	17.864	(3, 7920)	.000	.007
Assumption*WGT*ANC*ABIL	0.448	(12, 7920)	.944	.001
Assumption * WGT *DIF*ABIL	0.331	(12, 7920)	.984	.001
Assumption*WGT*ANC*DIF	0.007	(4, 7920)	1.000	.000
Assumption * WGT * ANC * DIF* ABIL	0.039	(12, 7920)	1.000	.000

[a] df=degrees of freedom

As can be noted in Table 11, for bias there is a significant four way interaction: assumption*weight*anchor*ability, two significant three way interactions: assumption*anchor*ability and assumption*weight*ability. There are three two way significant interactions: assumption*anchor, assumption*weight and assumption*ability. These interactions are presented in the following graphs of the marginal means, starting with the two way significant interactions.

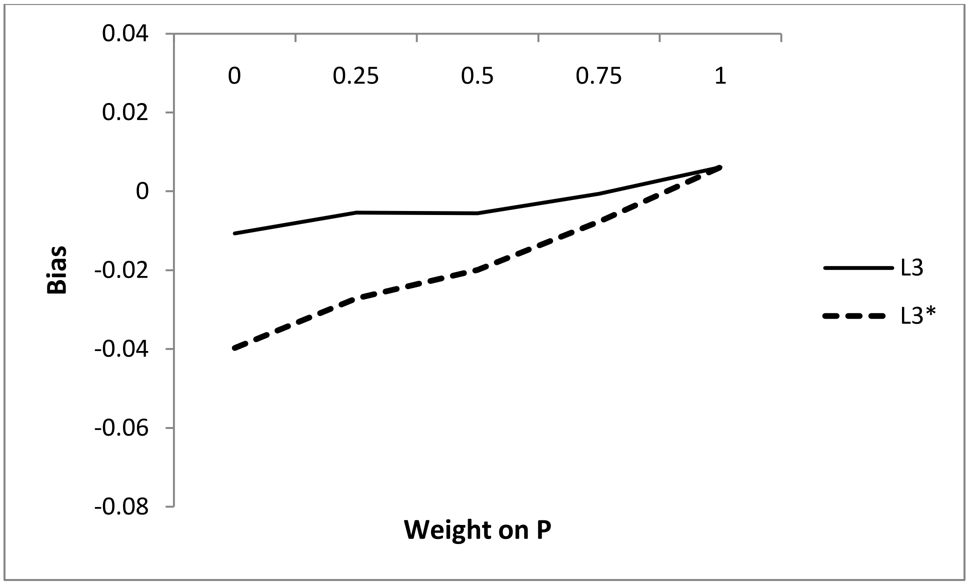


Figure 27. Graph of marginal means for the interaction assumption* weight for bias.

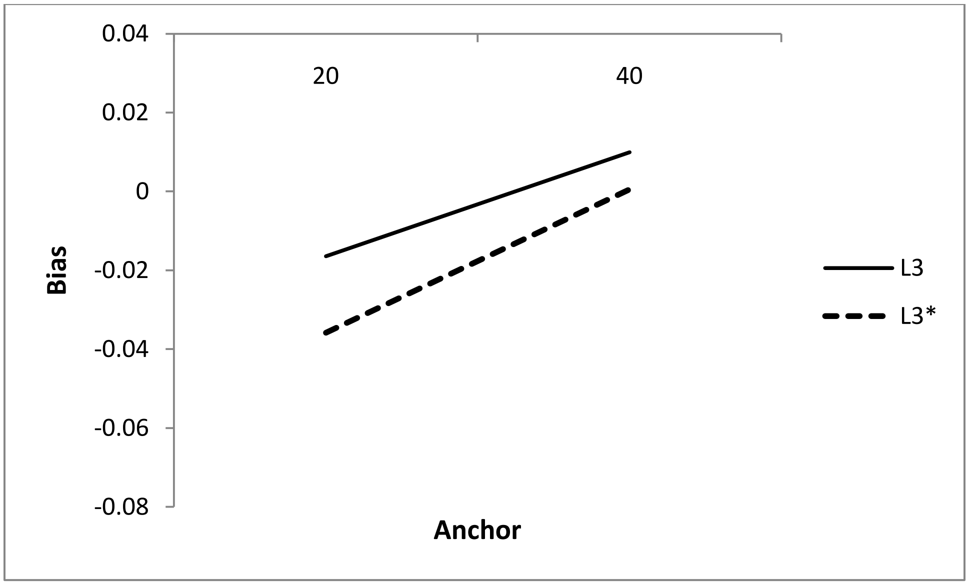


Figure 28. Graph of marginal means for the interaction assumption* anchor for bias.

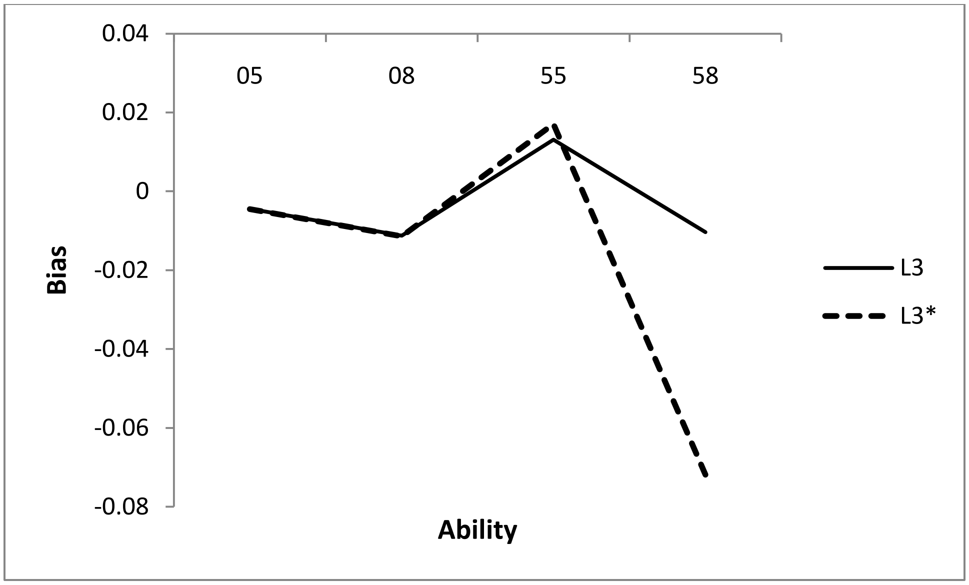
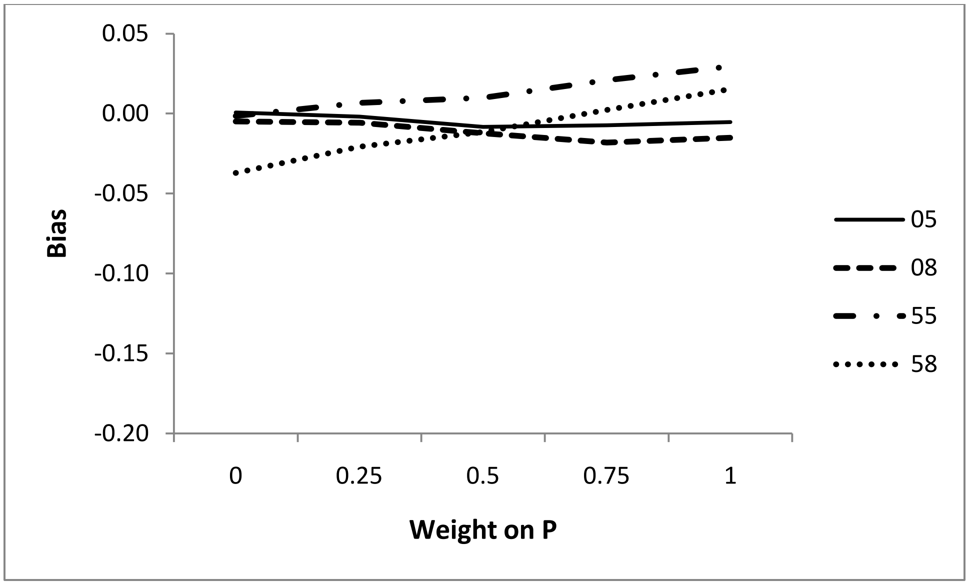


Figure 29. Graph of marginal means for the interaction assumption* ability for bias.



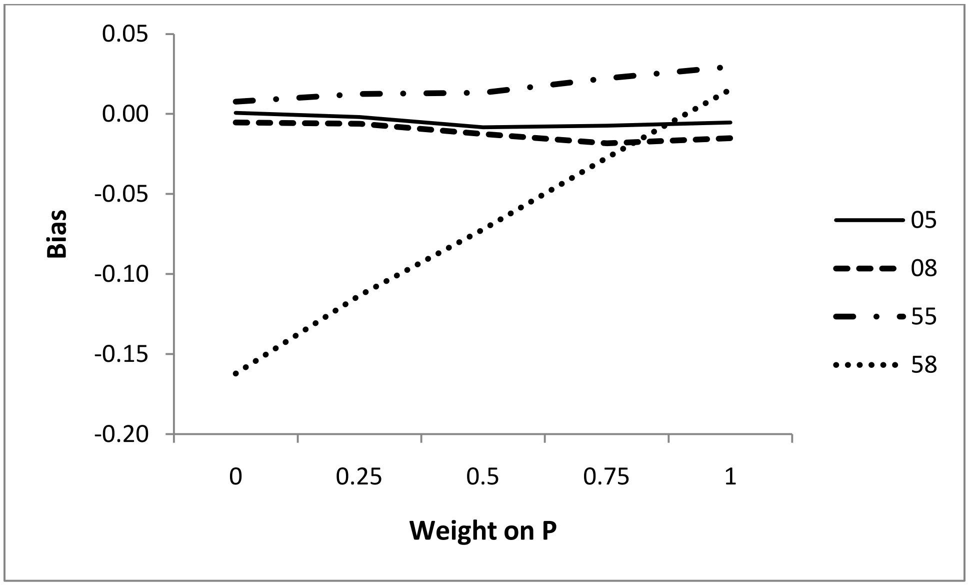


Figure 30. Graph of marginal means for the interaction assumption* weight*ability for bias: L3(top) and L3*(bottom).

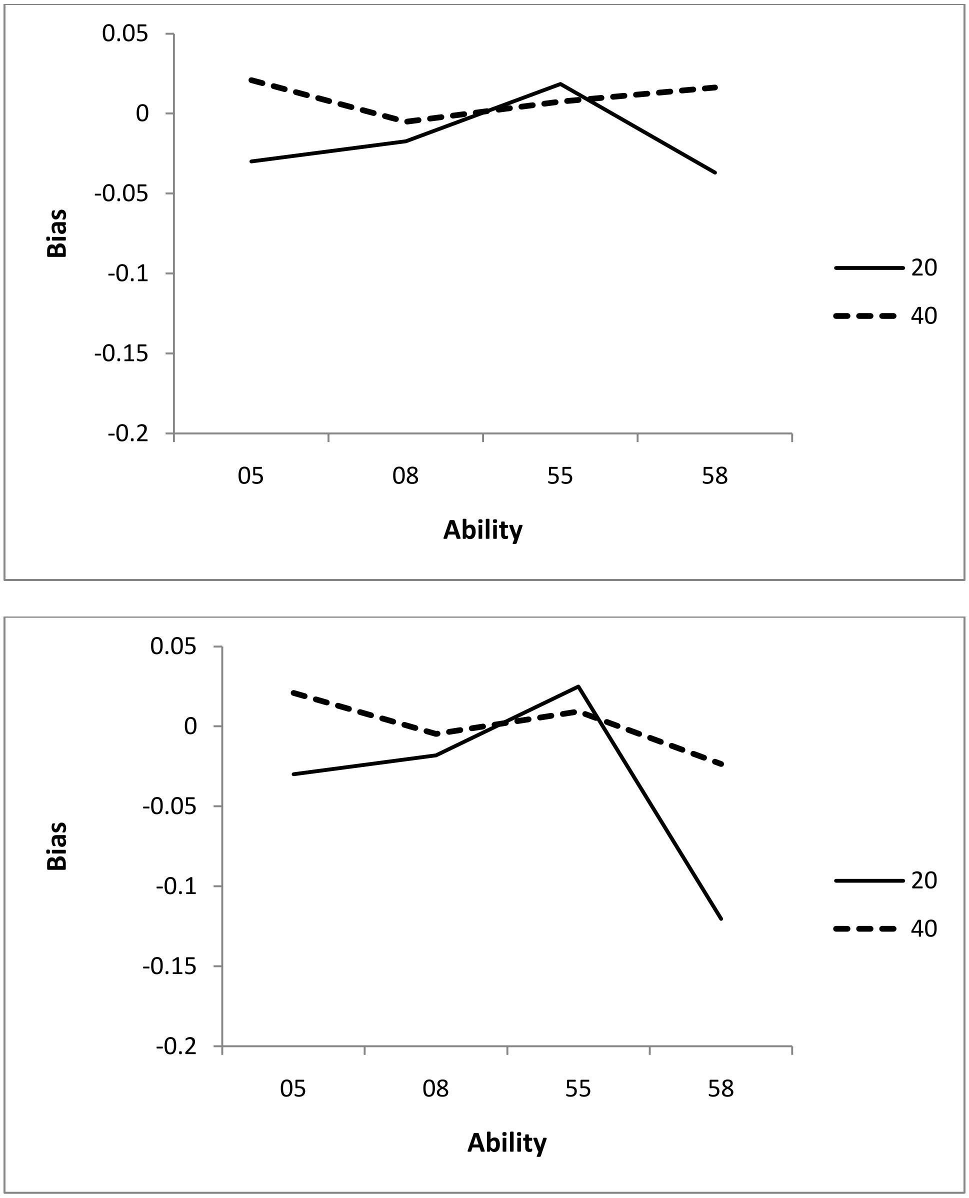
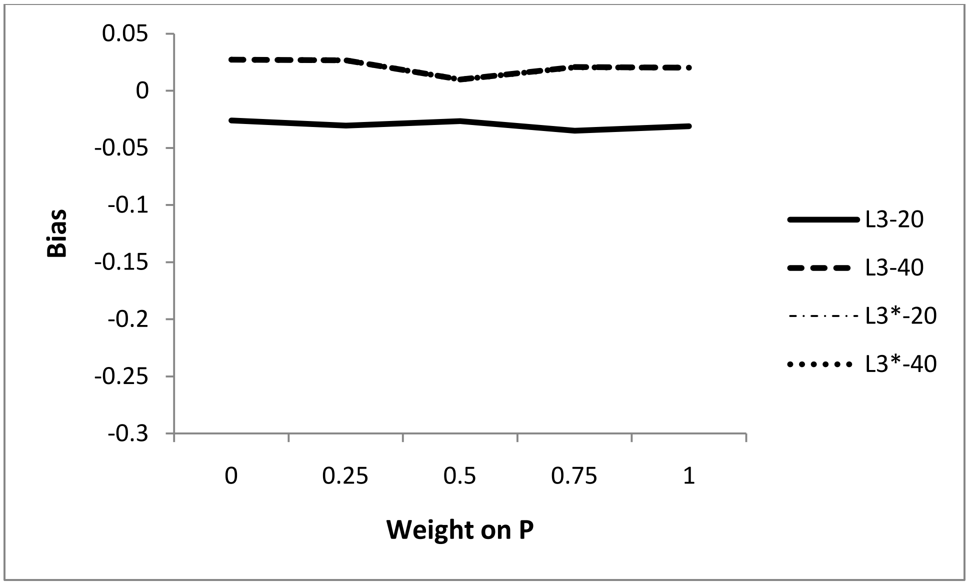


Figure 31. Graph of marginal means for the interaction assumption*ability*anchor for bias: L3(top) and L3*(bottom).



