Spring 2011

An evaluation of a new method of IRT Scaling

Shelley Ragland
James Madison University

An Evaluation of a New Method of IRT Scaling

Shelley Ragland

A dissertation submitted to the Graduate Faculty of

JAMES MADISON UNIVERSITY

In

Partial Fulfillment of the Requirements

for the degree of

Doctor of Philosophy

Assessment and Measurement

May 2010

Dedication

This manuscript is dedicated to all of my family members (two-legged and otherwise) who have supported me along this interesting journey. In order of appearance: Daddy, Beebo, Erica, Sarah, Rags, Bootsie, Gilbert, Sullivan, Jack-Jack, Thad, and TDR, Jr.

Acknowledgements

I would like begin by thanking everyone in JMU's program of the Center for Assessment and Research Studies. I am so grateful to have had the opportunity to work with such amazing students, faculty, and staff. The program allowed me to reach for my own particular star, but it also assured my awareness of many others.

I would like to thank the members of my committee (Dr. Josh Goodman, Dr. Robin Anderson, and Dr. Peter Pashley), each of whom has supported my work in this research at every step along the way. Their thoughtful contributions and guidance have been invaluable. I would also like to especially thank Dr. Ron Armstrong, for his work on the computer programming used in much of this research.

Finally, I would like to thank my advisor and committee chair, Dr. Christine DeMars. I have learned so many wonderful things from you, beginning with my first IRT class. Throughout this project you have remained responsive, supportive and understanding, providing the perfect balance of sympathetic listener and dissertation task-master. X and Y are quite delighted to have joined your troop.

Table of Contents

List of Tables

List of Figures

Abstract

In order to be able to fairly compare scores derived from different forms of the same test within the Item Response Theory framework, all individual item parameters must be on the same scale. A new approach, the RPA method, which is based on transformations of predicted score distributions was evaluated here and was shown to produce results comparable to the widely used Stocking-Lord (SL) method under varying conditions of test length, number of common items, and differing ability distributions in a simulation study. The new method was also examined using actual student data and a resampling analysis. Both the simulation study and actual student data study resulted in very similar transformation constants for the RPA and SL methods when 15 or 10 common items were used. However, the RPA method produced greater variance, especially when only 5 common items were used in the actual student data analysis compared to the SL method. The simulated and actual data research findings demonstrate that the RPA method is a viable option for producing the transformation constants necessary for transforming separately calibrated item parameter estimates prior to equating.

CHAPTER I

Introduction

The goal of modern scaling and equating procedures is to enable test users to fairly compare scores derived from different forms of the same linear test or scores derived from different items on a computerized adaptive test. When scaling and equating within the IRT context, it is essential to ensure that all individual item parameters are placed on the same scale. A new approach to ensuring a common scale based on transformations of predicted score distributions was presented by Ragland, Pashley and Armstrong (2009). Since this method (RPA) was shown to produce comparable model fit to other methods, it is now important to see how it compares with other scaling (or transformation) methods under various testing conditions.

The focus of this research is on the scaling procedure that precedes the equating process. Although actually equating scores from different forms will not be conducted in this study, it is important to consider the scaling within an appropriate equating framework. The equating design that will be used here involves equating two forms that share a common subset of items (common items), but does not assume that randomly equivalent groups of test takers were administered the two forms (i.e., a common-item nonequivalent groups design). Since equating can be conducted using classical test theory or item response theory, for the sake of brevity and simplicity, the equating context here will be within the item response theory (IRT) framework.

When IRT is used to calibrate items from different forms, various approaches may be used to place all items on the same IRT scale for equating purposes. One straightforward way of accomplishing this is to calibrate all items concurrently, say in

one BILOG calibration run. When concurrent calibration of all items is performed, the items from both forms are automatically placed on the same IRT scale and no further scaling is required.

Situations exist, however, that do not allow for the concurrent calibration of items. For example, individual item responses might not be available for both forms in situations where the original item responses are not available. Another situation that precludes the use of concurrent calibration occurs when items need to be placed on an existing IRT scale. For instance, there are times when one needs to equate examination results from multiple years of testing back to a previous administration. Additionally, if more than two forms (or item pools) need to be equated, concurrent calibrations of all items might become impractical or unwieldy as the number of items or examinee groups becomes very large. After reviewing the relevant literature, Kolen and Brennan (2004, chap. 6) concluded that concurrent calibration might not be the best approach to scaling items to a common scale for two reasons: violations of assumptions can result in less accurate estimates, and problems with parameter estimation are obscured with the calculation of only one estimate. With two estimates, different parameter estimates can be easily compared for similarities, but with only one estimate we may not know if a problem exists. Furthermore, separate calibration may be preferred in situations where the IRT model does not closely hold (Béguin, Hanson & Glas, 2000; Hanson & Béguin, 2002).

As an alternative to the concurrent calibration of test forms, separate test form IRT calibrations can be performed first, and then the resulting item parameters can be

placed on a common scale in a second step. This second step can be accomplished by noting that IRT parameters are indeterminate up to a linear transformation.

In the case of the 3PL model, the common item parameters from the two forms—Form R, the reference form and Form E, the equated form—should be related as follows:

$$a_{R_j} = \frac{a_{E_j}}{A}; \quad b_{R_j} = Ab_{E_j} + B; \quad c_{R_j} = c_{E_j}; \tag{1.1}$$

where the *a*, *b*, and *c*'s denote the discrimination, difficulty, and pseudo-guessing 3PL parameters, respectively; the indices R and E refer to Forms R and E; *j* indexes the common items; and *A* and *B* are the transformation constants. (Also note that the *A* and *B* are also called "scaling" constants in the equating literature.) Applying the transformation constants *A* and *B* to all Form E item parameters should place them on the same IRT scale as the Form R item parameters.

To date, two categories of procedures have been available to estimate the scaling constants *A* and *B*: the first employs the moments of the 3PL discrimination and difficulty parameters (Marco, 1977; Loyd & Hoover, 1980) and the second considers transformations of item characteristic curves (Haebara, 1980) or test characteristic curves (Stocking & Lord, 1983). Kolen and Brennan (2004) provide complete descriptions of these approaches, along with their strengths and weaknesses.

There are two approaches that use the moments of the *a* and *b* parameters: Mean/Mean and Mean/Sigma. The Mean/Mean method takes advantage of the statistical stability of the mean for each of the two item parameters by using the mean of the *a* parameters and the mean of the *b* parameters (Loyd & Hoover, 1980). The mean usually represents a distribution better than the standard deviation, and is therefore more stable. However, using the mean of the *b* parameters and the standard deviation of the *b*

parameters as introduced by Marco (1977), capitalizes on the stability of the item difficulty parameter, *b*, as compared to the discrimination parameter, *a*. While theoretically the ratio of the standard deviation of the *b*'s is equal to the inverse of the ratio of the mean of the *a*'s, the estimated *A* scaling constant will not be exactly the same due to estimation error in the item parameters. Both of these methods have been demonstrated to produce acceptable scaling constants (Hanson & Béguin, 2002; Kim & Cohen, 1998; Kim & Lee, 2006; Michaelides, 2006).

However, both the means and standard deviations of the item parameters may be affected by outliers (Baker & Al-Karni, 1991; Hu, Rogers & Vulmirovic, 2008). In order to minimize the effects of unusual difficulty parameter estimates, Haebara (1980) developed an approach that used the sum of the squared differences of the item characteristic curves. Stocking and Lord (1983) proposed a similar approach that used the squared difference of the test characteristic curve. The use of item characteristic curves and test characteristic curves in both these methods considers all of the parameters of the items.

In the RPA method under investigation here, the differences between the predicted score distributions (i.e., IRT estimated distribution of observed number-correct scores) are minimized instead of considering transformations of test characteristic curves. This procedure for deriving the IRT scale scaling constants *A* and *B* places the item parameters from alternative forms on the same scale by using the observed distribution of scores and the IRT estimated distribution of observed scores. It uses the probabilistic model that yields the estimated distribution of number correct scores given the distribution of ability and the item parameters, determines the actual distributions of

scores for the different forms, and minimizes the differences between those distributions to determine the scaling constants.

In contrast to the Stocking and Lord procedure, which finds scaling constants that minimize the difference between Form R and Form E test characteristic curves (under the assumption that they should be equivalent, aside from sampling and estimation error), in the RPA method the difference is minimized between the predicted score distributions for the common items from Form R and Form E (again, under the assumption that they should be equivalent, aside from sampling and estimation error).

The IRT parameter scaling approach introduced by Ragland et al. (2009) involves five basic steps.

1. Calibrate items from Form E based on its administration to a group of examinees, and estimate the latent traits of all examinees administered Form R. This can be done with a suitable 3PL estimation package.
2. Determine observed score frequencies from the administration of Form R.
3. Compute a predicted score distribution based on the latent traits estimated from Form R and the scaled IRT parameters of Form E given values for *A* and *B* (initially 1 and 0).
4. Derive *A* and *B* scaling constants that minimize a distance measure between score frequencies of Steps 2 and 3.
5. Apply the scaling constants to all Form E item parameters to place them on the same scale as Form R.

In addition to comparing the scaling constants generated by the RPA method with other transformation methods, it is important to examine them in various testing

conditions. For this research the test length, ability distribution and number of common items will be studied. In general, longer tests are more psychometrically sound than shorter tests under the same construction, but a balance must be made between test length and examinee motivation, attention, and fatigue. A test should be long enough to satisfy content domain specifications as well as desired psychometric rigor. These requirements often depend on the purpose of the test. Test construction often strives to create the shortest test possible that will meet these two different sets of requirements.

When the populations of test takers are sufficiently large and appropriately assigned to test forms they may be considered "randomly equivalent." No other assumptions need to be made, and more simple types of equating may be used (Braun & Holland, 1982). As those ability differences increase, however, more complex methods must be employed, and additional assumptions included. The degree of ability differences for students has been shown to be a factor in the accuracy of the equating process (Harris & Kolen, 1986; Loyd & Hoover, 1980; Marco, Peterson, & Stewart, 1983; Skaggs & Lissitz, 1988; Slinde & Linn, 1977, 1978), and may also impact the calculation of the transformation constants.

A growing body of research has begun to explore the characteristics of the common items that are used for generating the scaling constants (Gao, Zhu, Chen and Harris, 2008; Fitzpatrick, 2008; Meyers, Miller, & Way, 2009; Michaelides, 2006; Michaelides & Haertel, 2004; Sinharay & Holland, 2006; Yang & Houang, 1996). In particular, the number of common items used, their difficulty levels and distributions, sampling strategies for selection have been examined and will be addressed more thoroughly in Chapter II.

The purpose of this study is to compare the scaling coefficients generated by the new RPA method with the most commonly used IRT method, Stocking-Lord. Differences in test length, ability distributions and number of common items will be examined. A case of actual examination data will also be analyzed.

CHAPTER II

A Review of the Literature

Equating is an important process employed in the administration of large scale tests. When large numbers of examinees are tested, the need to provide test security and to prevent item over-exposure are facilitated by the use of different examination test forms, that, although containing different items, are constructed to be as nearly identical as possible with regard to item content and difficulty. Despite the tremendous effort made to ensure that different test forms are as similar as possible, differences do exist between forms, and for this reason, test equating must be conducted to ensure that scores across different forms are both fair and comparable. The importance of equating cannot be understated. "Only when tests are equated can it be fair to give them to different people and treat the scores as if based on the same test" (Holland & Rubin, 1982, p. 1).

The purpose of this chapter is to introduce the basic concepts of equating, including equating designs and equating properties. It will then describe the differences between equating contexts of classical test theory (CTT) and item response theory (IRT), with an emphasis on IRT, and justification for its use here, as that is the context for this study. The chapter will then describe in detail methods of estimating scaling constants in IRT that have been used and developed to date as precursors to equating. Because the scaling constants may eventually be used in an equating process, IRT true-score and IRT observed-score equating will be described. The chapter will conclude with the descriptions of test and item characteristics that have shown to be a factor in the equating process: test length, ability differences, common items, and sample size.

Equating Designs

A number of equating designs exist that may be used to accommodate the needs of test developers and the availability of appropriate examinees. The three most frequently used equating designs are the (1) single group, (2) randomly equivalent groups, and (3) common-item nonequivalent groups designs. Each design makes assumptions about the characteristics of the groups taking the test forms that are to be used in the equating.

The single group design is used when the same examinees are tested on both forms of the test. This method potentially introduces much less error than the other approaches because the same examinees take all the same items (Harris, 1991a; Zeng, 1991). Because examinees take the same versions of the same test, testing time is increased (doubled for two tests, tripled for three tests, and so on) so that using this method is not feasible when several forms are used. A testing effect of fatigue could lower the scores on the second test that was administered, but practice effects could raise the scores on the second test. For these reasons, two test forms are often spiraled or counterbalanced among the examinees so that half the examinees are randomly assigned Form R first and Form E second, while the other half are assigned Form E first and Form R second. The single group that is formed must be representative of the population to which scores will be generalized. For equating large scale nationally administered tests, finding an appropriate single group may be problematic (Muraki, Hombo, & Lee, 2000; Petersen, 2007).

Another commonly employed equating design uses randomly equivalent groups (Kim, Choi, Lee, & Um, 2008; Kolen, 2008; Yi, Harris, & Gao, 2008). The primary

advantage of this design is that examinees must only be tested once, thus reducing testing time. The primary disadvantage, however, is that very large representative samples must be selected and that all forms to be equated are administered at the same time (Kolen, 1988).

The equating design of interest for this study is the common-item nonequivalent groups design, which is another frequently used design in testing (Duong & Reckase, 2008; Gao, Zhu, Chen & Harris, 2008; Sinharay & Holland, 2007). It can be used when the ability of examinee groups is not known, when sample sizes are not large enough, or when sampling strategies are not robust enough to ensure randomly equivalent groups. For this approach, common items, also called anchor items, must be included on both forms of the test. When the items are included in examinees' score, they are known as *internal* common items; when they are not included in the score, they are *external* common items. External common items are sometimes administered separately as a different test.

One advantage of the common-item nonequivalent groups design is that each group is required to take only one test. Additionally, assumptions about the population characteristics required by other methods do not have to be met in the common-item nonequivalent groups design, which makes this design easier to implement from an administrative perspective. This design, however, does require the use of common items that are administered to both groups of examinees. Common items can introduce another facet of complexity to the equating process and will be discussed further in a later section. The overarching goal of equating in this situation is to separate group differences from form differences by using the jointly administered common items (Kolen, 1988).

Equating Properties

Several assumptions must be made to employ any equating methodology: assumptions about the data-collection design that applies to equating in general, assumptions about the use of IRT as a psychometric framework, and assumptions specific to the particular equating methods that will be used (von Davier & Wilson, 2007). For equating different versions, or forms, of the same test, certain properties must hold. When these properties do not hold, tests may be linked, but this is a much weaker relationship than equating.

Five properties have been identified that are important in equating (e.g., Dorans & Holland, 2000; Holland, 2007; Petersen, 2007). Some equating properties are based on observed variables, like the single score an examinee earns on one administration of the test, while others are based on unobservable variables, like the true score or latent ability of an examinee (Kolen & Brennan, 2004). These latter methods invoke test theory models.

The properties of symmetry, same test specifications, and same test reliability are based on observed scores. Symmetry requires that the transformation of a score on Form R to Form E must have an inverse transformation to convert a Form E score to the Form R scale. The same test specifications regarding content coverage must be present in order to conduct equating. The same test reliability for tests to be equated affects the equity of the tests directly as well as that of the same test specifications. It should be a matter of indifference to the examinee which test to take (Lord, 1980, chap. 13). However, Dorans and Holland (2000) suggest that nearly equal reliabilities for tests will suffice, and that higher levels of reliability should be of greater importance than equal levels of reliability.

Common equating methods based on observed scores include mean, linear, and equipercentile equating.

The equity property and group invariance property are based on unobserved variables. Lord's equity property, also known as first-order equity (Lord, 1980, chap. 13), is based on true scores from classical test theory. A true score is defined to be the expected value of the examinee's observed score (Lord & Novick, 1968). Examinees with the same true score would have the same observed score means, standard deviations and distributions on both forms under Lord's equity. However, this requires the forms to be identical, which rarely occurs in practice. As an alternative, Morris (1982) defined a weak equity property, or second-order equity that requires only that the expected scores on test forms to be identical. The group invariance property states that the equating procedure will be the same regardless of the groups used to conduct the equating. For example, Dorans and Holland (2000) showed that when the population invariance assumption did not hold for subpopulations based on gender, language at home or ethnicity, the linking functions were not similar.

The equity assumption may be violated in practice. Tong and Kolen (2005) tested the robustness of equating methods with regard to the first- and second-order equity properties and found that as the level of difficulty increased between the different forms, the first and second order equity properties did not hold. Sinharay and Holland (2007) suggest that first- and second-order equity may not hold when the specifications regarding the common items are changed.

Since there are many ways to equate test forms, the decision as to which method is best depends on the how data are collected, the sample sizes involved, and the nature of

the statistical assumptions that are appropriate for the situation. Harris and Crouse (1993) suggest that the definition of equating, which also can vary from one situation to another, should direct the decision of which equating method(s) to use. Classical test theory methods and item response theory methods have their own strengths and weaknesses and may sometimes be combined to produce the best equating method. It is important to be aware of all aspects of the testing situation when choosing an equating method.

A comparison of item parameter estimates using the CTT and IRT frameworks was conducted by Fan (1998). While acknowledging that theoretical differences between the two approaches were substantial, he sought to demonstrate empirically how noteworthy the differences between the two methodologies actually are in practice. Using data from a large-scale state testing program, he correlated item and ability estimates for IRT models and the CTT model and found that they were similar. To answer a separate research question, he found that the invariance of the person statistics and item statistics were quite comparable for CTT and IRT. If the invariance of the parameter estimates is truly comparable, how will equating functions differ for the two approaches?

Classical Test Theory

Classical test theory (CTT) has been used extensively in test development in a number of areas including test equating. In CTT, linear and equipercentile methods (Angoff, 1971) have been used extensively in equating. Item response theory (IRT) has become more widely used as technological developments have made the use of high-speed computing and estimation software more readily available. For the purposes of equating, both types of methods are still employed. Even when IRT is used for test

development and scoring, classical methods of equating may still be used (Kolen & Brennan, 2004, chap 6).

Linear Equating

Linear equating is one of the simpler forms of equating and is given in equation 2.1. It defines the differences between scores on two tests with a straight line. Z-scores are used to make the conversion to yield the following linear transformation of the observed scores on Form E to those on Form R.

$$l_y = y = \frac{\sigma(R)}{\sigma(E)}x + \left[\mu(R) - \frac{\sigma(R)}{\sigma(E)}\mu(E)\right] \quad (2.1)$$

By including the standard deviations of the scores of the two groups, relative levels of the difficulty of the forms can differ for examinees of different abilities. When the standard deviations of the two groups are identical, linear equating reduces to mean equating. Scores on the new Form E vary only by the difference in the means of Form R and E. Linear equating has been shown to be appropriate in some situations (Angoff, 1971; Skaggs, 2005; Zeng & Cope, 1995). In fact, when score distribution differences are "sufficiently trivial" Angoff (1971) stated that linear equating is preferable to equi-percentile equating. However, there are some limitations to linear equating. One, this method is dependent on the characteristics of the groups used in the test administration. Two, extreme ends of the score scale may be equated out-of range. Three, the true equating relationship may not be linear.

Equipercentile Equating

Equipercentile equating is a non linear equating method that preserves the order of examinee performance on two forms. It is used to equate scores on the equated from to

the reference form so that the equipercentile equivalent of scores on the equated form has the same cumulative frequency distribution as the scores on the reference form (Braun & Holland, 1982). This form of equating is a mathematical procedure that relates two continuous functions

$$e_Y(x) = G^{-1}[F(x)], \tag{2.2}$$

where G^{-1} is the inverse of the cumulative distribution function G (Braun & Holland, 1982). Figure 1 depicts equipercentile equating, where a score of 2.0 on Form R is equivalent to a score of 2.83 on Form E (Kolen & Brennan, 2004, chap2.).

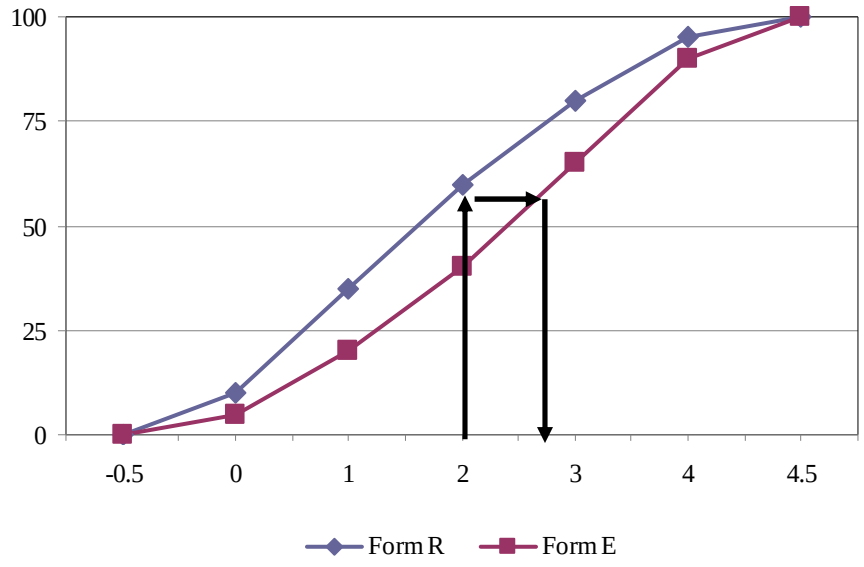

Figure 1. Equipercentile Equating of Form R to Form E

Test scores, and thus their corresponding percentiles, are not usually continuous and are reported as integers. However, it is conventional to accept percentiles as continuous for many test professionals (Kolen & Brennan, 2004, chap. 2). A percentile is the percentage of scores in the frequency distribution that are lower or equal to a value of interest. Equipercentile equating is conducted with the percentile ranks of the integer

scores, with a range from 0 to the number of items on the test. Additionally, score distributions may not resemble continuous distributions because of irregularities that naturally occur when reporting sample statistics. Smoothing is a way to adjust an empirical distribution to make it more similar to underlying population distribution.

Equipercentile equating is a more general conversion than linear because the equipercentile curve is more flexible than the line used in linear equating, and this allows for the reference form to be more difficult at the lower and higher ends of the score distribution, but easier in the middle range (Kolen & Brennan, chap. 2). Because equipercentile equating uses the distribution of the examinees, equated scores will always be within the range of actual scores. Disadvantages of equipercentile equating include greater mathematical complexity, introduction of systematic error when smoothing must be used, and potential difficulties with conversions to scale scores, especially when using this method over time. As forms become easier or more difficult, adjustments at the top and bottom of the scale could lead to values that are out of the scale range.

Identity Equating

Identity equating is used when scores on Form E are considered to be equivalent to scores on Form R without any mathematical adjustment. It may also be thought of as the case of *not* equating. It is often useful to use the identity equating function for comparisons with other forms of equating like mean equating and linear equating, as it provides a baseline reference. However, identity equating is best used when other forms of equating introduce more error than the identity equating. This situation is more likely to occur when only small sample sizes are available or when score distributions are difficult to approximate.

Item Response Theory

IRT provides a theoretical framework that may be integrated into a wide array of practical test applications: test development, item analysis, test equating, and test scoring. It is based on the notion that the probability of a correct response on any item is based on the examinee's ability, and the item's characteristics (Hambleton, Swaminathan, and Rogers, 1991, chap. 1 The IRT model may take various forms, which are discussed next, and certain assumptions must be met to appropriately use IRT equating (Cook & Eignor, 1991).

A general form of the IRT 3PL model is given in equation 2.3.

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \qquad (2.3)$$

In this model, the probability of a correct response on item *i* is given by the item's difficulty, b_i, discrimination, a_i, and a pseudo-guessing parameter, c_i. Here θ represents the underlying latent ability of interest, and D is a constant, usually set to1.7 to ease comparison between the normal and logistic ogives (Crocker and Algina, 1983,chap 15). Other IRT models, such as the 2PL model and 1PL model may be derived from this general form. In the case of the 2PL model, the possibility of guessing is not included, so the *c* parameter is set to 0. The 1PL model is further simplified by constraining all item discriminations to be equal.

Assumptions

Most commonly used IRT models assume than only one latent trait is measured by a test. When this assumption of unidimensionality does not hold, more complicated models should be used. Research suggests that the violation of this assumption may not strongly impact equating. When assessing the degree to which violations of the assumption of unidimensionality affected equating the Graduate Record Examination (GRE) verbal scale, Dorans and Kingston (1985) found that there were two highly correlated verbal dimensions. Comparisons of the equating based on the assumption of only one dimension with that of assuming two dimensions revealed an asymmetry that is not desirable in equating. However, this asymmetry was slight, and for the most part, the two equatings were similar.

In three studies conducted using the Law School Admissions Test (LSAT) data, strict adherence to unidimensionality was not found to be necessary. Camilli, Wang, and Fesq (1995) examined the LSAT data to assess the multidimensionality. They found that two main factors, inductive reasoning and deductive reasoning, emerged from factor analyses. In a comparison of six consecutive administrations of the LSAT, these factors consistently appeared. However, when unidimensional equating procedures were used to equate sets of items measuring these two dimensions, the results were not dramatically different from the results of equating sets with only homogeneous items. They make an important distinction between functional dimensionality, “which depends on the testing situation and the use of test scores”, and statistical dimensionality, “which is defined as the requirement for conditional (local) independence”. The presence of functional unidimensionality may be sufficient for equating purposes.

De Champlain (1996) showed that the differences in equating functions for three ethnic groups, for whom underlying latent trait compositions were different, were small and generally occurred at the range of very low scores. For African-American and Caucasian examinees, he found the two traits from the Camilli et al.'s (1995) study where deductive reasoning was distinct from the combined inductive reasoning and reading comprehension. For Hispanic examines, more than two underlying traits were present. More importantly, from an equating perspective, he found that minority groups of African-Americans and Hispanic examinees were not penalized when the equating function from the Caucasian group was used, even though these groups indicated different cognitive requirements than the Caucasian examinees for the deductive reasoning, inductive reasoning and reading comprehension elements of the test.

In two simulation studies based on the multidimensional structure of the LSAT data, Bolt (1999) demonstrated that IRT true-score equating was preferable to classical methods as long as the correlation between dimensions was high (≥ 0.7). His concern was that violation of unidimensionality would result in lack of equity for scores. A simulation study, unlike one using real data, is able to isolate patterns of multi-dimensionality that may not be easily discerned in practice. In the second study Bolt (1999) simulated greater difficulties between test forms and the difference between the test form difficulties differed across the levels of the two latent traits. For most of the examinees, the equatings were very similar; differences occurred for small numbers of examinees who had more unusual ability distributions (i.e., very high ability on one dimension, and very low ability on the other dimension).

Model Fit

The assessment of model fit was of paramount importance to the multi-dimensionality studies described previously (Bolt, 1999; Camilli et al., 1985; De Champlain, 1986) in order to evaluate dimensionality. A primary cause of poor model fit occurs when unidimensional assumptions are made for multidimensional structures (Loyd & Hoover, 1980). Yet, even for unidimensional tests, model fit is important because violation of unidimensionality is not the only reason for poor model fit. Model fit, for IRT equating, ensures that any linear transformation of the ability scale will also fit the data (except for sampling variation that introduces random error), as long as the item parameters have been similarly transformed (Lee & Ban, 2007; von Davier & Wilson, 2007). Model fit was demonstrated by Kolen (1981) to directly affect the equating outcomes. The equating using the 1PL model was not as good as that using the 3PL model. Kolen suggested that this was due to the fact that the 1PL model did not include the guessing parameter which was likely part of the data structure. Petersen, Cook, and Stocking (1983) also showed that poor model fit impacted the equating results for the verbal portion of the SAT test. The verbal portion of the test has less consistent equating across methods than the mathematics portion, and it demonstrated more model misfit. Generalizations from equating studies may only be confidently made when the appropriate model is applied (Hanson, 1996; Tong & Kolen, 2005).

Comparison between Classical Methods and IRT Methods

Comparisons between the two frameworks and their capacity for equating have yielded mixed results. Kolen (1981) demonstrated that the three-parameter logistic model performed more consistently than the classical methods of linear and equipercentile, and other IRT models, especially the one-parameter model. However, the equipercentile

method was better than the one-parameter model with regard to cross-validation. Cook, Dunbar, and Eignor (1981) compared the classical methods of linear, equipercentile, and frequency estimation equipercentile methods with IRT methods in terms of how well the methods agreed: in particular, how well the traditional methods agreed with the IRT method. They found that the methods all produced comparable results with some interaction between method and ability level.

In a comprehensive equating study, Gialluca, Crichton, Vale & Ree (1984) found that IRT methods performed much better (in terms of RMSEs) than classical methods of linear and equipercentile when power tests were equated and the tests or subtests used in the equatings were parallel. Power tests are those that have no time limit and may include extremely difficult items. However, speeded tests, those that do have strict time limits and are often composed of easier items, had better equating results with classical methods. Both power tests and speeded tests were simulated in different lengths: a short subtest with 15 items, and a long version with 30 items. In both cases, equating longer tests produced less error than shorter tests for power tests, but no difference in the speeded tests. Further inspection of these results lead the authors to consider item difficulty, although it had not been explicitly manipulated in their simulations. The IRT equating was found to perform slightly better when tests were more difficult.

Wolkowitz (2009) compared the classical equating methods of chained equipercentile equating and Tucker's linear equating with IRT methods including the 1-, 2-, and 3-parameter logistic models and the multiple choice model for formula scoring and number-right scoring. Formula scoring most often penalizes the examinee for making guesses by subtracting a fraction of point from the examinee's total score. In number-

right scoring, no such deductions are made: only the items answered correctly are counted towards the examinee's total score. As The College Board considered changing its SAT Reasoning Test from formula scoring to number-right scoring the most appropriate equating was needed to make fair comparisons between those students. The classical method, Tucker linear equating, was found to have the smallest absolute mean bias statistics compared to the other methods for equating number-right scored tests to formula-scored tests.

Part of the difficulty in assessing differences between classical and IRT equating methods lies in the interaction between many factors that are involved in equating. "The comparison of IRT and conventional methods is influenced substantially by many factors, such as the reliability of the tests to be equated, the properties of the common items, the ability levels of the samples, and the types of tests to be equated" (Skaggs, 1990).

When IRT is used throughout a testing program for item parameter estimation, examinee ability estimation, and other psychometric functions, it seems logical to extend its use to develop scaling constants and the equating process. IRT has been shown to be robust to violations of assumptions that must be made for implementation. Additionally, IRT is widely used by many testing organizations for high-stakes large scale tests. It is for these reasons that IRT has been chosen to be used in this study.

IRT Transformation Constants

When equating forms for two groups of examinees with differing ability, the IRT property of invariance implies that two separately calibrated sets of IRT parameters will differ only by a linear transformation given by the following equation:

$$\theta_{R_i} = A\theta_{E_i} + B, \qquad (2.3)$$

where θ_{R_i} and θ_E are the ability values for an examinee *i* on forms E and R respectively, and *A* and *B* are the constants that provide the linear transformation.

The stability of the scaling constants is an important consideration if they are to be used with confidence in equating. Baker (1996) conducted a simulation study to examine the sampling distributions of the scaling constants. He found that the sampling distributions of the scaling constants were bell-shaped, symmetric, and generated no outliers or unusual characteristics. The sampling distributions were, in fact, "well-behaved" which gives practitioners assurance in the use of these statistics.

Moment Methods

The Mean/Mean and Mean/Sigma methods are less mathematically complex methods used to determine scaling constants, yet they have been shown to be more prone to error than characteristic curve methods. Moment methods tend to be less stable (Baker & Al-Karni, 1991; Hanson & Béguin, 2002; Kim & Cohen, 1992) and less accurate (Béguin, Hanson, & Glas, 2000). Yet many equating experts recommend using several different methods, including Mean/Mean and Mean/Sigma for comparison to select the best approach (Kolen & Brennan, 2004, Ch 6.; Harris & Crouse, 1993).

The Mean/Mean transformation method was introduced in a vertical equating study using the Rasch model (Loyd & Hoover, 1980). Vertical equating is used when scores are compared on a single dimension for examinees with expected levels of differing abilities, i.e., differences between 3rd, 4th, and 5th graders. Previous research (Gustafsson, 1979; Slinde & Linn, 1978; Slinde & Linn, 1979) had found the use of the Rasch model to be unsatisfactory for equating primarily when used with differing ability groups, which is often the case in vertical equating. Loyd and Hoover (1980) found

sizeable inconsistencies when ability distributions differed dramatically, but they attributed these problems to a lack of model fit for the Rasch model, violation of the assumption of unidimensionality, and content differences across grades. They suggested that the Mean/Mean method may be more appropriate for horizontal equating, and that Mean/Mean should only be used in vertical equating situations with "extreme caution."