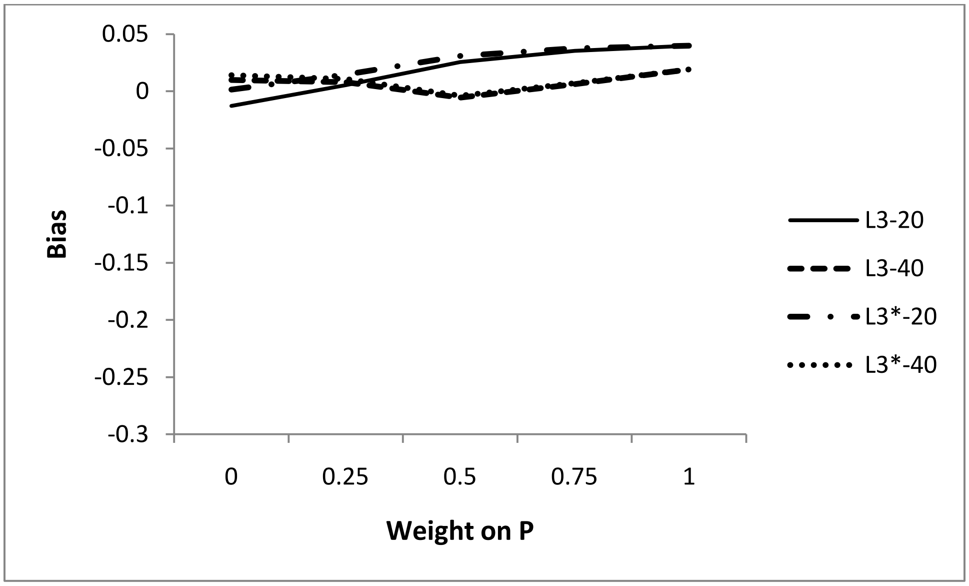
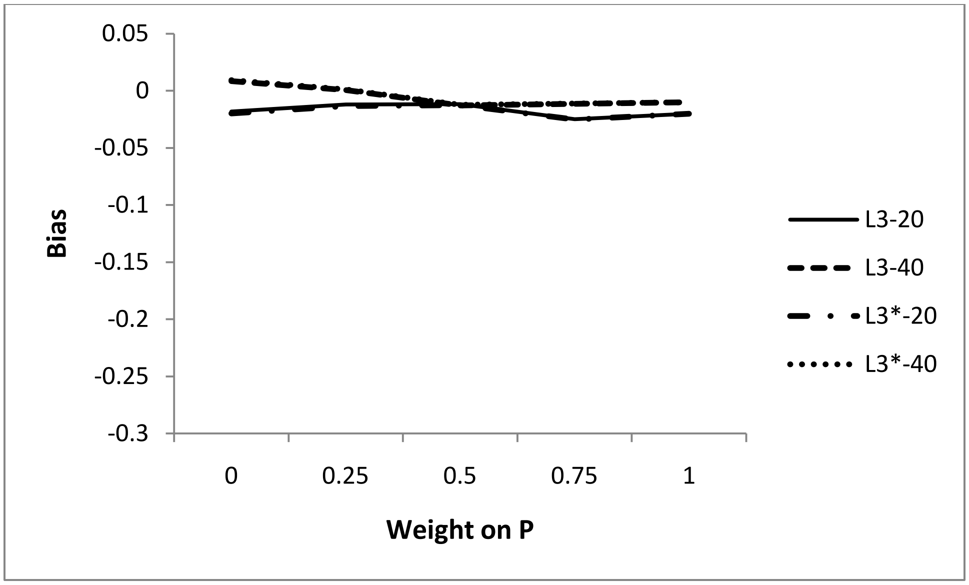


Figure 32a. Graph of marginal means for the interaction assumption*weight*anchor*ability for bias: N(50%, 15%) (top) and N(55%, 15%) (bottom).



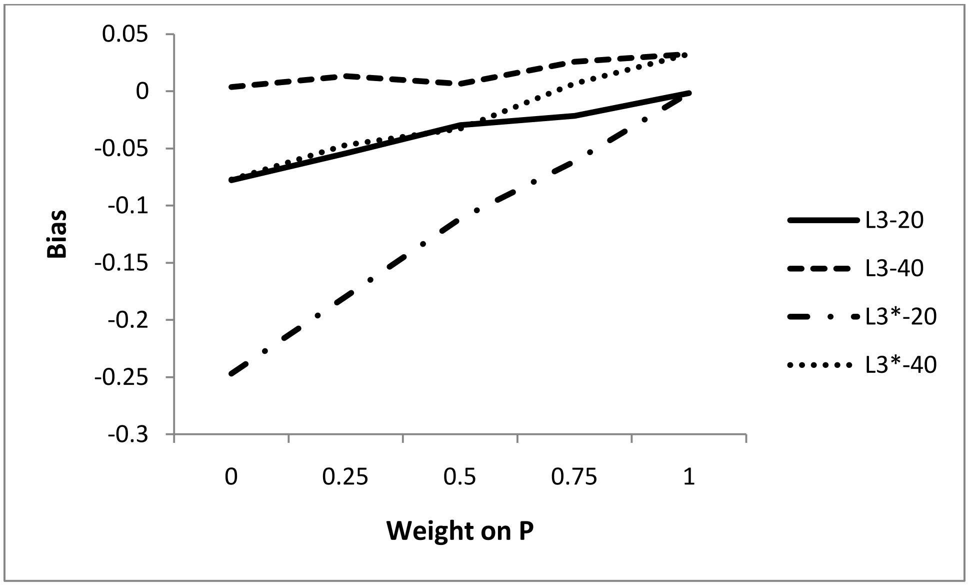


Figure 32b. Graph of marginal means for the interaction assumption*weight*anchor*ability for bias: N(50%, 18%) (top) and N(55%, 18%) (bottom).

Figures 32a and 32b show how ability distribution interacts with anchor length, weight and assumption. For the condition N(55%, 18%), L3* with an anchor of 20% has the largest bias for the w=0 case. This bias decreases as the *w* weight gets closer to 1. In the other three conditions for this interaction, bias is closer to 0 and L3 and L3* behaves similarly.

Figure 31 shows how assumption interacts with ability and anchor. L3* gets a higher bias for the condition N(55%, 18%) and shorter anchor, whereas for the other conditions L3 and L3* behaves similarly and bias is close to 0.

Figure 30 shows how assumption interacts with weight and ability. For the condition N(55%, 18%), L3* has a larger bias for weight 0 and bias decreases as weight gets closer to 1. For other conditions of ability, L3 and L3* are similar and close to 0.

Figure 29 illustrates the interaction of assumption and ability. For the condition N(55%, 18%), L3* shows a larger bias and for the other conditions L3 and L3* behave similarly and are close to 0.

Figure 28 illustrates the interaction of assumption and anchor. L3* shows a larger bias for anchor length 20% although the magnitude of the difference is rather small.

Figure 27 illustrates the interaction of assumption and weight. For weight 0, L3* shows the largest bias and bias gets closer to that of L3 as *w* approaches 1. However, the magnitude of the difference is rather small.

As it can be noted in Table 12, there is a significant (statistically and practically) three way interaction for RMSE: assumption*ability*anchor. The two way interactions assumption* anchor and assumption*ability are also significant. The marginal means for RMSE for these two interactions are presented in figures 33 to 35.

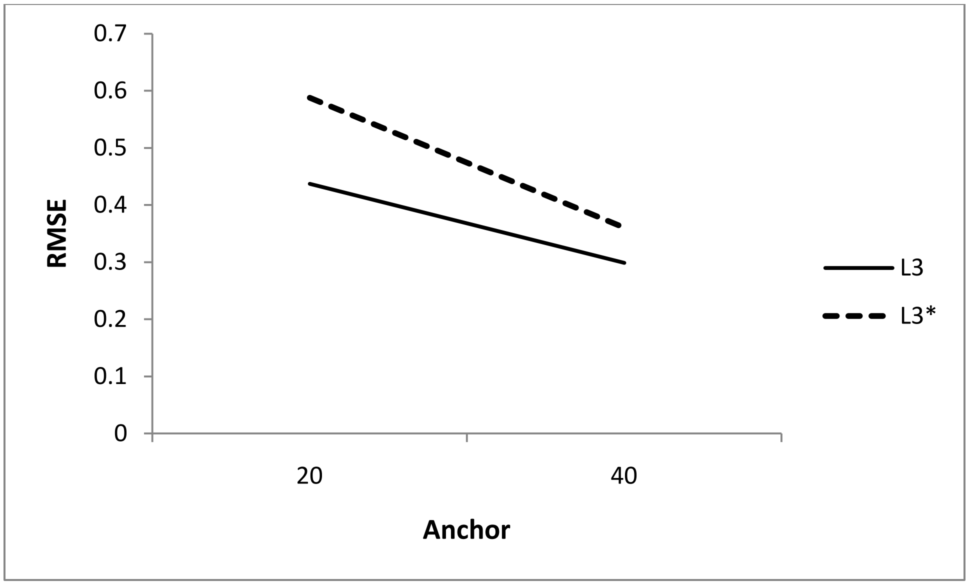


Figure 33. Graph of marginal means for the interaction assumption* anchor for RMSE.

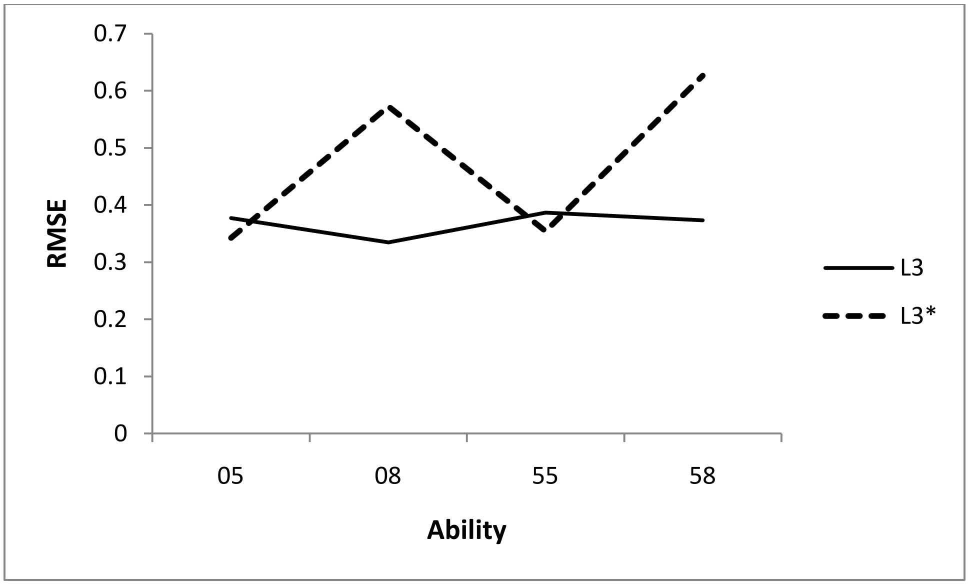
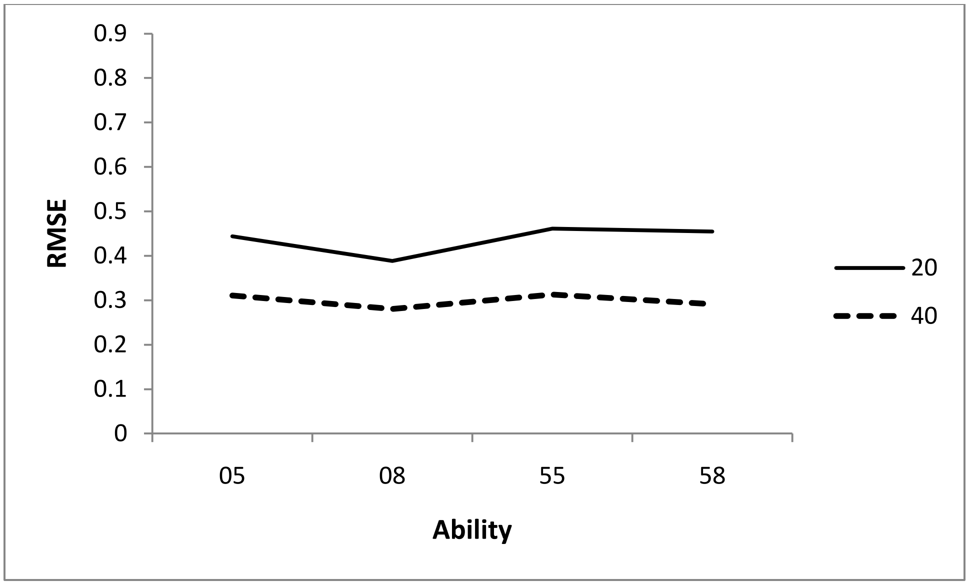


Figure34. Graph of marginal means for the interaction assumption* ability for RMSE.



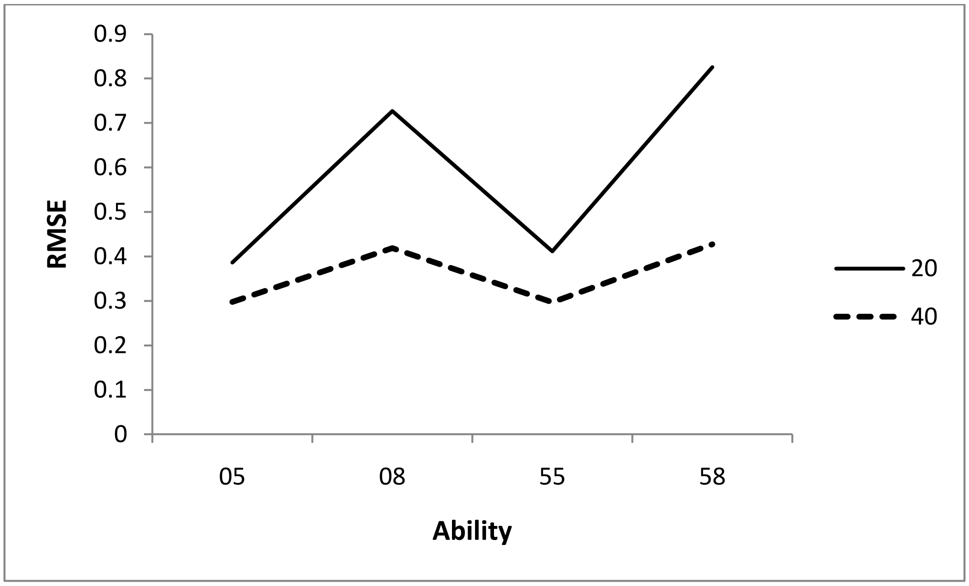


Figure 35. Graph of marginal means for the interaction assumption*ability*anchor for RMSE: L3(top) and L3*(bottom).

Figure 35 illustrates the interaction of assumption, ability, and anchor for RMSE. For the conditions with an 18% SD, RMSE is much higher for L3* for the anchor length 20%.

Figure 34 illustrates the interaction of assumption and ability for RMSE. Conditions with 18% SD show a higher RMSE for L3*.

Figure 33 illustrates the interaction of assumption and anchor. For anchor length 20%, L3* features a higher RMSE.

Bias and RMSE in the Samples at Score Level

To further explore the differences between L3 and L3* in the samples, bias and RMSE were computed at score level. This was accomplished by applying the equating functions obtained from L3 and L3* to each score X level from 0 to 81, and this was done for the five different *w* weights. The obtained values can be then compared to the true values of the corresponding synthetic populations and a bias can be obtained at each score level for each sample. Each of this bias at score level was averaged across the 100 samples corresponding to each condition.

In a similar fashion, an RMSE at each score level was computed across the 100 samples.

The following figures show the bias and RMSE at score level for each condition. These figures are based on samples of size 2000 given that sample size was dropped from the original ANOVAs and that the previously reported ANOVAs were conducted with sample size 2000. These figures are based on the case *w*=0.5. Figures for other *w* values are not reported because the results were similar.

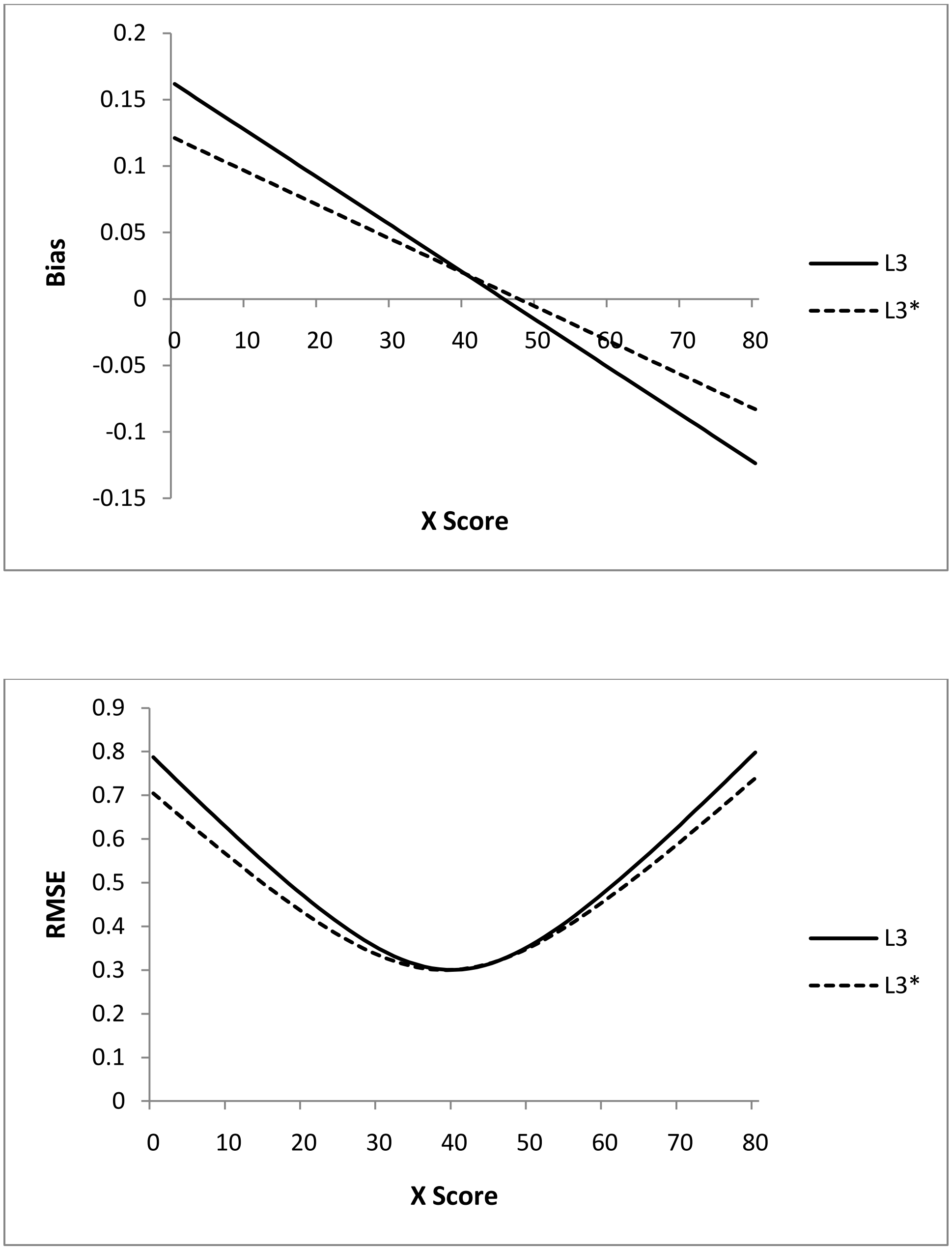
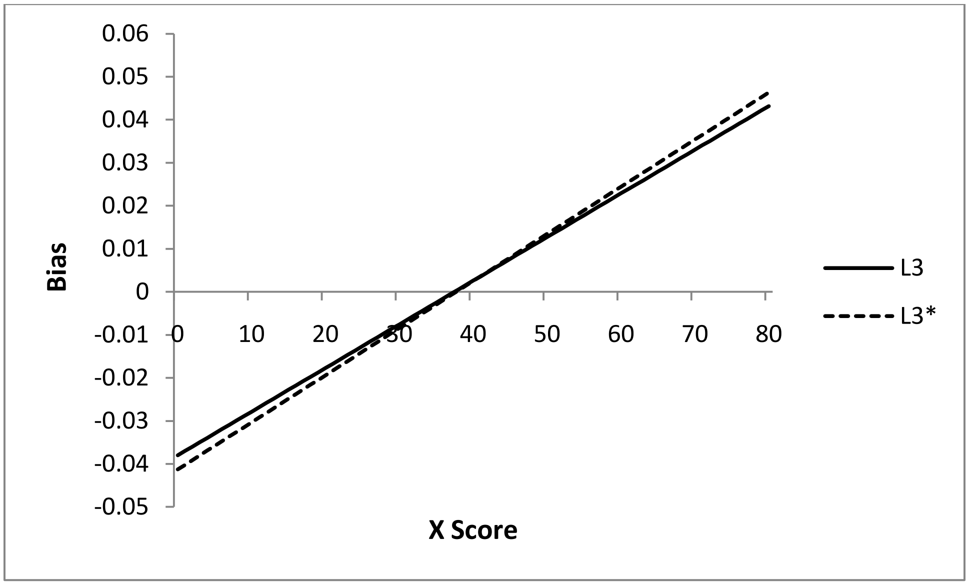


Figure 36. Bias (top) and RMSE (bottom) at score level for 100 samples, anchor 40%, easier test Y and ability N(50%, 15%).



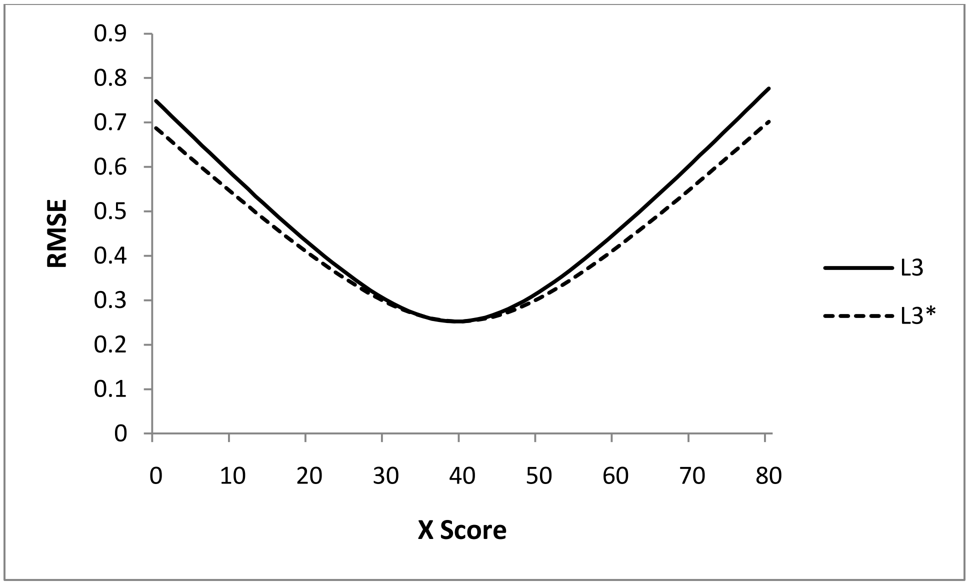


Figure 37. Bias (top) and RMSE (bottom) at score level for 100 samples, anchor 40%, more difficult test Y and ability N(50%, 15%).

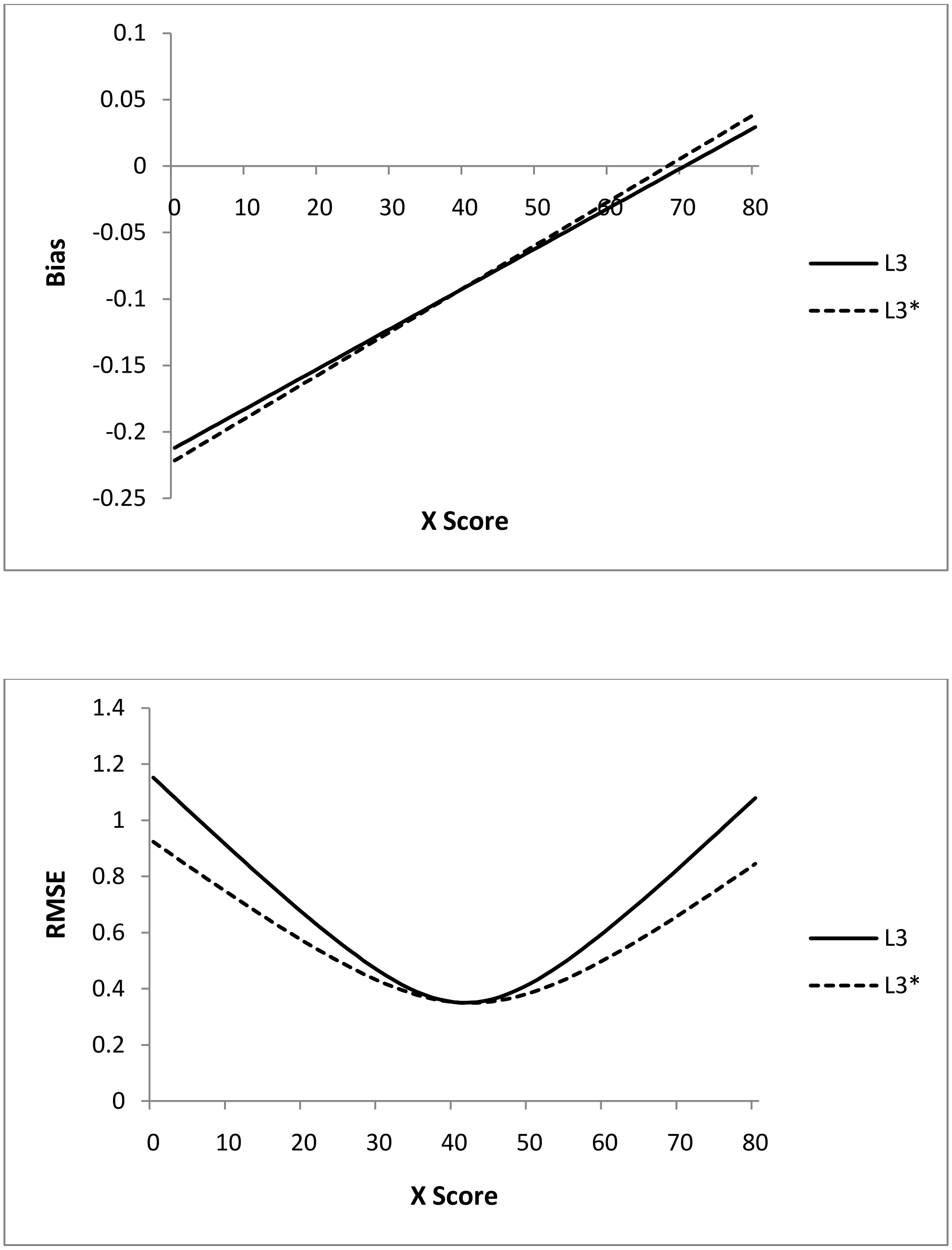


Figure 38. Bias (top) and RMSE (bottom) at score level for 100 samples, anchor 20%, easier test Y and ability N(50%, 15%).

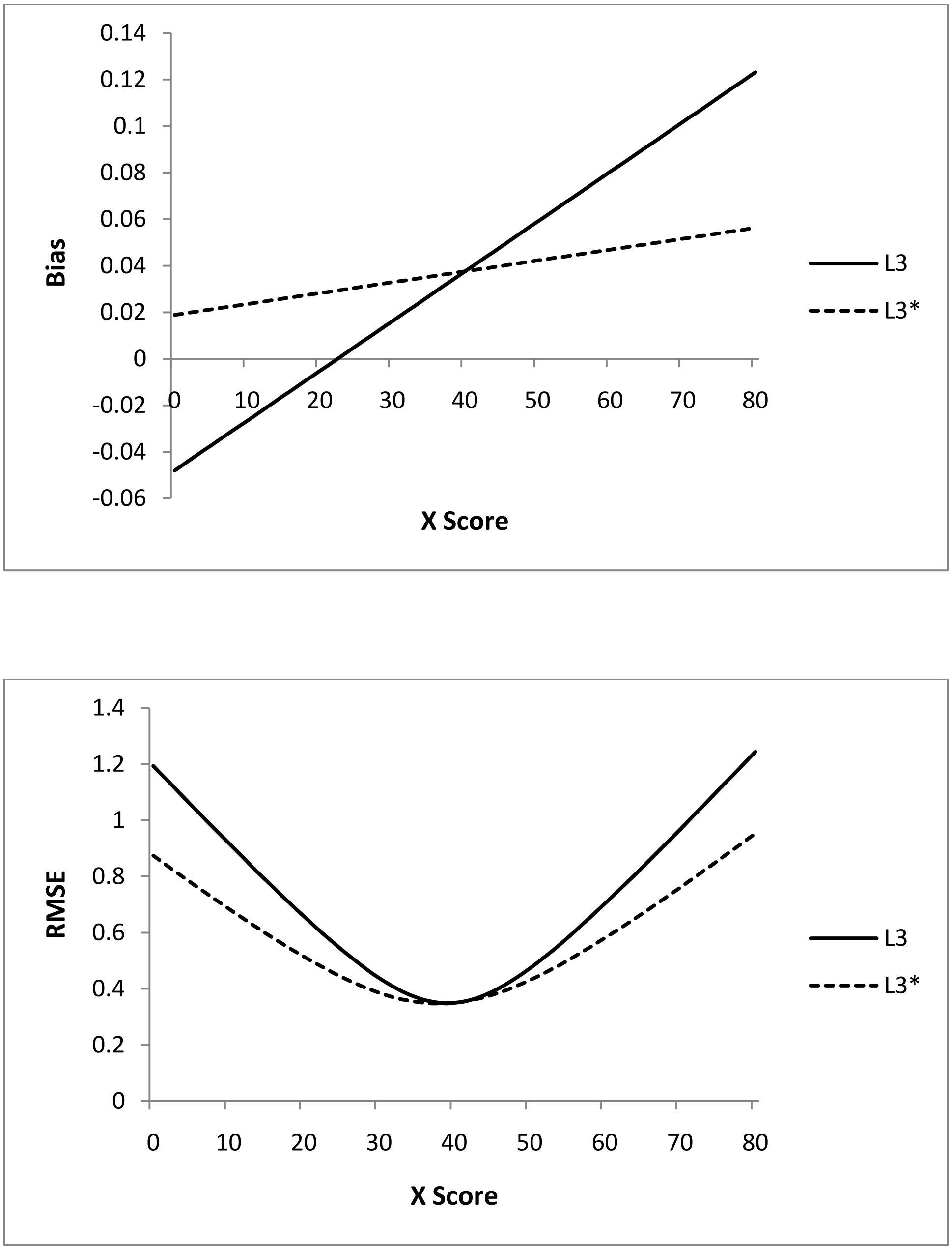


Figure 39. Bias (top) and RMSE (bottom) at score level for 100 samples, anchor 20%, more difficult test Y and ability N(50%, 15%).