Mean/Sigma was developed by Marco (1977) for transforming item parameters and examinee abilities to the same scale as part of his research to find solutions to intractable testing problems of that time. He noted that before using the items for any situation (equating, pre-equating, or test information), they must be scaled appropriately. His scaling method was not the primary focus of his research, yet this method has been applied equating designs in several studies (Gao, Zhu, Chen & Harris, 2008; Hanson & Béguin, 2002; Keller, Kim, Nering & Keller, 2007; Michealides & Haertel, 2004).

Characteristic Curve Methods

Haebara (1980) introduced a method for use in horizontal equating that considered all item parameters from the IRT model simultaneously. He used an optimization process that minimized the loss function that is the sum of the squared differences for the common items that appear in subset *V* as in equation 2.4 . It then must be summed over values of theta before it can be minimized.

$$Hdiff\,(\theta_i) = \sum_{j:V} [\,p_{ij}\,(\theta_{Ji}; \hat{a}_{Jj}, \hat{b}_{Jj}, \hat{c}_{Jj}) - p_{ij}\,(\theta_{Ji}; \frac{\hat{a}_{Ij}}{A}, A\hat{b}_{Jj} + B, \hat{c}_{Ij})]^2 \qquad (2.4)$$

The Stocking-Lord (1983) method of transforming item parameters using the test information curve was used originally to overcome the shortcomings of the moment methods. Other researchers (Bejar & Wingersky, 1981; Linn, Levine, Hastings, &

Wardrop, 1981) had made modifications to the Mean/Mean and Mean/Sigma methods in an effort to improve them, but they were limited due to the heavy reliance on the *b* parameters. The Stocking-Lord method, given in equation 2.5, capitalizes on the use of more information from the entire test characteristic curve. *SLdiff* is then summed over theta values and then minimized. The Haebara method differs from the Stocking-Lord method because it uses item-level information and the Stocking-Lord approach uses test-level information (von Davier & von Davier, 2007).

$$SLdiff(\theta_i) = [\sum_{j:V} p_{ij}(\theta_{Ji}; \hat{a}_{Jj}, \hat{b}_{Jj}, \hat{c}_{Jj}) - \sum_{j:V} p_{ij}(\theta_{Ji}; \frac{\hat{a}_{Ij}}{A}, A\hat{b}_{Jj} + B, \hat{c}_{Ij})]^2 \quad (2.5)$$

Comparison of Moment Methods with Characteristic Curve Methods

Baker and Al-Karni (1991) compared the Mean/Mean (Loyd & Hoover, 1980) method with the Stocking-Lord (1983) test characteristic curve method, and concluded that the Stocking-Lord method was preferred. In a simulation study that included a horizontal equating, a vertical equating, and an IRT parameter recovery study, they found that the Stocking-Lord procedure yielded slightly better results, but the Mean/Mean method did produce "acceptable equating coefficients." They recommended the use of the Stocking-Lord method for unusual combinations of examinee ability, item difficulty, and discrimination, but acknowledged that the Mean/Mean method was easier to compute.

Hanson and Béguin (2002) also showed that characteristic curve methods were preferable when they demonstrated that the Mean/Mean and Mean/Sigma methods of generating transformation constants produced larger mean squared error than either the Haebara or Stocking-Lord characteristic curve methods. In a simulation study they varied

estimation program (MULTILOG and BILOG-MG), sample size (3,000 and 1,000 examinees), number of common items (10 and 20 for the 60 item test forms), and sampling design (equivalent and nonequivalent groups). They used four methods of item parameter scaling including characteristic curve methods of Stocking-Lord and Haebara and the two moment methods of Mean/Mean and Mean/Sigma. They also found that concurrent calibration was preferable to separate estimation, in that it produced smaller standard errors, although they hypothesized that the concurrent estimation may have been better because the common item parameter estimates were based on larger samples.

Kim and Lee (2006) studied the differences in the IRT linking methods of Mean/Mean, Mean/Sigma, Haebara, and Stocking-Lord for the 3-PL model and polytomous IRT models. They, too, found that generally the characteristic curve methods were more accurate than the moment methods for mixed format tests (tests that include both dichotomously and polytomously scored items). In their study, the Haebara method had the lowest MSE across all four methods.

RPA Method

The RPA method was introduced to incorporate observed examinee data into the transformation process. Both moment and characteristic curve methods use item characteristics from the common items on separate forms to find the optimal A and B constants that minimize some distance measure between the two distributions. Ragland et al. (2009) examined transformation constants obtained by minimizing the difference between the observed and predicted score distributions. Five different distance measures (likelihood ratio G-statistic, chi-square statistic, Kolmogorov-Smirnoff distance, absolute value difference, and squared difference distance) were compared. These transformation

constants were additionally compared with the transformation constants from the Stocking-Lord procedure. Differences were comparable for all measures for both simulated and student data. They suggest that the advantage in using observed score data may be to clarify the equating process in applied practice and to improve IRT observed score equating accuracy.

IRT True-Score vs. IRT Observed-Score Equating

Under IRT, estimates are made for both item characteristics and examinee abilities. The examinee ability scale, θ, is arbitrarily set to have a mean of 0 and a standard deviation of 1. The moment methods and curve methods described above are used to place the θs from the different forms onto the same scale. Scores are not usually reported on this scale, however, for several reasons according to Kolen and Brennan (2004): examinees with the same number-correct score can have different θ estimates, difficulty and cost in obtaining ability estimates, and differential amounts of measurement error. First, examinee abilities in IRT estimation are based on the entire response string, not just the number of items answered correctly. From a psychometric perspective, this improves the precision of the estimate. However, test takers and other test score consumers (policy makers, teachers) may not understand why different IRT scores are given for the same number of correct responses. For example, an examinee who answers 40 of the more discriminating items correctly could have a higher estimated ability than another examinee who also answered 40 items correctly that were less discriminating. Second, IRT estimates must be obtained with specialized software. Obviously this is not a problem for large scale testing organizations, but this can be a limiting factor in the use of IRT. Third, the estimates for ability tend to be much more

accurate for examinees in the middle range of ability, and much more prone to measurement error for examinees at the extreme ends of the spectrum. Consequently, even when tests are developed and equated with IRT methods, number-correct scoring may be employed for score reporting. Two methods have been developed to put the equated results on the number-correct metric.

IRT True-Score Equating

The number-correct true score, also called the "expected number of correct answers" (Lord & Wingersky, 1984) is defined to be the sum of the probabilities of correct response for all items on the reference form (Form R), conditional on θ. (It should not be confused with the number of correct responses, which is often called the raw score or number-correct observed score.) Note that the lower bound for true scores under the 3-PL model is not 0, but instead, it is the sum of the lower asymptotes; the upper bound is the number of items on the test. For example, a 40 item test with 5 response options, assuming *c* to be .20 across all items, would have number-correct true scores that ranged from 8 to 40. With this method, assuming the items on Form R and Form E have been calibrated on the same scale, θ can be estimated based on the examinee's responses on Form E. Then the examinee's true score can be calculated on Form R. In operational practice, sometimes a shortcut is used. Select a true score of interest, usually an integer. Then find the ability level, θ_i that corresponds to that true score on Form E, and then find the true score on Form R that corresponds to that θ_i (Kolen & Brennan, 2004).

Lord and Wingersky (1984) describe true score equating mathematically by defining the true score on Form R to be

$$\xi \equiv \sum_{i=1}^{n} P_i(\theta), \qquad (2.6)$$

a monotonic increasing function of θ. Similarly, the true score on Form E is defined to be

$$\eta \equiv \sum_{i=1}^{m} P_j(\theta). \quad (2.7)$$

A table of corresponding true scores may be generated by using item parameters in the above two equations for all scores of interest, usually integer scores that will be reported. The process of determining the corresponding true scores must be estimated iteratively with computer methods (Han, Kolen & Pohlman, 1997; Kolen and Brennan, 2004). In the shortcut described above, this equating is simply applied to the observed scores.

The simplicity and elegance of true-score equating is limited by one serious drawback: true scores can never be explicitly known; they can only be estimated based on the estimated θ. To avoid estimating true scores, and to avoid the issue that examinees with the same observed score may have different true scores, the true score equating relationship is used in practice with number-correct observed scores. The procedure is called true-score equating even when it is applied to observed scores. Lord and Wingersky (1984) demonstrated that the empirical results were very similar between IRT true-score and observed-score equating, which is reassuring, as there is "no clear theoretical justification" for using true score equivalency tables with observed scores.

IRT Observed-Score Equating

The IRT observed-score equating method uses IRT estimation to produce a distribution of number-correct true scores on two forms to be equated, and then uses equipercentile equating to enable score comparability. In fact, Lord and Wingersky (1984) refer to it as the "IRT equipercentile observed-score method." This distinction as IRT observed-score is crucial because classical methods of equating are often called

observed-score methods. Braun and Holland (1982) use the term "observed-score equating" to refer to most of the then-employed methods of equating at ETS. They distinguish it from Levine's (1955) true-score equating and all IRT equating. When von Davier and Wilson (2007) used the terms true-score and observed-score, they were referring to IRT true-score equating and traditional non-IRT equating methods (Tucker and chain equipercentile). However, Han, Kolen and Pohlman (1997) used true-score and observed score when they compared equating with IRT true-score and observed-scores.

When Lord (1980, chap. 13) introduced what is now most commonly known as IRT observed-score equating, he called it, "raw-score 'equating' with an anchor test" because he considered this an approximation of equating. Observed-score equating uses the distributions of observed scores that have been generated by the IRT model and the θ distribution. One way to describe this process is to think through a simplified example which is provided in Table 1. Consider the case of a three-item test for an able examinee that has the probability of correctly answering item 1, 2, and 3 of .80, .65, and .50, respectively. The score distribution for this person is formed by calculating the probability of all possible scores that could be earned on Form R.

Table 1. Probabilities of Response Patterns

	Response pattern			Probability of response pattern			
	Item 1	Item 2	Item 3				
Test Score =0	0	0	0	0.20	0.35	0.45	0.0315
Test Score =1	1	0	0	0.80	0.35	0.45	0.1260
	0	1	0	0.20	0.65	0.45	0.0585
	0	0	1	0.20	0.35	0.55	0.0385
Test Score =2	1	1	0	0.80	0.65	0.45	0.2340
	0	1	1	0.20	0.65	0.55	0.0715
	1	0	1	0.80	0.35	0.55	0.1540
Test Score =3	1	1	1	0.80	0.65	0.55	0.2860

The probability of scoring 0 on the test could only occur if all items were answered incorrectly: P(0)=(.20) (.35) (.45) = .0315. The probability of scoring 3 on the test could only occur if all items were answered correctly: P(3)=(.80) (.65) (.55) = .286. There are three ways that this examinee could score 1 on the test by answering either the first or second or third item correct P(1):=(.80) (.35) (.45) = .1260 + (.20) (.65) (.45) = .0585 + (.20) (.35) (.55) = .0385 =.223. Finally, there are three ways that this examinee could score 2: P(2)=(.80) (.65) (.45) = .2340 + (.20) (.65) (.55) = .0715 + (.80) (.35) (.55) = .2860 =.4595. A separate value must be estimated for each examinee based on the examinee's ability, and then averaged across examinees to estimate the score distribution in the population. Although this example uses brute force to illustrate the approach, the algorithm produced by a computer takes advantage of previous calculations in a recursive manner to simplify the calculations. In practice, after the item parameters are estimated the distribution of number-correct observed scores is calculated for each ability level with the following function (Han, Kolen, Pohlman, 1997; Kolen & Brennan, 2004; Lord & Wingersky, 1984):

$$
\begin{aligned}
f_r(x|\theta_i) &= f_r(x|\theta_i)(1-p_{ir}), && x = 0 \\
&= f_{r-1}(x|\theta_i)(1-p_{ir}) + f_{r-1}(x-1|\theta_i)(p_{ir}), && 0 < x < r \\
&= f_{r-1}(x-1|\theta_i)(p_{ir}), && x = r
\end{aligned}
\qquad (2.8)
$$

where $f_{r-1}(x|\theta_i)$ is the distribution function of the first r items for the i^{th} level of ability. As each item is added to the response string, the predicted frequency for each x is updated. When r = test length, $f_r(x|\theta_i)$ = the predicted frequency of observed score x on the total test for θ_i. At each x, the frequencies are integrated over the θ distribution to yield the expected observed score distribution.

In IRT observed score equating, the item parameters and θs are transformed to the same metric using one of the IRT equating methods, such as the Stocking-Lord method. Then, using these equated item parameters, the cumulative distribution of the scores is produced for the population of examines who took Form R. Similarly, the predicted score distributions of Form E for the same population of examines who took Form E is estimated. Equipercentile equating is then used.

Comparisons between IRT True-Score and IRT Observed-Score Equating

The Lord and Wingersky (1984) study was done with actual student data, LOGIST software, and for the 90-item test, 40 items were common to the two forms being equated. They were able to randomly select approximately 2,670 examinees from the test administration of the unidimensional SAT verbal test, thus using equivalent groups. They concurrently calibrated the item parameters from the common items to put the parameters on the same metric.

The conditions of the Lord and Wingersky (1984) study: actual data, LOGIST software, test length, number of common items, sample size, population group ability distributions, and calibration methods are important to note because much of the equating literature focuses on varying the factors that influence the equating results. For this reason as well, it can be difficult to directly compare studies (Gialluca et al., 1984).

A comparison of IRT true-score equating and IRT-observed score equating with the classical equipercentile method in the Han, Kolen and Pohlman (1997) study indicated that neither method consistently produced smaller equating errors, although IRT-true score equatings were more stable than the other equating methods and IRT

observed-score equating was more stable than the equipercentile. Whether stability is a sufficient criteria for equating accuracy depends on how the equating results are used.

When using bootstrap estimation methods to compare the standard errors of equating, Tsai, Hanson, Kolen, and Forsyth (2001) found that all five of the methods they compared produced acceptable standard errors of equating (< .01 standard deviation units). They evaluated IRT true-score and IRT observed-score equatings based on item parameters from separate calibrations with an IRT true-score and IRT observed-score equatings that were concurrently calibrated and an IRT chained true-score equating method for common-item nonequivalent groups. They also compared large samples ($n \approx 1500$) with small samples ($n \approx 500$). They found that the IRT observed-score equatings produced smaller standard errors than the IRT true-score method, even for the small samples.

Hendrickson and Kolen (2003) compared IRT true-score equating with IRT observed-score equating and traditional classical equipercentile equating for the Medical College Admissions Test (MCAT). Because they were studying the practical implications of changing equating methods for the administration of an actual examination, true parameters could not be determined. They chose to evaluate the differences between the three methods' resulting equating functions. In particular, they examined how the reported scores would vary using the different methods, which could impact decisions of student selection to medical school.

The equating methods were used for the three sections of the MCAT: Biological Sciences, Physical Sciences and Verbal Reasoning. The authors concluded that changing the equating methodology from the equipercentile method to an IRT method would

impact the scores. The impact would be less substantial for the Biological Sciences than Physical Sciences and Verbal Reasoning regardless of IRT model chosen (one-, two- or three-parameter). Using IRT methods for Physical Sciences and Verbal Reasoning produced different results: one-parameter and two-parameter models produced much lower equivalents than the equipercentile methods (11 score points for the one-parameter and 9 points for the two-parameter), but the scores from the three-parameter models were lowered by only 5 points on average. Although standard errors of equating are oftentimes greater at extreme ends of the score distribution, the differences across these test sections affected score points across the entire range of scores.

This study, too, emphasized the great frustration in comparing equating methods in that the best method, or true equating function, is not known. The decision to change from equipercentile equating to an IRT-based equating methodology cannot be based on knowledge of which approach most accurately represents the true equating of the population. In examining the differences between methods, the authors could only observe which IRT method was closest to the equipercentile.

Hendrickson and Kolen (2003) also pointed out that observed score and true score methods were consistent across all three test sections. Small changes that occurred were due to choice of IRT model, not equating method. There was only one score point equivalent difference between the methods at points 12, 13, 21, and 32 for the 55-item Verbal Reasoning test. Findings were similar for the Biological Sciences and Physical Sciences tests.

Tong and Kolen (2005) compared IRT true-score equating and IRT-observed score equating with a classical equipercentile method. This study was similar to the Han,

Kolen and Pohlman (1997) study, but different equating criteria were used: differences between ability distributions, first-order equity and second-order equity. First-order equity is present when examinees have the same expected score on alternative forms after controlling for examinee true score (Lord, 1980, chap. 13). Second-order equity is present when examinees have the same conditional standard error of measurement on alternative forms after controlling for examinee true score (Morris, 1982). Their examination of these methods indicated that the extent to which the difficulty parameters differed affected the degree to which the three equating properties held. In both actual data and simulation, they found that when the raw score distributions were almost the same, all three methods lead to acceptable equating functions for all of their criteria, but when the raw score distributions differed, the IRT true score method performs best with regard to the first order property. The equipercentile method and IRT observed score method performed equally well when they used the same distributions and second order equity criteria. As difficulty differences increased between the different forms, the first and second order equity properties did not hold.

Test and Item Characteristics

Test Length

Different test lengths are often employed for a variety of reasons in operational testing depending on the purpose of the test, the age of the students, or curricular or content needs. Equating studies that incorporate actual student data often must use the test lengths that are already established. Lord and Wingersky (1984) used a 90-item test in their research; Fitzpatrick (2008) used tests with between 60 and 75 items, and Gao et

al. used 30-items tests. Although the impact of test length does not appear to have been systematically studied, the ranges of 30 -60 items seem realistic.

Ability Differences

Differences in examinee ability have been examined in equating research to determine their impact on the equating results. Although much of the earlier research on differences between ability distributions primarily focused on vertical equating situations (Loyd & Hoover, 1980; Marco, Peterson, & Stewart, 1983; Skaggs & Lissitz, 1988; Slinde & Linn, 1977, 1978), subsequent research considered IRT equating methods for horizontal equating.

Distinctions are often made between vertical equating and horizontal equating (Felan, 2002; Kolen, 1988; Skaggs & Lissitz, 1982). Equating alternative forms of a test designed to possess the same psychometric properties is often identified as horizontal equating. It is used when test forms display only slight differences in difficulty.

Horizontal equating may also be viewed as the situation where ability distributions are similar for examinees. The question of similarity, of course, is a matter of degree. Examinees who participate in different test administrations may be systematically different in a number of ways, including ability level. Students who take an examination in the fall, for example, may be of lower ability than those who test in the spring, due to additional learning time and experience of students who take the examination (Schmitt, Cook, Dorans & Eignor, 1990). Equating studies that simulate examinee ability levels often use normally distributed abilities with means of -0.5, 0, and +0.5 to represent low, medium, and high ability levels (Gustafsson, 1979; Holmes, 1982; Skaggs & Lissitz, 1988; Slinde & Linn, 1978).

Vertical equating is used to describe situations where scores will be compared on a single dimension for examinees with purposefully selected differing abilities (Baker, 1984; Skaggs & Lissitz, 1986). For example, the differences that exist between forms due to content differences on the same basic trait of mathematics between 3rd, 4th, and 5th grade math would need to be vertically equated to be compared. As another example, consider when comparisons are made for scores of fall students in a lower grade with the scores of fall students in the next grade, even larger differences are to be expected. As the differences between ability distributions of the groups increase, one moves from horizontal equating to vertical equating. The purpose of this research is to examine ability differences between groups that may be larger than what has been previously examined in other research, yet still close enough in ability distributions to be considered horizontal equating.

An important caveat about the term “vertical equating” must be addressed. Many equating experts argue that the requirements for equating are not met when differences in test difficulty are so great that an entirely different latent trait is actually being measured (Holland, 2007; Kolen & Brennan, 2004, chap. 9; Patz & Yao, 2007). Such is often the case where elementary and secondary achievement scores are to be compared across multiple grades, which is when vertical scaling is usually employed. The term “vertical scaling” is used instead to indicate a weaker relationship, or linkage, than that of equating. This type of linking may be based on common items in a test, like that used in horizontal equating, but the links between examinees weakens as tests become increasingly different in difficulty (Patz & Yao, 2007). Generalizations from these linkings must be made very cautiously. There are others (Baker & Al-Karni, 1991;

Camilli, Yamamoto & Wang, 1993; Cook & Douglass, 1982; Harris & Hoover, 1987; Loyd & Hoover, 1983; Muraki, Hombo & Lee, 2000; Skaggs & Lissitz, 1982; von Davier & von Davier, 2007), however, who perform vertical equatings and simply acknowledge that this type of equating is different from horizontal equating.

Gialluca et al. (1984) examined different ability levels of the groups of examinees and simulees. They simulated ability distributions from comparable samples of military examinees to form two distinct ability groups. They found that equating similar ability groups resulted in smaller RMSEs than equating groups with different ability levels.

Harris and Kolen (1986) compared linear, equipercentile, and IRT 3PL true score equating for a 40-item test for examinees of two different ability groups. Ability level was determined from self-reported high school grade point average. The authors found similar equating functions between the higher and lower ability groups for several pairings of forms, and across all equating methods. They concluded that population differences between groups alone should not determine the type of equating procedure.

In a study of equating sensitivity to different examinee sampling approaches, Schmitt, Cook, Dorans, and Eignor (1990) compared classical equating methods (Tucker, Levine, equipercentile, chained equipercentile curvilinear) with the IRT 3PL method for sampling strategies involving representative sampling, matched sampling and target sampling. They hypothesized that purposeful sampling of examinees could improve the quality of equating. Classical methods were less affected by differences in ability distributions of the samples. However, the Biology test they used in the study was "somewhat multidimensional" which would have affected the IRT parameter estimates and resultant equatings.

Gao, Zhu, Chen, and Harris (2008) studied the effects of different numbers of common items on equating functions for different ability groups. They found that dissimilar distributions of ability had the greatest impact using the Mean/Sigma transformation approach prior to equating than either the Haebara or Stocking-Lord method. Similar equating functions emerged with Mean/Mean, Haebara, and Stocking-Lord procedures for both IRT true and observed score methods across different sample sizes using common item sets of 5, 10, and 15 items for the 30-item test. They also suggested that IRT scale transformation and equating methods might be susceptible to interactions between examinee ability distributions and the properties of the common items, although this hypothesis was not tested.

Common Items

Common items are required to estimate the scaling coefficients needed to equate different groups who take different forms of the same test. Yet the use of these items introduces another level of complexity to the equating process. Research in the role of common items in equating has generally adhered to some basic guidelines with regard to the number of common items as well as the distribution of their parametric properties and their correlation with other test items. The ability distributions of the populations whose scores are to be equated have been shown to interact with common item characteristics. Whether the items are external or internal to the test may also impact the equating results.

An ETS task force was charged with the task of developing common guidelines for the large number of testing programs that fell under the purview of that agency. Dorans, Kubiak, and Melican (1998) prepared the report that was for that purpose based on Huddleson's (1957) earlier work in this area. They noted that many of the guidelines

remained intact. The primary responsibility for the task force was to clarify the earlier requirements that had left much "room for interpretation." They addressed six areas in particular for the characteristics of internal common items: number of items, location of items on the test, amount of changes that can be made to the item, content specifications, statistical requirements, and the conditions of test administration. These authors made no distinction between classical equating methods and IRT-based methods, although Cook and Petersen (1987) had discussed problems related to equating methods when circumstances were less than optimal for classical and IRT equating. Cook and Petersen (1987) compared and contrasted different requirements for the two approaches and found that guidelines for one type of equating did not necessarily apply to the other.

Dorans, Kubiak, and Melican (1998) reiterated Huddleston's (1957) guideline of 20 items or 25% of the number of items on the test, whichever is larger, for common items, as long as this number is sufficiently large to guarantee content representation and statistical specifications. Angoff (1971) also adhered to this guideline. They noted, however, that this rule of thumb could be relaxed with tests that are highly reliable (>.90). Fitzpatrick (2008) points out that a number of other researchers adhered to this principal as well (McKinley & Reckase, 1981; Peterson, Marco, & Stewart, 1982; Vale, Maurelli, Gialluca, Weiss, & Ree, 1981; Wingersky, Cook, & Eignor, 1987).

Budescu (1985) found that the number of common items necessary in linear equating to achieve the desired equating efficiency was a function of test reliability as well as a high correlation between the set of common items, U, and both the reference test, R, and the test to be equated, E. When comparing classical and IRT equating methods for the impact of the number of common items, Yang and Houang (1996) found

that, in general, using more items resulted in more accurate equating. They concluded that the larger of 20 items or 20% of the number of items on the test was sufficient for either classical or IRT equating. Fitzpatrick (2008) experimented with using smaller number of common items (5, 10, and 15 for a 60-item test) and found a notable lack of stability when fewer than 15 common items were used. She concluded that using fewer common items for the high-stakes achievement tests she had analyzed "might not be a good idea." Gao, Zhu, Chen and Harris (2008) found that using common item sets of 5, 10, and 15 items for the 30-item test provided adequate equating functions for their 30-item test, but they, too, noted that more items produced better equating results.

The location of common items with respect to the reference form and equated form is less important for unspeeded tests than speeded ones (Dorans et al., 1998). Of course, items that correspond to a common prompt must remain together and content dependencies must be recognized and avoided (Kolen & Brennan, 2004, chap. 8). Dorans et al. (1998) stated that common practice was to avoid using any item as a common item that has been reached by less than 90% of the examinees (for speeded tests). They additionally noted that common practice was to avoid using the first item on any test as a common item to allow examinees "start-up time." Harris (1991b) investigated the effects of different item orders, or scrambling, of all test items on equating results for a classical equating method (equipercentile equating with cubic spline smoothing) and with an IRT method (true score equating). While there was little difference between the two, they both produced different results for different scrambling configurations. Results from a study conducted by Meyers, Miller, and Way (2009) support the other findings that changes in item position significantly change equating, They emphasized that this occurred because

the item positions impacted item difficulty. These authors further postulated that test construction practices could mitigate equating inequalities. Leary and Dorans (1985) summarized the literature for item placement and noted the shift in focus as this research progressed from simple main effects of item order on test performance to interactions between item order and examinee characteristics (both psychological and biological) to a more recent focus on the stability of item parameters.

Dorans et al. (1998) found it difficult to enumerate the nature of the changes that can be made to an item when using it as a common item because there are so many types of variation. They cautioned against making any changes that would be likely to affect examinee performance. While this requirement has intuitive appeal, no research was found that explored this topic further.

On the other hand, content specification has been extensively reviewed in the equating literature. Klein and Jarjoura (1985) concluded that it was quite important to use content-representative items for common items. When describing guidelines for common items, Kolen and Brennan (2004, chap. 8) specifically refer to the content of the common items, which should be proportional to the content of the total test. Common items need to be content representative to minimize the difficulty difference between them and the total test which will result in less equating error (Gao, Hanson, & Harris, 1999).

The statistical properties of the common items should parallel those of the total test (Dorans et al., 1998; Petersen et al., 1989; von Davier, Holland, & Thayer, 2004). Kolen and Brennan (2004, chap. 8) also stated that the means and standard deviations of the difficulty of common items and those of the total test should be similar.

Unfortunately for many testing situations, like those used for credentialing exams, it can be difficult to accumulate sufficient appropriate items to follow these guidelines closely (Bené, 2008). Sinharay and Holland (2006) questioned whether the strict adherence to previously established guidelines was truly necessary. They compared the condition of statistical representativeness of tests with external common items (minitests) and found that the spread of item difficulties could be less extensive than the fully representative set (miditests) to function sufficiently for equating. They cautioned that their results would be most suitable for tests with external common items compared to those with internal ones for two reasons: first, their research was done only with external common items and second, test construction problems might arise in balancing internal common items (more closely distributed around average difficulty) with non-common items to meet test level difficulty targets overall.

Even for multidimensional IRT equating, the strict rules for selection of common items need not all be required. Duong and Reckase (2008) studied the effects of number of common items, dimension coverage, and difficulty coverage. The results of their simulation study showed that the number of common items and the dimension coverage were more influential than the range of difficulty of the common items. They, too, found that errors increased when examinee ability distributions were less similar.

Another study which focused on multidimensional IRT equating also found several interactions between key characteristics of common items. Lu (2008) compared the effects of different methods for selecting common items. This simulation study used multidimensional data. The findings indicated that none of the methods emerged as

superior, primarily due to large variations in the estimation of item parameters prior to equating.

Sample Sizes

The importance of equating is becoming more evident with the administration of tests to fewer test takers. NCLB legislation has required testing for all K-12 students, and accommodations must be made for students with special needs. Many certification tests are given to smaller numbers of examinees. These groups are often much smaller than desired for equating, yet the scores for these test takers, too, may need to be equated to fairly make comparisons across cohorts or test forms.

Small sample sizes can be problematic for equating of all types, but pose additional problems when IRT parameter calibration is employed as a precursor to equating. The impact of small sample sizes on linear equating in the CTT framework was investigated by Parshall, Houghton, and Kromrey (1995) when they compared random samples of size 15, 25, 50, and 100 for randomly equivalent groups, using n=500 as a basis for comparison. They found that the standard errors of equating increased monotonically as sample sizes became smaller, and for score points farther from the average raw score. The increases were not linear, and were more noteworthy for the lower end of the score range. They noted four important characteristics from their study that may have contributed to the regular patterns of the low standard errors and small bias: at least 50% of the items were common to the randomly generated forms for the five separate content areas they studied (one area had 69% common items), the population samples were nearly identical in ability level, only a single linking was conducted, and

test takers were selected from a single administration of a test. When these four conditions are not present in actual data, results may be quite different.

Sample sizes of the magnitude used in the Parshall et al. study are not possible with IRT equating because larger numbers of test takers are needed for the estimation of item and ability parameters in IRT. When Mislevy, Sheehan, and Wingersky (1993) explored way to equate tests with small sample sizes, they experimented with samples of 100. However, they used collateral information (expert judgment and statistical specifications) that had been based on administration of the same forms to 5000 examinees. While this may be helpful in equating small samples to larger groups or pre-existing data, it does not answer the question of what sample sizes may be considered minimum for an initial equating.

Kolen and Whitney (1981) found that sample sizes of 200 were not sufficient to produce adequate equatings for the IRT 3PL model. Although sufficiently large samples ($n \approx 1200$) were available for the five common-item forms used in the five general areas in the administration of the GED test, only around 170 to 198 test takers had taken the forms to be equated. The LOGIST software they used in their study did not converge for the equating forms, (i.e. was unable to produce viable parameter estimates) so the guessing parameter was fixed. The resulting parameter estimates were much more extreme than the comparable 1PL estimates and impacted the equating results dramatically with unacceptable levels of bias and standard errors too high to be used.

In a simulation study, Cui and Kolen (2007) compared the differences in computing standard errors of equating using parametric and nonparametric bootstrap estimation. While their focus was on the best computational method to use in different

situations, their research design varied sample sizes as well to include 300, 1000, and 3000 simulated test takers, suggesting that 300 may be considered a minimum sample size. However, they used BILOG to estimate the item parameters, whereas Kolen and Whitney (1981) used LOGIST to recover item parameters.

In general, it is appropriate to assume that larger sample sizes produce better results: improved accuracy of parameter estimation with less error. The minimum sample size requirements, however, must be met in each situation, and the studies just described have demonstrated its importance. Therefore, sample size will not be a factor directly manipulated in this study. It will be necessary to ensure sample size is sufficient. For this research, the student data sample well-exceeded the minimum of 300 (n=1,898), but it is important to note the problems that could incur in the event that it does not. The simulation study generated 2,000 simulees for each ability distribution.

Error in Equating

Because equating is a statistical procedure, error is introduced. “Error” in statistics generally refers to the differences that result from having taken a sample of some outcome of interest. It is sometimes referred to as “noise,” as it interferes with the trait of interest. Groups are sampled from a population of examinees, items are sampled from many possible items that could be administered, and examinee responses are samples on a particular date and time set aside for testing. Errors may be of two types: random and systematic.

Random errors are arbitrary and can fluctuate from item to item and from one examinee to another. They occur as a result of the necessity of sampling. Random error may be introduced from sampling groups from a population of examinees.

In equating, systematic errors may arise as the result of four situations (Kolen & Brennan, chap. 7). First, there may be occasions when systematic error is deliberately introduced to adjust for wild fluctuations in observed data. This process, smoothing, is designed to reduce random error by a greater amount than the systematic error, and is especially beneficial when only smaller samples are available. Second, systematic error can be present if assumptions of the equating method(s) are violated. For IRT equating, the assumption of unidimensionality is often made. Violations of this assumption introduce error across the board for examinees that may possess much more or much less of a secondary trait. Third, systematic error can result from improperly applied data collection protocols. If spiraling test forms to ensure random distribution is not followed closely, assignment of forms may not truly be random. Finally, systematic error may be a factor of systematic differences between the groups who take the different forms.

Error in equating does not occur uniformly across score or ability distributions, and is usually more severe at the extreme ends of these distributions because of the scarcity of data at these points (Jaeger, 1981). In order to accurately define equating error, one must know the true equating function, which is rarely known in practice.

The standard error of equating is used to describe the amount of random error present in an equating relationship. The standard error of equating "may be conceived of as the standard deviation of equated scores over hypothetical replications of an equating procedure in samples from a population or populations of examinees" (Kolen & Brennan, chap. 7, p. 232).