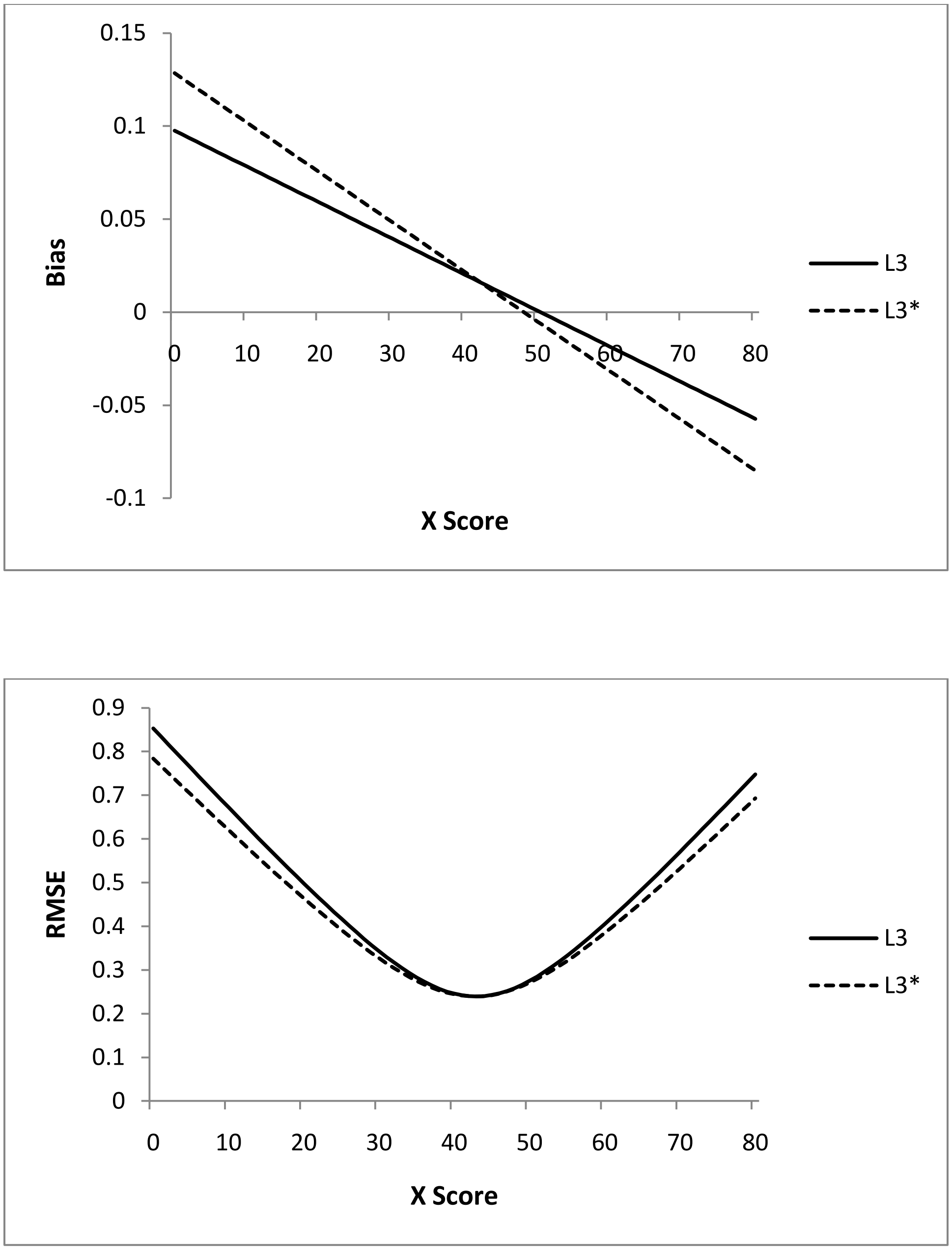
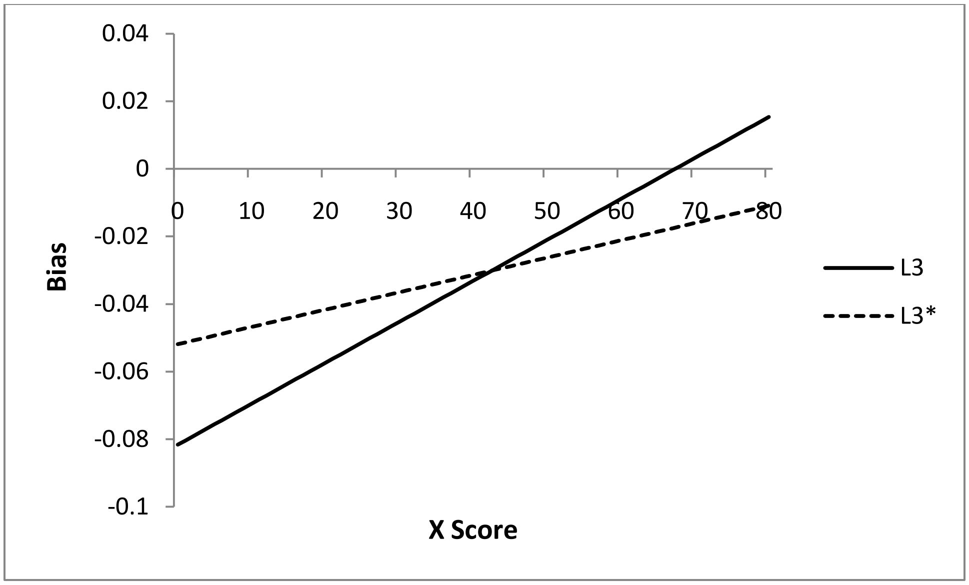


Figure 40. Bias (top) and RMSE (bottom) at score level for 100 samples, anchor 40%, easier test Y and ability N(55%, 15%).



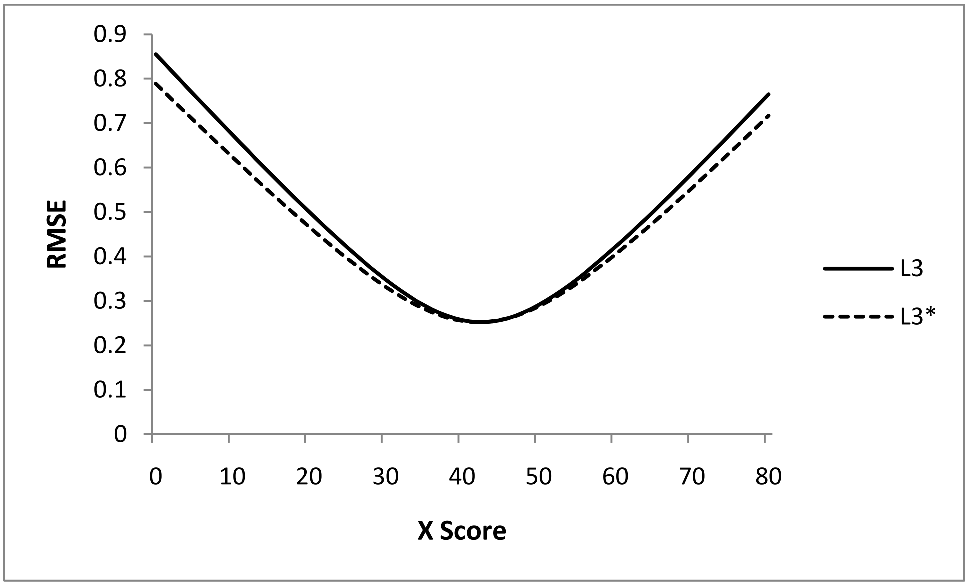
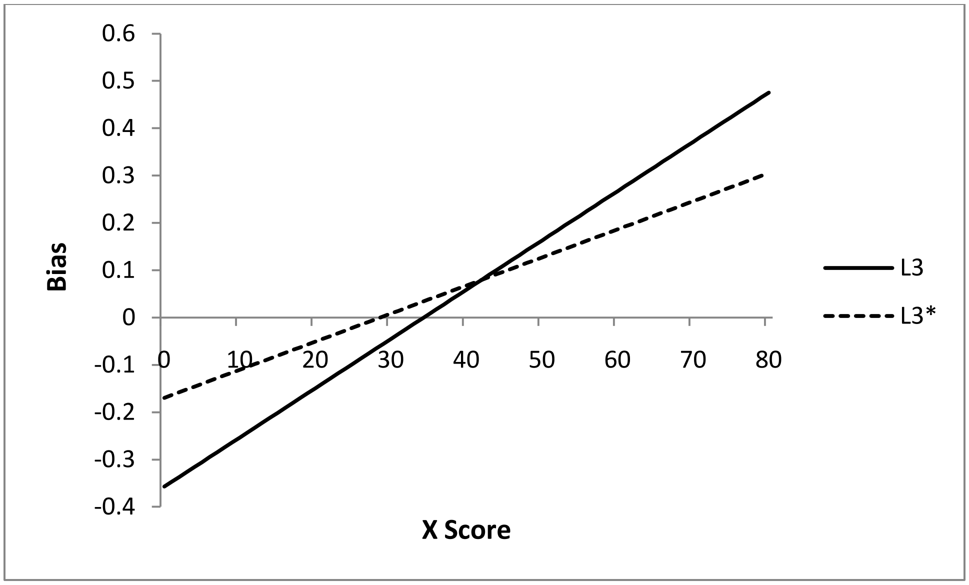


Figure 41. Bias (top) and RMSE (bottom) at score level for 100 samples, anchor 40%, more difficult test Y and ability N(55%, 15%).



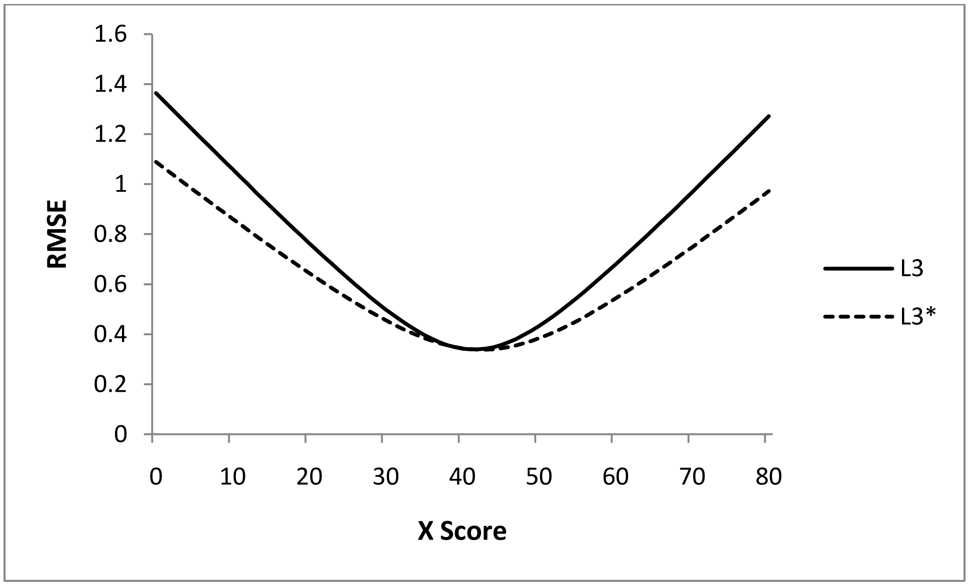


Figure 42. Bias (top) and RMSE (bottom) at score level for 100 samples, anchor 20%, easier test Y and ability N(55%, 15%).

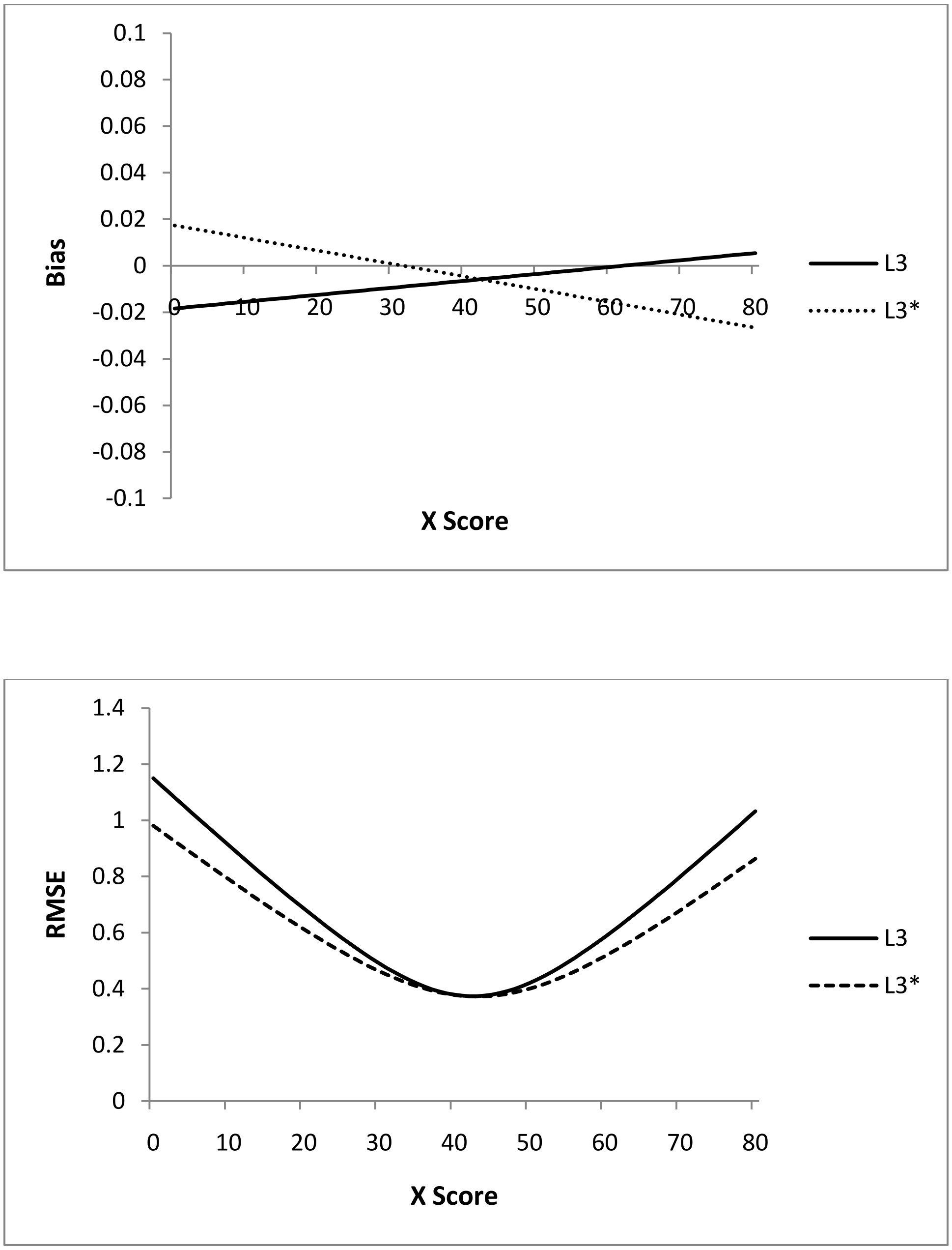


Figure 43. Bias (top) and RMSE (bottom) at score level for 100 samples, anchor 20%, more difficult test Y and ability N(55%, 15%).

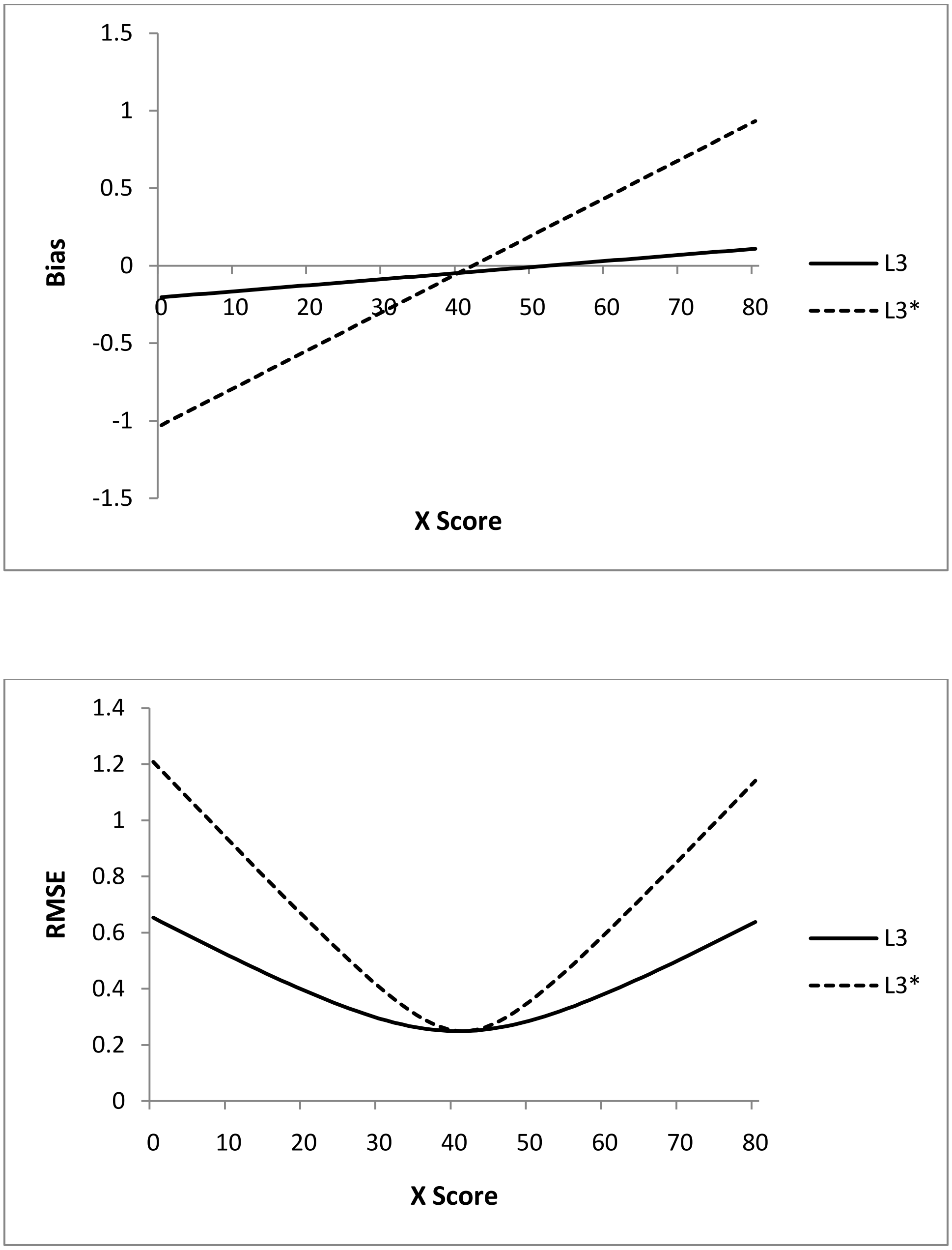


Figure 44. Bias (top) and RMSE (bottom) at score level for 100 samples, anchor 40%, easier test Y and ability N(50%, 18%).

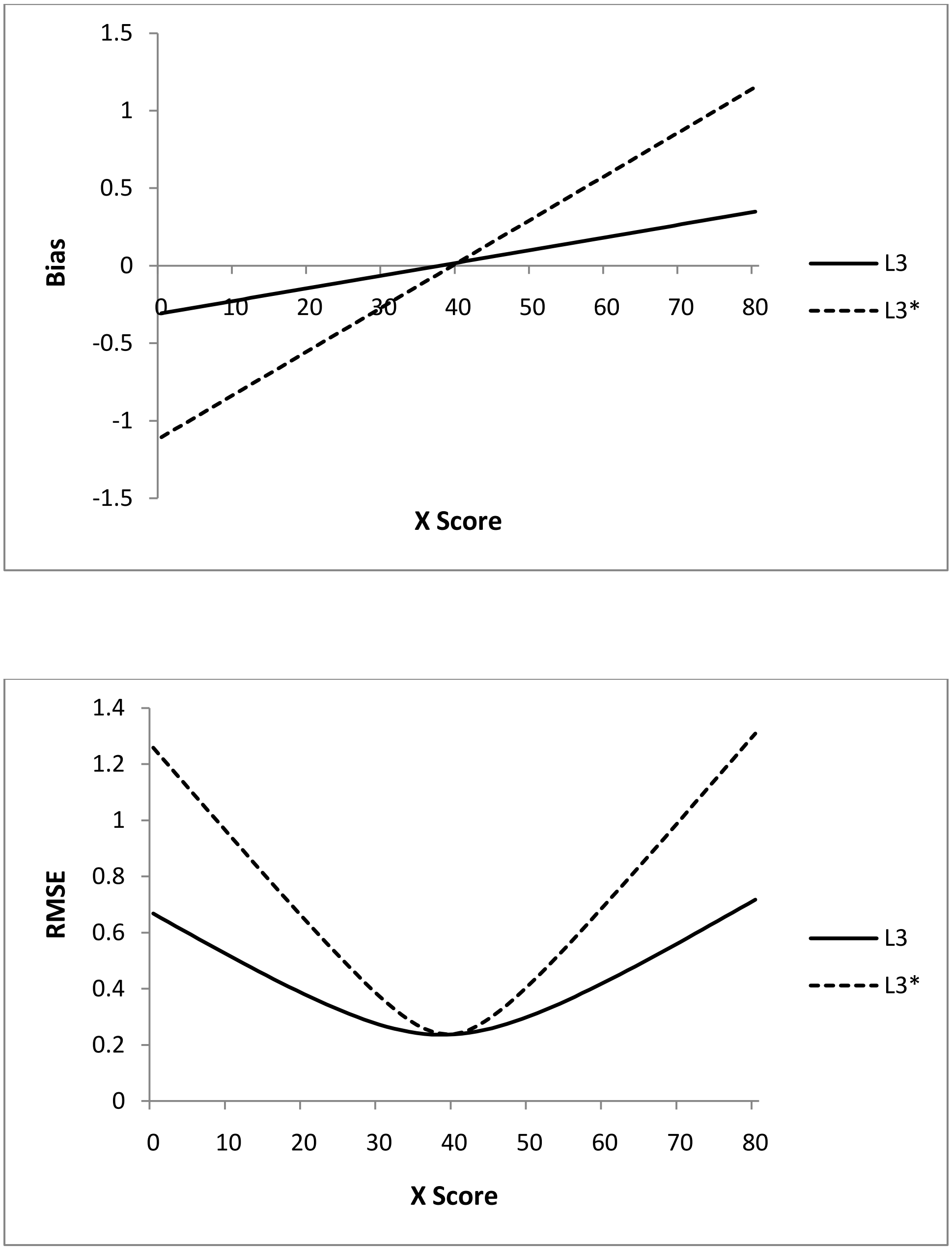


Figure 45. Bias (top) and RMSE (bottom) at score level for 100 samples, anchor 40%, more difficult test Y and ability N(50%, 18%).

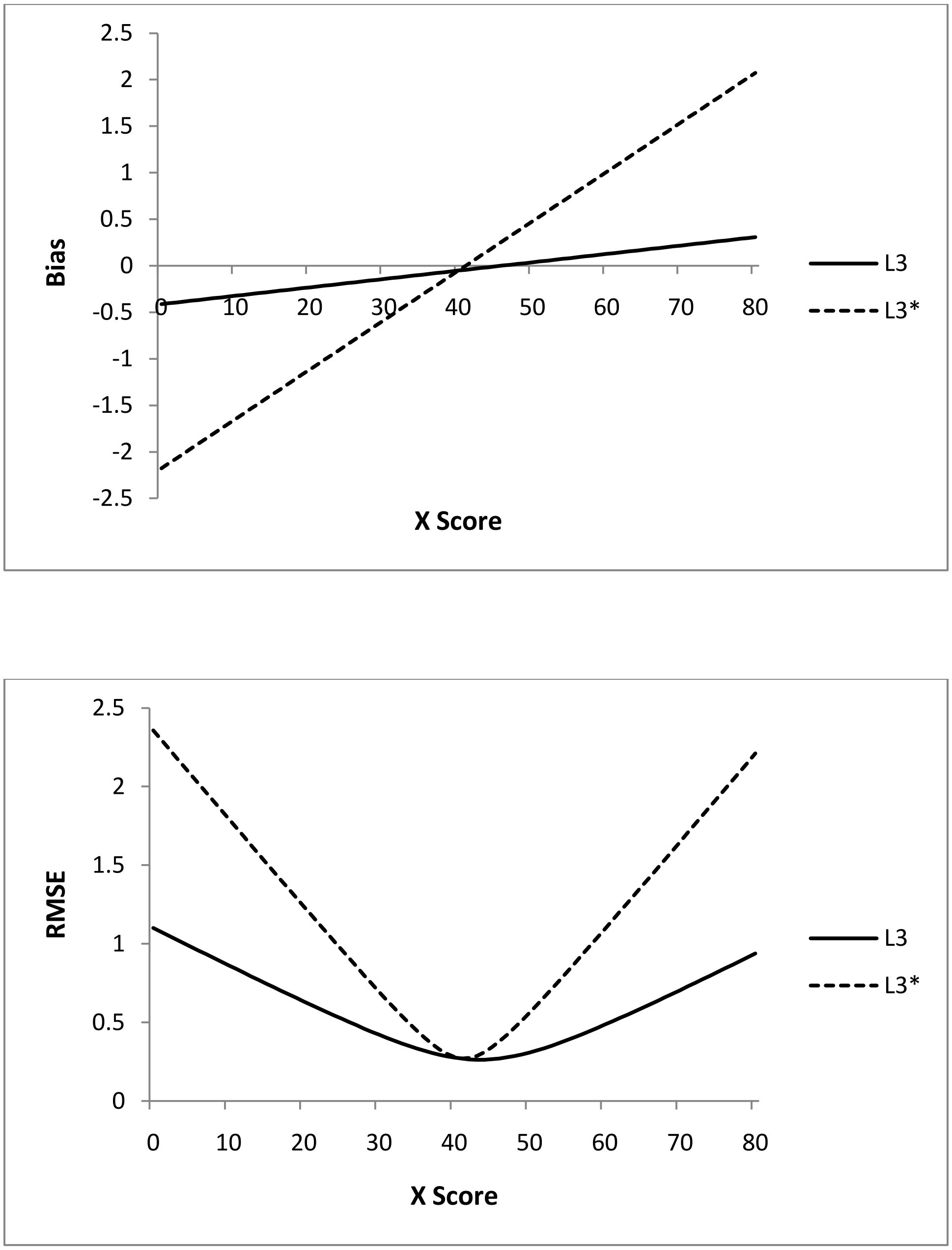


Figure 46. Bias (top) and RMSE (bottom) at score level for 100 samples, anchor 20%, easier test Y and ability N(50%, 18%).

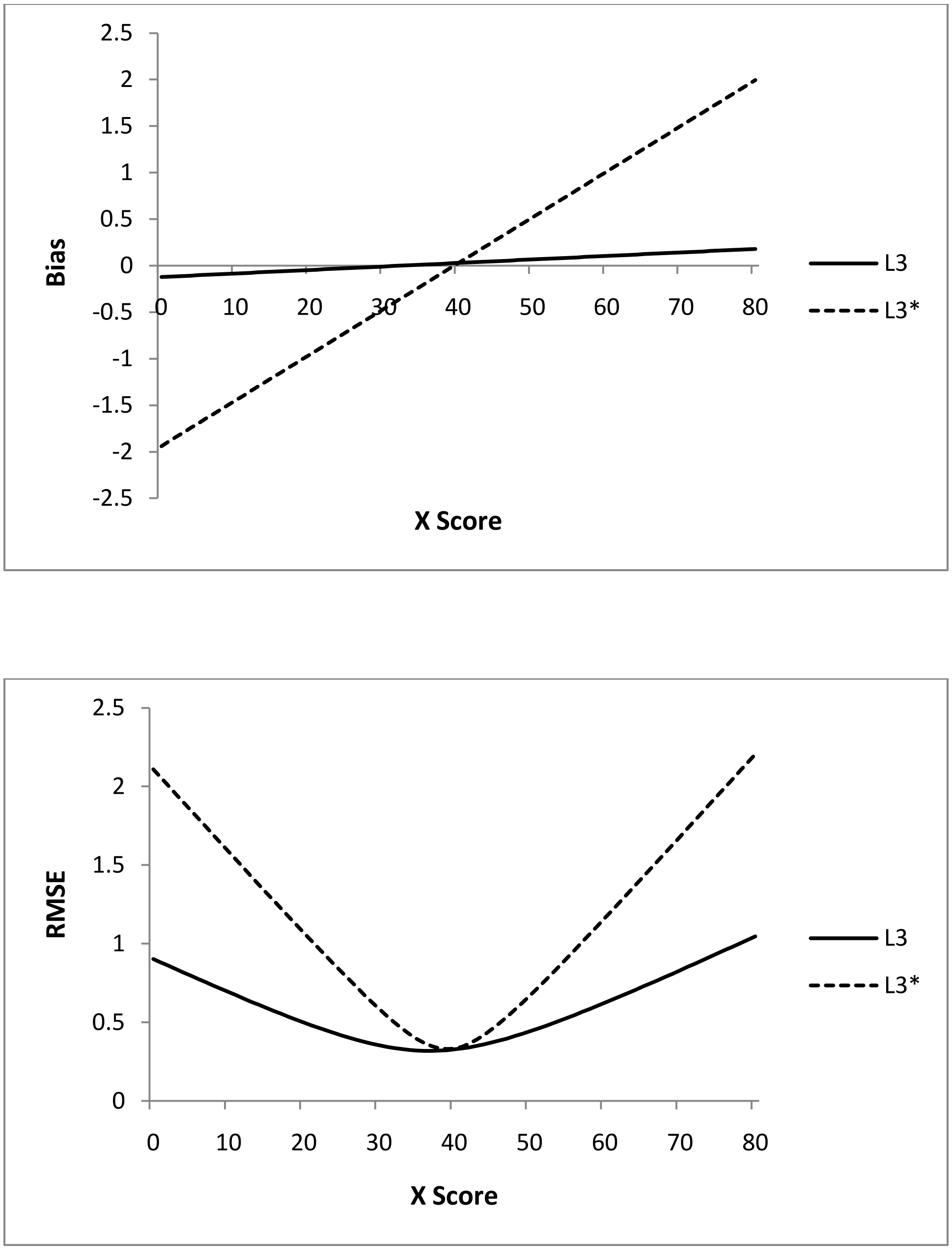
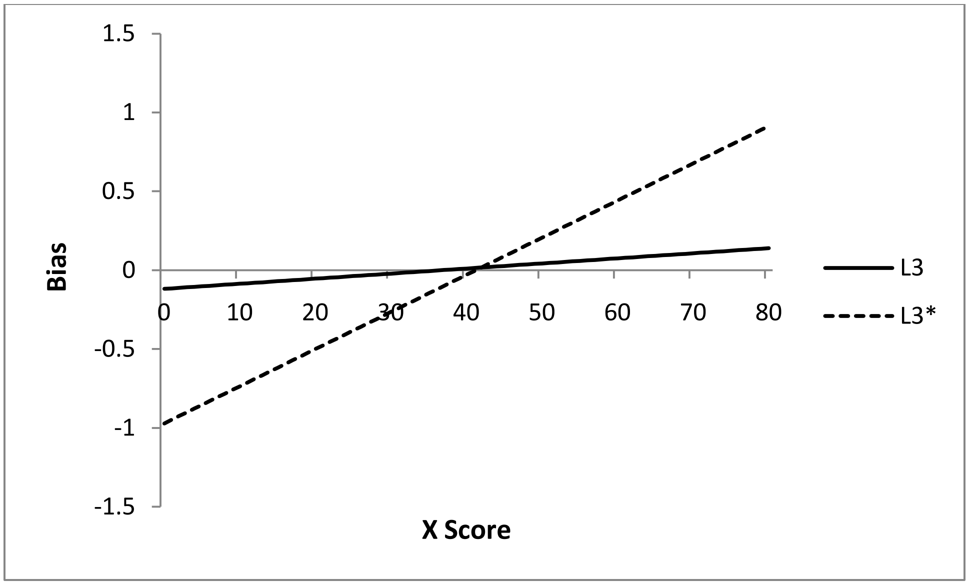


Figure 47. Bias (top) and RMSE (bottom) at score level for 100 samples, anchor 20%, more difficult test Y and ability N(50%, 18%).



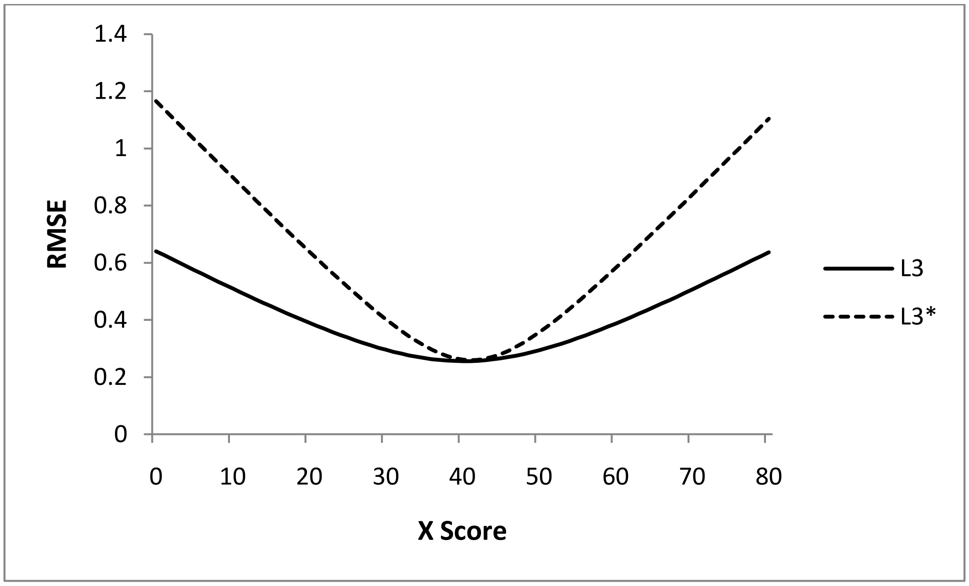
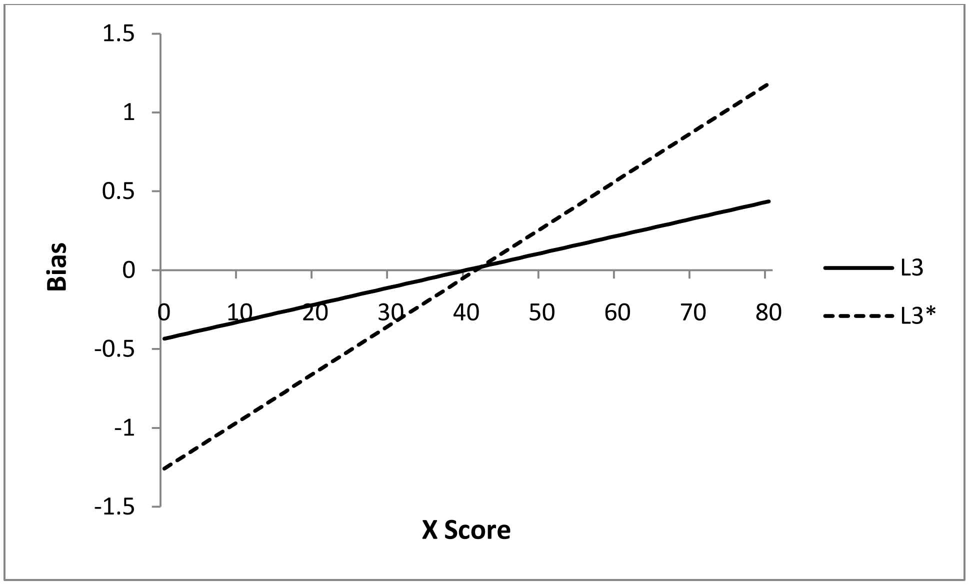


Figure 48. Bias (top) and RMSE (bottom) at score level for 100 samples, anchor 40%, easier test Y and ability N(55%, 18%).