For some equating functions, theoretical estimators for asymptotic standard errors have been derived to assess the magnitude of equating. Lord (1982) derived the first

formula for the asymptotic standard error of a true score IRT equating with external common items using the delta method which is based on a Taylor expansion. Liou and Cheng (1995) presented a simplified version of asymptotic standard error which is easier to implement for more complicate equating situations (chained equipercentile equating, smoothed equipercentile equating and equating using the frequency estimation method). Ogasawara (2000, 2001a, 2001b, 2001c) derived theoretic expressions for several other IRT equating methods. He derived asymptotic standard errors of equating coefficient estimates for the moments methods of mean/sigma and mean/mean methods (Ogasawara, 2000). He also derived asymptotic standard errors of the estimates of equating coefficients using the characteristic curve methods of Haebara and Stocking and Lord (Ogasawara, 2001c), as well as asymptotic standard errors for IRT true score equatings (Ogasawara, 2001a).

When asymptotic standard errors are mathematically intractable, or when the underlying assumptions do not hold, bootstrap methods have been shown to accurately represent standard errors across the spectrum of scores (Cui & Kolen, 2008; Michaelides & Haertel, 2004; Tsai, et al., 2001). The bootstrap is a computer intensive sampling procedure from a given dataset to repeatedly estimate a parameter of interest (Efron & Tibshirani, 1985). The standard deviation of the bootstrap estimates is an estimate of the standard error.

Most formulas for the standard error of equating only take into account the sampling variation of examinees. Michealides and Haertel (2004) demonstrated that additional error variance is introduced when the common items used for equating have been sampled from a larger population. Comparing their analytic formula with a

bootstrap procedure, they found that the equating error for these methods was quite similar when the difficulty parameters were assumed to be bivariately normally distributed. This finding confirmed the accuracy of their asymptotic formula.

Criteria for Accuracy

There are many ways to evaluate whether an equating function has produced a desirable result. Jaeger (1981) introduced indices for determining when to use linear or equipercentile equating methods under the CTT framework. As noted earlier, Harris and Crouse (1993) suggest that the definition of equating, which also can vary from one situation to another, should direct the decision of which equating method(s) to use. Their survey of the literature included mostly classical equating studies, although some IRT-based equatings were included. They list nine descriptions of equating criteria that were extensively used, and point out the strengths and weakness of each. Standard errors, one of the nine categories, are an analytical approach used to estimate the amount of random error. Harris and Crouse (1993) point out that the advantage to using standard errors is that they are easy to apply and interpret, but a disadvantage is that they ignore systematic error. Another category of equating criteria is that of overall summary indices, which often include the root mean square given in equation 2.9

$$RMS = \left(\frac{\sum_i f_i (A_i - B_i)^2}{\sum_i f_i} \right)^{1/2} \qquad (2.9)$$

where A_i is the equivalent of a raw score of i on the equated test and B_i is the true value, f_i is the frequency of a raw score of the equated test, and i indexes the score range. Like standard errors, RMS indices are easy to interpret.

Because equating is “situation specific,” Harris and Crouse suggest that further research is needed to determine which criteria are most appropriate under explicit conditions.

Visual inspection of differences between statistics of interest has been used quite often in equating studies. Lord and Wingersky (1994) compared IRT true-score equating with IRT-observed-score equating and used visual inspection of observed scores with the estimated observed scores they derived from the two different approaches. They also commented that accuracy for different methods is difficult to assess, and did not consider stability of equating results to be a sufficient measure of accuracy. When Baker (1996) described the characteristics of the sampling distributions of equating coefficients, he used visual inspection of the plots of those distributions as a criterion for accuracy. He also observed the differences between summary statistics to determine their properties. Jodoin, Keller, and Swaminathan (2003) explored how IRT equating in high stakes testing affected classification outcomes for students using linear, fixed common item parameters and concurrent IRT calibration methods. Classification consistency was easily assessed with κ, but visual inspection of the b-parameter plots was used for assessing equating.

Since Harris and Crouse’s (1993) summary of criteria for evaluating the accuracy of equating, no clear consensus has arisen for what criterion to use in any given situation. Kolen and Brennan (2004, chap. 8) encourage the consideration of “practical circumstances” when determining criteria to be used in the evaluation of any equating. Kolen (1981) had earlier noted that “no demonstrably superior criterion” exists for

comparison of equating results, and that condition continues to pose difficulties for equating practice.

Research Questions

Two studies will be conducted to explore some of the factors that influence the new transformation method: test length, number of common items, and differences in the ability distributions of the populations to be scaled. Both a simulation study and actual student data sets will be used.

Study 1: Simulation Study

Research Question 1: How will test length affect the accuracy of the transformation constants of the new scaling method compared with the Stocking-Lord method? Tests need to be of sufficient length to properly cover content and to ensure psychometric soundness, yet no longer than necessary to prevent examinee fatigue, as well as to minimize the high costs of item development. It is important to distinguish minimum levels of acceptable numbers of items. Fairly long tests of 150 items with 38 common items have been used (Tsai, et. al, 2001), medium tests of 90 items with 40 common items (Lord & Wingersky, 1984) to shorter tests of 30 to 15 items (Gialluca, et al., 1984). It is likely that differences in test length will not be large in this study.

Research Question 2: How will the number of common items affect the accuracy of the transformation constants of the new scaling method compared with the Stocking-Lord method? Related to the question of test length impact on equating, the number and proportion of common items is also an important factor in the accuracy of equating. Earlier rules of thumb regarding the nature and number of common items have been relaxed in recent studies, and equating results have been mixed. Fitzpatrick (2008) noted a lack of stability when fewer than 15 common items were used, but Gao, Zhu, Chen and

Harris (2008) found that using common item sets of 5, 10, and 15 items for their 30-item test provided adequate equating functions, although using more items produced better equating results.

Research Question 3: How will differences in the ability distributions of the populations to be scaled influence the accuracy of the resulting transformation constants? Ability distributions have been shown to have a tremendous impact on equating results Gao, Zhu, Chen, & Harris, 2008). The simulation study will compare three different ability distributions. Scaling results will most likely be more favorable when examinee ability distributions are more similar.

Study 2: Student Data

Research Question 4: How will the transformation constants of the new method and the Stocking-Lord method differ when preparing to equate actual student data? The forms developed for these data were not constructed with strict psychometric guidelines for use in a high-stakes, large-scale testing situation, but instead were developed by college faculty interested in assessing their students' learning. An informal comparison will be made between the simulation study and the student data.

CHAPTER III

Methodology

Two studies were conducted to investigate the predicted observed score transformation method. Study I was a simulation study, with varying test length and number of common items for four ability group distributions. Study II utilized real student data from a low-stakes, standardized test administered to college freshmen and sophomores. An overview of the studies is provided in Figure 2. In Figure 2, for the simulation study, the first four steps were replicated 100 times.

The benefits of simulation studies are discussed by Davy, Nering, and Thompson (1997). First, they point out that a primary benefit of simulation studies is that the true underlying values for both examinees and items are known because the researcher has defined them. Second, simulation studies allow researchers to confirm theories empirically, as opposed to mathematical proofs. Third, simulations allow impossible conditions to be present, like the situation of "erasing" examinee memory over repeated testings. Fourth, they point out that simulations are much less time consuming, expensive, and difficult to conduct than data collection from human participants.

Davy, Nering, and Thompson (1997) go on to point out the often overlooked weaknesses of simulation studies. Simulated data studies may only be generalized to a population of interest to the extent they reflect that population. They suggest that, for many situations, simulated data are poor reflections of the actual counterpoint. For these reasons, this study included both a simulation component and an actual data set.

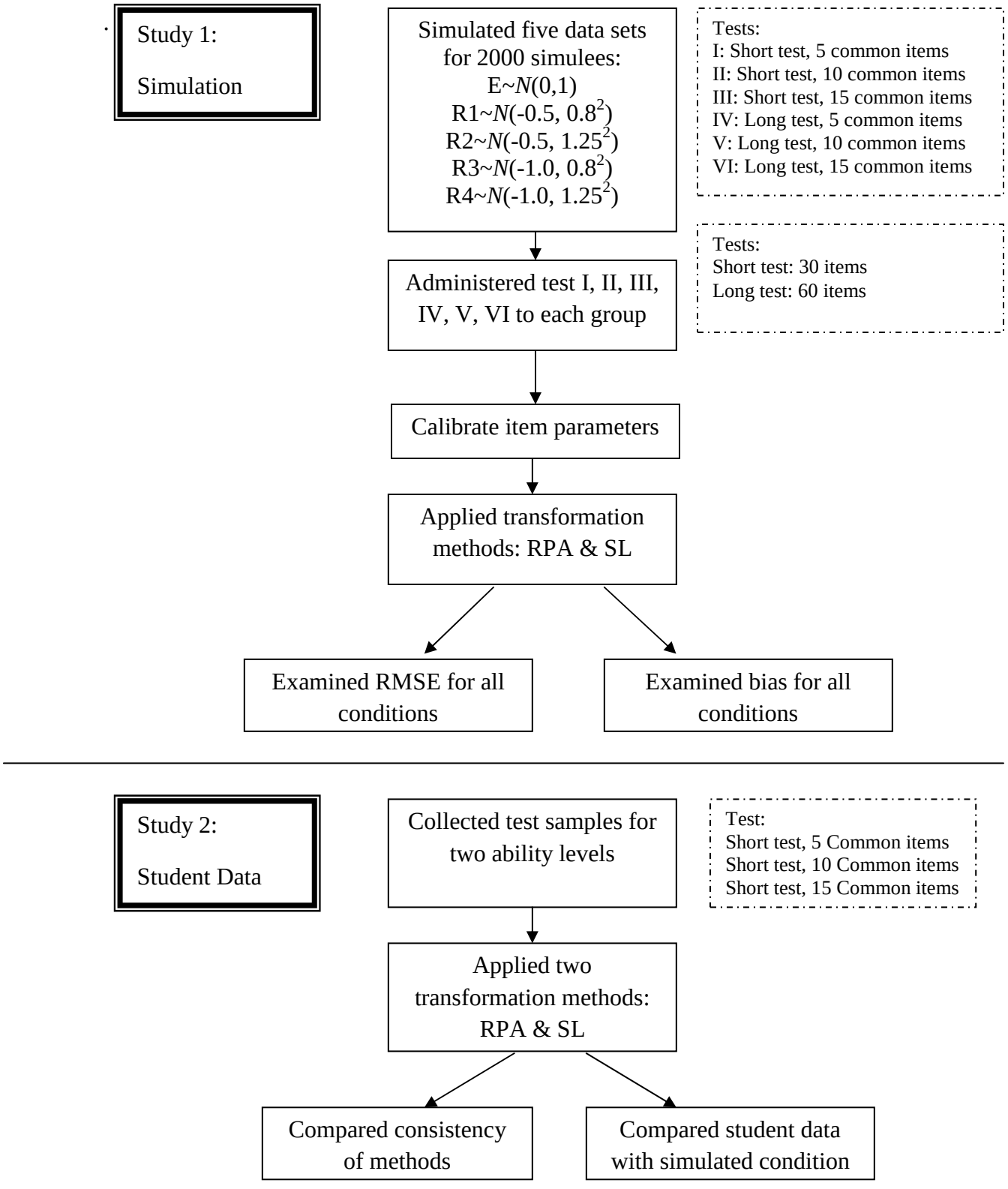

Figure 2. Description of research studies.

Overview of Study I: Simulation Study

The purpose of this study is to compare results of IRT scaling using the transformation constants estimated by the RPA method, and the IRT characteristic curve method Stocking-Lord (SL) with varying test length, number of common items, and differences in distributions of examinee abilities. This section describes the simulation conditions in detail.

Simulation Conditions

All study conditions were fully crossed resulting in 48 total conditions: 2 transformation methods by 2 test lengths by 3 different sets of common items (5, 10, or 15) and by 4 ability level comparisons. Table 2 shows the design of the simulation data analysis.

Table 2. Simulation Study Design

Test Length	Number of Common Items	Equated Form Ability Distribution							
		R1 ~N(-0.5, 0.8^2)		R2 ~N(-0.5, 1.25^2)		R3 ~N(-1.0, 0.8^2)		R4 ~N(-1.0, 1.25^2)	
		RPA	SL	RPA	SL	RPA	SL	RPA	SL
30-item	5								
	10								
	15								
60-item	5								
	10								
	15								

100 replications for each of the 24 conditions

Test Length

Two test lengths were selected to compare the RPA method with the SL method. The shorter test with 30 items is a relatively short test, and the longer test with 60 items is considered of medium length. The 30-item test can be used for comparison with the student data used in Study II.

Number of Common Items

The conventional rule of thumb described in Chapter II for the number of common items is at least 20 items. As noted, several IRT equating studies have used fewer items. For this simulation study, the number of common items used was 5, 10, or 15, which is below that rule of thumb especially when just 5 items are used. It is expected that using only 5 common items will produce greater amounts of error in the scaling constants; what is of particular interest here is the matter of degree.

Examinee Ability Distributions

The equating group used in this simulation study was drawn from a $N(0, 1)$ population. Four different comparison groups for scaling (reference groups) were also generated: R1~$N(-0.5, 0.8^2)$, R2~$N(-0.5, 1.25^2)$, R3~$N(-1.0, 0.8^2)$, R4~$N(-1.0, 1.25^2)$. They were each compared with the equating group. As compared to the equating group, they may be conceptualized as lower and narrow (R1), lower and wider (R2), much lower and narrow (R3), much lower and wider (R4), and are depicted in Figure 3.

Item parameters

Student data were used to model the generating parameters for the simulation. Descriptive statistics are provided for all items used for the student data in Table 3, and

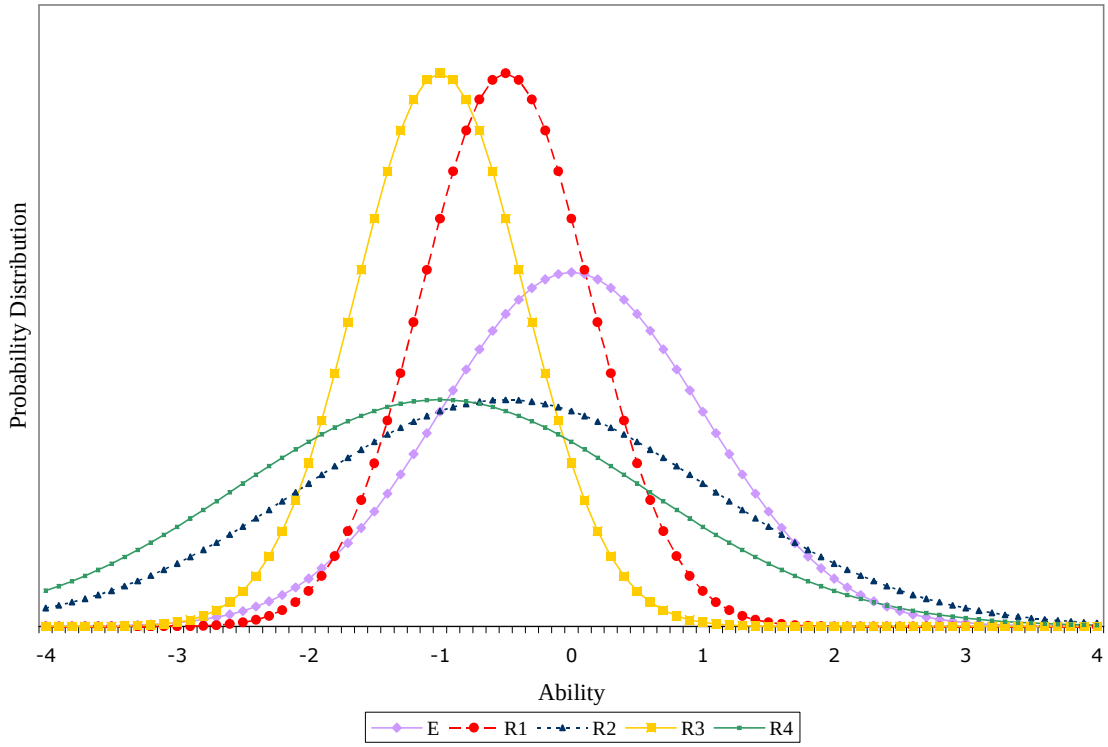

Figure 3. Distributions of Examinee Groups with Differing Abilities

are provided only for the common items in Table 4. The dichotomously scored responses were based on 1,898 examinees who responded to all items on the test.

Table 3. Characteristics of the Test Data Used to Generate the Simulated Data

Cronbach's α	.74
Number of items	30
Sample size	1,898
Raw score mean and standard deviation	20.69, 4.35
Raw score skewness and kurtosis	-0.57, 0.26
Percent score mean and standard deviation	69%, 15%
IRT Item difficulties (mean, min, max)	-0.70, -3.27, 2.18
IRT Item discriminations (mean, min, max)	0.72, 0.21, 1.56
Guessing parameters (mean, min, max)	0.24, 0.11, 0.37

Table 4. Characteristics of Common Items Used to Generate the Simulated Data

All IRT Item difficulties (mean, min, max)	-0.64, -3.27, 2.18
15 common-item set	-0.71, -2.54, 2.18
10 common-item set	-0.69, -2.54, 1.02
5 common-item set	-1.12, -2.54, 0.63
All IRT Item discriminations (mean, min, max)	0.74, 0.27, 1.56
15 common-item set	0.70, 0.29, 1.55
10 common-item set	0.78, 0.29, 1.55
5 common-item set	0.79, 0.29, 1.55
All Guessing parameters (mean, min, max)	0.24, 0.11, 0.44
15 common-item set	0.24, 0.11, 0.32
10 common-item set	0.25, 0.11, 0.37
5 common-item set	0.24, 0.11, 0.32

These item parameters were used to generate items for the simulated data. They ranged from -3.27 to 2.18 for item difficulty with a mean of -0. 64 which indicates that this test is somewhat easy for these students. Item discriminations ranged from 0.27 to 1.56 with a mean of 0.74 indicating, that overall, the item discriminations are a bit lower than those that might be seen in large-scale standardized tests. The guessing parameter estimates ranged from 0.11 to 0.44 with a mean of 0.24, with most of the parameters close to the 0.25 that would be expected by purely random guessing for the four response options. These item parameter estimates were then treated as the true parameters for simulating data for the 30 item test. For the 60-item test, the same *b*-parameters were used from the student data for 30 of the items, and each was duplicated and slightly adjusted to create an additional 30 items. The adjustment to each item was drawn from a normal distribution $N(0, 0.05)$ to provide an additional slight variation to the difficulty parameters. The discrimination parameters remained the same. Like before, these item

parameter estimates were then treated as the true parameters for simulating data for the 60-item test.

Items to be designated as common items were selected by sorting the difficulty parameters in order of lowest to highest. Common items were selected at even intervals from that continuum. For the 30-item test with 15 common items every other item was chosen beginning with the second item; with 10 common items, every third item was chosen beginning with the second item; and with 5 items, every sixth item was chosen beginning with the second item. Tables 5 and 6 present the item parameters and the items that were common to both forms for the 30-item test and the 60-item test, respectively. As can be seen in the tables, literally all items were common to both forms. However, the items not designated *common* were only used in the calibration stage and were not used in estimating the scaling constants.

Table 5. Item Parameters Used for Generating Simulated Data for 30 Item Test

Item number	Item discrimination	Item difficulty	Guessing parameter	Number of common items		
				15	10	5
1	0.266	-3.268	0.321			
2	0.527	-1.934	0.316	•	•	•
3	0.907	-0.632	0.439			
4	0.459	-0.388	0.257	•		
5	0.595	0.312	0.366		•	
6	0.768	-0.976	0.307	•		
7	0.598	-1.351	0.225			
8	0.291	-2.537	0.248	•	•	•
9	0.453	-1.606	0.227			
10	0.655	-2.190	0.231	•		
11	0.562	-0.421	0.196		•	
12	0.403	-0.982	0.276	•		
13	0.455	-2.218	0.229			
14	1.548	0.628	0.317	•	•	•
15	1.555	1.752	0.137			
16	0.445	-2.125	0.194	•		
17	0.757	-2.265	0.207		•	
18	0.938	0.024	0.273	•		
19	1.234	1.359	0.189			
20	0.852	-1.588	0.114	•	•	•
21	0.773	-1.280	0.226			
22	1.158	-0.259	0.255	•		
23	0.700	0.060	0.206		•	
24	0.634	-0.782	0.199	•		
25	0.675	0.256	0.179			
26	0.742	-0.160	0.202	•	•	•
27	1.048	-0.307	0.229			
28	0.492	2.184	0.170	•		
29	1.205	1.015	0.333		•	
30	0.610	0.477	0.173	•		

Table 6. Item Parameters Used for Generating Simulated Data for 60 Item Test

Item number	Item discrimination	Item difficulty	Guessing parameter	Number of common items		
				15	10	5
1	0.266	-3.268	0.321			
2	0.266	-3.302	0.321			
3	0.527	-1.934	0.316	•	•	•
4	0.527	-1.935	0.316			
5	0.907	-0.632	0.439			
6	0.907	-0.658	0.439			
7	0.459	-0.388	0.257	•		
8	0.459	-0.364	0.257			
9	0.595	0.312	0.366		•	
10	0.595	0.285	0.366			
11	0.768	-0.976	0.307	•		
12	0.768	-1.068	0.307			
13	0.598	-1.351	0.225			
14	0.598	-1.406	0.225			
15	0.291	-2.537	0.248	•	•	•
16	0.291	-2.562	0.248			
17	0.453	-1.606	0.227			
18	0.453	-1.604	0.227			
19	0.655	-2.190	0.231	•		
20	0.655	-2.155	0.231			
21	0.562	-0.421	0.196		•	
22	0.562	-0.375	0.196			
23	0.403	-0.982	0.276	•		
24	0.403	-0.971	0.276			
25	0.455	-2.218	0.229			
26	0.455	-2.155	0.229			
27	1.548	0.628	0.317	•	•	•
28	1.548	0.601	0.317			
29	1.555	1.752	0.137			
30	1.555	1.715	0.137			
31	0.445	-2.125	0.194	•		
32	0.455	-2.125	0.194			
33	0.757	-2.265	0.207		•	
34	0.757	-2.285	0.207			
35	0.938	0.024	0.273	•		
36	0.938	0.101	0.273			
37	1.234	1.359	0.189			
38	1.234	1.420	0.189			

Table 6. Item Parameters Used for Generating Simulated Data for 60 Item Test Continued

39	0.852	-1.588	0.114	•	•	•
40	0.852	-1.571	0.114			
41	0.773	-1.280	0.226			
42	0.773	-1.235	0.226			
43	1.158	-0.259	0.255	•		
44	1.158	-0.259	0.255			
45	0.700	0.060	0.206		•	
46	0.700	0.029	0.206			
47	0.634	-0.782	0.199	•		
48	0.634	-0.783	0.199			
49	0.675	0.256	0.179			
50	0.675	0.295	0.179			
51	0.742	-0.160	0.202	•	•	•
52	0.742	-0.140	0.202			
53	1.048	-0.307	0.229			
54	1.048	-0.320	0.229			
55	0.492	2.184	0.170	•		
56	0.492	2.222	0.170			
57	1.205	1.015	0.333		•	
58	1.205	0.984	0.333			
59	0.610	0.477	0.173	•		
60	0.612	0.424	0.173			

The simulated responses to items were then generated. Ability levels, θs, were randomly sampled from the appropriate Reference or Equating distribution. First, the 3PL model was used to calculate the probability of a correct response for an examinee based on item and ability parameters. Then a number was randomly drawn from a uniform U (0, 1) distribution. When that random draw was less than or equal to the probability of a correct response, the item was scored correct.

BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996) (version 3.0) was used to estimate the 3PL item parameters in the calibration potion of the simulation study using data that had been prescored in SAS. The FLOAT option was used to estimate the means of the prior distributions for the item parameters. The maximum number of EM

cycles was set to 50, and the maximum number of Gauss-Newton cycles that followed the EM cycles was 20. The EMPIRICAL option was used to estimate the density of the ability distribution on 15 quadrature points.

For each replication under each condition, the BILOG parameters were estimated separately. Next, the common item groups were selected under the different conditions of 15 common items, 10 common items, and 5 common items as detailed in tables 5 and 6 for both test lengths. Then the common item parameter estimates were exported to an Excel spreadsheet for each of the 100 replications to calculate the RPA and SL transformation constants. Finally, all A and B transformation constants for both methods were harvested with a SAS program to calculate summary statistics and RMSEs.

Simulation Scaling

Sample sizes of 2000 were simulated for each group. The RPA transformation was conducted using a Visual Basic algorithm written for Excel (developed by Armstrong and described in Ragland et al., 2009). The likelihood ratio G-statistic was used as the criterion for minimizing the difference between the observed and predicted score distributions. The Stocking-Lord transformation constants were also computed with another Visual Basic Algorithm.

Evaluation Criteria

The performance of the scaling methods was based on the accuracy of the equating coefficients, *A* and *B*. Two criteria were used to quantify the accuracy: (1) root mean square error (RMSE) and (2) bias. Their equations are given below.

$$bias = \frac{\Sigma(\hat{A}-A)}{N} \tag{3.1}$$

$$RMSE = \sqrt{\frac{\Sigma(\hat{A}-A)^2}{N}}, \tag{3.2}$$

where $\hat{A}$ is the estimated value of *A* and *N* is the number of replications. Substituting $\hat{B}$ and *B*, bias and RMSE was defined in the same way for *B*. Note that for the lower group with the narrowly spread distribution, the true values of *A* and *B* are 1.25 and 0.625, respectively; for the lower group with more widely spread distribution the true values of *A* and *B* are 0.8 and 0.4, respectively; for the much lower group with the narrowly spread distribution, the true values of *A* and *B* are 1.25 and 1.25; and for the much lower group with the more widely spread distribution, the true values of *A* and *B* are 0.8 and 0.8 respectively.

Overview of Study II: Student Data

The purpose of this study is to compare results with actual data using the equating coefficients generated by the RPA method and the SL method. This section describes the factors used in the study including the participants, instruments, experimental design, computing software, scaling procedure, and proposed statistical analyses.

Participants

JMU students participate in two university-wide assessment days during their undergraduate experience: as incoming freshmen, and after the completion of 45-70 credit hours when they are sophomores. JMU is a mid-size public institution with a total undergraduate enrollment for 2007-2008 of 16,414. Most students (70%) are Virginia residents, the student body has a larger proportion of females (60%) than males (40%), and minority enrollment is small (11%). Roughly 4,000 students are in each entering class.

Freshmen students are randomly assigned to tests by the last two digits of their JMU student id numbers in August, and then they retake those tests approximately 18 months later. All freshmen and sophomores are required to participate in assessment day, but the scores are used only for program evaluation. Thus, this is a low-stakes test for students, and motivation can be a threat to the validity of the scores.

To compare the transformation constants in an applied setting, two different ability groups were selected, freshmen and sophomores, where the freshmen students typically score approximately 0.5 standard deviations lower than the sophomores. Entering freshmen of 2007 and 2008 (N=2,049) were compared with the sophomore sample. For this analysis, the sophomore test was considered the reference form, Form R. The reference form was given to 1,898 sophomores in 2007 and 2008.The freshmen test, Form E, was equated to it. Because both forms contained exactly the same items, and scaling was conducted for research purposes, common items in sets of 5, 10, and 15 items were selected based on those used in the simulation study.

Instrument

The Global Experience Exam is a test of student knowledge of global issues, covering economic, social, political, and cultural areas. The 32-item multiple choice, dichotomously scored exam is administered by pencil and paper on assessment day. Only 30 of the items were used in the analysis, because one item had a negative item-total correlation and another had a low item-total correlation.

Proposed Statistical Analyses

As in Study I, the RPA transformation constants were calculated using the Visual Basic algorithm written for Excel (developed by Armstrong and described in Ragland et al., 2009), and the SL transformation constants were computed with that same program. Transformation constants were calculated for each equating method. Comparisons were also made between these actual data and the simulation condition of comparable proportions (sample size, test length, number of common items) from Study I.

BILOG-MG (version 3.0) was used to estimate item parameters and SAS was used for data management and descriptive statistics. Because the true values of *A* and *B* were unknown, the estimates were compared to each other instead of to the true values. Instead of bias, the average difference between each pair of methods was calculated.

A resampling procedure, also known as the bootstrap procedure, was also conducted on the student data. The resampling procedure was used to estimate the sampling distribution of the transformation constants of the observed student data. SAS statistical software was used to generate samples of 5, 10, or 15 common items from the 30-item test. There was no replication of an item within a sample, but items were replicated across samples. For example, if item 5 was chosen in the first random sample,

it could not serve as another common item in that set, but it was eligible for selection for the next sample. The item parameters were those originally estimated from the test.

The selection of common items for transforming scores to be on the same scale is rarely done by random selection in practice. Stable, reliable, well-placed items from appropriate content areas or objectives are purposefully selected to ensure the most stable results. A limited number of these items may exist in practice, and it is important to carefully select and place the common items on the test forms.

However, from a theoretical perspective, random selection of items as common items for transformation provides an opportunity to examine what range of transformation constants to expect from a particular data set. The resampling procedure is often used to examine the properties of the mean or variance of an estimator of interest for a single data set of interest. This analysis provides another estimate of the true *A* and *B* constants as well as the standard deviation of the distribution of those constants. Therefore, while a single data set has only one estimate for *A* and one estimate for *B*, a single data set that has been repeatedly sampled has a *distribution* of estimates, which can be summarized by the mean and standard deviation of *A* and *B*.

CHAPTER IV

Results

The results in this chapter are presented in sections corresponding to each research question posed in Chapter II. First, the impact of test length on the accuracy of the transformation constants of the RPA method compared with the SL method will be examined. Second, the influence of the number of common items will be examined to see how that factor affects the accuracy of the two methods. Third, the transformation constants derived from differing ability distributions of populations to be scaled will be compared. Then the interactions among these factors will be described. Finally, the relationship between the transformation constants generated through the simulation study will be informally compared with actual student data. For this study, a resampling (bootstrap) analysis was done to create an empirical distribution of transformation constants based on random selection of possible common item sets of size 15, 10, or 5. Since the true values of *A* and *B* are not known, the resampling analysis provides an estimate that may be used to evaluate the single constants that are calculated from one application of the RPA and SL methods.

Study 1: Simulation Study

The focus of the research questions was on the main effects of test length, number of common items, and differing ability distributions, and the interactions between these factors and the transformation methods. Boxplots for each reference group show the magnitude of the A or B transformation constant and are presented in Figures 4 through 11 and the bias and RMSE are shown in Tables 7 through 10. To address each research question, the bias and RMSE were averaged over the other factors. The bias and RMSE

were calculated for each transformation constant under each condition, and are presented in Tables 7 -10.

The boxplots are given by each reference group first for the A transformation constant and then the B transformation constant. For Figures 4 through 11 the scale of the y-axis was re-centered to match the true parameter values in each group, and the range was adjusted to accommodate different levels of variance. Other summary data, namely means, standard deviations, minimum and maximum for each condition are presented in Appendix A.