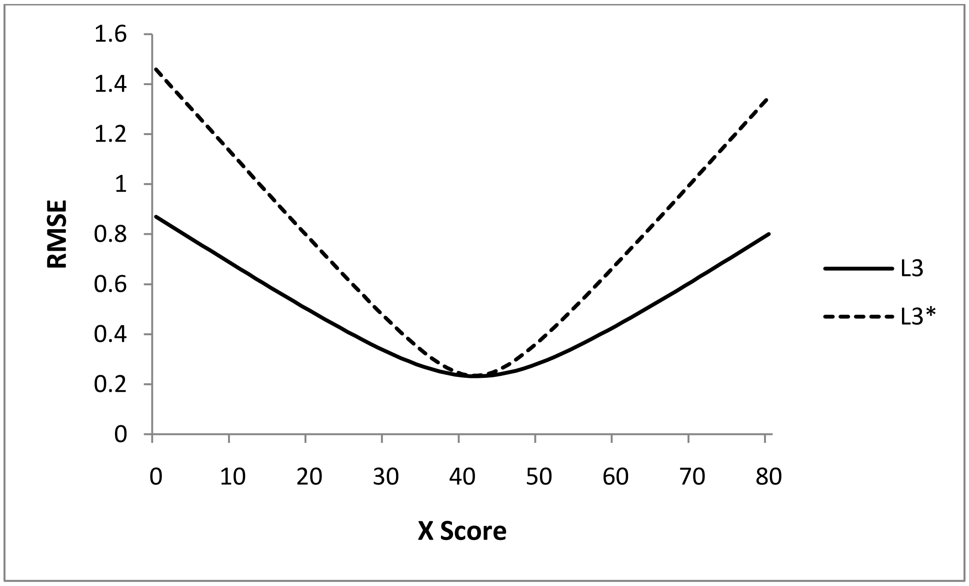


Figure 49. Bias (top) and RMSE (bottom) at score level for 100 samples, anchor 40%, more difficult test Y and ability N(55%, 18%).

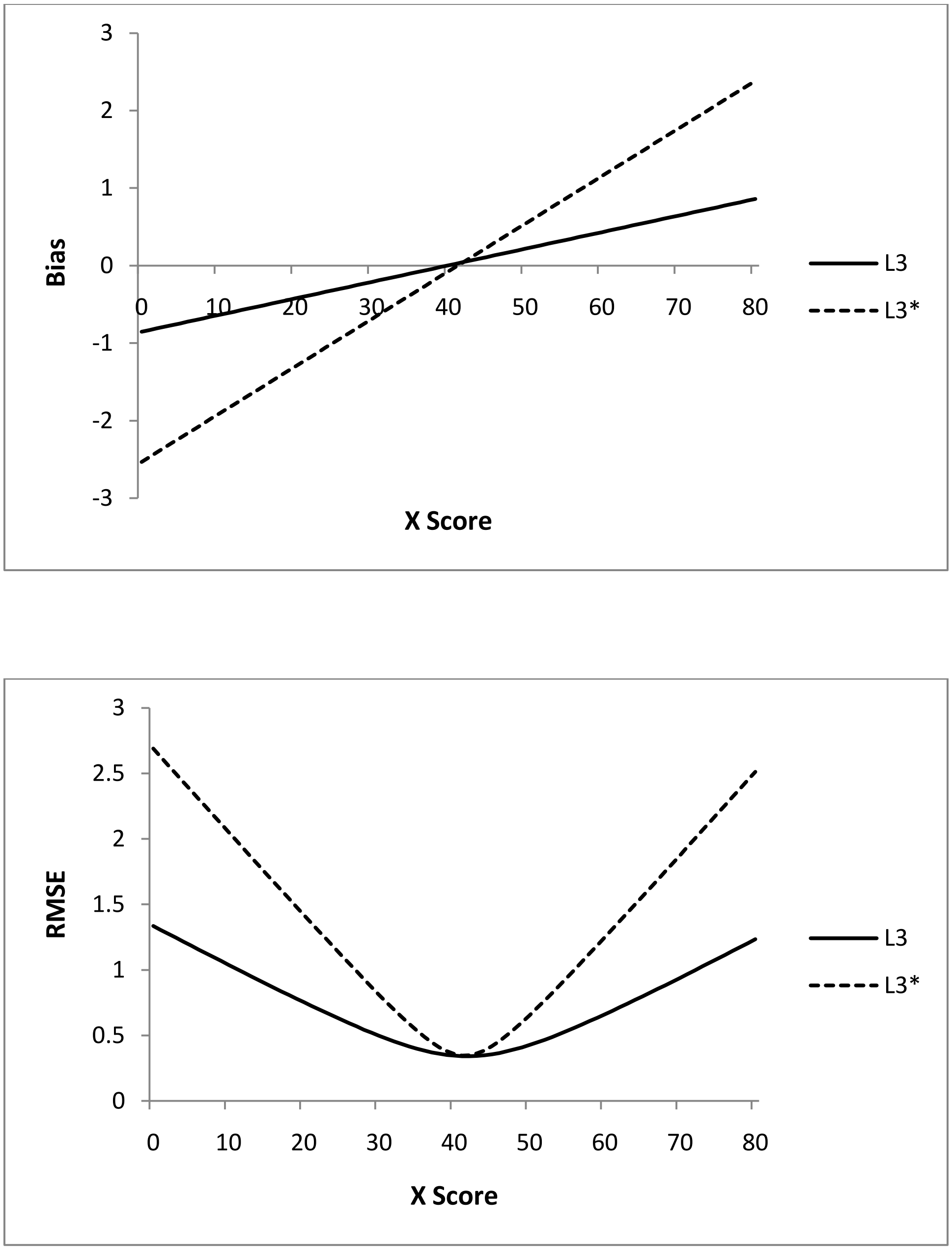
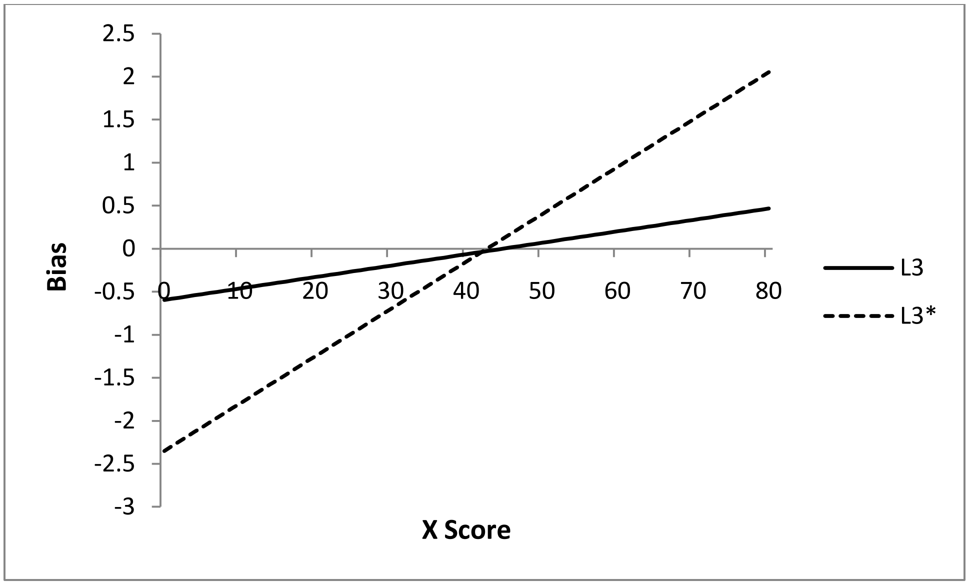


Figure 50. Bias (top) and RMSE (bottom) at score level for 100 samples, anchor 20%, easier test Y and ability N(55%, 18%).



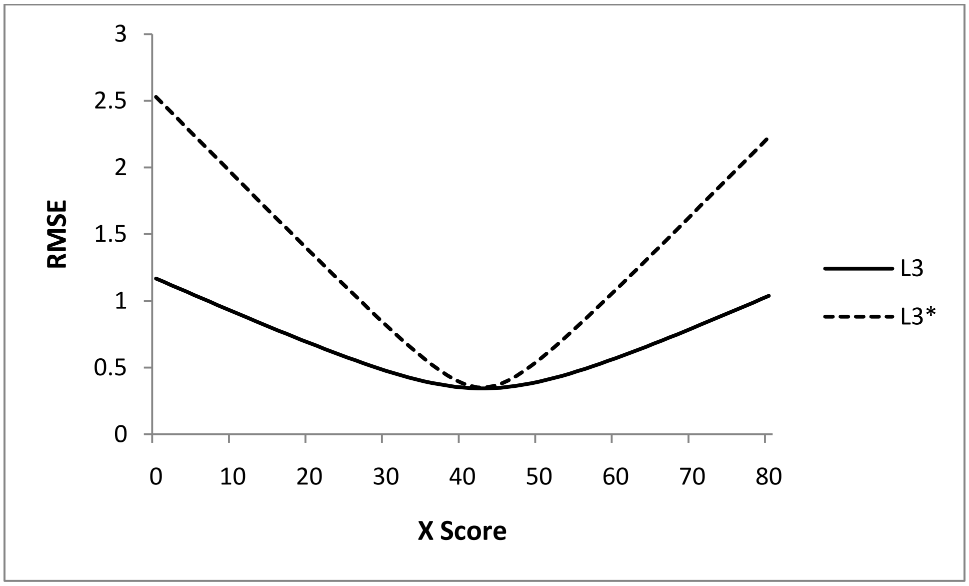


Figure 51. Bias (top) and RMSE (bottom) at score level for 100 samples, anchor 20%, more difficult test Y and ability N(55%, 18%).

In general, figures 36 to 51 show that both L3 and L3* have a larger bias and RMSE towards the extremes of the score range and a smaller bias and RMSE towards the middle of the score range. However there are important differences among conditions.

For conditions of equal ability distribution between P and Q, i.e. for condition N(50%, 15%) both methods seem to perform similarly.

For conditions in which Q has ability distribution N(55%, 15%) both methods seem to perform similarly except for the condition with anchor 20% and easier test Y, in which L3* seems to perform slightly better than L3.

In all the remaining conditions, in which SD is 18%, L3* is outperformed by L3 in the extremes of the score range, especially for anchor test 20%.

An Illustrative Example

To illustrate the use of L3 and L3* equating with a real data example these methods were applied to two 36-item forms X and Y. The data sets were obtained from the software CIPE referred by Kolen and Brennan (2004). For these forms the anchor A is formed by every third item (items 3, 6, 9, …, 36). Since scores on A are contained in X and Y, A is an internal anchor. However, according to Chapters Three and Four, the formulas derived from L3 and L3* can be applied to situations with either external or internal anchors.

Descriptive statistics for test X administered to a group P of 1,655 examinees and test Y administered to a group Q of 1,638 examinees are presented in Table 13.

Table 13
Descriptive Statistics for a Real Data Example

Group	Test	Mean	SD	Variance
P	X	15.8205	6.5298	42.6383
P	A	5.1063	2.3767	5.6489
Q	Y	18.6728	6.8805	47.3418
Q	A	5.8626	2.4522	6.0135

Note that mean A in P is 5.1063 which is about 42.6% of anchor length, whereas mean A in Q is 5.8626 which is 48.86% of anchor length. On the other hand, SD of A in P is 2.3767, which is 19.81% of anchor length, whereas SD A in Q is 2.4522, which is 20.44% of anchor length. Therefore the mean difference is about 6% and the SD difference is about 0.6%. Given the very small difference in SD, if the results of this study are applicable to a shorter anchor and shorter tests X and Y, the equated scores form L3 and L3* would be expected to be similar.

Reliabilities for tests A, X and Y are shown in Table 14.

Table 14
Reliabilities of Tests A, X and Y for a Real Data Example

Group	Test	Reliability
P	X	.842
P	A	.609
Q	Y	.860
Q	A	.630

After the application of an equating procedure analogous to the one described previously in this chapter for the equating at sample level, the equated scores from L3 and L3* were computed. It is not possible to compute true equated scores for this case; therefore the difference between the equated scores from L3 and L3* instead of bias was computed. Similarly, an RMSE cannot be computed, but it is possible to compute an RMSD based on the difference between the equated scores from L3 and L3*. The following two figures present the resulting difference and the corresponding RMSD. These difference and RMSD were computed based on the scores of the 1655 examinees in group P.

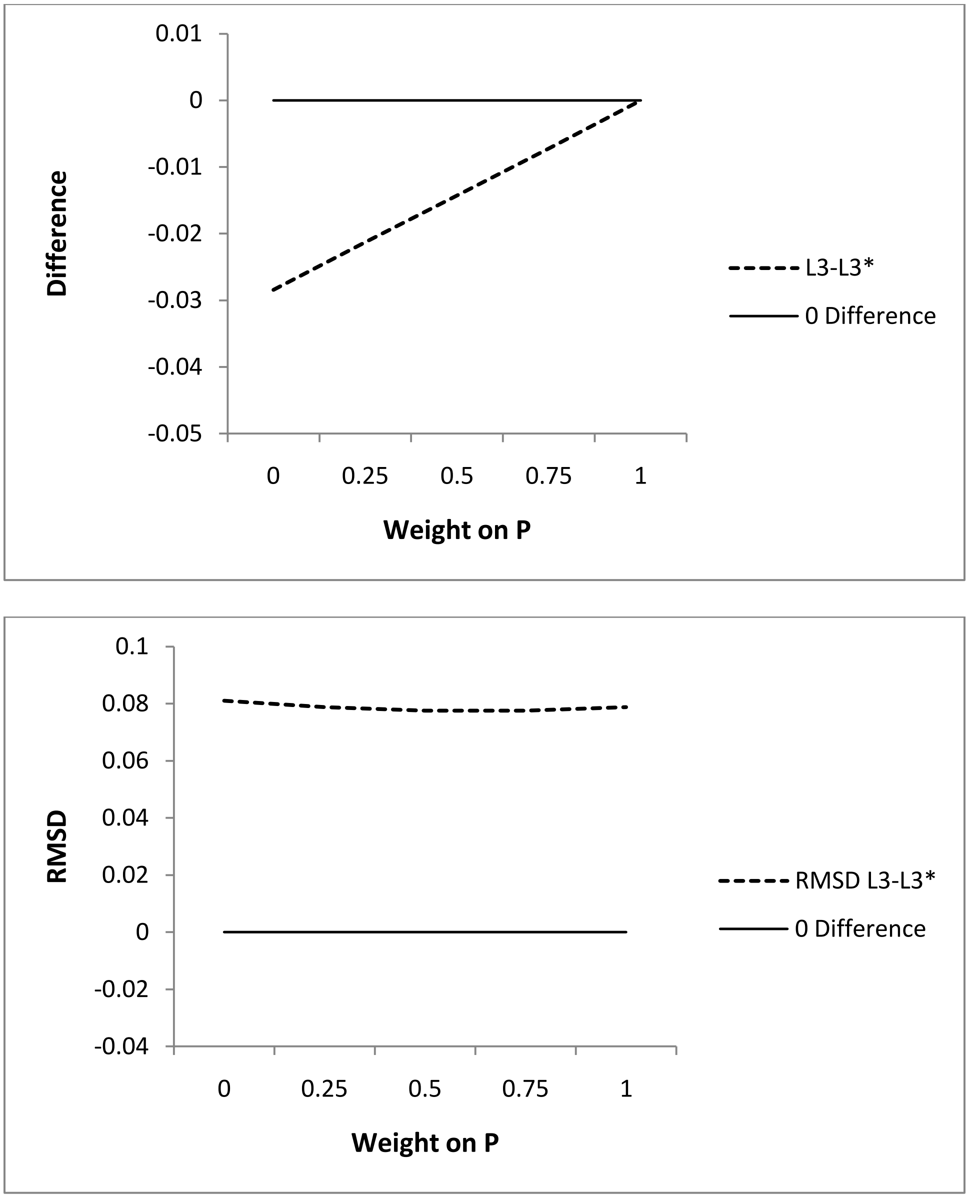


Figure 52. *Difference (top) and RMSD (bottom) for L3-L3* for a real data example.*

In addition, it is possible to compute the equated score for each score in the score range i.e. from 0 to 36. Table 15 shows the Y equivalent scores (equated scores) based on L3 and L3* for different *w* weights.

Table 15

Y Equivalent Scores for L3 and L3 for Five w Weights for a Real Data Example*

	w=0		w=0.25		w=0.50		w=0.75		w=1	
Score	L3	L3*	L3	L3*	L3	L3*	L3	L3*	L3	L3*
0	-0.23	-0.02	-0.23	-0.02	-0.23	-0.03	-0.23	-0.03	-0.23	-0.04
1	0.78	1.00	0.78	1.00	0.78	0.99	0.78	0.99	0.78	0.98
2	1.79	2.02	1.79	2.02	1.79	2.01	1.79	2.01	1.79	2.01
3	2.80	3.04	2.80	3.04	2.80	3.03	2.80	3.03	2.80	3.03
4	3.81	4.06	3.81	4.06	3.81	4.06	3.81	4.05	3.81	4.05
5	4.81	5.09	4.82	5.09	4.82	5.08	4.82	5.07	4.82	5.07
6	5.82	6.11	5.83	6.11	5.83	6.10	5.83	6.09	5.83	6.09
7	6.83	7.13	6.84	7.13	6.84	7.12	6.84	7.12	6.84	7.11
8	7.84	8.15	7.85	8.15	7.85	8.14	7.85	8.14	7.85	8.13
9	8.85	9.17	8.86	9.17	8.86	9.16	8.86	9.16	8.85	9.15
10	9.86	10.19	9.87	10.19	9.87	10.18	9.87	10.18	9.86	10.18
11	10.87	11.21	10.88	11.21	10.88	11.20	10.88	11.20	10.87	11.20
12	11.88	12.23	11.89	12.23	11.89	12.23	11.89	12.22	11.88	12.22
13	12.89	13.26	12.90	13.26	12.90	13.25	12.90	13.24	12.89	13.24
14	13.90	14.28	13.91	14.28	13.91	14.27	13.90	14.26	13.90	14.26
15	14.91	15.30	14.91	15.30	14.92	15.29	14.91	15.29	14.91	15.28
16	15.92	16.32	15.92	16.32	15.93	16.31	15.92	16.31	15.92	16.30
17	16.93	17.34	16.93	17.34	16.94	17.33	16.93	17.33	16.93	17.32
18	17.94	18.36	17.94	18.36	17.94	18.35	17.94	18.35	17.94	18.35
19	18.95	19.38	18.95	19.38	18.95	19.38	18.95	19.37	18.95	19.37
20	19.96	20.40	19.96	20.40	19.96	20.40	19.96	20.39	19.96	20.39
21	20.97	21.43	20.97	21.43	20.97	21.42	20.97	21.41	20.96	21.41
22	21.98	22.45	21.98	22.45	21.98	22.44	21.98	22.43	21.97	22.43
23	22.99	23.47	22.99	23.47	22.99	23.46	22.99	23.46	22.98	23.45
24	24.00	24.49	24.00	24.49	24.00	24.48	24.00	24.48	23.99	24.47
25	25.01	25.51	25.01	25.51	25.01	25.50	25.01	25.50	25.00	25.49
26	26.02	26.53	26.02	26.53	26.02	26.52	26.02	26.52	26.01	26.52
27	27.03	27.55	27.03	27.55	27.03	27.55	27.03	27.54	27.02	27.54
28	28.04	28.57	28.04	28.57	28.04	28.57	28.04	28.56	28.03	28.56
29	29.05	29.60	29.05	29.60	29.05	29.59	29.05	29.58	29.04	29.58
30	30.06	30.62	30.06	30.62	30.06	30.61	30.06	30.61	30.05	30.60
31	31.07	31.64	31.07	31.64	31.07	31.63	31.07	31.63	31.06	31.62
32	32.08	32.66	32.08	32.66	32.08	32.65	32.07	32.65	32.07	32.64
33	33.08	33.68	33.09	33.68	33.09	33.67	33.08	33.67	33.08	33.66
34	34.09	34.70	34.10	34.70	34.10	34.69	34.09	34.69	34.08	34.69
35	35.10	35.72	35.11	35.72	35.11	35.72	35.10	35.71	35.09	35.71
36	36.11	36.74	36.12	36.74	36.12	36.74	36.11	36.73	36.10	36.73

The following three figures illustrate the previous conversion table for three values of w: 0, 0.5 and 1.

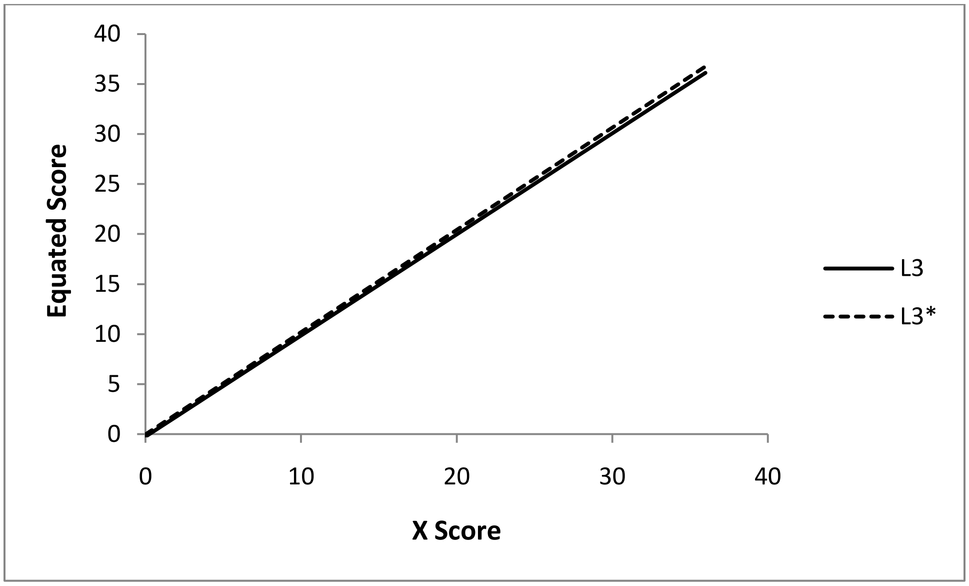


Figure 53. Equated scores for a real data example for w=0.

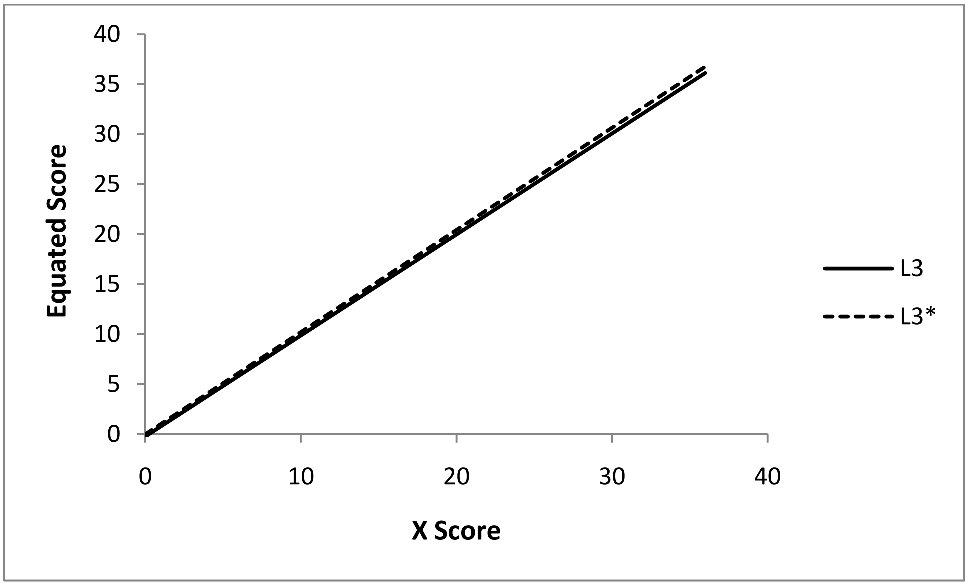


Figure 54. Equated scores for a real data example for w=0.5.

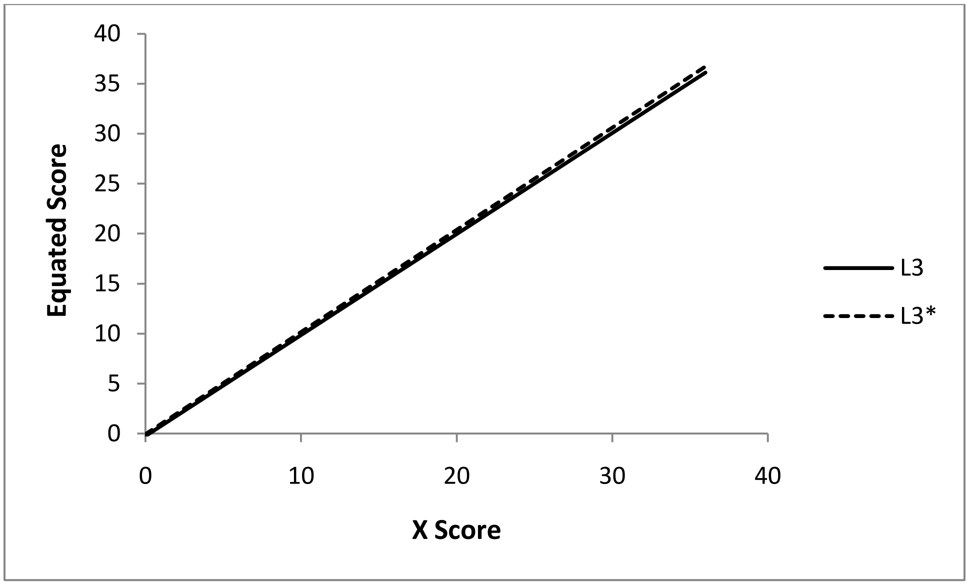


Figure 55. Equated scores for a real data example for w=1.

As can be noted, the expectation of obtaining very similar equated scores based in L3 and L3* for this example was confirmed. Given the characteristics in ability distribution of the groups P y Q in this real data example, this example confirms some of the results of the current study.

Chapter Five – Discussion

The administration of multiple forms is a current practice in many testing programs. Therefore, equating becomes a matter of necessity. Having various methods for equating is an advantage for test administrators since the results of such methods can be compared and evaluated in order to make a final decision before test scores are reported. Usually there is not a best method, but some methods can be better applied based on characteristics of the subjects taking the test and characteristics of the test itself.

That is why it is important to study under what conditions an equating method is likely to produce more accurate results and how an equating method compares to others.

This study focused on the comparison of two different sets of assumptions for the Levine Observed Score method of linear equating and how accurately these two sets of assumptions recover the true equating function. This was attempted through the design of a simulation study in which five factors were manipulated: anchor length, test difficulty, ability distribution, mixture of populations and sample size. The data generation and the computation of the equating functions were developed in the context of the Non Equivalent Groups with Anchor Test design (NEAT) whereby population P takes test X and population Q takes test Y and both populations take an anchor test A. Some assumptions are needed to estimate the equating function using only information from X in P, Y in Q and from the anchor A in both P and Q.

Data for a population P with 100,000 subjects and four populations Q of the same size were generated so that the four populations Q had the following characteristics regarding ability distribution: no difference, difference only in mean ability, difference only in variability and difference in mean ability and variability. The data generation also simulated the administration of two tests X and Y to those five populations. These two tests had a fixed length of 80 items but varied in two characteristics: anchor length and difficulty. Two conditions for anchor length were manipulated: 20% and 40% of test length. In addition, two conditions for test difficulty were manipulated: a test Y easier than test X and a test Y more difficult than test X. This was done to try to replicate reasonable settings in real testing situations. By combining population P and populations Q with certain chosen percentages, synthetic populations were created resulting in 80 cases corresponding to 16 conditions. The equating was conducted for those 80 cases (true equating, and equating under L3 and L3*).

Parallel to this simulation for synthetic populations, a sampling plan was implemented by drawing 100 pairs of samples of 500, 1000 and 2000 subjects from populations P and populations Q. The equating was conducted at sample level and bias and RMSE were computed. This sampling plan was conducted because in practice the equating does not happen at population level but at sample level.

This chapter discusses the results of this study. This discussion is organized in five sections: equating at population level, equating at sample level and equating in the samples at score level; then a discussion about meeting the assumptions is presented and a final section discusses the limitations of the study and future research.

Equating at Population Level

Although in the NEAT design two different populations take test X and test Y, the equating occurs in a single population (Kolen and Brennan, 2004) known as synthetic population. A common way to produce this synthetic population was proposed by Braun & Holland (1982) whereby P and Q were combined to produce the synthetic form $S = wP + (1 – w)Q$, where w varies from 0 to 1.

In this study a population P and four populations Q were combined and crossed with four test conditions: two conditions for anchor length and two conditions for difficulty of the test Y as well as five mixtures of populations. The result was 80 synthetic populations for which the three equating functions were computed and compared: a true equating function, an equating function form the traditional assumption L3 and an equating function from the alternative assumption L3*.

For the conditions in which P and Q had a similar ability distribution, namely N(50%, 15%), where those percentages refer to anchor length, no differences were observed between the two assumptions. L3* slightly outperformed L3 for one of the conditions for which Q had an ability distribution of N(55%, 15%); specifically when test Y was easier than test X and anchor length was 20% of test length. For the remaining conditions associated to N(55%, 15%) the two methods performed similarly.

However, a clear difference in performance between L3 and L3* was observed for conditions in which the SD of population Q was 18%. In these conditions L3 outperformed L3*, especially when anchor length was 20% of test length. For L3*, RMSE values as high as 0.97 were observed for this condition for w=0 and RMSE decreased to 0.75 for w=1, whereas the corresponding values for L3 were 0.34 and 0.29.

It is important to take into consideration the concept of difference that matters (DTM), which addresses the fact of whether or not the difference between two equating functions has important consequences for reported scores. Recall that DTM is dependant of the test and its use. For example on the SAT the DTM is 5 reported-score points because SAT scores are reported and rounded in steps of 10 points.

In the case of this study, assuming that the scores are reported in steps of 1 point, the DTM would be 0.5 and therefore values for RMSE such as 0.97 and 0.75 are greater than DTM for L3* whereas they are smaller than DTM for L3 (0.34 and 0.29 respectively).

Equating at Sample Level

The combination of the factors ability distribution (four conditions), anchor length (2 conditions) and test difficulty (2 conditions) produced 16 conditions. For each of the 16 conditions 100 samples of size 500, 1000 and 2000 from population P and 100 samples of size 500, 1000 and 2000 from population Q were randomly extracted in a bootstrap with replacement fashion whereby after a subject is selected for the sample he is returned to the population so that he could be selected again. This produced 300 pairs of samples P and Q for each of 16 conditions i.e. in total 4800 pairs of samples were generated.

For each of these pairs of samples five equating functions were computed under the L3 assumption and five equating functions were computed under the L3* assumption. Each of these five equating functions in each case was produced using a different *w* weight, where *w*: 0, 0.25, 0.50, 0.75, and 1.