Table 7. RPA and SL Bias and RMSE for Transformation Constant A for 60-item Test

Common Items		RPA			SL		
	Group	Mean	Bias	RMSE	Mean	Bias	RMSE
15	R1	1.2645	0.0145	0.0598	1.2578	0.0078	0.0597
10		1.2546	0.0046	0.0596	1.2538	0.0038	0.0608
5		1.2660	0.0160	0.1001	1.2649	0.0149	0.0994
15	R2	0.8095	0.0095	0.0300	0.8118	0.0118	0.0335
10		0.8090	0.0090	0.0334	0.8127	0.0127	0.0374
5		0.8160	0.0160	0.0550	0.8177	0.0177	0.0588
15	R3	1.2743	0.0243	0.0600	1.2730	0.0230	0.0598
10		1.2636	0.0136	0.0679	1.2721	0.0221	0.0704
5		1.3265	0.0765	0.1490	1.3254	0.0754	0.1413
15	R4	0.8184	0.0184	0.0382	0.8179	0.0179	0.0397
10		0.8177	0.0177	0.0397	0.8205	0.0205	0.0427
5		0.8313	0.0313	0.0721	0.8307	0.0307	0.0760

Table 8. RPA and SL Bias and RMSE for Transformation Constant A for 30-item Test

Common Items		RPA			SL		
	Group	Mean	Bias	RMSE	Mean	Bias	RMSE
15	R1	1.2658	0.0158	0.0577	1.2634	0.0134	0.0582
10		1.2767	0.0267	0.0712	1.2708	0.0208	0.0699
5		1.2523	0.0023	0.0997	1.2614	0.0114	0.0931
15	R2	0.8105	0.0105	0.0376	0.8100	0.0100	0.0389
10		0.8100	0.0100	0.0377	0.8107	0.0107	0.0407
5		0.8143	0.0143	0.0532	0.8163	0.0163	0.0570
15	R3	1.2856	0.0356	0.0767	1.2808	0.0308	0.0757
10		1.2892	0.0392	0.0893	1.2833	0.0333	0.0874
5		1.2251	-0.0249	0.1048	1.2560	0.0060	0.1102
15	R4	0.8202	0.0202	0.0451	0.8195	0.0195	0.0465
10		0.8236	0.0236	0.0496	0.8221	0.0221	0.0506
5		0.8305	0.0305	0.0650	0.8345	0.0345	0.0742

Table 9. RPA and SL Bias and RMSE for Transformation Constant B for 60-item Test

Common Items		RPA			SL		
	Group	Mean	Bias	RMSE	Mean	Bias	RMSE
15	R1	0.6271	0.0021	0.0593	0.6225	-0.0025	0.0603
10		0.6189	-0.0061	0.0742	0.6177	-0.0073	0.0723
5		0.6222	-0.0028	0.0721	0.6188	-0.0062	0.0731
15	R2	0.3916	-0.0084	0.0399	0.3941	-0.0059	0.0398
10		0.3883	-0.0117	0.0423	0.3921	-0.0079	0.0424
5		0.3837	-0.0163	0.0560	0.3858	-0.0142	0.0541
15	R3	1.2490	-0.0010	0.0673	1.2467	-0.0033	0.0677
10		1.2301	-0.0199	0.0839	1.2378	-0.0122	0.0823
5		1.2766	0.0266	0.1246	1.2713	0.0213	0.1041
15	R4	0.7942	-0.0058	0.0458	0.7943	-0.0057	0.0460
10		0.7926	-0.0074	0.0498	0.7972	-0.0028	0.0514
5		0.7893	-0.0107	0.0619	0.7881	-0.0119	0.0591

Table 10. RPA and SL Bias and RMSE for Transformation Constant B for 30-item Test

Common Items		RPA			SL		
	Group	Mean	Bias	RMSE	Mean	Bias	RMSE
15	R1	0.6353	0.0103	0.0526	0.6332	0.0082	0.0515
10		0.6406	0.0156	0.0580	0.6361	0.0111	0.0579
5		0.6345	0.0095	0.0903	0.6341	0.0091	0.0832
15	R2	0.3982	-0.0018	0.0405	0.3976	-0.0024	0.0401
10		0.3955	-0.0045	0.0421	0.3956	-0.0044	0.0414
5		0.4009	0.0009	0.0516	0.4008	0.0008	0.0494
15	R3	1.2792	0.0292	0.0711	1.2716	0.0216	0.0658
10		1.2781	0.0281	0.0833	1.2685	0.0185	0.0795
5		1.2308	-0.0192	0.1094	1.2499	-0.0001	0.0899
15	R4	0.8035	0.0035	0.0405	0.8019	0.0019	0.0397
10		0.8050	0.0050	0.0431	0.8016	0.0016	0.0431
5		0.8077	0.0077	0.0665	0.8084	0.0084	0.0609

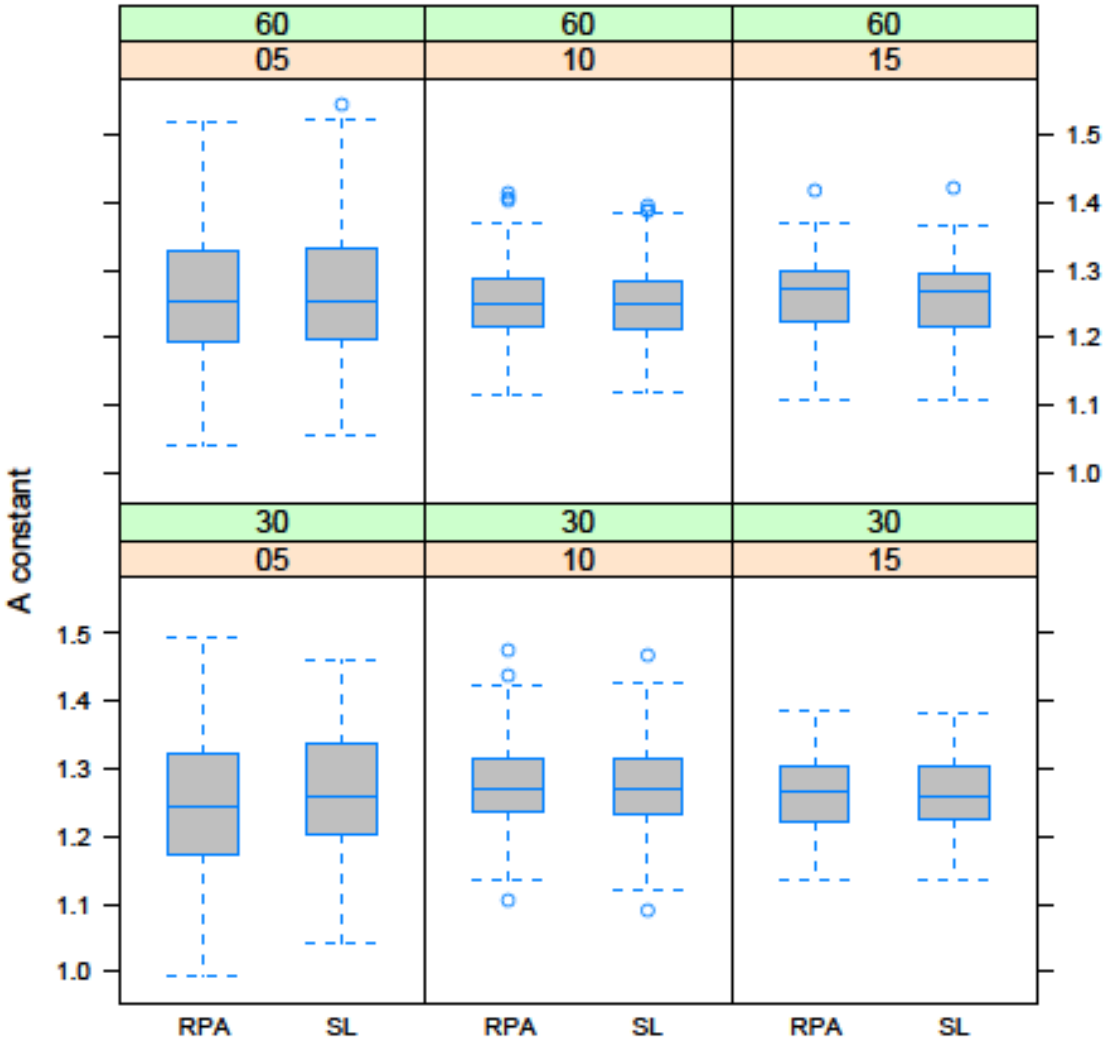

Figure 4. Box plots for RPA and SL methods for Transformation Constant *A* for Reference Group R1 for 60 and 30-item tests with 5, 10 or 15 common items.

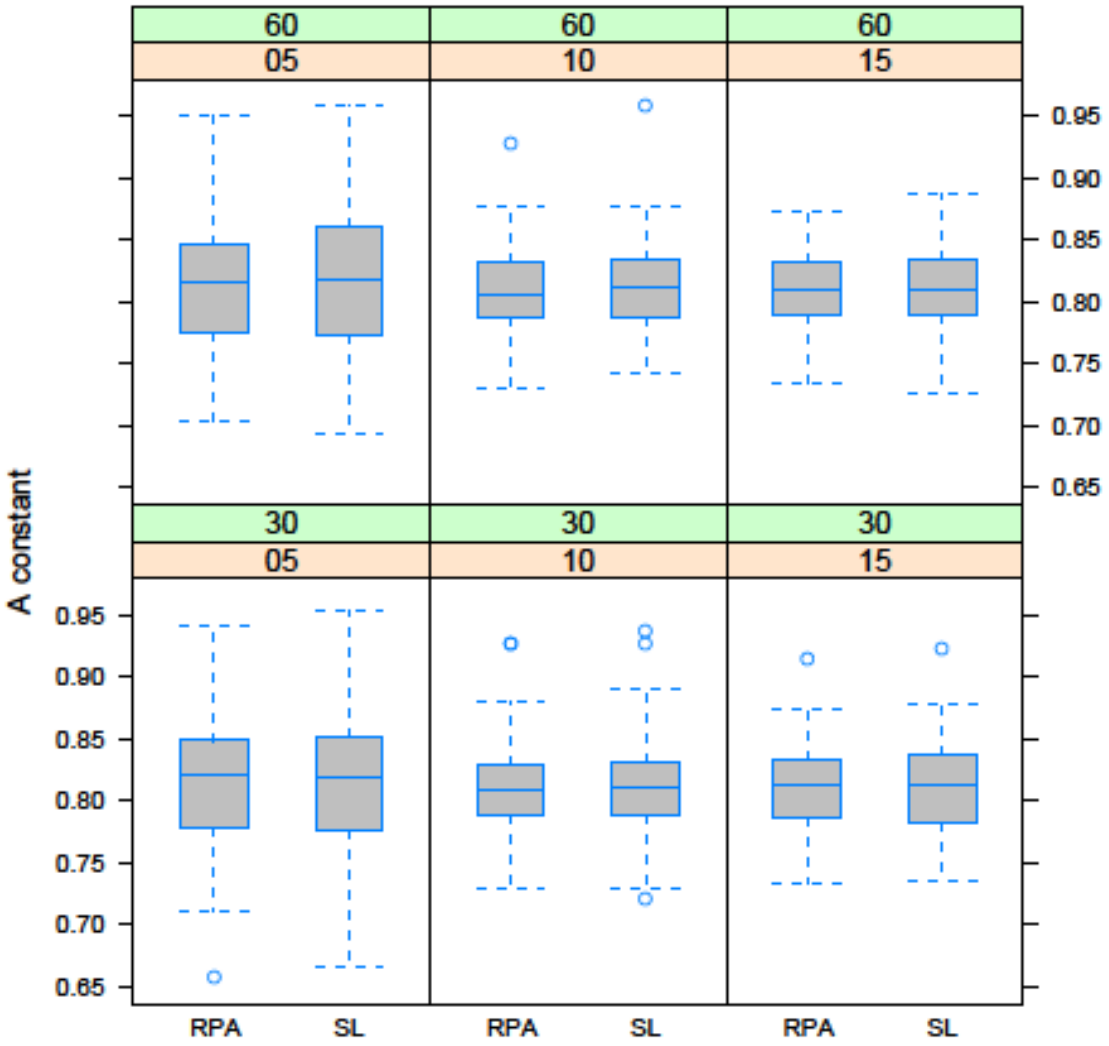

Figure 5. Box plots for RPA and SL methods for Transformation Constant *A* for Reference Group R2 for 60 and 30-item tests with 5, 10 or 15 common items.

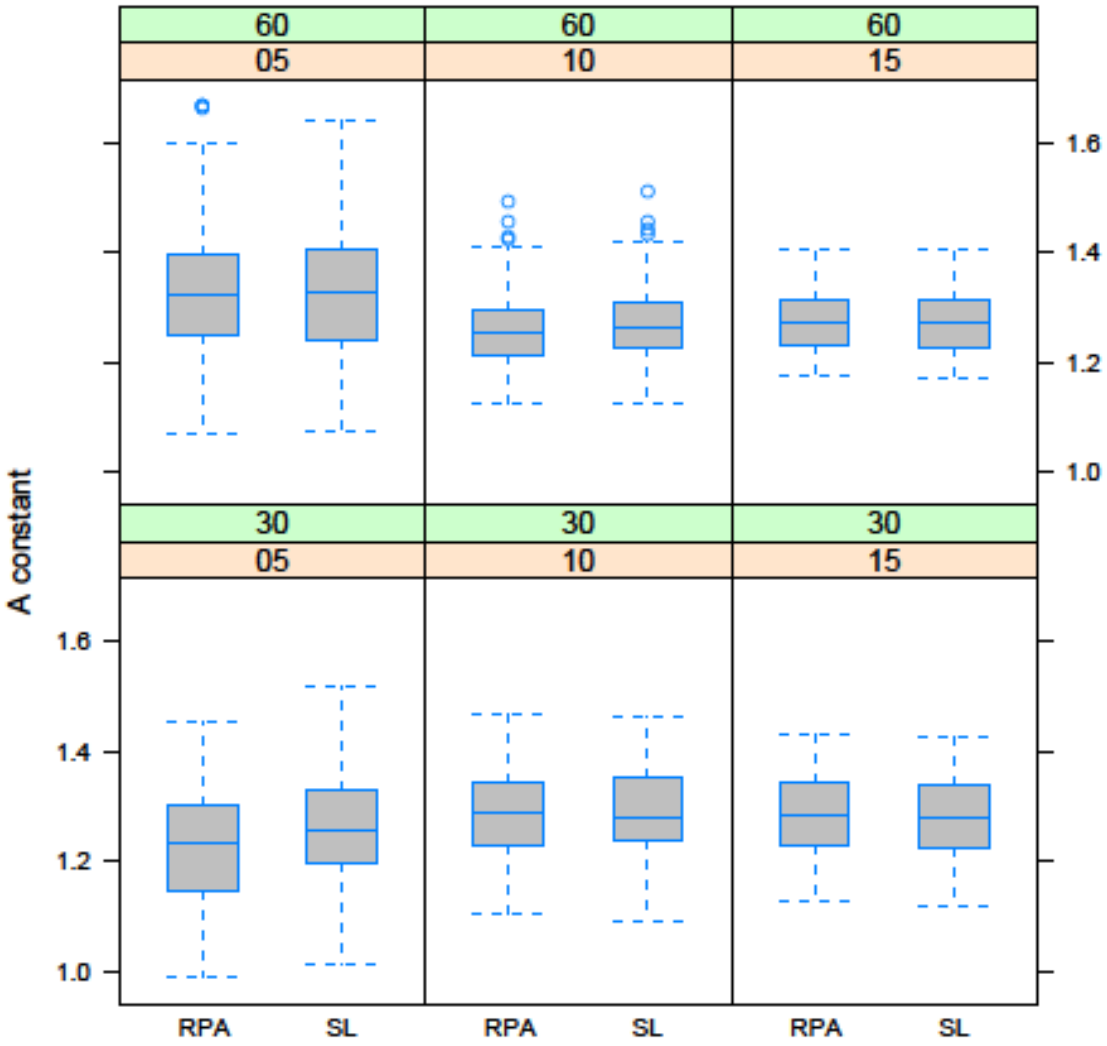

Figure 6. Box plots for RPA and SL methods for Transformation Constant *A* for Reference Group R3 for 60 and 30-item tests with 5, 10 or 15 common items.

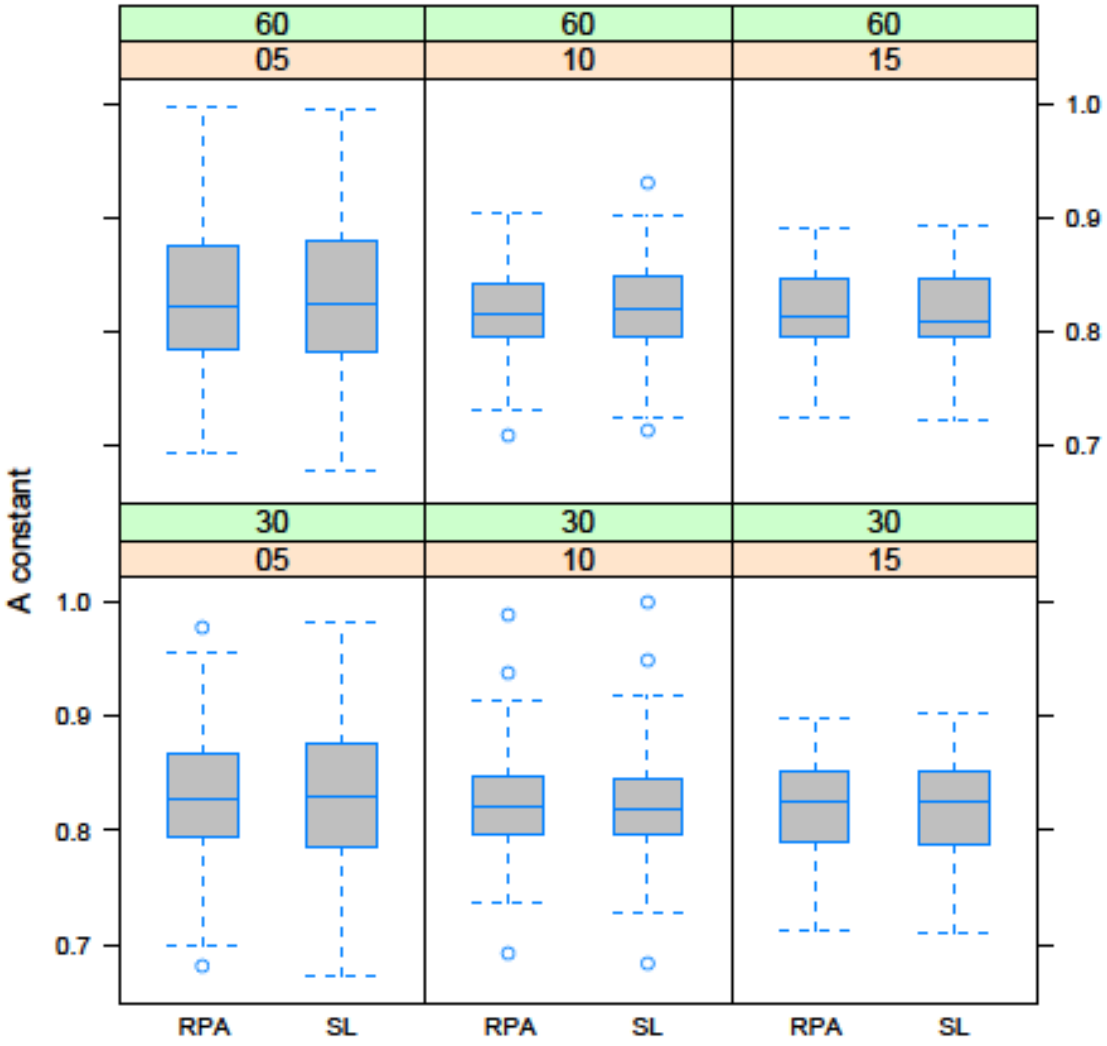

Figure 7. Box plots for RPA and SL methods for Transformation Constant *A* for Reference Group R4 for 60 and 30-item tests with 5, 10 or 15 common items.

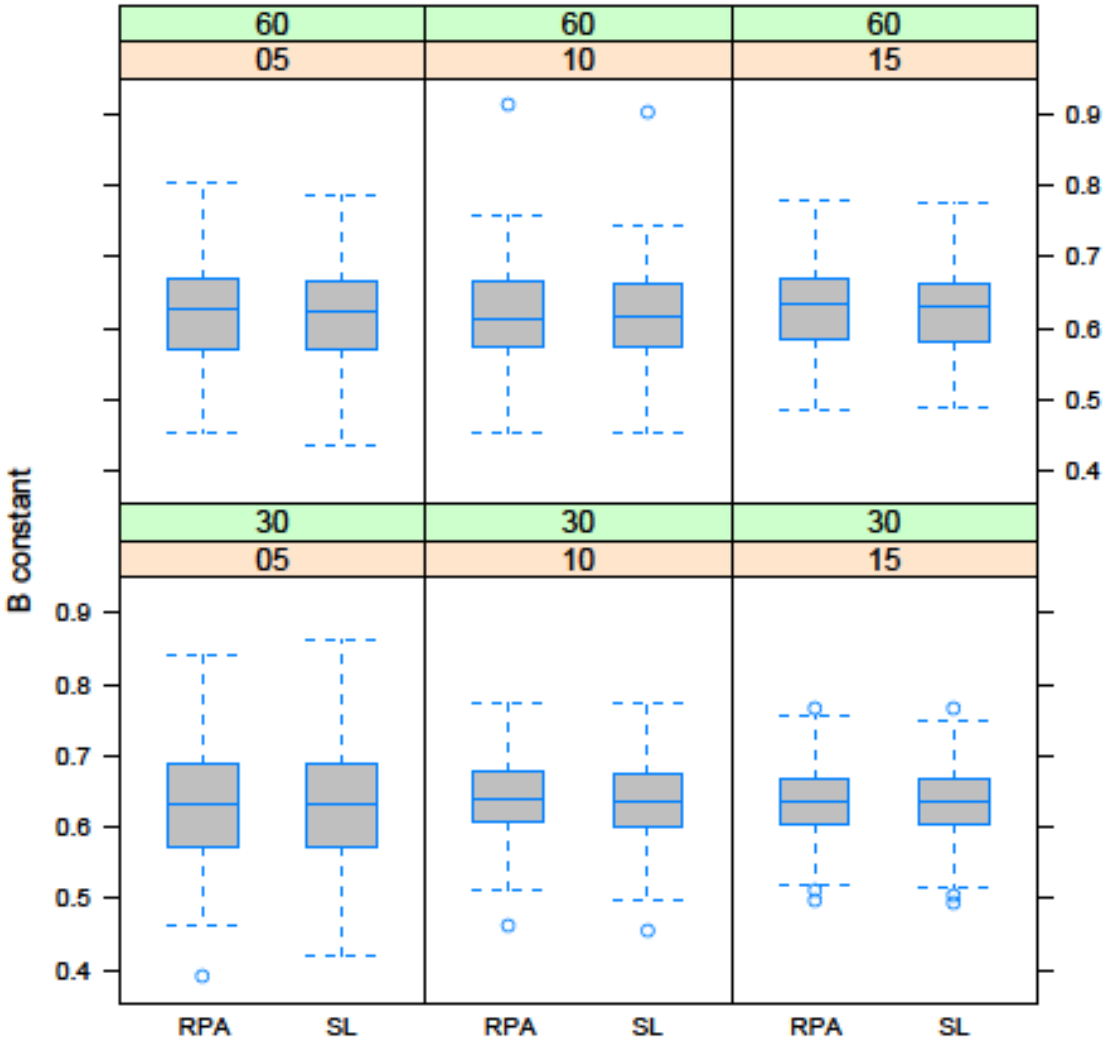

Figure 8. Box plots for RPA and SL methods for Transformation Constant *B* for Reference Group R1 for 60 and 30-item tests with 5, 10 or 15 common items.

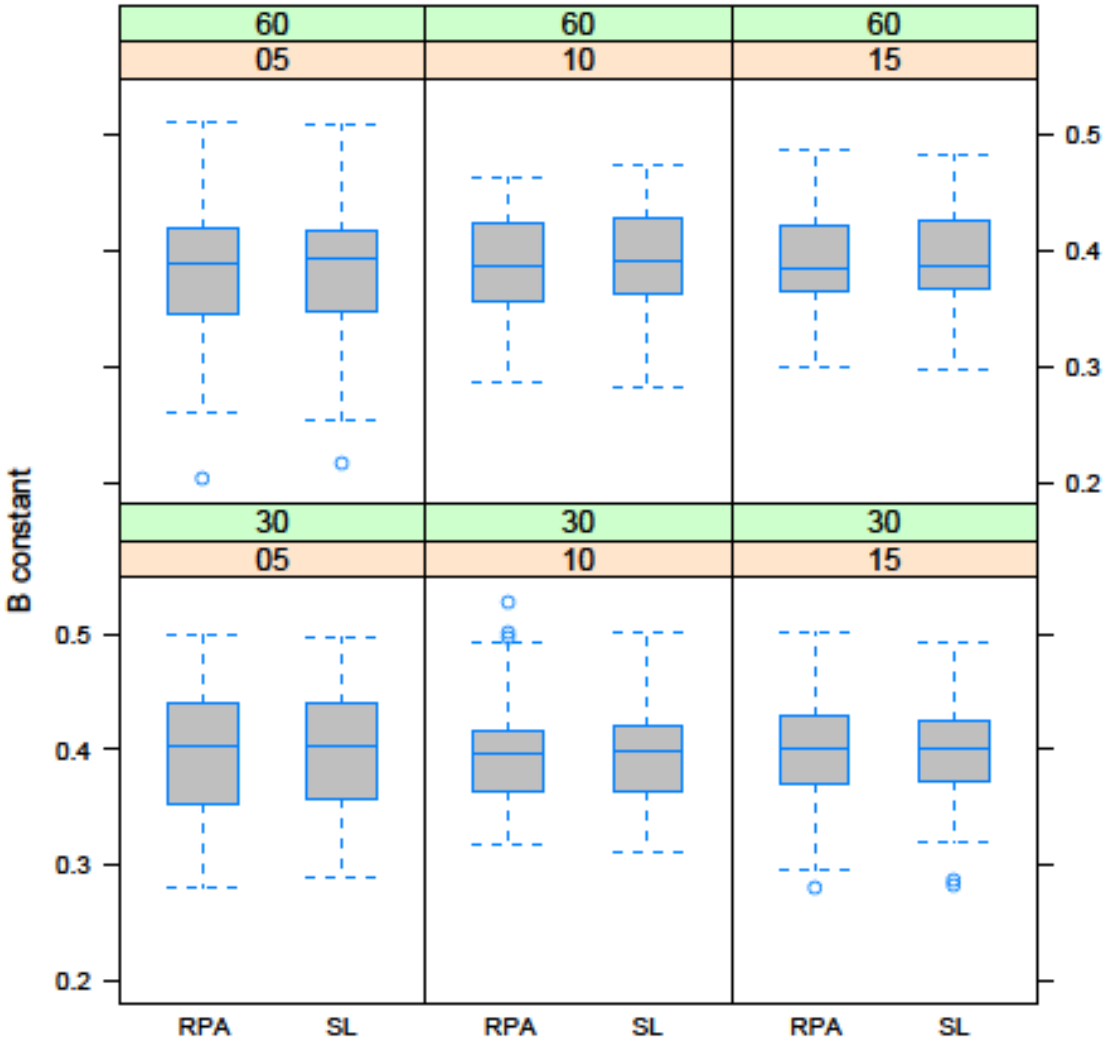

Figure 9. Box plots for RPA and SL methods for Transformation Constant *B* for Reference Group R2 for 60 and 30-item tests with 5, 10 or 15 common items.

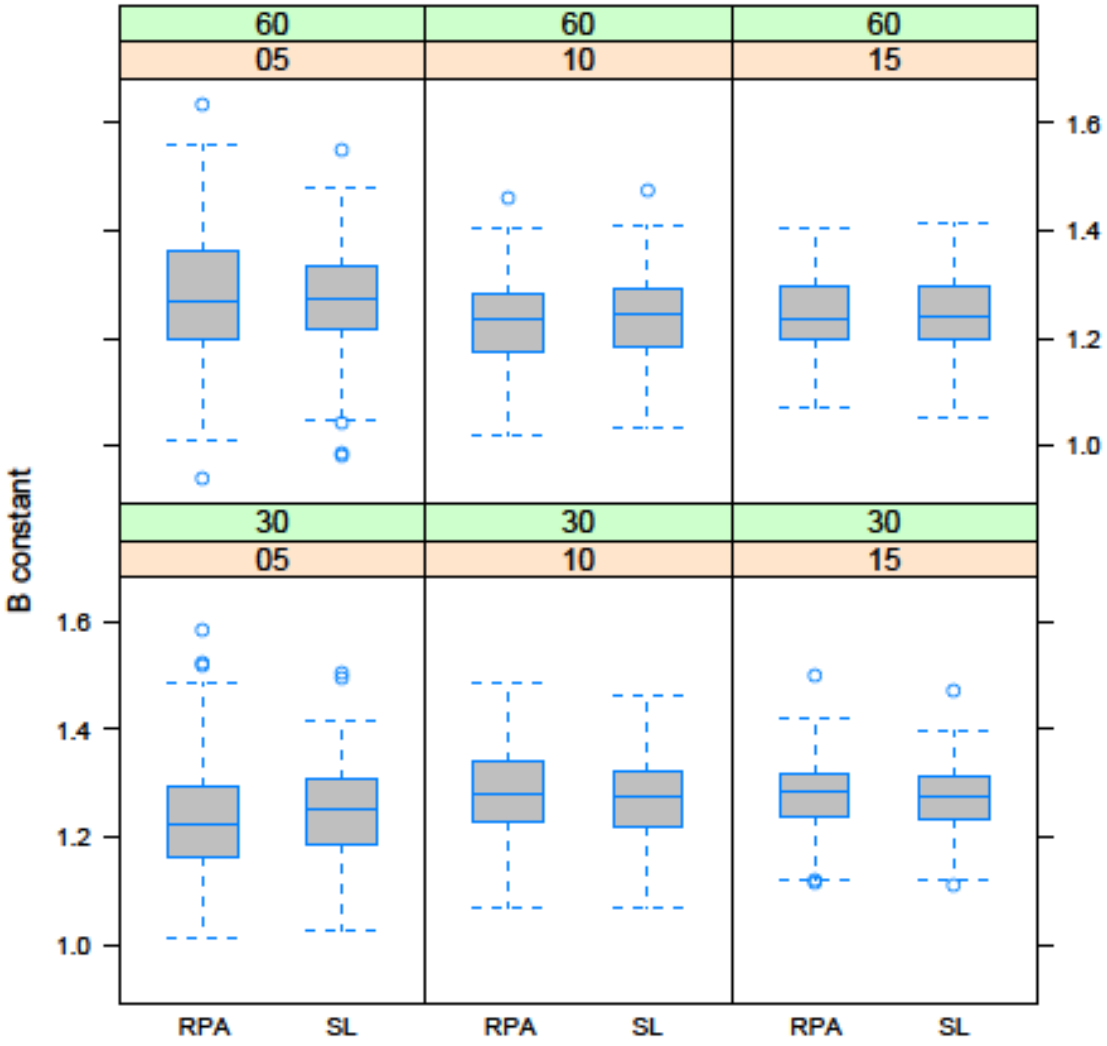

Figure 10. Box plots for RPA and SL methods for Transformation Constant *B* for Reference Group R3 for 60 and 30-item tests with 5, 10 or 15 common items.

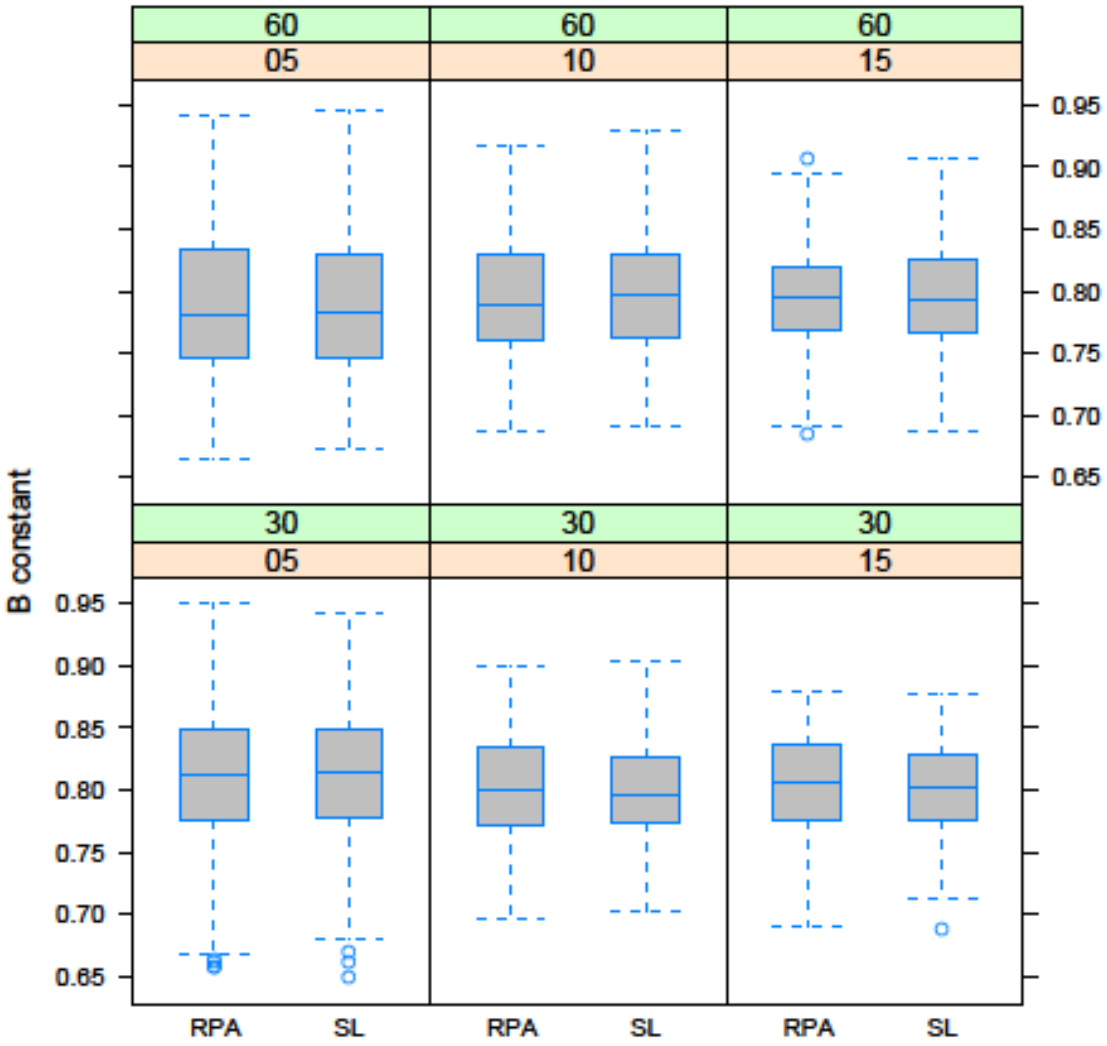

Figure 11. Box plots for RPA and SL methods for Transformation Constant *B* for Reference Group R4 for 60 and 30-item tests with 5, 10 or 15 common items.

Test Length

Two test lengths were used in the simulation study to reflect tests that are commonly used in practice. A 30-item test is on the end of the spectrum that would be considered a short test, and the 60-item test could be considered a medium test. Each equating and reference group was simulated to take both tests. Overall, test length had little impact on bias or RMSE for both the *A* constant and *B* constant.

The bias and RMSE for each of the transformation constants for each method is provided in the following figures. In Figure 12, the average bias across all reference groups and numbers of common items for transformation constant *A* is presented. For both methods, about the same amount of bias was produced in the 60-item test condition as in the 30-item test condition. This is also evident from the boxplots in Figures 4-7; the medians do not appear to vary by test length. The average bias for the 60-item test was virtually the same for the RPA method (0.0210) and the SL method (0.0215). For the 30-item test, the average bias was again similar for the RPA method (0.0170) and the SL method (0.0191).

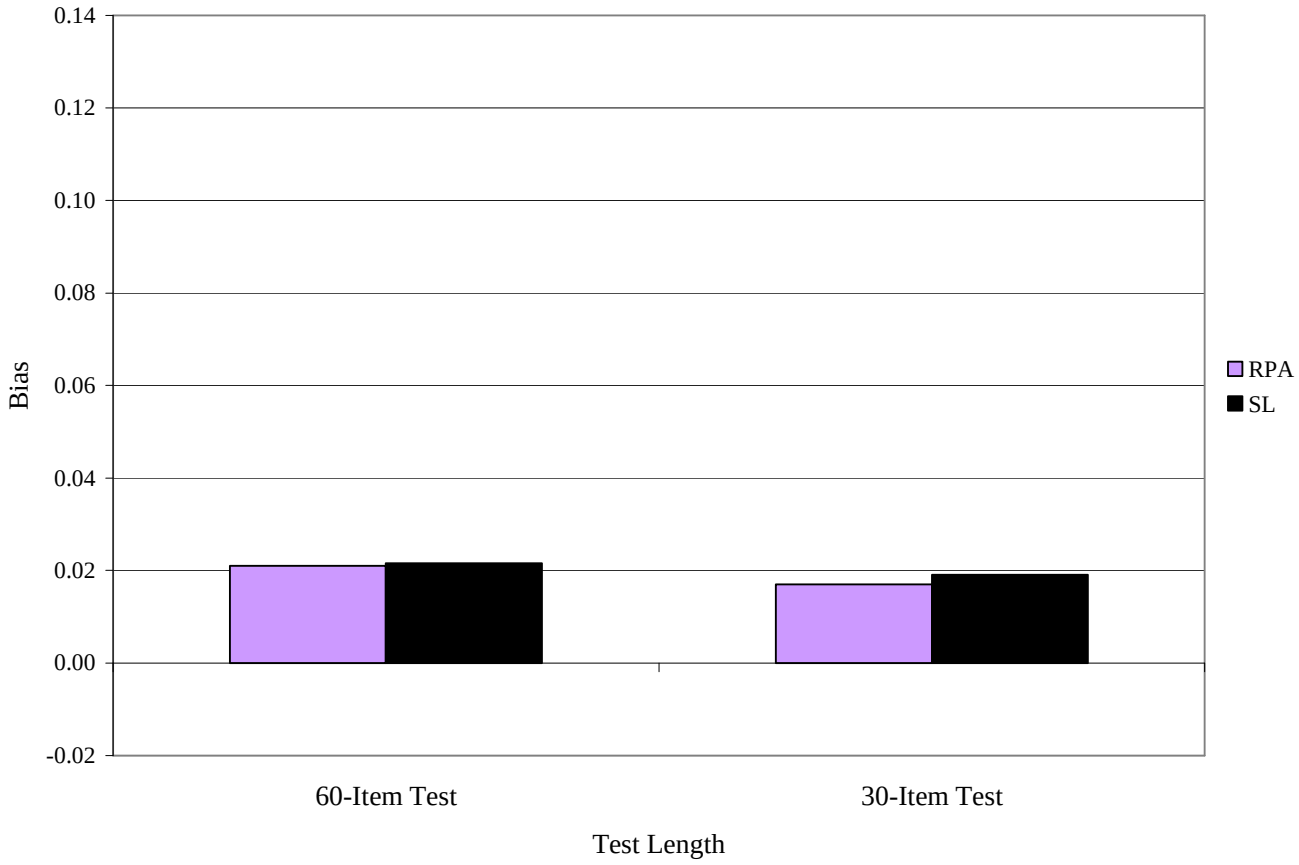

Figure 12. Bias for Transformation Constant *A* for 60-Item Test and 30-Item Test

The RMSE was also averaged within each test length condition. The amount of RMSE was nearly the same for the 60-item test and the 30-item test. Figure 13 shows that the amount of RMSE for the RPA method (0.0712) was identical to the SL method (0.0712) for the 60-item test and that the amount of RMSE for the RPA method (0.0693) was similar to the SL method (0.0702) for the 30-item test. The boxplots in Figures 4-7 also show little difference between test lengths.

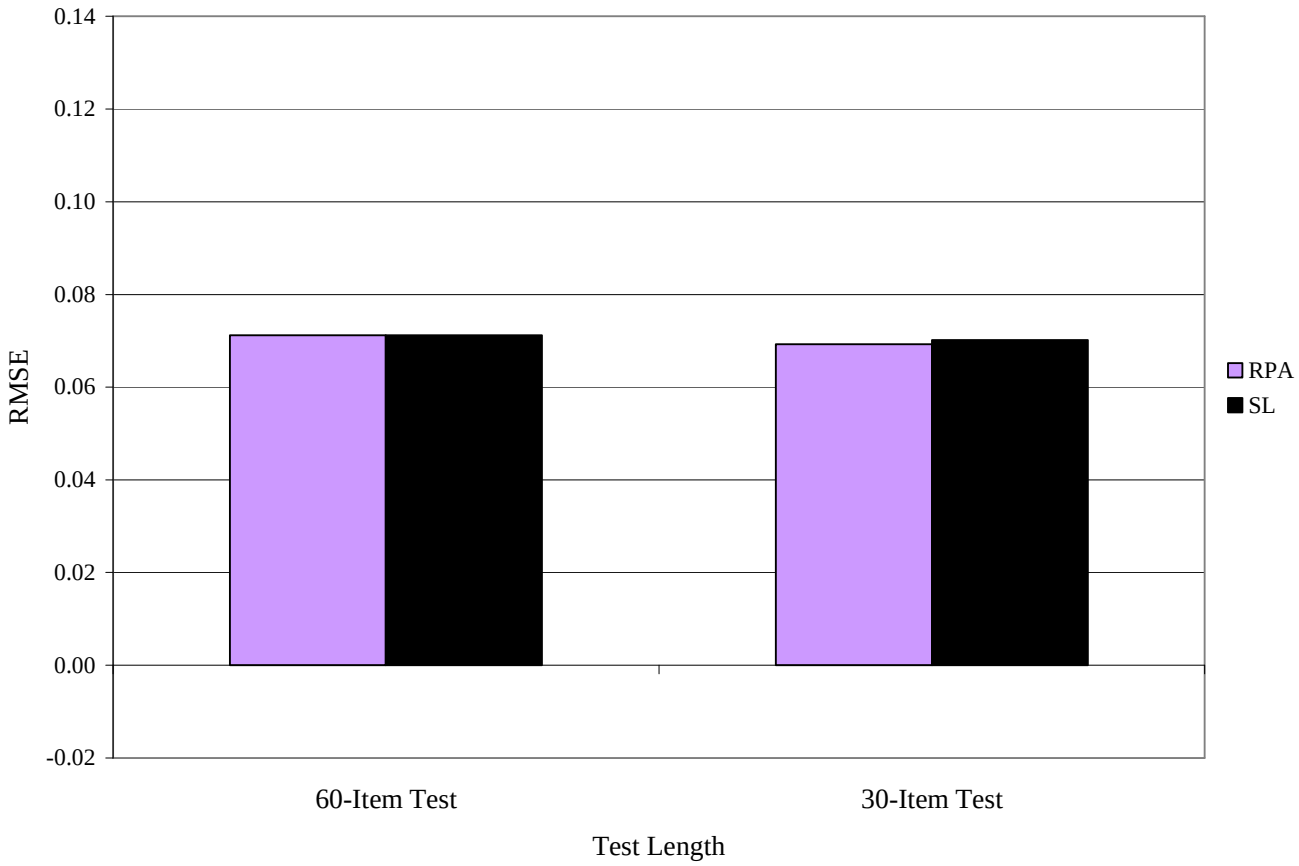

Figure 13. RMSE for Transformation Constant *A* for 60-Item Test and 30-Item Test

In Figure 14, it can be seen that slightly less bias was produced in the 60-item test condition than in the 30-item test for the *B* transformation constant. On the 60-item test the average bias was -0.0051 for the RPA method and -0.0049 for the SL method. The average bias for the 30-item test for the RPA method was 0.0070, and the average bias for the SL method was 0.0062 for the 30-item test. The boxplots in Figures 8-10 illustrate the differences between the RPA and SL methods for each of the reference groups. As seen for transformation constant *A*, the medians do not vary much by test length for transformation constant *B*.

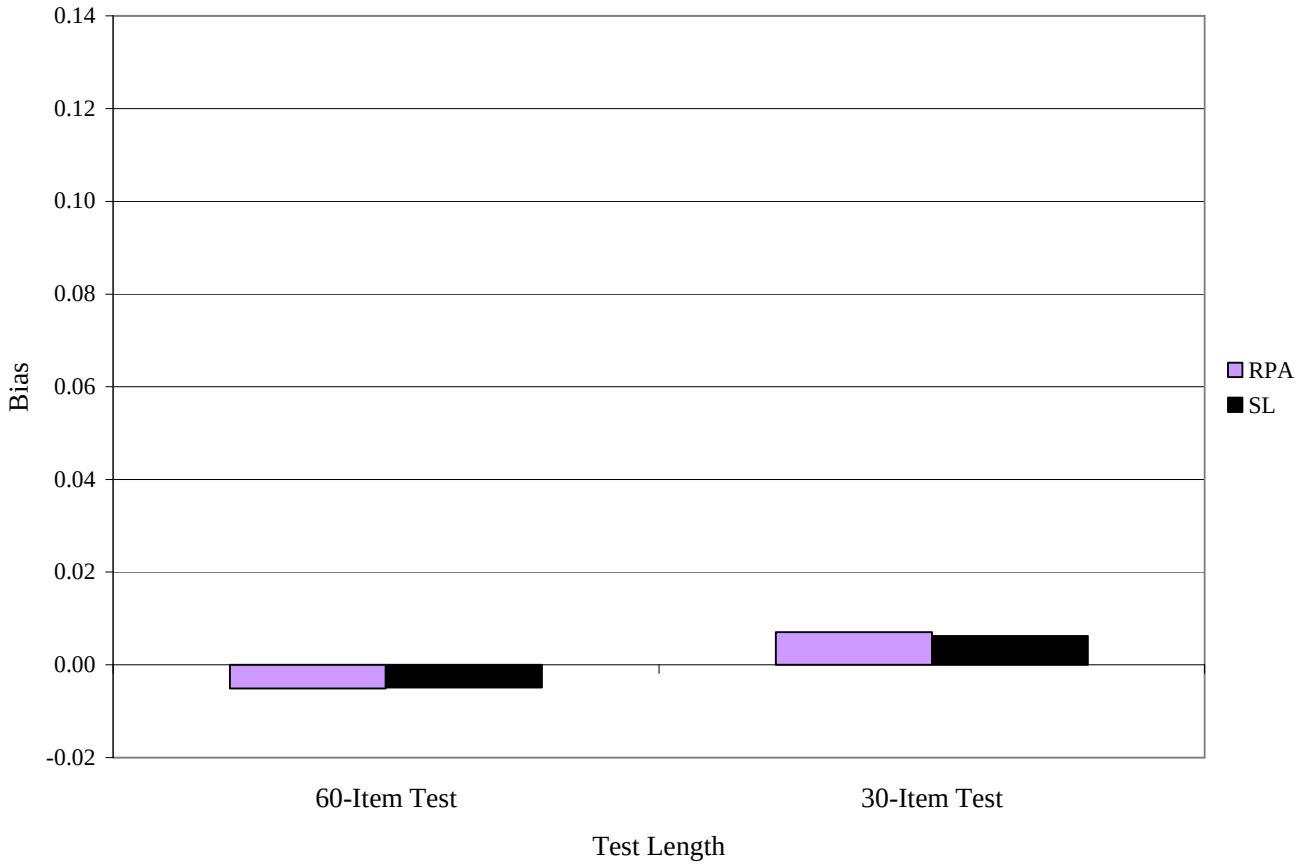

Figure 14. Bias for Transformation Constant *B* for the 60-Item Test and the 30-Item Test

Figure 15 shows the amount of RMSE for the RPA method and the SL method for the *B* transformation constant. The amount of RMSE was quite similar for the RPA method (0.0684) and the SL method (0.0652) for the 60-item test condition as well as the 30-item test condition, where the RPA method RMSE was 0.0660 and the SL method RMSE was 0.0609. The boxplots in Figures 8-11 also show little variability between the two test lengths.

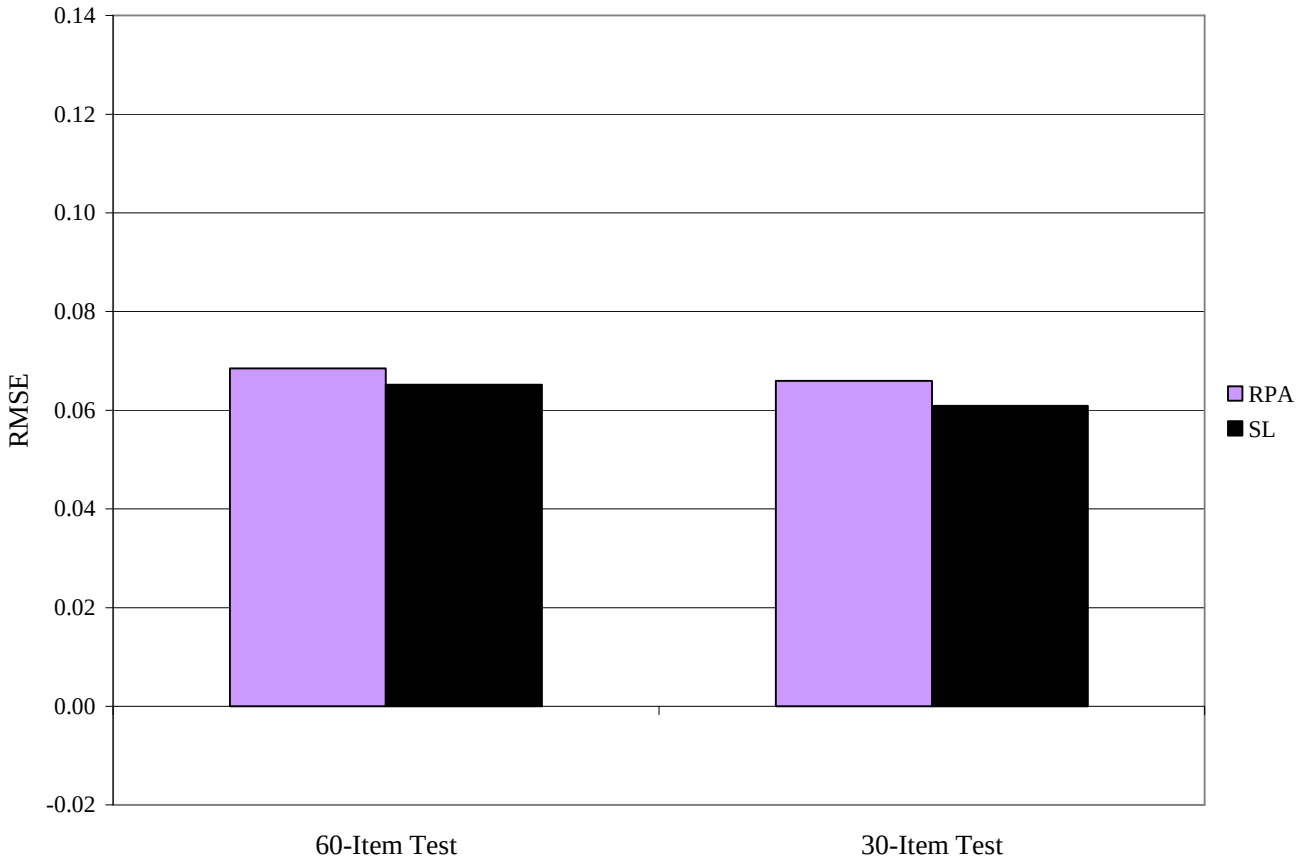

Figure 15. RMSE for Transformation Constant *B* for the 60-Item Test and the 30-Item Test

Number of Common Items

Three different sets of common items were used in this study: 5, 10, and 15 common items. It is generally expected that more common items will produce more accurate transformation constants. The following figures illustrate the differences between the two methods of estimating the *A* transformation constant. Figure 16 shows the average bias within each common-item condition. The RPA method produced slightly more bias than the SL method when 15 common items were used but slightly less bias when 5 items were used. The bias was largest for 5 common items for both methods where the RPA method (0.0203) was slightly lower than the SL method (0.0259).

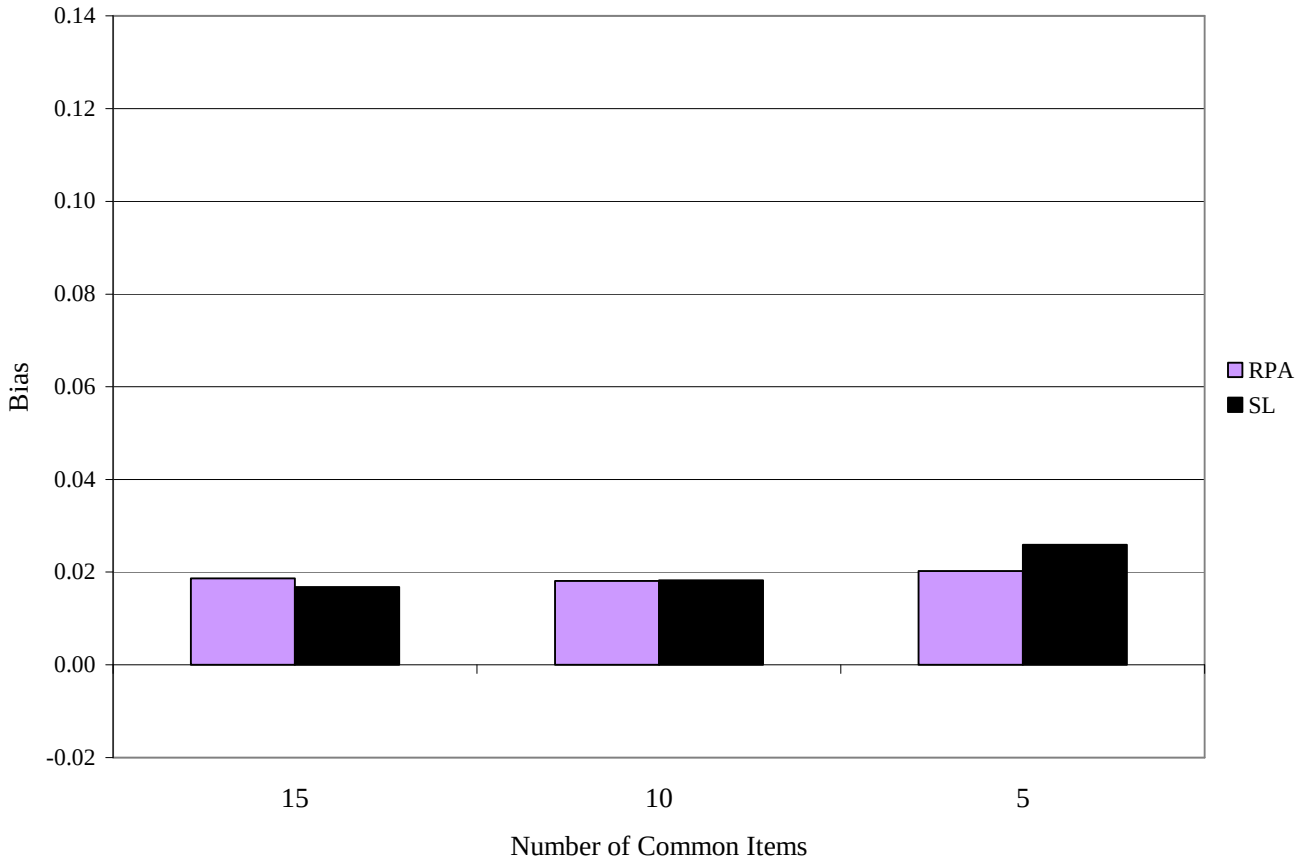

Figure 16. Bias for Transformation Constant *A* for 15 Common Items, 10 Common Items, and 5 Common Items

Figure 17 shows the RMSE for the RPA method and the SL method for the *A* transformation constant. The amount of RMSE for the RPA method was similar to the SL method in all common item conditions. It was greatest for both methods in the 5 common-item condition where the RPA method RMSE was 0.0925 and the SL method RMSE was 0.0927. This result is also reflected in the boxplots in Figures 4-7. The variability of the distributions is much greater for the 5 common-item condition than the 10 or 15-common item conditions for all reference groups.

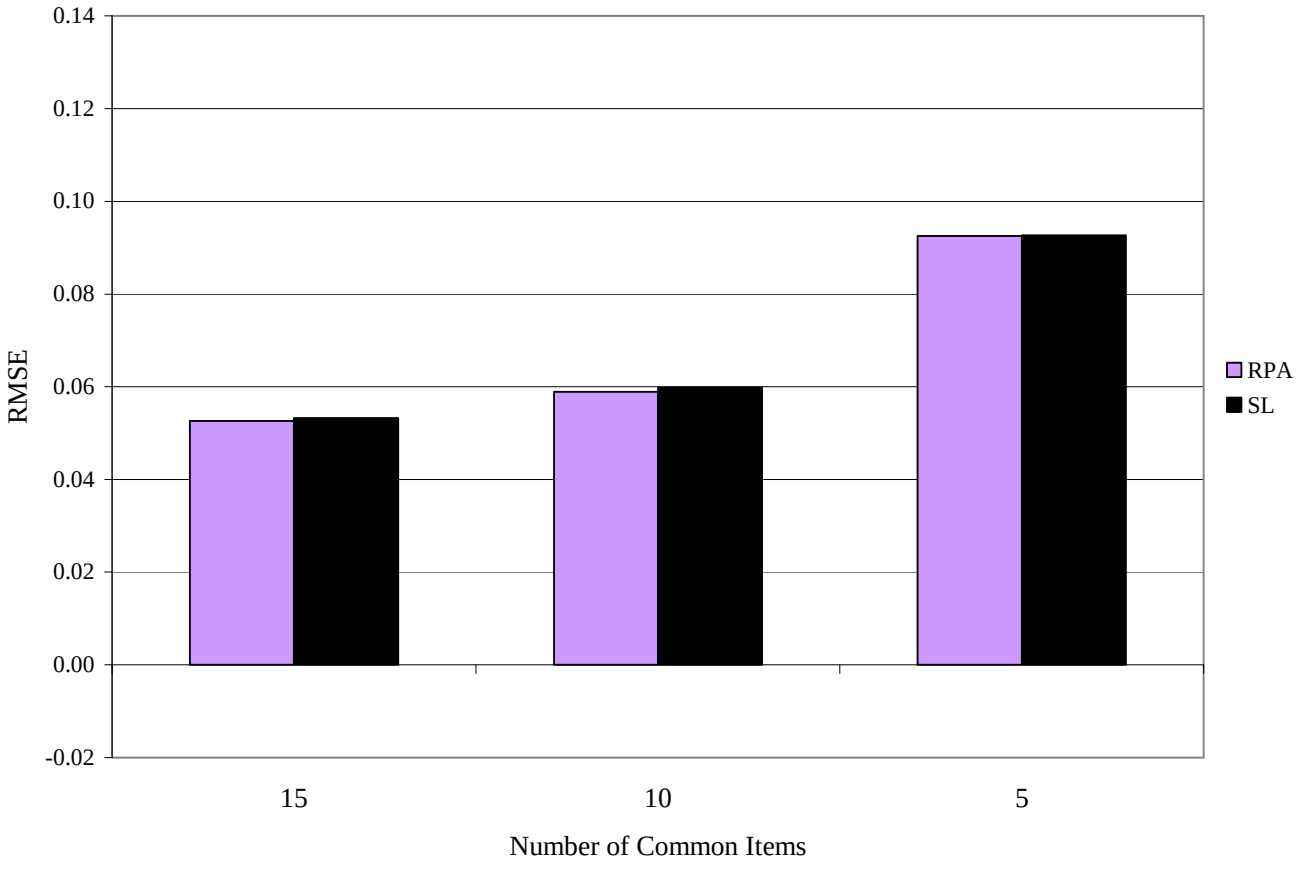

Figure 17. RMSE for Transformation Constant *A* for 15 Common Items, 10 Common Items, and 5 Common Items

Figure 18 shows the average bias within each of the three common-item conditions for the *B* transformation constant. For all common item size groups, the RPA produced about the same amount of bias as the SL method. The amount of bias was extremely small for all levels of common items ranging from the -0.0001 for the RPA method for 10 common items to 0.0035 for the RPA method for 15 common items. The boxplots in Figures 8-11 also show the medians do not vary by number of common items for all reference groups.

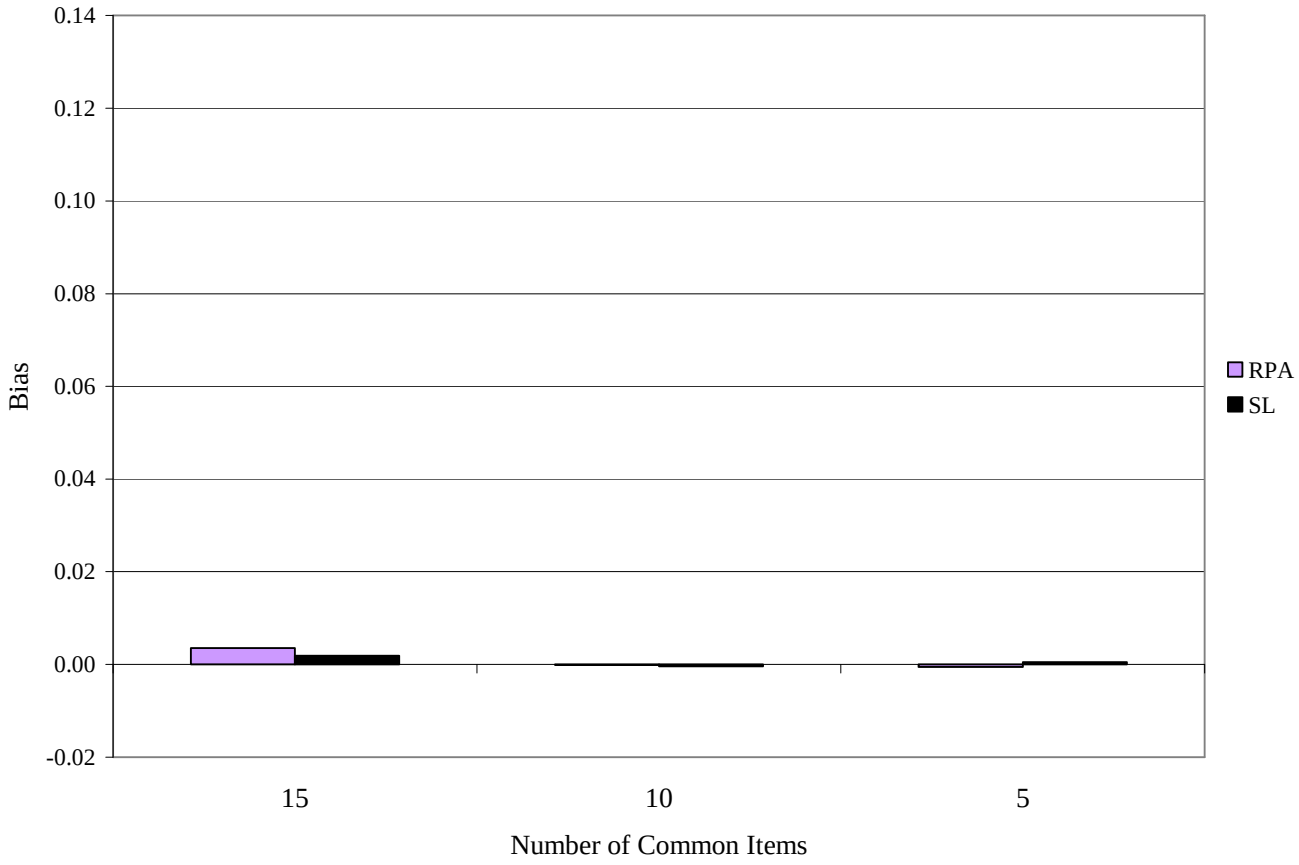

Figure 18. Bias for Transformation Constant *B* for 15 Common Items, 10 Common Items, and 5 Common Items

Figure 19 shows the average RMSE across the three common-item conditions for the *B* transformation constant. For 15 and 10 common item size groups, the RPA method produced about the same RMSE as the SL method. The largest RMSEs occurred in the 5 common item condition where the RPA method (0.0828) was higher than the SL method (0.0739). Although not as dramatically as for transformation constant *A*, the boxplots in Figures 8-11 show that the variability of the distributions for the *B* transformation constant is much greater for the 5 common-item condition than the 10 or 15-common item conditions.

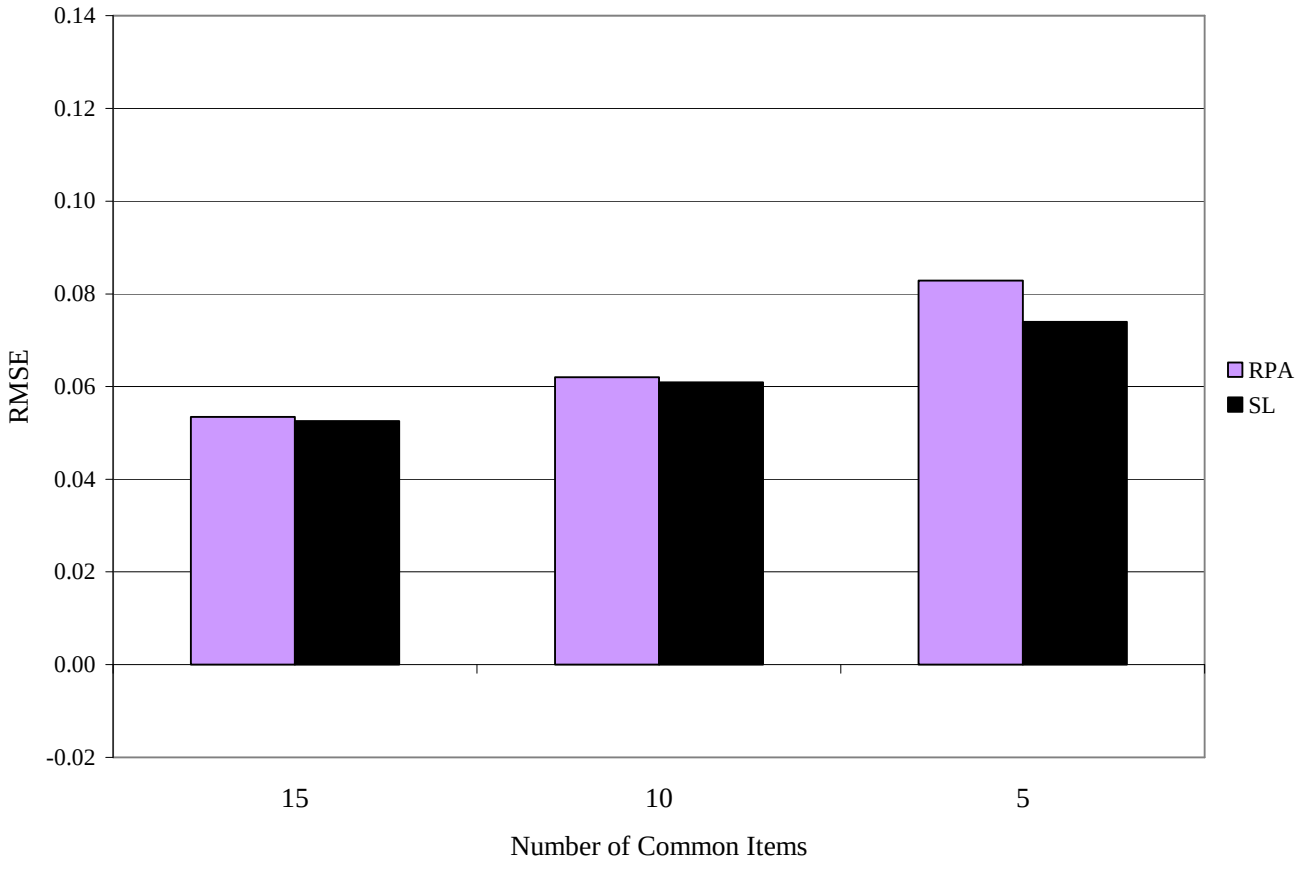

Figure 19. RMSE for Transformation Constant *B* for 15 Common Items, 10 Common Items, and 5 Common Items

Ability Differences

Four different reference groups were used in this study for comparison with the equating group that was distributed N(0, 1): R1~N(-0.5, 0.8^2), R2~N(-0.5, 1.25^2), R3~*N*(-1.0, 0.8^2), R4~*N*(-1.0, 1.25^2). All reference groups produced small, positive amounts of bias for the *A* transformation constant. Figure 20 shows that the amount of bias for all reference groups was similar for the RPA method and the SL method. The largest difference was between the RPA method (0.0274) and the SL method (0.0318) for reference group R3.