Two repeated measures ANOVAs were conducted, one for bias and another for RMSE, and sample size were dropped since no interactions that included sample size were significant (statistically and practically). For bias, the ANOVA results showed that ability distribution interacted with anchor length, weight and assumption (L3 or L3*) and that for the condition N(55%, 18%), L3* with a anchor of 20% had the largest bias for the w=0 case. This bias decreased as weight got closer to 1. In the other three conditions for this interaction, bias was closer to 0 and L3 and L3* behaved similarly. Other interactions were significant for bias and confirmed how L3* was outperformed by L3 for conditions in which the SD was 18% of anchor length.

The ANOVA for RMSE showed that the interaction of assumption, ability, and anchor was significant. For the conditions with an 18% SD, RMSE is much higher for L3* for the condition anchor length 20%. Other interactions also confirmed the disadvantage of L3* for conditions with 18% SD and shorter anchor (20%).

These results at sample level confirm those at population level whereby L3* appears to be more affected than L3 by a greater variability in populations Q and a shorter anchor test.

Equating in the Samples at Score Level

To further explore the differences between L3 and L3* in the samples, bias and RMSE were computed at score level. The computation in the samples of size 2000 of bias and RMSE at score level show that both L3 and L3* have a larger bias and RMSE towards the extremes of the score range and smaller bias and RMSE towards the middle of score range.

For conditions of equal ability distribution between P and Q, i.e. for condition N(50%, 15%) both methods seem to perform similarly.

For conditions in which Q has ability distribution N(55%, 15%) both methods seem to perform similarly except for the condition with anchor 20% and easier test Y, in which L3* seems to perform slightly better than L3

However, for all the remaining conditions, in which SD is 18%, L3* is outperformed by L3 in the extremes of the score range, especially for anchor test 20%.

The results obtained with this analysis at score level appear to indicate that L3* is more affected than L3 in the extreme portions of the range of scores, especially for conditions in which SD is 18% and when anchor test is shorter.

Meeting the assumptions?

Given that in this simulation study the parameters in the population are known, it is natural to question whether the observed differences between L3 and L3* reside on the violation of assumptions, assumptions that otherwise are untestable in practical situations. For this purpose it is valuable to explore the degree to which the L3 and L3* assumptions are met in the data for this study.

Recall that the assumption that characterizes L3 is that the error variances $\sigma^2_{e_A/S}$, $\sigma^2_{e_X/S}$, and $\sigma^2_{e_Y/S}$ are the same for any S of the synthetic form. Table 16 presents the error variances for anchor test A and Table 17 presents the error variances for test X and test Y. These error variances were known in this study because of how the data was generated but they are, of course, unknown in real data situations.

It can be observed in Table 16 that the error variances corresponding to the cases where the original populations P and Q had equal ability distribution are very similar. For example for N(50%, 15%), anchor length 40% and test Y easier than test Y, which corresponds to the first data row of the Table 16, the error variances are 7.55, 7.53, 7.54, 7.52 and 7.51 for the w weights of 0, 0.25, 0.5, 0.75 and 1 respectively. On the other hand, there is a tendency of more variation in error variances as the mean and SD increases. For example for N(55%, 18%), anchor length 40% and test Y easier than test Y, which corresponds to the fourth row of Table 16 from the bottom up, the corresponding error variances are 7.14, 7.25, 7.33, 7.41 and 7.51, so it can be argued that there is some violation of the assumption of equal error variances for this case.

Table 16

Error Variances for Anchor Test A across 16 Conditions and Five w Weights

Anchor Length %	Test Y Diff	Ability	w				
			0.00	0.25	0.50	0.75	1.00
40	e	05	7.55	7.53	7.54	7.52	7.51
40	d	05	7.58	7.56	7.58	7.53	7.46
20	e	05	3.90	3.93	3.92	3.94	3.94
20	d	05	3.88	3.88	3.88	3.89	3.89
40	e	55	7.40	7.43	7.48	7.47	7.51
40	d	55	7.42	7.41	7.46	7.47	7.46
20	e	55	3.83	3.88	3.88	3.91	3.94
20	d	55	3.82	3.85	3.86	3.88	3.89
40	e	08	7.21	7.28	7.37	7.41	7.51
40	d	08	7.17	7.23	7.29	7.39	7.46
20	e	08	3.76	3.83	3.85	3.90	3.94
20	d	08	3.74	3.77	3.80	3.85	3.89
40	e	58	7.14	7.25	7.33	7.41	7.51
40	d	58	7.10	7.19	7.30	7.42	7.46
20	e	58	3.69	3.78	3.82	3.88	3.94
20	d	58	3.69	3.74	3.78	3.85	3.89

[a]Ability 05: N(50%, 15%) 55: N(55%, 15%) 08: N(50%, 18%) 58:N(55%, 18%)

[b]Test Y Diff e: Test Y is easier d: Test Y is more difficult

For tests X and Y in Table 17, a similar situation occurs. For conditions with equal ability distribution, like those of the first four data rows of Table 17, the error variances are very close to each other. For example for the condition N(50%, 15%), anchor length 40% and test Y easier than test Y, which corresponds to the first data row of the Table 17, the error variances for Y are 18.06, 18.07, 18.13, 18.06 and 18.01. On the other hand, for conditions with greater variability in Q relatively to P, greater difference in error variances are observed. For example, for N(55%, 18%), anchor length 40% and test Y easier than test Y, which corresponds to the fourth row of Table 17 from the bottom up, for test Y the error variances are 16.91, 17.23, 17.54, 17.76 and 18.01, therefore it can be argued that there is some violation of the assumption of the equal error variances for this case.

Table 17

Error Variances for Tests X and Y across 16 Conditions and Five w Weights

Anchor	Test		w									
Length	Y		X					Y				
%	Diff	Ability	0.00	0.25	0.50	0.75	1.00	0.00	0.25	0.50	0.75	1.00
40	e	05	18.26	18.34	18.39	18.42	18.48	18.06	18.07	18.13	18.06	18.01
40	d	05	18.21	18.23	18.25	18.28	18.19	18.09	18.10	18.15	18.16	18.12
20	e	05	18.31	18.26	18.24	18.20	18.15	18.16	18.11	18.09	18.06	18.05
20	d	05	18.27	18.25	18.26	18.12	18.13	18.29	18.29	18.24	18.17	18.06
40	e	55	17.99	18.10	18.26	18.34	18.48	17.48	17.58	17.77	17.89	18.01
40	d	55	17.93	18.04	18.10	18.20	18.19	18.23	18.20	18.21	18.17	18.12
20	e	55	17.86	17.96	18.11	18.11	18.15	17.42	17.55	17.65	17.87	18.05
20	d	55	18.12	18.08	18.14	18.08	18.13	18.27	18.30	18.25	18.15	18.06
40	e	08	17.54	17.80	18.07	18.33	18.48	17.32	17.51	17.75	17.87	18.01
40	d	08	17.62	17.80	17.97	18.11	18.19	17.21	17.45	17.68	17.93	18.12
20	e	08	17.39	17.57	17.80	17.96	18.15	17.31	17.50	17.70	17.88	18.05
20	d	08	17.39	17.53	17.68	17.86	18.13	17.44	17.66	17.86	17.93	18.06
40	e	58	17.05	17.41	17.77	18.13	18.48	16.91	17.23	17.54	17.76	18.01
40	d	58	17.20	17.46	17.74	17.99	18.19	17.43	17.57	17.78	17.96	18.12
20	e	58	17.32	17.56	17.79	17.98	18.15	16.79	17.12	17.41	17.71	18.05
20	d	58	17.24	17.43	17.64	17.81	18.13	17.34	17.57	17.78	17.89	18.06

[a]Ability 05: N(50%, 15%) 55: N(55%, 15%) 08: N(50%, 18%) 58:N(55%, 18%)

[b]Test Y Diff e: Test Y is easier d: Test Y is more difficult

On the other hand, recall that the characteristic assumption for L3* is that the ratios $\frac{\rho_{X/S}}{\rho_{A/S}}$ and $\frac{\rho_{Y/S}}{\rho_{A/S}}$ are constant as functions of S of the synthetic form (Holland & Walker, 2006); that is, the ratios of the square roots of the reliabilities $\frac{\rho_{X/S}}{\rho_{A/S}}$ and $\frac{\rho_{Y/S}}{\rho_{A/S}}$ are population invariant.

Those ratios are shown in Table 18 for each of the five *w* weights. It can be noted in Table 18 that there is clear difference between the first eight conditions and the second eight conditions. For the first 8 conditions i.e. conditions in which the ability distribution of the original P and Q are either equal or differ in mean ability only, the ratios of square root of reliabilities $\frac{\rho_{X/S}}{\rho_{A/S}}$ are extremely similar. Take for example the first data row of Table 18 for X

Table 18

Ratios of Square Root of Reliabilities

Anchor	Test		w									
Length	Y		X/A					Y/A				
%	Diff	Ability	0.00	0.25	0.50	0.75	1.00	0.00	0.25	0.50	0.75	1.00
40	e	05	1.087	1.087	1.086	1.086	1.084	1.087	1.086	1.085	1.085	1.084
40	d	05	1.089	1.088	1.089	1.090	1.087	1.086	1.087	1.089	1.089	1.087
20	e	05	1.224	1.229	1.226	1.229	1.226	1.226	1.229	1.226	1.228	1.226
20	d	05	1.223	1.220	1.217	1.217	1.218	1.220	1.219	1.216	1.217	1.218
40	e	55	1.086	1.083	1.084	1.083	1.084	1.086	1.084	1.084	1.083	1.084
40	d	55	1.085	1.083	1.084	1.086	1.087	1.087	1.085	1.085	1.086	1.087
20	e	55	1.217	1.218	1.216	1.221	1.226	1.220	1.220	1.217	1.221	1.226
20	d	55	1.219	1.215	1.211	1.213	1.218	1.219	1.214	1.212	1.214	1.218
40	e	08	1.060	1.065	1.070	1.077	1.084	1.061	1.065	1.070	1.076	1.084
40	d	08	1.058	1.062	1.070	1.078	1.087	1.059	1.064	1.071	1.078	1.087
20	e	08	1.159	1.175	1.188	1.205	1.226	1.160	1.175	1.186	1.205	1.226
20	d	08	1.158	1.169	1.181	1.197	1.218	1.159	1.170	1.181	1.196	1.218
40	e	58	1.059	1.062	1.068	1.076	1.084	1.059	1.063	1.068	1.075	1.084
40	d	58	1.057	1.061	1.068	1.079	1.087	1.058	1.062	1.069	1.078	1.087
20	e	58	1.151	1.166	1.182	1.202	1.226	1.153	1.167	1.184	1.203	1.226
20	d	58	1.154	1.165	1.177	1.196	1.218	1.154	1.166	1.178	1.197	1.218

[a]Ability 05: N(50%, 15%) 55: N(55%, 15%) 08: N(50%, 18%) 58:N(55%, 18%)

[b]Test Y Diff e: Test Y is easier d: Test Y is more difficult

[c]X/A: $\frac{\rho_{X/S}}{\rho_{A/S}}$ Y/A: $\frac{\rho_{Y/S}}{\rho_{A/S}}$

in which those ratios are 1.087, 1.087, 1.086, 1.086 and 1.084. A similar situation occurs for the ratios corresponding to test Y. In the same first data row of Table 18, those ratios are, 1.087, 1.086, 1.085, 1.085 and 1.084.

However, for the cases with 18% SD which are the second 8 cases of Table 18, much greater differences are observed in such ratios, with the largest difference observed for cases with anchor 20%. Take for example the case N(55%, 18%), anchor length 40% and test Y easier than test Y, which corresponds to the second to last row of Table 18. For test X the ratios of square root of reliabilities are 1.151, 1.166, 1.182, 1.202, and 1.226 and for test Y such ratios are 1.153, 1.167, 1.184, 1.203, and 1.226.

Whether these departures from the assumption for L3* are "big" cannot be determined since there is not a classification for these departures, but it is clear that there is a different pattern between the first 8 cases and the second 8 cases and that the cases with 18% SD and especially those with anchor test 20% were where the biggest differences in the equating between L3 and L3* were observed through this dissertation study.

Now, recall from Chapter Four that an accumulation of out of range scores occurred to a greater extent in the conditions that have an 18% SD and it was accentuated for conditions in which the test Y was easier and that this accumulation of out of range scores on the right side of the distribution at some extent modeled a ceiling effect in which very able subjects are not allowed to score higher than the maximum score in the test. This ceiling effect might be a confounding factor regarding the observed differences between the equating under L3 and L3*. This will be discussed in the next section.

On the other hand, even if there was a confounding effect from the ceiling effect observed in the 18% SD conditions, the fact is that L3 appears to be more robust to violations of assumptions because L3 showed a smaller RMSE and outperformed L3* in the conditions with 18% SD.

Limitations of the Study and Future Research

This study is limited to the use of normal distributions for the scores of tests A, X and Y. For future research other types of distributions could be used to compare the performance of L3 versus L3*.

Another limitation of this study is that the reliabilities were computed as the ratio of true score variance to observed score variance. An area of future research should be a different data generation that includes item responses, which can open the possibility to use other methods for the computation of reliabilities.

A third possible limitation of this study is the ceiling effect that was observed in some conditions in which there was an accumulation of out of range scores, especially for conditions with SD 18% and mean 55%. These conditions were precisely where the largest differences between L3 and L3* were observed. Although on one hand the ceiling effect is something that occurs in reality and it was seen as a reasonable setting for the study, on the other hand it is not possible to determine if the observed difference between L3 and L3* for certain conditions is due solely to the difference in the manipulated conditions in the study or the ceiling also effect also an effect. Given the conditions of the study, the difference was observed but it would be valuable to disentangle this possible interaction between the conditions of the study and the ceiling effect. Therefore an area of future research is a design in which the effects of ceiling effect and the effects of the differences in standard deviation of ability distribution between population P and populations Q can be isolated.

Another area of future research should be a variation of the values *a, b, c* and *d* for the data generation, in order to understand the effect of such values in the relative performance of L3 and L3*.

References

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

Braun, H. I., & Holland, P. W. (1982). Observed-score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 9-49). New York: Academic Press.

Carvajal, J., Walker, M., & Oh, H. J. (2008). A comparison of Angoff's and Holland's Assumptions in the Levine Observed Score Equating Function. Paper presented at the annual meeting of the National Council on Measurement in Education (NCME). New York, NY.

Cook, L. L., & Eignor, D. R. (1991). An NCME Instructional Module: IRT equating methods. Educational Measurement: Issues & Practice., 10(3), 37-45.

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

Dorans, N. J. (1990). Equating Methods and Sampling Designs. *Applied Measurement in Education,* 3, 3-17.

Dorans, N. J., and Holland P.W. (2000). Population Invariance and the Equatability of Tests: Basic Theory and the Linear Case. *Journal of Educational Measurement*, 4, 281-306.

Harris, D. J., & Kolen, M.J. (1990). A comparison of two equipercentile equating methods for common item equating. *Educational and Psychological Measurement*, 50, 61-71.

Holland, P. W. (2004). *Three methods of linear equating for the NEAT design*. Unpublished manuscript. Princeton, NJ: Educational Testing Service.

Holland P. W. & Dorans, N. J. (2006). Linking and Equating. In R.L. Brennan (Ed.), *Measurement* (pp. 187- 220). Conneticut: Praeger.

Holland P. W. & Walker, M. (2006). *Notes on Angoff's estimates of the true score slope in Levine's Observed Score Equating function*. Unpublished Manuscript. Princeton, NJ: Educational Testing Service.

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

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. (2004). *Equating Test Scores (without IRT).* Princeton, NJ: Educational Testing Service.

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

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

Mekhael, M., Wenmin, Z., & Miao, J. (2009). The Robustness of Tucker and Levine Observe Equating Methods in the Nonequivalent Group Anchor-Test Design. Paper presented at the annual meeting of the National Council on Measurement in Education (NCME). San Diego, CA.

Puhan, Gautam. (2010). A comparison of Chained Linear and Poststratification Linear Equating Under Different Testing Conditions. *Journal of Educational Measurement*, 47: 54-75.

Suh, Y., Mroch, A., Kane, M., & Ripkey, R. (2009). An Empirical Comparison of Five Linear Equating Methods for the NEAT Design. *Measurement*, 7, 147-173.