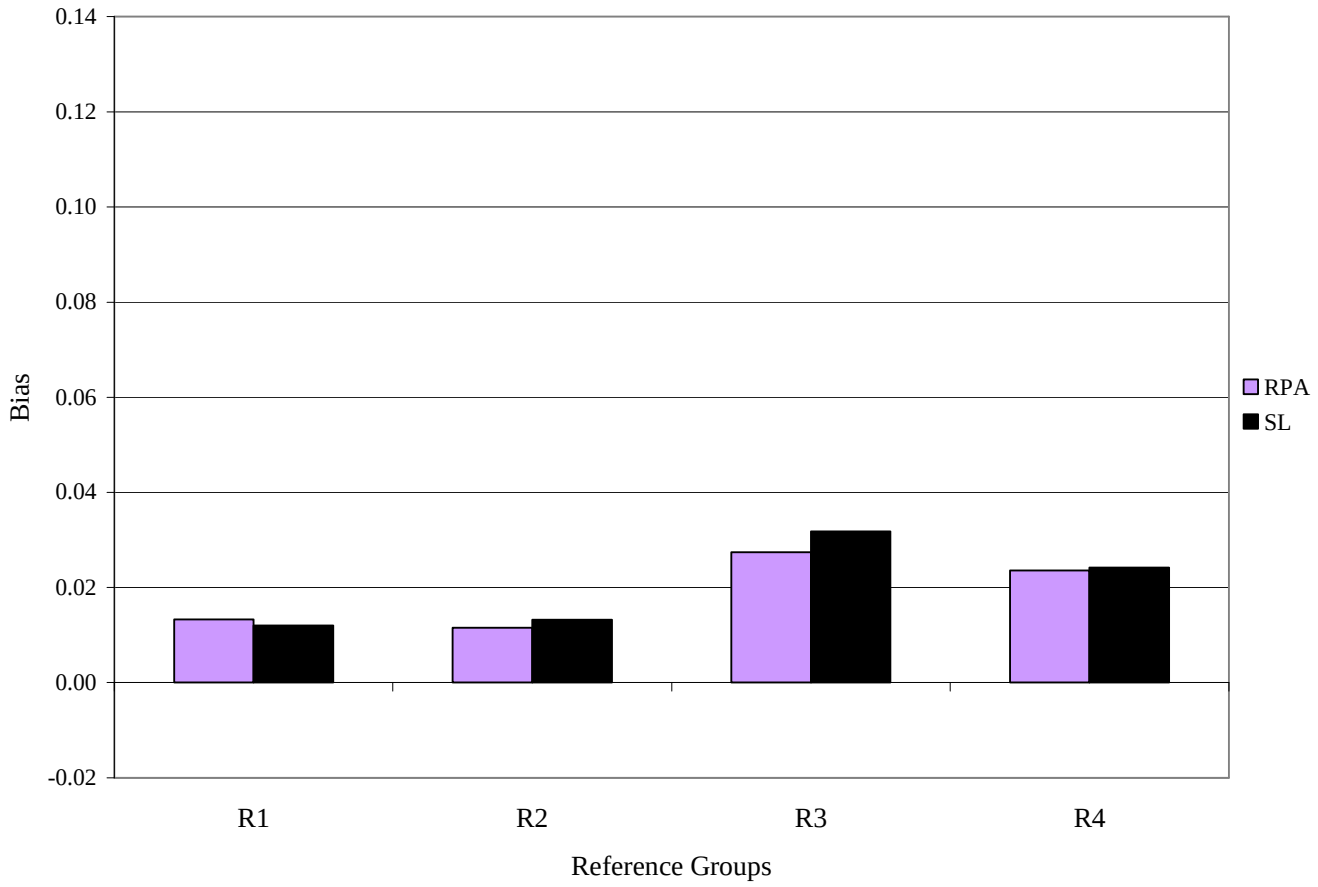

Figure 20. Bias for Transformation Constant *A* for Ability Difference Groups R1, R2, R3, and R4

The RMSE for the two transformation methods were also very similar for the RPA method and the SL method for each reference group. However, more variability was seen when comparing the different reference groups with each other, as shown in Figures 4-7. Reference groups R1 and R3 had higher RMSEs than reference groups R2 and R4. The R3 reference group had the greatest RMSE for the RPA method (0.0959) and the SL method (0.0949). The boxplots in Figures 4-7 show the differences in variability between the reference groups. Groups R1, and R3 are more variable than groups R2 and R4.

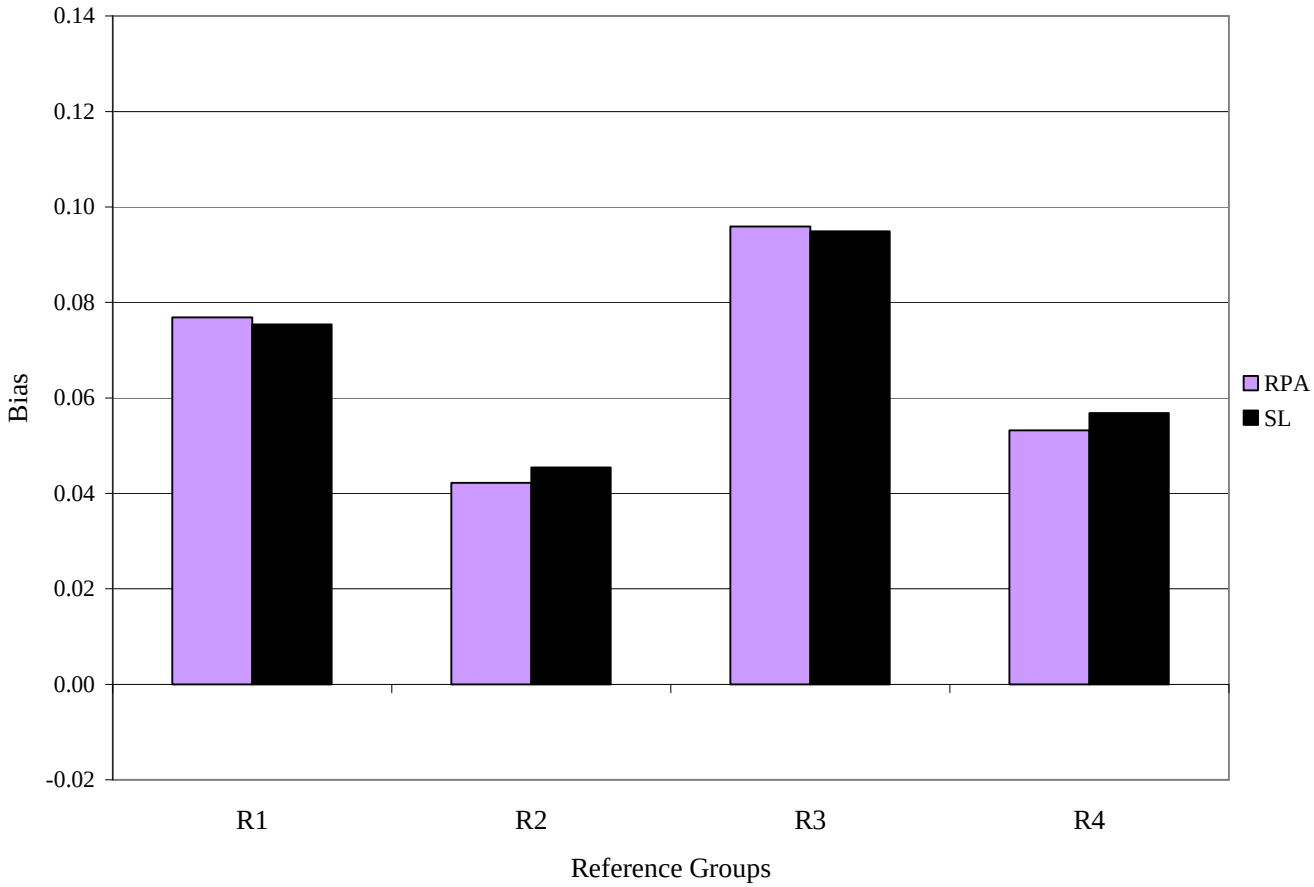

Figure 21. RMSE for Transformation Constant *A* for Ability Difference Groups R1, R2, R3, and R4

Figure 22 shows the difference between the RPA method and the SL method for the different reference groups that were used for transformation constant *B*. The methods performed comparably for all the reference groups. Reference groups R1 and R3 produced small, positive amounts of bias, and reference groups R2 and R4 produced small negative bias of somewhat smaller magnitude. For all groups, the amount of bias was quite small, ranging from -0.0013 for the RPA method in reference group R4 to the largest of 0.0076 for the SL method in reference group R3.

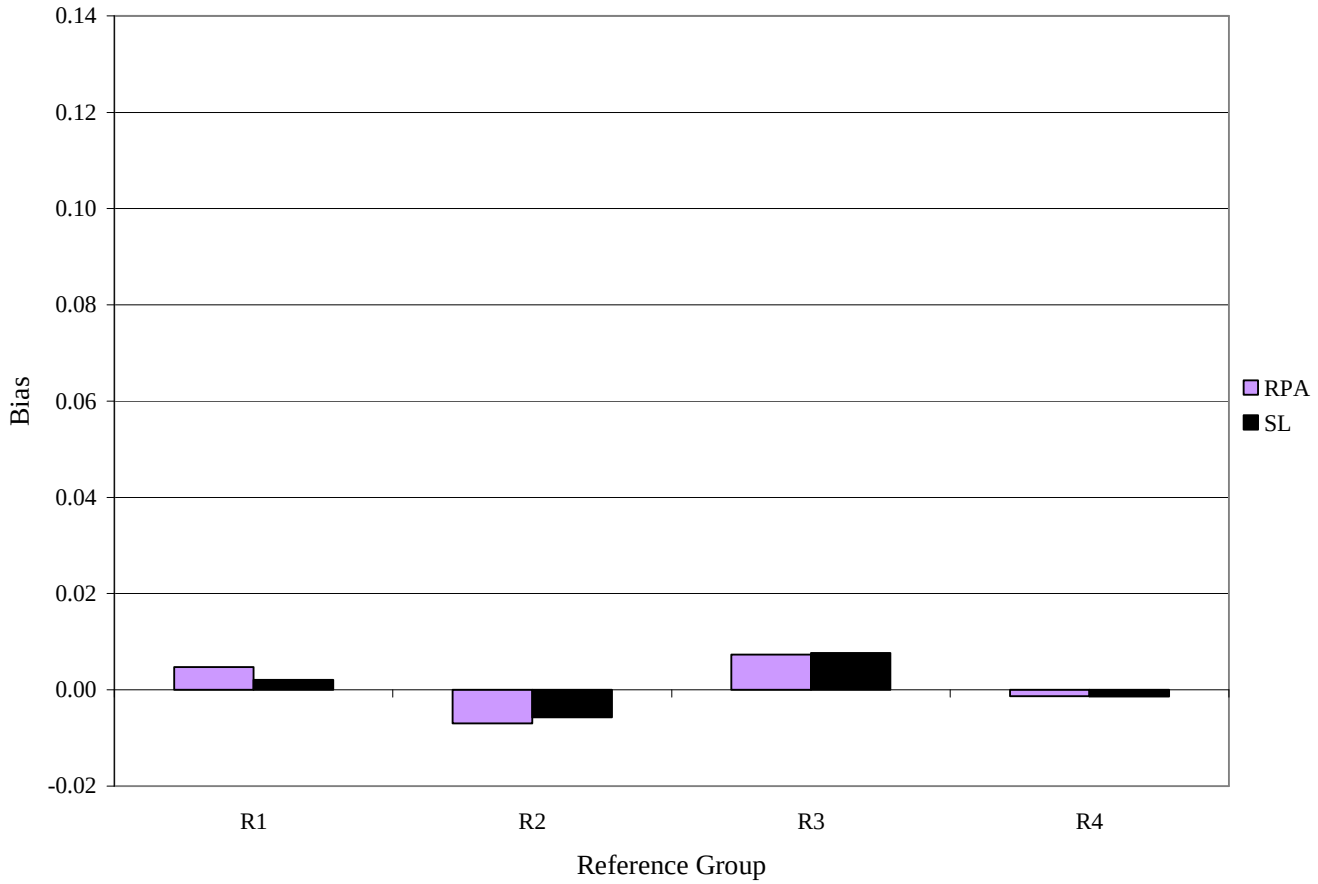

Figure 22. Bias for Transformation Constant *B* for Ability Difference Groups R1, R2, R3, and R4

Figure 23 shows the RMSE for the *B* transformation constant. Like the RMSE for transformation constant *A*, the methods performed comparably for all the reference groups and the RMSE was somewhat greater for the R1 and R3 reference groups than the R2 and R4 reference groups. The largest RMSE, for reference group R3, was 0.0922 for the RPA method and 0.0826 for the SL method.

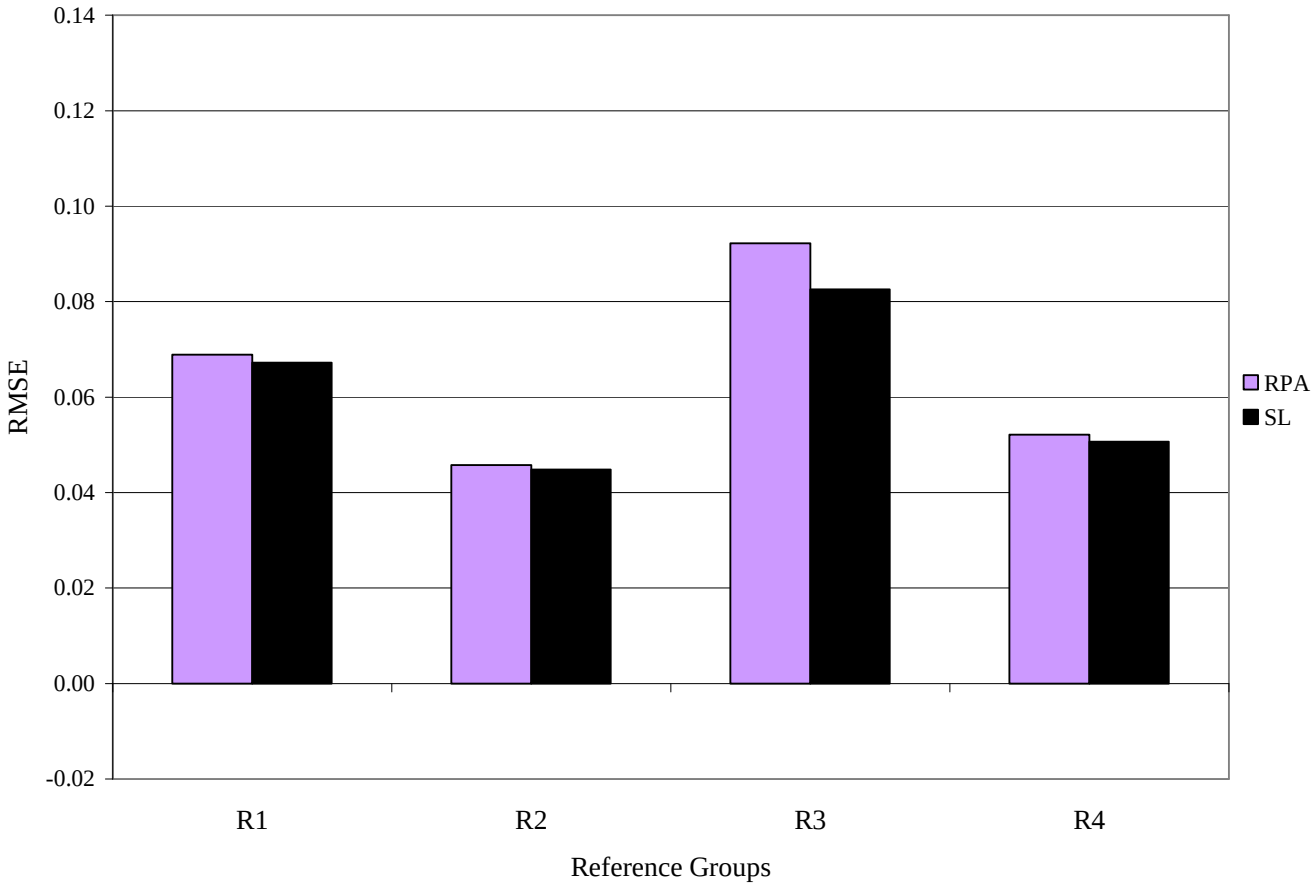

Figure 23. RMSE for Transformation Constant B for Ability Difference Groups R1, R2, R3, and R4

Test Length x Number of Common Items

Figures 24 and 25 show the interaction of test length by number-of-common-items for the *A* transformation constant. In Figure 24, the RPA method generally produced similar or slightly larger amounts of bias at nearly all levels of the common items. However, in the 30-item test, the RPA method (0.0056) produced less bias in the 5 common-item condition than the SL method (0.0171). The largest amount of bias occurred in the longer test with 5 common items for both the RPA method (0.0350) and the SL method (0.0347). Figures 4-6 show a similar pattern with medians for 15 and 10 common items much more similar than for the 5 common items on the 30-item test.

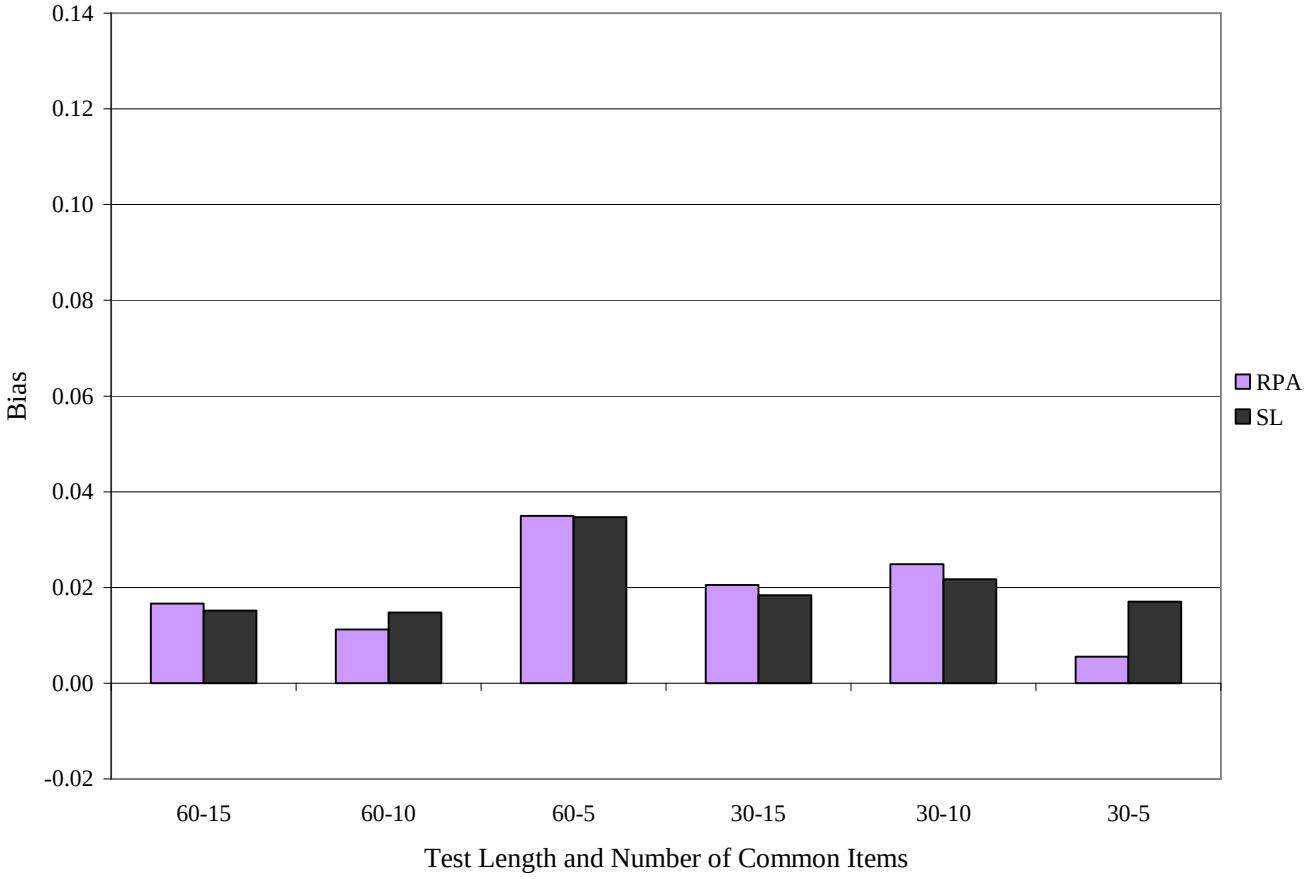

Figure 24. Bias for Transformation Constant *A* for the 60-Item Test and the 30-Item Test and 15 Common Items, 10 Common Items, and 5 Common Items

Figure 25 shows the RMSE for the *A* transformation constant. The RPA method produced almost exactly the same amount of RMSE as the SL method for these factors. The RMSEs decreased as the number of common items increased, and there was very little difference between the longer test and the shorter test. The lowest RMSE was for the RPA method (0.0488) for the longer test and 15 common items and the highest RMSE was for the RPA method (0.1005) for the longer test and 5 common items. The boxplots in Figures 4-7 also illustrate this trend: little difference between test lengths, but increasing variability as fewer common items are used.

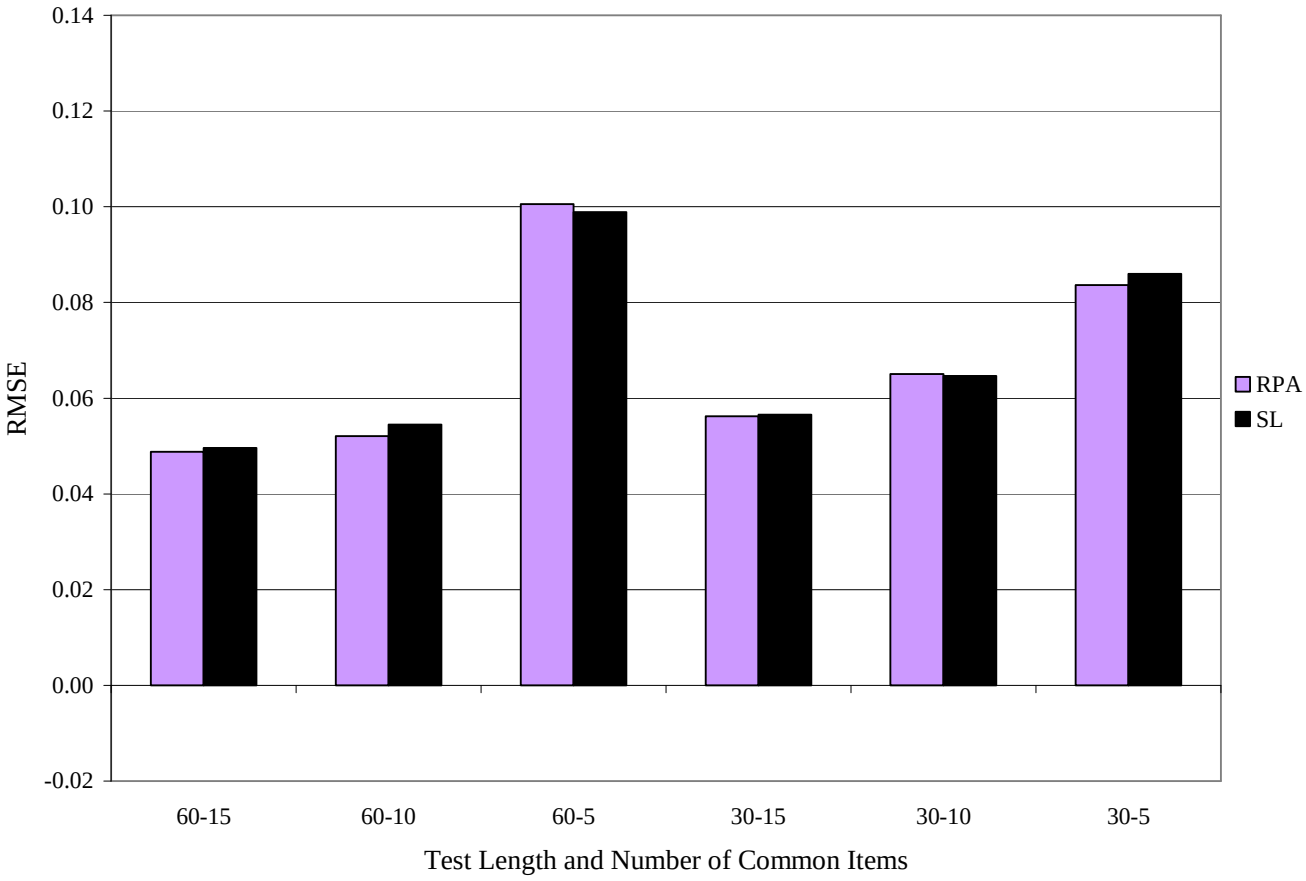

Figure 25. RMSE for Transformation Constant *A* for the 60-Item Test and the 30-Item Test and 15 Common Items, 10 Common Items, and 5 Common Items

Figure 26 presents the bias for the transformation constant *B*, which was similar for both methods in all conditions. The bias was smaller for the 60-item test for both the RPA method and the SL methods for 15 and 5 common items groups and increased when 10 common items were used. For each of these common-item groups the bias was negative. The 30-item test showed about the same amount of positive bias for the 15 and 10 common-item conditions and was smaller when only 5 common items were used. The boxplots in Figures 8-11 also show how similar the distributions of sample transformation constant *B* were, with the slightly negative bias for the longer test.

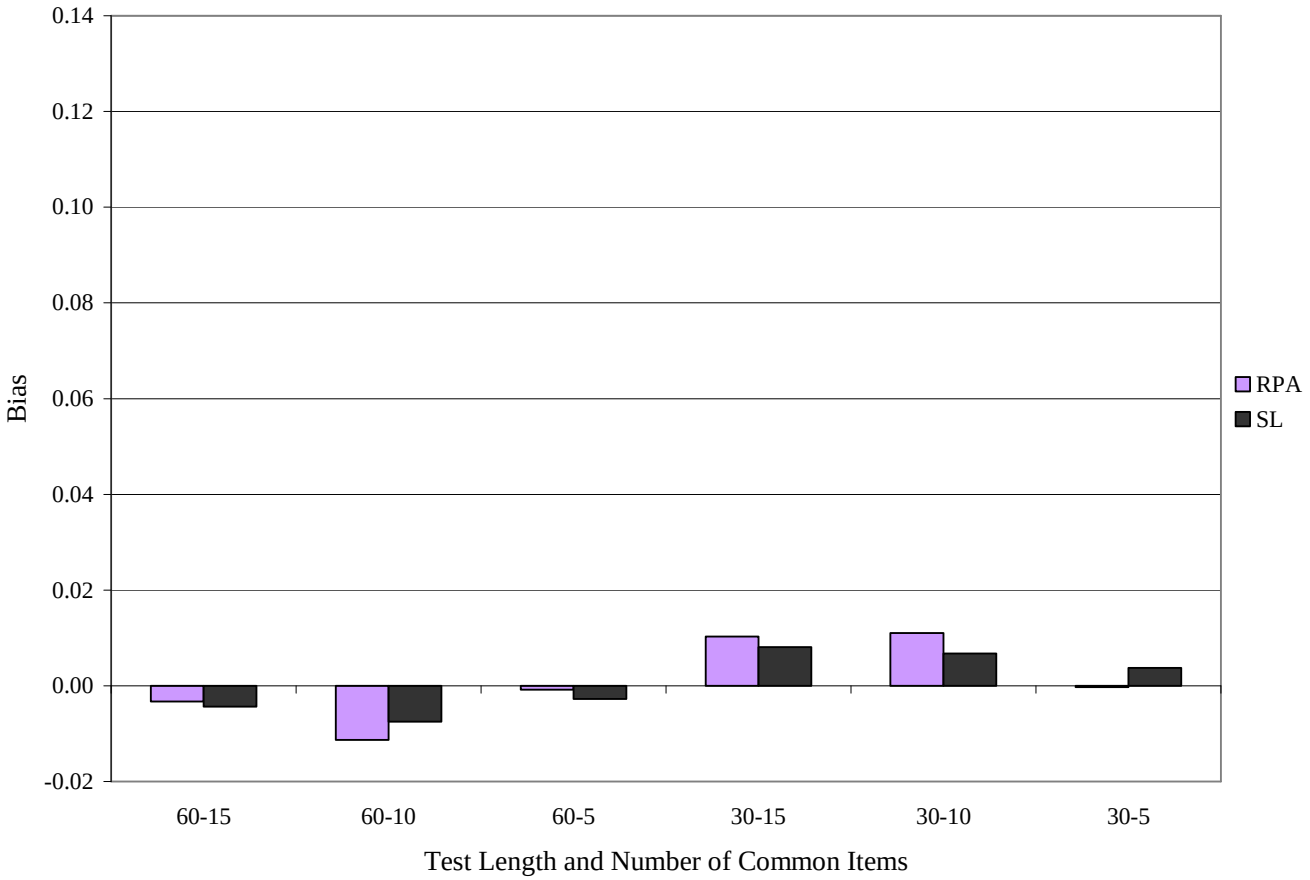

Figure 26. Bias for Transformation Constant *B* for the 60-Item Test and the 30-Item Test and 15 Common Items, 10 Common Items, and 5 Common Items

Figure 27 shows the RMSE by test length and the number of common items for transformation constant *B*. No substantial differences were present between the longer test and the shorter test. When only 5 common items were used, the RMSE was higher for both test lengths.The RPA method and the SL method produced similar amounts of RMSE for all of the conditions, although the RPA method produced slightly more RMSE than the SL method when 5 common items were used.

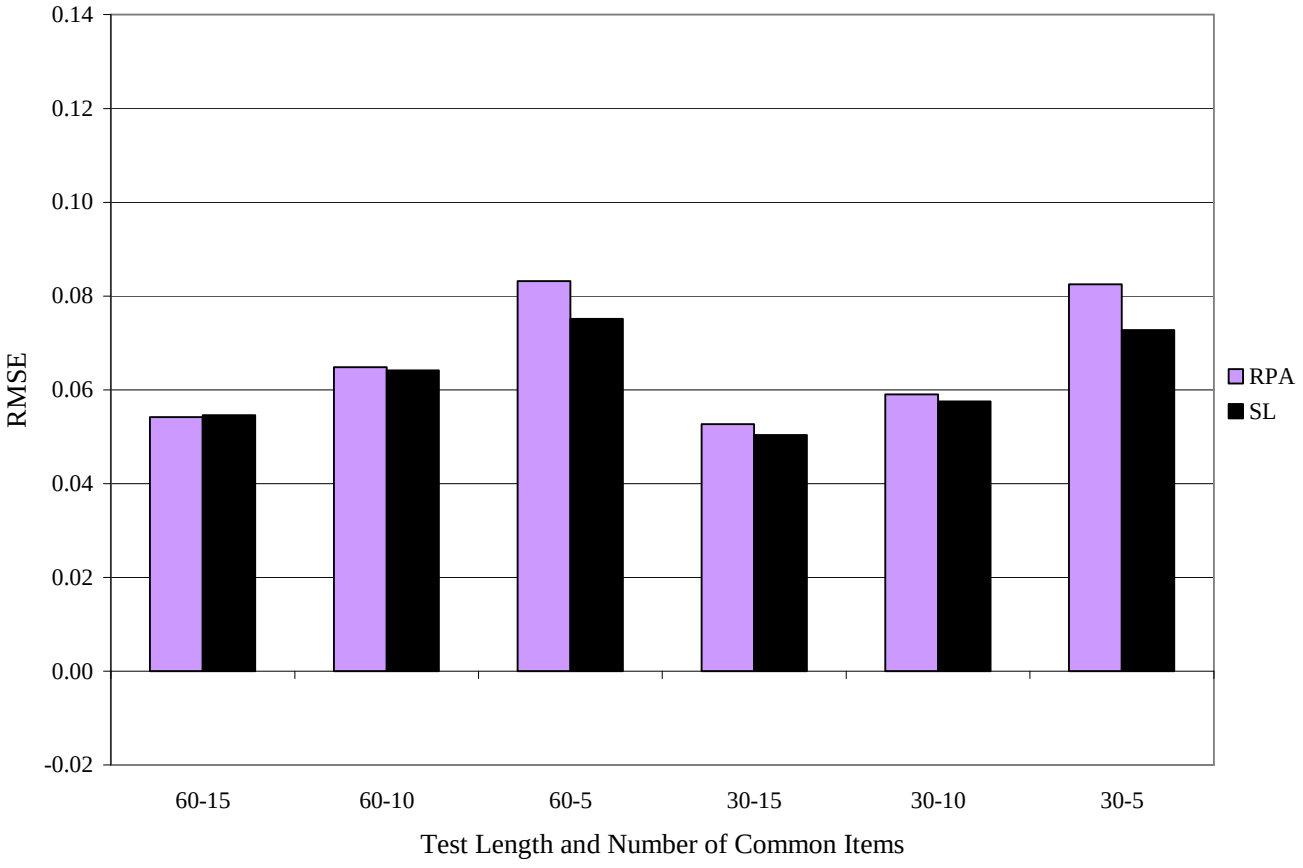

Figure 27. RMSE for Transformation Constant *B* for the 60-Item Test and the 30-Item Test and 15 Common Items, 10 Common Items, and 5 Common Items

Test Length x Ability Differences

The bias produced by the RPA method and SL method was similar for transformation constant *A* for most of the conditions when averaged over the number of common items. Figure 28 shows that the bias for the 60-item test was less than the bias for the 30-item test for all reference groups except reference group R3. R2 showed the least amount of bias overall. For all groups and both test lengths, the RPA and SL methods showed similar amounts of bias, although bias for the R3 reference group on both test lengths was slightly smaller for the RPA method than the SL method.

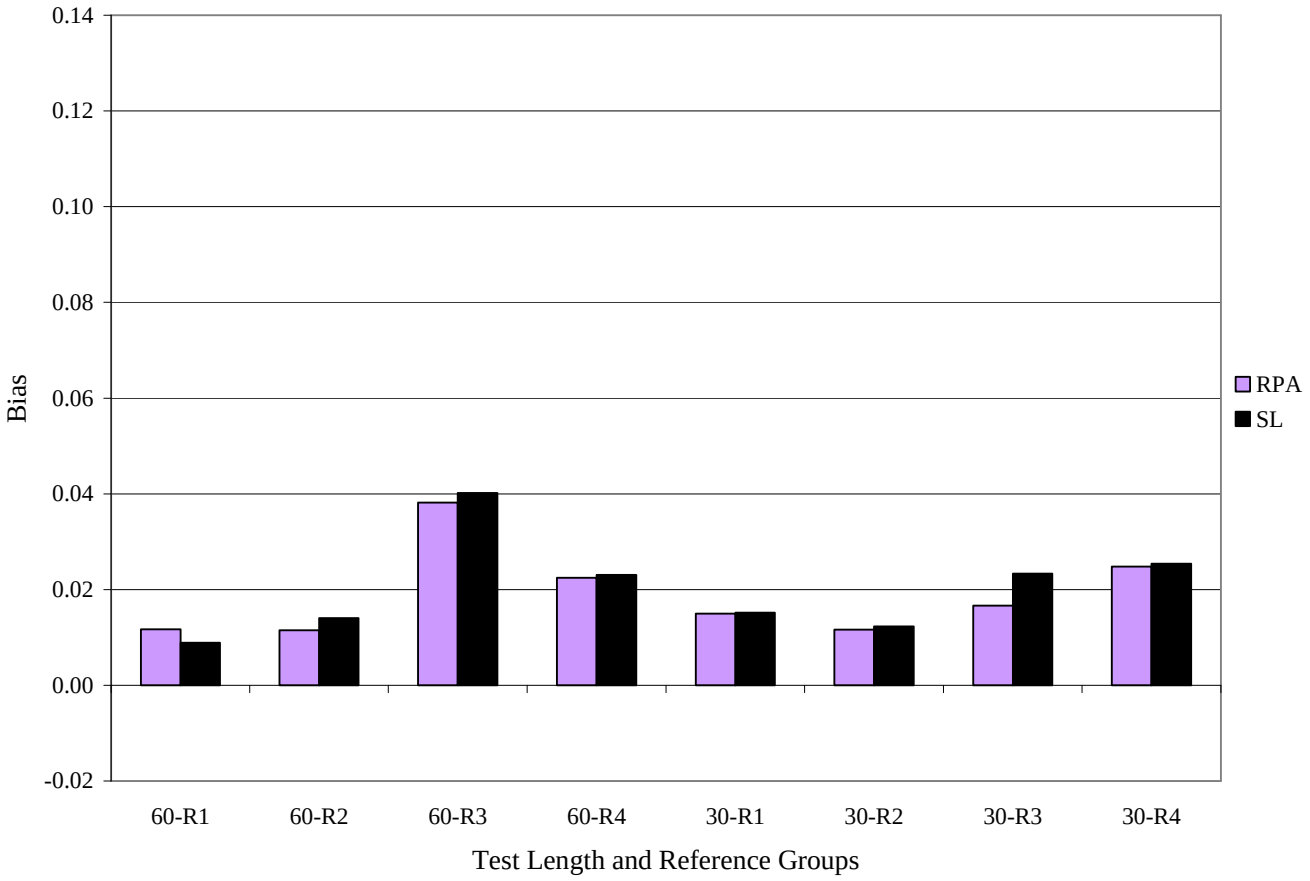

Figure 28. Bias for Transformation Constant *A* for the 60-Item Test and the 30-Item Test and Reference groups R1, R2, R3, and R4

The RMSE was quite similar for the RPA method and SL method for all conditions as seen in Figure 29. The test length did not affect the RMSE as strongly as did the reference group. Reference group R2 had the smallest RMSE for both test lengths, but with no noticeable differences between test lengths. The R3 reference group showed the most RMSE in the 60-item test for both the RPA method (0.1006) and the SL method (0.0975). The boxplots in Figure 4 and 6 illustrate the high level of variability seen in reference groups R1 and R3, with values ranging from 1.0 to 1.5 and 1.0 to 1.6, respectively. Reference group R2 only ranged from 0.65 to 0.95 for its transformation constant *A*.

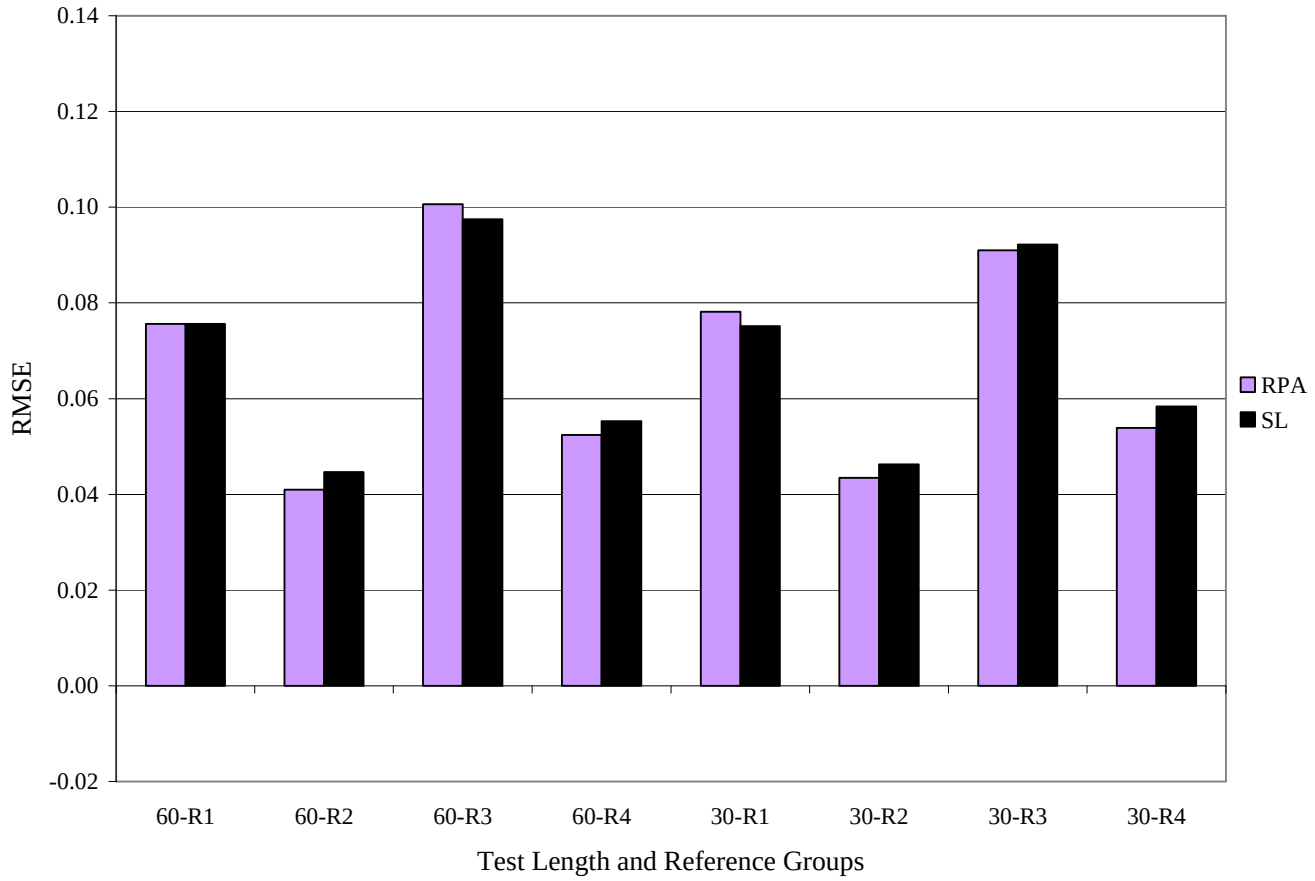

Figure 29. RMSE for Transformation Constant *A* for the 60-Item Test and the 30-Item Test and Reference Groups R1, R2, R3, and R4

Figure 30 shows that the bias for transformation constant *B* was low for all the reference groups. The bias was extremely small for both methods and occurred in both positive and negative directions. The largest amount of bias overall was produced in the 30-item test for reference groups R1 and R3, but reference group R2 had a relatively large amount of bias on the longer test. The small bias may be noted in Figures 8-11 for transformation constant B by observing how closely the distributions center around their true values.

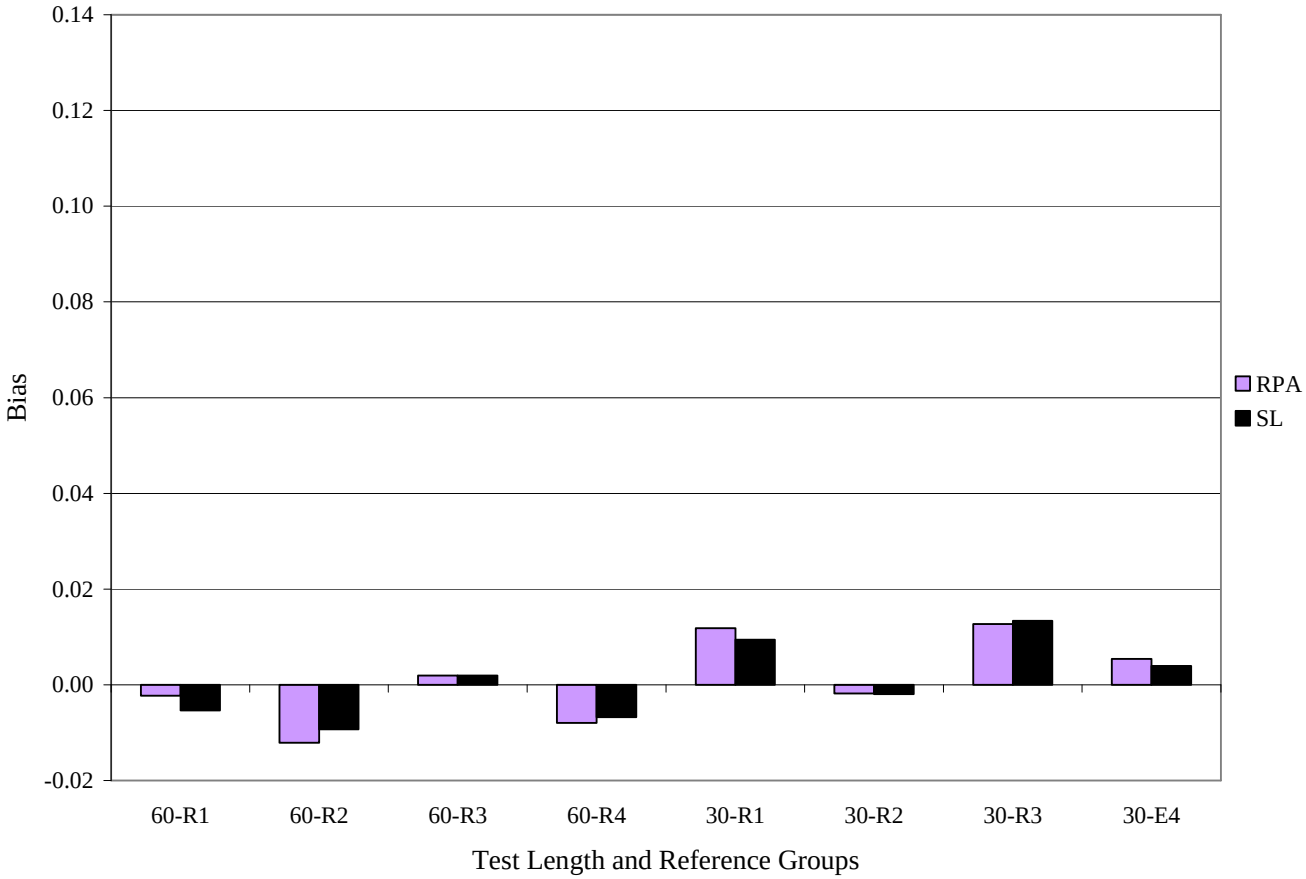

Figure 30. Bias for Transformation Constant *B* for the 60-Item Test and the 30-Item Test and Reference Groups R1, R2, R3, and R4

Figure 31 shows that the RPA method and SL method performed comparably in each of the simulation conditions in the amount of RMSE produced for transformation constant *B*, although the RPA method was higher then the SL method for reference group R3 for both test lengths. Reference groups R2 and R4 produced the smallest amount of RMSE, with little difference between the longer and shorter tests.

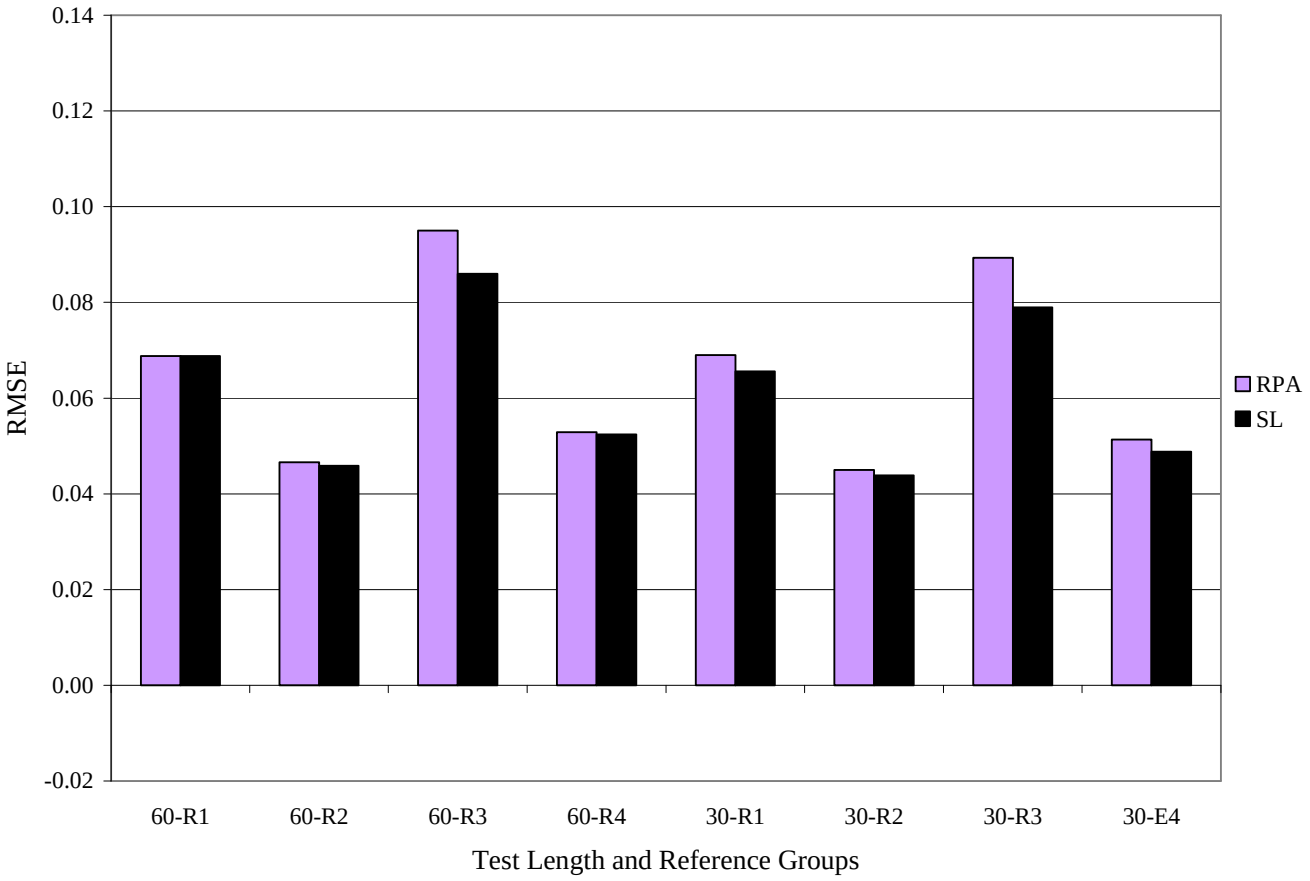

Figure 31. RMSE for Transformation Constant *B* for the 60-Item Test and the 30-Item Test and Reference Groups R1, R2, R3, and R4

Number of Common Items x Ability Differences

Figure 32 shows the differences between the RPA and SL methods for the *A* transformation constant when averaged over test length. For nearly all reference groups, the two methods performed similarly, and most conditions resulted in small amounts of bias. Increasing the number of common items did slightly decrease bias for reference groups R2 and R4, but for R3 the greatest bias occurred in the 5 common-item condition for the SL method.

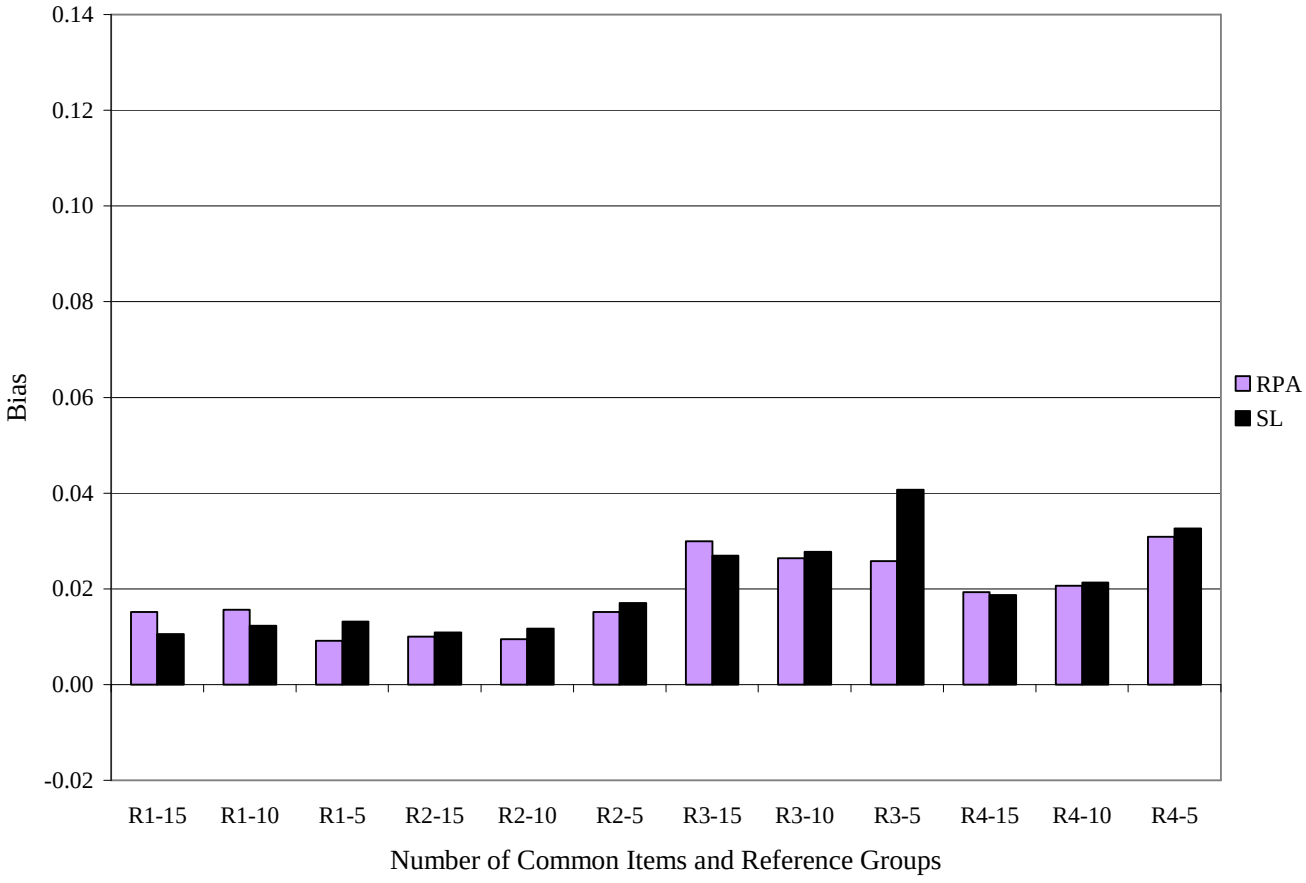

Figure 32. Bias for Transformation Constant *A* for the 15 Common Item, 10 Common Item and 5 Common-Item Test for Ability Difference Groups R1, R2, R3, and R4

Figure 33 shows the RMSE for the RPA and SL methods for the *A* transformation constant when averaged over test length. The two methods performed similarly, and the magnitude of the RMSE was relatively small for most conditions. However, increasing the number of common items did decrease the RMSE across all reference groups. Reference groups R2 and R4 produced less RMSE than reference groups R1 and R3. This can also be seen in the boxplots in Figures 4-7. The variability generally decreases as number of common items increases, and in a similar pattern for each of the reference groups.

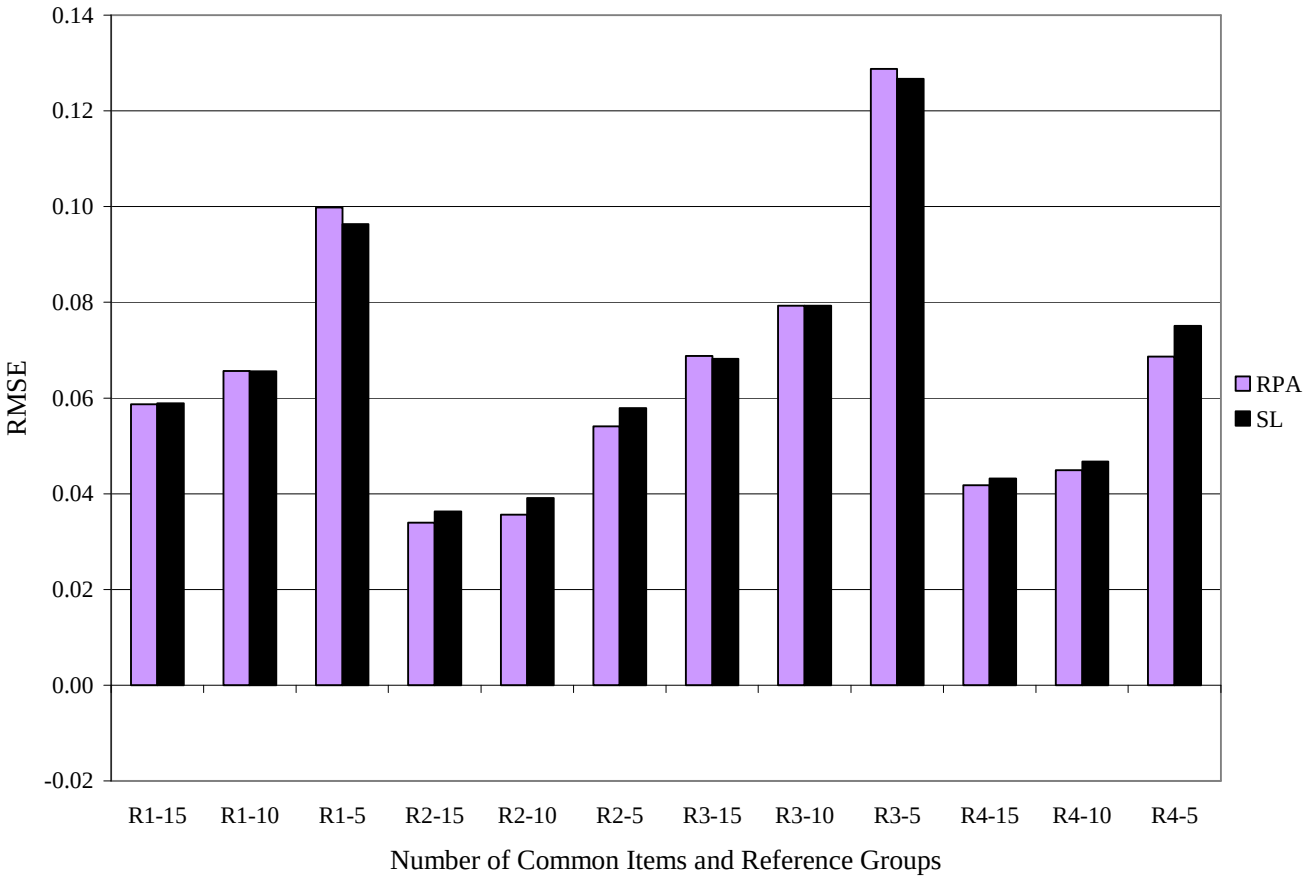

Figure 33. RMSE for Transformation Constant *A* for the 15 common Item, 10 Common Item and 5 Common-Item Test for Ability Difference Groups R1, R2, R3, and R4

Figure 34 shows the differences in bias between the RPA and SL methods for the *B* transformation constant when averaged over test length. For all reference groups, very small amounts of bias resulted. Reference groups R2 and R3 produced more bias for their common-item groups in opposite directions, while groups R1 and R4 produced less bias for their groups, but again in opposite directions. The number of common items did not strongly impact the bias of the transformation constant estimates.

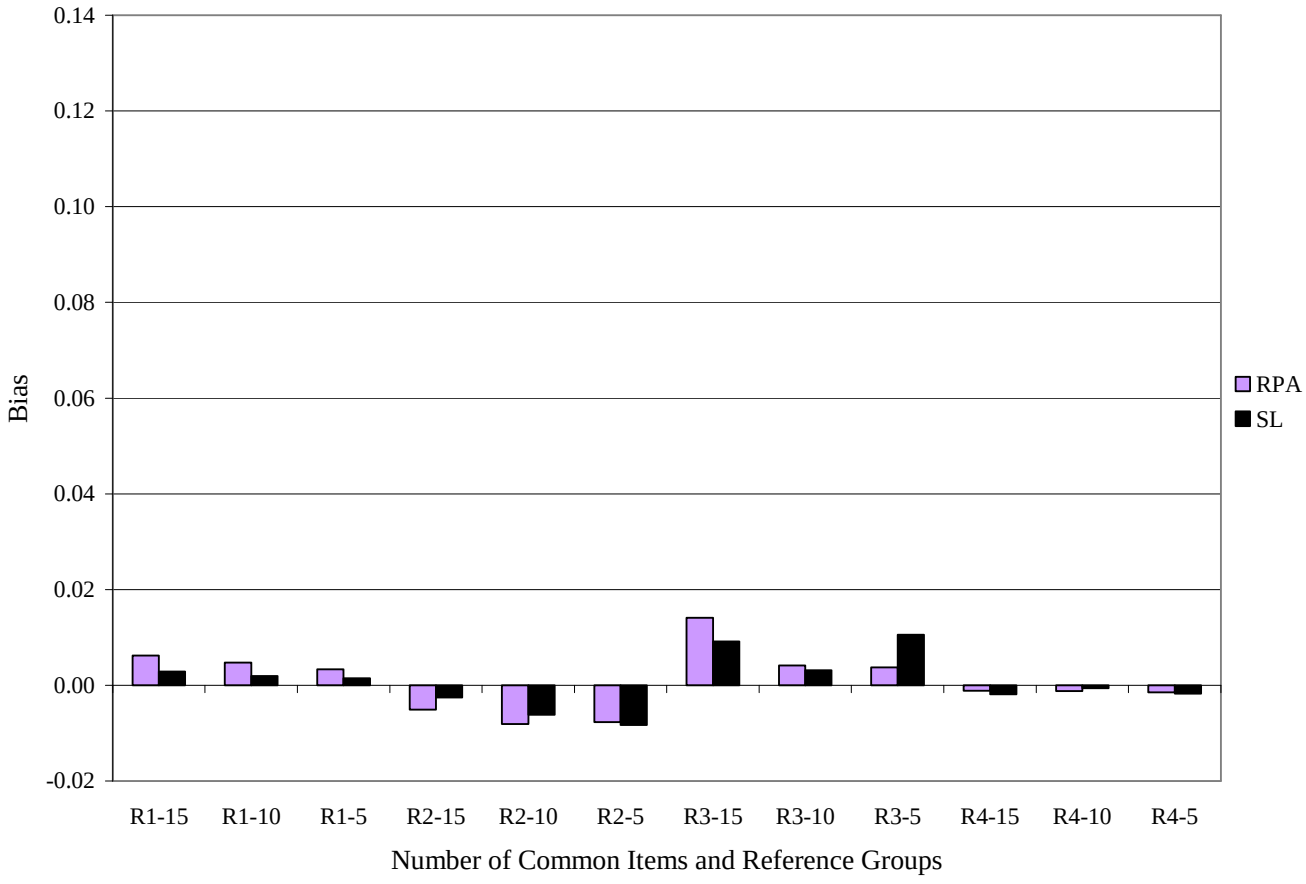

Figure 34. Bias for Transformation Constant *B* for 15 Common Items, 10 Common Items and 5 Common Items and Reference Groups R1, R2, R3, and R4

Figure 35 shows the differences between the RPA and SL methods for the *B* transformation constant when RMSE was averaged over test length. For most reference groups, the two methods performed similarly. The R2 and R4 reference groups yielded lower amounts of RMSE than R1 and R3. Increasing the number of common items did decrease the RMSE for the both methods in all reference groups. The largest difference between the RPA method (0.1172) and the SL method (0.0972) occurred when only 5 common items were used in reference group R3.

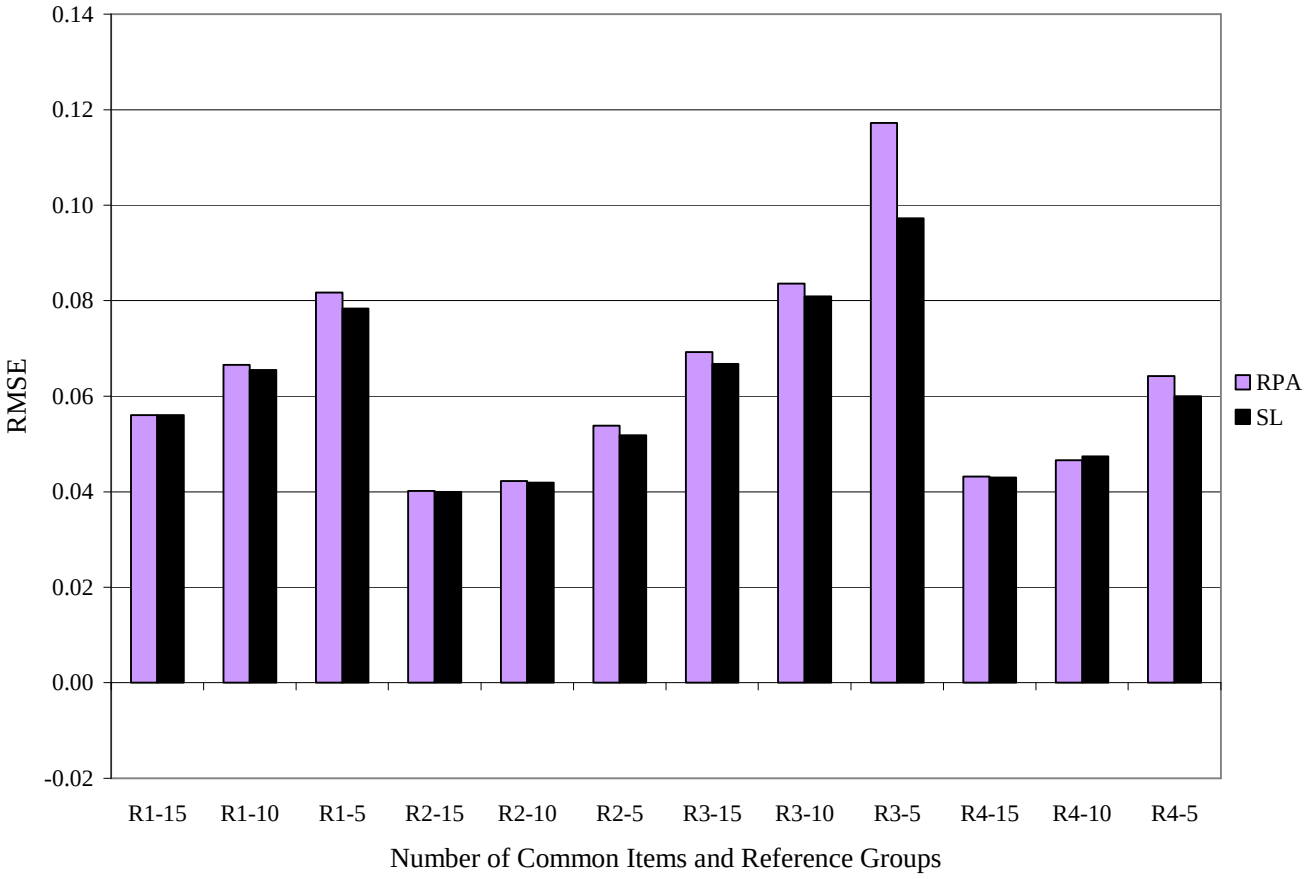

Figure 35. RMSE for Transformation Constant *B* for 15 Common Items, 10 Common Items and 5 Common Items and Reference Groups R1, R2, R3, and R4

Study 2: Student Data

Research Question 4 focused on how the transformation constants of the RPA method and the SL method would compare when preparing to equate actual student data. The actual student data most closely resembles that of reference group R1, which was distributed R1~$N(-0.5, 0.8^2)$ with an equating group distributed $N(0.0, 1.0^2)$. Over previous administrations, the comparable student reference and equating groups tended towards similar distributional differences.

The A and B transformation constants were calculated using the Excel macros used for the 100 replications in the simulation study. But for the real student data, the item parameters were only calibrated once for the reference group and once for the equating group. For this calibration, the same common items used in the simulation study were used in the student data study.

Because the true values of the transformation constants were unknown, the *A* and *B* estimates were compared to each other instead of to the true values. Instead of bias, the average difference between each pair of methods was calculated. These values are presented in Table 11.

Table 11. RPA and SL Transformation Constants for Actual Student Data for 15, 10, and 5 Common Items

	RPA		SL		Difference	
	A	*B*	*A*	*B*	*A*	*B*
15	0.9789	-0.3670	0.9699	-0.3615	0.0090	-0.0055
10	0.9204	-0.4500	0.9120	-0.4370	0.0084	-0.0129
5	0.6800	-0.4129	0.8475	-0.5104	-0.1675	0.0978
			Average Difference		-0.0500	0.0264

Transformation Constant *A* Comparison with Actual Data

Similar to the findings in Study 1, the transformation constant *A* was comparable when generated using either the RPA method or the SL method. The resulting constants were most similar when 15 or 10 common items were used. In these cases, the absolute difference between them was quite small (<0.010). Even when averaging all three levels of common items, the average difference was small (-0.0500). However, when examining only the 5 common-item condition, the differences between the RPA and SL method were much larger: the RPA method *A* constant (0.6900) was much smaller than the SL method *A* constant (0.8475), resulting in a much higher absolute difference (-0.1675).

Transformation Constant *B* Comparison with Actual Data

Similar to transformation constant *A*, the transformation constant *B* was comparable when generated using either the RPA method or the SL method. Again, the resulting constants were most similar when 15 or 10 common items were used. In these cases, the absolute difference between them was quite small as well (<0.015), and only slightly larger than that seen for transformation constant *A*. In addition, when averaging the three levels of common items, the average difference was small (0.0264), smaller than the *A* transformation constant's average difference. For the *B* transformation constant as well, the 5 common-item condition, resulted in much larger differences between the RPA method and the SL method.

Resampling Analysis

As described in Chapter 3, a resampling procedure was conducted using the student data. This procedure was used to estimate the sampling distribution of the transformation constants of the observed student data. SAS statistical software was used to generate samples

of 5, 10, or 15 common items from the 30-item test with no replication of an item within a sample. Items were, however, replicated across samples. All item parameters were those that had been originally estimated from the test. 100 replications were conducted using SAS proc surveyselect. The item parameters were only estimated once, but different sets of common items were sampled for each replication. The item parameters *a*, *b*, and *c* were placed in the Excel spreadsheet developed by Armstrong to calculate the RPA and SL *A* and *B* transformation constants. The means and standard deviations for the 100 replications are presented in Table 12.

Table 12. RPA and SL Transformation Constants for Resampled Student Data for 15 Common Items, 10 Common Items, and 5 Common Items

		RPA		SL		
		Mean	Std Dev	Mean	Std Dev	Difference of Means
Transformation Constant *A*	15	0.9257	0.0628	0.9387	0.0505	-0.0130
	10	0.9122	0.0916	0.9327	0.0731	-0.0206
	5	0.9614	0.1587	0.9598	0.1096	0.0016
Transformation Constant *B*	15	-0.4068	0.0589	-0.4136	0.0598	0.0068
	10	-0.4128	0.0875	-0.4240	0.0900	0.0113
	5	-0.4264	0.1586	-0.4243	0.1349	-0.0021

Transformation Constant *A* in Resampling Analysis

Overall, the means for transformation constant *A* ranged from the RPA constant for 10 common items (0.9122) to the RPA constant for 5 common items (0.9614). The RPA

constants were similar to the SL transformation constants for all common-item sets, with the largest difference occurring for 10 common items (0.0206).

The means for the transformation constants produced by the RPA method were more variable than those produced by the SL method. The range between common-item groups for the RPA method was 0.05, but for the SL method it was 0.03. Transformation constant *A* was slightly lower for the RPA method than the SL method for the 15 and 10 common-item groups, which is not what was found in the single student data study. It was about the same for the RPA method as the SL method for the 5 common-item group, and the mean was smaller for the RPA method than the SL method for the 5 common-item group in the student study.

The standard deviations overall, were much more varied than the means. They ranged from 0.05 for the SL method for 15 common items to 0.16 for the RPA method for 5 common items. The trend for both methods was for the standard deviation to increase as the number of common items decreased. The standard deviations from the resampling replications were smallest in the 15 common-item group for both the RPA and SL methods for transformation constant *A*. However, the RPA method produced larger standard deviations than the SL method for transformation constant *A* for all levels of common items, with especially larger standard deviation when only 5 common items were used.

Transformation Constant *B* in Resampling Analysis

The general range for transformation constant *B* was from the low RPA constant for 5 common items (-0.4264) to the high RPA constant for 15 common items (-0.4068). The RPA constants were similar to the SL transformation constants for all common-item sets, and the largest difference between them occurred when 10 common items were used.

The means for the transformation constant *B* produced by the RPA method were slightly more variable than those produced by the SL method. The difference between common-item groups for the RPA method was 0.02, but for the SL method it was 0.01. Transformation constant *B* was slightly lower in absolute value for the RPA method than the SL method for the 15 and 10 common-item groups, which is not what was found in the single student data study. It was about the same for the RPA and SL methods for the 5 common-item group, and the mean was smaller for the RPA method than the SL method for the 5 common-item group in the student study.

The standard deviations overall, were also much more varied than the means for transformation constant *B*. They ranged from 0.06 for the RPA and SL methods for 15 common items to 0.16 for the RPA method for 5 common items. Like transformation constant *A*, the standard deviation increased as the number of common items decreased. The standard deviations from the resampling replications were smallest in the 15 common-item group for both the RPA and SL methods for transformation constant *B*. However, the RPA method produced larger standard deviations than the SL method for transformation constant *B* for the 10 common-items and 5 common-item conditions.

CHAPTER V

Discussion

The purpose of this study was to determine whether the RPA method of generating transformation constants produced suitable constants for use in scaling IRT parameter estimates prior to conducting equating. In particular, this research explored whether several conditions (test length, number of common items, and ability differences) would affect the accuracy of the resulting transformation constants. Bias and RMSE were used to evaluate accuracy. Additionally, the RPA method was compared to the SL method on all these factors to see how its accuracy compared to a widely-accepted scaling method.

Bias and RMSE were defined in Chapter III, and figures were presented in Chapter IV illustrating the differences in the amounts of bias and RMSE under the study conditions. It is now important in this chapter to address what these differences mean in terms of the resulting transformation constants. Transformation constants with large amounts of bias will result in inaccurate estimates of the transformation process. In practice, only one transformation constant estimate is calculated, and its difference from the true value is unknown. The distribution of estimated values for bias collected in the 100 replications of the simulation study provides a range from within which the single estimator would be expected to occur. Any single estimate may be higher or lower than the true value, and since the values in the 100 replications were averaged, many differences between those values can cancel each other out. For this reason, a small amount of bias is not sufficient by itself to demonstrate the quality of an estimate.

RMSE was also used in this research because it indicates how widely the estimates vary within the 100 replications. Because the RMSE is calculated with a value that has been

squared, it is always positive. When bias and RMSE are analyzed together, a better prediction of the behavior of the estimate can be made. Transformation constants with small amounts of bias and large RMSEs would be more unstable than those with small amounts of bias and small RMSEs. The resulting equated scores would be different depending on each estimate of the transformation constant, and it would be difficult to have confidence in the accuracy of the results.

The comparisons of bias and RMSE were made through two separate studies. The first was a simulation study, and the second used actual student data. The results of the studies will be discussed separately in the following sections, followed by a general discussion. Finally, this paper will conclude with the limitations of this research and recommendations for future research.

Study 1: Simulation Study

The intent of Study 1 was to answer the research questions through computer-simulated data. The first research question sought to determine how test length affected the accuracy of the transformation constants of the RPA method and the SL method. As expected, the differences in test length were not large in this study for either transformation constant, *A* and *B*. The longer test (60 items) resulted in slightly more bias and RMSE for both the RPA method and the SL method in terms of bias and RMSE for transformation constant *A*, and slightly less bias for transformation constant *B*. However, the amounts for both methods were extremely small. While the shorter test (30 items) resulted in less RMSE and bias for both the RPA method and the SL method for transformation constant *A* and slightly less RMSE and bias for transformation constant *B*, again, in absolute terms, the differences were negligible. Therefore, from the results of this study, it is reasonable to expect that with a minimum

number of 30 items, accurate transformation constants can be calculated by either the RPA method or the SL method.

Research Question 2 focused on how the number of common items affected the accuracy of the transformation constants of the RPA method compared to the SL method. Although Fitzpatrick (2008) noted a lack of stability when fewer than 15 common items were used, that was not found in this study. In most of the conditions in this simulation, 10 common items were sufficient for producing reasonably accurate transformation constants. However, with only 5 common items the RMSE increased and does align with the Fitzpatrick (2008) findings. Very little difference in either bias or RMSE was seen when comparing the RPA method with the SL method for both transformation constants.

Although Gao, Zhu, Chen, and Harris (2008) found common-item sets of 5, 10, and 15 items for their 30-item test provided adequate equating functions, they demonstrated that using more items produced better equating results. That finding was not strictly replicated in this study. For transformation constant *A*, the amount of bias was about the same for the 5, 10, and 15 common-items for transformation constant *A*, and the RMSE increased in the expected manner, i.e., decreasing as the number of common items increased. For transformation constant *B*, the amount of bias was slightly larger for the 15 common-item condition than the other two. This unusual result must be carefully considered within the framework of how very small the biases were at all levels of common items. The largest bias for the *B* transformation constant was only 0.0292 for the RPA method and 0.0216 for the SL method. In the case where the true value of transformation constant B is 1.25, a bias of 0.02 is less than 2 percent higher than the true value. Given the small values of bias, RMSE is a more meaningful indicator of accuracy here. However, the RMSE for the *B* transformation constant

did increase as the number of common items decreased, which was consistent with the Gao et al. (2008) findings.

Closer examination of the items selected for the 15 common-item condition and the 10 common-item condition reveals that the range of difficulty parameters was greater in the 15 common-item condition. The range for 15 common items (-2.5370 to 2.1840), 10 common items (-2.5370 to 1.0150) and 5 common items (-2.5370 to 0.6280) represent widely different common items. The items selected for these conditions were intended to be fairly distributed across the test for each of the three common item conditions, yet the pseudo-random selection resulted in a more narrow range for the 10-item condition that may have produced less bias. The simulation study did not exactly replicate the earlier research of Gao et al. (2008), as the 10-common item group did not produce transformation constants greatly different from the 5-common-item group. However, this was most likely due to the selection of one single set of items, and not the repeated sampling done in the resampling analysis. In the actual student data resampling analysis, the addition of more common items did behave in the way described by Gao et al. (2008). In that situation, using more items produced better equating results as both the RPA and SL produced smaller standard deviations when more common items were used.

Sinharay and Holland (2006) demonstrated that the spread of item difficulties could be less extensive than the fully representative set (miditests) that had long been deemed necessary to function sufficiently for equating. The 10 common-item spread of item difficulties (-2.5370 to 1.0150) which was less extensive than the 15 common-item set produced more bias for transformation constant *A*, and the 5 common-item spread of item difficulties (-2.5370 to 0.6280) which was even more restrictive set produced less bias for

transformation constant A. The more restricted range in this study resulted in more variable bias, but the RMSEs were stable and increased in the expected way.

For the interaction between test length and the number of common items, the 10-common-item condition also resulted in slightly more bias for the *A* transformation constant in the shorter test, but not in the longer test. This pattern was not repeated in the RMSE for transformation constant *A* as it was higher for the 5 common-item condition regardless of test length. The RMSE for transformation constant *A* was similar for the shorter and longer tests when 10 or 15 common items were used, but higher RMSE resulted for the longer test when only 5 common items were used. The bias for transformation constant *B*, was largest when 10 common items were used for both the shorter and longer test. For the RMSE on transformation constant B, the values increased as the number of common items decreased, for both test lengths.

The third Research Question, " How will differences in the ability distributions of the populations to be scaled influence the accuracy of the resulting transformation constants?" was studied by using four reference groups [R1~$N(-0.5, 0.8^2)$, R2~$N(-0.5, 1.25^2)$, R3~$N(-1.0, 0.8^2)$, R4~$N(-1.0, 1.25^2)$]. As expected, ability distributions did show an impact on the resulting transformation constants and accuracy of the transformation constants was more favorable when examinee ability distributions are more similar.

Small, positive amounts of bias were produced for all four reference groups for transformation constant *A*, with the smallest produced by R2. The RMSE was slightly larger for reference group R3 which had the smaller variance of 0.8^2 and was one full standard deviation lower than the equating group. The bias for transformation constant *B* was also smaller, however, in the negative direction for reference groups R2 and R4. The pattern of the

RMSEs were similar to those for transformation constant *A*, compared to the less variable groups, R1 and R3.

Mean differences in ability distributions affected the transformation constants less than did the variability of those distributions. As Gialluca et al. (1984) demonstrated in their simulation study, equating ability distributions from comparable samples of similar ability groups resulted in smaller RMSEs than groups with different ability levels. In this study, it is important to define "more comparable samples" for it can imply the group closest in mean ability, or the group with the more similar variance. Reference groups R2 and R4 produced smaller RMSEs, even though reference group R1 was closer in mean ability to the equating group than group R4.

Differences in ability did impact bias and RMSE differentially depending on the number of common items used. For transformation constant *A*, the bias and RMSE were very close for the RPA method and the SL method for each reference group for each set of common items, with the exception of the R3 reference group, where the bias resulting from the RPA method was much smaller than the SL method for transformation constant *A* when only 5 common items were used. Reference groups R1 and R2 showed fewer differences in bias between different numbers of common items than groups R3 and R4. More consistent with other findings, the RMSE did increase as the number of common items decreased across all reference group distributions.

For transformation constant *B*, again, the bias and RMSE were very close for the RPA method and the SL method with the exception of the RMSE for reference group R3 with only 5 common items. The RPA method produced higher RMSE than the SL method. Reference groups R2 and R4 had bias in the negative direction; group R4 had the least bias overall.

Reference group R3 had the smallest amount of bias when 10 common items were used, and it was in the positive direction. More consistent with other research, however, the RMSE for transformation constant *B* did increase as the number of common items decreased across all reference group distributions. Reference group R3 may be considered an example of a reference group that is too different from the equating group to produce desirable scaling results. The difference in mean of one full standard deviation and a more narrowly distributed group of students resulted in RMSEs that were too high for their use in scaling.

Study 2: Student Data

As Study 2 used actual examinee data, the factors could not be manipulated for all of the conditions that were used in Study 1. Therefore, Research Question 4 was examined in Study 2, comparing the RPA method with the SL method. Study 2 confirmed that the RPA method and the SL method perform comparably for both the *A* and *B* transformation constants when 15 or 10 common items were used. As in the simulation study, the 15 common-item condition and 10 common-item condition resulted in very similar transformation constants *A* and *B* for the RPA method and the SL method. As is often the case in using actual data, the true values are unknown, but it was of interest in this study to see how similar the resulting constants were using the separate transformation methods.

An additional resampling study was conducted to provide a more detailed look at the student data. Of particular interest were the sampling distributions for the 100 replications of random selection of 5, 10, or 15 common items from the 30-item student test. The RPA and SL methods performed comparably to each other in the resampling analysis, but the RPA method produced greater variance, especially when only 5 common items were used.

The resampling study was most similar to the student data when 10 common items were used. For transformation constant *A*, the RPA method constants were 0.9122 (resampling study) and 0.9204 (student data) and the SL method constants were 0.9327 (resampling study) and 0.9120 (student data). The notion of the true value of transformation constant *A* occurring within this range is further supported by the resampling study's values for 15 common items of 0.9257 (RPA method) and 0.9387 (SL method).

Compared to the single-common-item-set estimates for transformation constant *A*, the estimates resulting from the 5 common items were larger in the resampling study, where the RPA method yielded a constant of 0.9614 and the SL method produced 0.9598. These most closely match the higher estimates in the single study when 15 common items were used for the RPA method (0.9789) and SL method (0.9699). None of the resampling results are comparable to the single student data set estimates of 0.6800 for the RPA method and 0.8475 for the SL method. Transformation constant B performed similarly to A in the resampling study. The results for the 5 common-items indicate that 5 common items should not be used to generate the transformation constants in a single student data set as they are quite different from the larger common-item sets and much more variable. The greater variability in the sampling distribution indicates a larger range of estimates that could be chosen, and increases the chance of choosing one unacceptably far from the true estimate.

General Discussion

The RPA method was shown to perform comparably to the SL method in nearly every condition addressed in this study. Either method could be used in a variety of testing situations. Even when test length is limited to 30 test items, the RPA method and the SL method resulted in small amounts of bias and RMSE, with at least 10 common items. The 30-

item tests in these studies were of sufficient length for either method to be used to successfully transform item parameters from separately calibrated estimations prior to equating.

Reducing the number of common items to a set as small as 5 cannot ensure appropriate transformation constants for scaling. In this study, although the bias was not prohibitively large, the high RMSE indicates too much instability in the transformation constants. It would be especially ill-advised to use only 5 common items in situations where the ability distributions are much lower and more variable than the reference population because while the bias was not prohibitively high, the RMSEs became fairly large.

There are other issues that arise in test construction that would require more than just 5 common items to be used. Kolen and Brennan (2004, chap 6.) demonstrate how outlier parameter estimates can influence the magnitude of the transformation constants and the associated equating. Items that behave unusually from one administration to the next should not be used as common items for transforming or equating. As it is difficult to predict this sort of erratic behavior, ample items should be selected initially to prepare for this possibility. In addition, content considerations should be given high priority when selecting common items. Many large scale tests have several subscales, and common items across all subsets would be needed. Five common items would not likely adequately cover several domains.

The ability distribution differences between the reference group and the equating group are also important to consider. Earlier research demonstrated that equating results were more stable when more similar groups were used. This research explored that concept further by using reference groups that were similar in mean ability, but with differing variance. Not

only should the reference and equating group have similar means, they should also be similarly distributed across the ability distribution.

Limitations and Future Research

One of the strengths of this research was the inclusion of both simulation data and actual student data in comparing the RPA method with the SL method. The simulation allowed for control of key variables of interest, and the actual student data contained inconsistencies and characteristics that may occur when using real test items to examine human students. The simulation parameters were generated from data similar to that collected in the actual student data, in order to more carefully mirror real world application. (Albeit, the student data used here were only one possible testing demographic, i.e. collegiate students.) Most importantly, the simulation allowed for systematic manipulation of several conditions to determine how the transformation methods might function under different testing situations. Therefore, the results of this research should be applicable to many testing frameworks. Nevertheless, there are limitations to this study that should be considered in subsequent research.

Despite the strengths of conducting simulation studies, there are several limitations as well. While both the simulation data and actual student were based on a regularly administered test, the items used here may be somewhat easier and less discriminating than might be seen on other large-scale tests. The student population used on these tests, which is primarily white, college-enrolled, and of at least moderate socio-economic status, may not be representative of other testing populations.

Simulated data, even when based on student data, fit the IRT model perfectly, and do not provide the variability that occurs naturally in large assessments. Nor can all the variables of interest be included in a single simulation study. For example, in this study, the number of common items was addressed, but their placement within a test, and the extent to which they must be similar across separate administrations, have been shown to be another important characteristic of common items and were not included here.

The foremost area for further research would be to use the RPA method transformation constants in an actual equating. Ragland et al. (2009) suggested that this method may be preferable to other methods in IRT observed-score equating situations because minimizing the distance between observed and predicted score distributions for scaling purposes provides a viable alternative to minimizing the differences between test characteristic curves (as is done with the SL method). While the SL method is well suited as a precursor to IRT true score equating (as the distance between test characteristic curves is minimized), the approach introduced here should yield better results when IRT observed score (or equipercentile) equating is employed (as the distance between observed and predicted test score distributions is minimized). Now that the method has been demonstrated to produce well-behaved transformation constants for a variety of conditions, it would be appropriate to use them in actual observed-score equating.

Other research should also continue to examine the importance of common item characteristics. As demonstrated in this study, the number of common items alone is not sufficient to guarantee accurate transformation constants. The range of difficulty parameters can also impact the quality of the transformation constants. A growing body of research has begun to address this topic (Duong & Reckase, 2008; Fitzpatrick, 2008; Gao et al., 2008, Lu,

2008; Sinharay & Holland, 2006), and more research is needed. It is of primary consequence to ensure that a sufficient number of common items are included without using far more than are actually necessary. When common items are repeated over several test administrations, they become susceptible to over exposure, resulting in artificially high parameter estimates, thus limiting the quality of the transformation constants and the resultant equating.

A final suggestion for future research would be to replicate these results with additional applications to actual student data. As mentioned earlier, although actual student data were used, the test administered used slightly easier and less discriminating items than might be seen on other large-scale tests. These data were drawn from fairly stable collegiate populations, with several years of testing with documented stability and reference groups over time. More highly variable populations with less established trends might present additional challenges.

In conclusion, this study was the first to investigate the RPA method of calculating transformation constants to be used in IRT equating. It compared favorably with the currently commonly used SL method, as it produced similar transformation constants in a number of varied situations. This research has demonstrated that the RPA method is a viable option for generating the transformation constants necessary for transforming separately calibrated item parameter estimates prior to equating. The RPA method also provides a practical advantage over the SL method. Because the transformation constants are derived from the observed score frequencies, they are more rooted in empirical data, not theoretical indices. Transformation methods like the RPA are a necessary precursor to sound equating, which allows student test scores across different forms of the same test to be compared fairly.

Appendix A. Summary Statistics for Transformation Constants

Table 12. Summary Statistics for Transformation Constant *A* for 60-item Test

Common Items		RPA				SL			
	Group	Mean	Std Dev	Min	Max	Mean	Std Dev	Min	Max
15	R1	1.2645	0.0583	1.1093	1.4166	1.2578	0.0594	1.1084	1.4213
10		1.2546	0.0597	1.1159	1.4128	1.2538	0.0610	1.1203	1.3931
5		1.2660	0.0993	1.0403	1.5175	1.2649	0.0988	1.0549	1.5451
15	R2	0.8095	0.0286	0.7335	0.8723	0.8118	0.0315	0.7259	0.8865
10		0.8090	0.0324	0.7297	0.9282	0.8127	0.0353	0.7429	0.9589
5		0.8160	0.0528	0.7028	0.9506	0.8177	0.0564	0.6941	0.9586
15	R3	1.2743	0.0551	1.1757	1.4026	1.2730	0.0555	1.1723	1.4048
10		1.2636	0.0668	1.1244	1.4907	1.2721	0.0671	1.1246	1.5075
5		1.3265	0.1285	1.0672	1.6670	1.3254	0.1201	1.0757	1.6399
15	R4	0.8184	0.0336	0.7253	0.8907	0.8179	0.0356	0.7224	0.8935
10		0.8177	0.0358	0.7098	0.9051	0.8205	0.0376	0.7135	0.9311
5		0.8313	0.0653	0.6939	0.9973	0.8307	0.0699	0.6785	0.9959

Table 13. Summary Statistics for Transformation Constant *A* for 30-item Test

Common Items		RPA				SL			
	Group	Mean	Std Dev	Min	Max	Mean	Std Dev	Min	Max
15	R1	1.2658	0.0558	1.1340	1.3854	1.2634	0.0570	1.1366	1.3837
10		1.2767	0.0663	1.1044	1.4751	1.2708	0.0671	1.0907	1.4665
5		1.2523	0.1001	0.9922	1.4956	1.2614	0.0929	1.0421	1.4616
15	R2	0.8105	0.0362	0.7339	0.9153	0.8100	0.0378	0.7346	0.9230
10		0.8100	0.0366	0.7299	0.9268	0.8107	0.0395	0.7201	0.9364
5		0.8143	0.0515	0.6568	0.9413	0.8163	0.0549	0.6663	0.9528
15	R3	1.2856	0.0683	1.1285	1.4312	1.2808	0.0694	1.1186	1.4274
10		1.2892	0.0807	1.1055	1.4648	1.2833	0.0812	1.0916	1.4614
5		1.2251	0.1023	0.9870	1.4519	1.2560	0.1106	1.0116	1.5168
15	R4	0.8202	0.0405	0.7124	0.8992	0.8195	0.0424	0.7091	0.9019
10		0.8236	0.0438	0.6916	0.9891	0.8221	0.0457	0.6840	0.9995
5		0.8305	0.0577	0.6819	0.9780	0.8345	0.0660	0.6717	0.9822

Table 14. Summary Statistics for Transformation Constant *B* for 60-item Test

Common Items		RPA				SL			
	Group	Mean	Std Dev	Min	Max	Mean	Std Dev	Min	Max
15	R1	0.6271	0.0596	0.4858	0.7777	0.6225	0.0606	0.4900	0.7752
10		0.6189	0.0743	0.4521	0.9141	0.6177	0.0723	0.4552	0.9039
5		0.6222	0.0724	0.4544	0.8042	0.6188	0.0732	0.4347	0.7877
15	R2	0.3916	0.0392	0.2994	0.4868	0.3941	0.0395	0.2981	0.4833
10		0.3883	0.0409	0.2868	0.4635	0.3921	0.0419	0.2817	0.4746
5		0.3837	0.0538	0.2038	0.5114	0.3858	0.0525	0.2156	0.5102
15	R3	1.2490	0.0676	1.0710	1.4053	1.2467	0.0680	1.0529	1.4133
10		1.2301	0.0819	1.0208	1.4596	1.2378	0.0818	1.0342	1.4738
5		1.2766	0.1224	0.9403	1.6352	1.2713	0.1024	0.9801	1.5479
15	R4	0.7942	0.0456	0.6838	0.9068	0.7943	0.0459	0.6876	0.9074
10		0.7926	0.0495	0.6868	0.9176	0.7972	0.0515	0.6901	0.9294
5		0.7893	0.0612	0.6638	0.9412	0.7881	0.0582	0.6717	0.9455

Table 15. Summary Statistics for Transformation Constant B for 30-item Test

Common Items		RPA				SL			
	Group	Mean	Std Dev	Min	Max	Mean	Std Dev	Min	Max
15	R1	0.6353	0.0518	0.4973	0.7674	0.6332	0.0511	0.4946	0.7653
10		0.6406	0.0561	0.4608	0.7740	0.6361	0.0571	0.4538	0.7722
5		0.6345	0.0902	0.3901	0.8423	0.6341	0.0831	0.4203	0.8608
15	R2	0.3982	0.0407	0.2798	0.5008	0.3976	0.0402	0.2822	0.4924
10		0.3955	0.0420	0.3186	0.5267	0.3956	0.0414	0.3111	0.5013
5		0.4009	0.0519	0.2819	0.5001	0.4008	0.0496	0.2895	0.4980
15	R3	1.2792	0.0651	1.1146	1.4992	1.2716	0.0624	1.1107	1.4734
10		1.2781	0.0787	1.0675	1.4875	1.2685	0.0777	1.0695	1.4632
5		1.2308	0.1082	1.0146	1.5836	1.2499	0.0903	1.0257	1.5052
15	R4	0.8035	0.0405	0.6893	0.8789	0.8019	0.0398	0.6874	0.8778
10		0.8050	0.0430	0.6953	0.8995	0.8016	0.0433	0.7031	0.9038
5		0.8077	0.0664	0.6563	0.9499	0.8084	0.0606	0.6484	0.9426

Appendix B SAS Macros for Simulation Study

SAS Code for Generating Data

```
%LET maxnum=30;
%LET rep=1;
%LET path=C:\Diss\Data\sim\files\S30items\;
libname lib1 "&path.";
 data paras;
  input a1-a&maxnum. b1-b&maxnum. c1-c&maxnum.;
  cards;
   .266    .527   .907   .459   .595   .768   .598   .291   .453    .655
   .562    .403   .455  1.548  1.555   .445   .757   .938  1.234   .852
   .733   1.158   .700   .634   . 675   .742  1.048   .492  1.205   .610
 -3.268 -1.934 -0.632 -0.388  0.312 -0.976 -1.351 -2.537 -1.606 -2.190
 -0.421 -0.982 -2.218  0.628  1.752 -2.125 -2.265  0.024  1.359 -1.588
 -1.280 -0.259  0.060 -0.782  0.256 -0.160 -0.307  2.184  1.015  0.477
   .321    .316   .439   .257   .366   .307   .255   .248   .227   .231
   .196    .276   .229   .317   .137   .194   .207   .273   .189   .114
   .226    .255   .206   .199   .179   .202   .229   .170   .333   .173  ;
 run;

%macro makedata(maxnum);
%do rep=1 %to 100;
 data students;
 length group $2;
   do id=8001 to 10000;
            group="R4";
    theta=-1+sqrt(1.5625)*rannor(0);
    output;
  end;
 run;

data score;
  set students;
  if _n_=1 then set paras;
  d=1.7;
  array a a1-a&maxnum.;
  array b b1-b&maxnum.;
  array c c1-c&maxnum.;
  array r i1-i&maxnum.;
   do i=1 to &maxnum.;
   r(i)= c(i) + (1-c(i))/(1 + exp(-d * a(i)*(theta - b(i))));
     x = rand('uniform');
              if x < r(i) then r(i) = 1;
              else r(i) = 0;
   end;
  total=sum (of i1 -i&maxnum.);
  rep=&rep;
 keep rep group id theta total i1-i&maxnum;
run;
```

```
proc means data=students;
  var theta;
run;

data _null_; set score(where=(group="R4"));
  file "&path.R4&maxnum.R&rep..dat";
  put id 1-4 @6 (i1-i&maxnum) (1.0);
run;

proc datasets nolist;
  delete students score;
run;

data _null_;
  file "&path.\BGR4code.BAT" mod;
  put  "&path.blm1   &path.R4&maxnum.R&rep";
  put  "&path.blm2   &path.R4&maxnum.R&rep";
  put  "&path.blm3   &path.R4&maxnum.R&rep";
run;

data _null_;
  file "&path.\R4&maxnum.R&rep..blm";
   if _n_=1 then do;
   put
    "                          "/
    ">COMMENT data are prescored;"/
    ">GLOBAL DFNAME='R4&maxnum.R&rep..dat' NPARM=3 SAVE;"/
    ">SAVE PARM='R4&maxnum.R&rep..ITM' SCORE='R4&maxnum.R&rep..ml';"/
    ">LENGTH NITEMS=(30);"/
    ">INPUT NTOTAL=30, NIDCHAR=4;" /
    ">ITEMS INAMES=(Q01(1)Q30);"/
    ">TEST;"/
    "(4A1, 1X, 30A1)"/
    ">CALIB FLOAT, CYCLES=50, NEWTON=20, EMPIRACAL, TPRIOR PLOT=0.1;"/
    ">SCORE METHOD=1;";
  end;
  run;
  %end;
%mend makedata;
%makedata(30);
```

SAS Code for Generating *A* and *B* Transformation Constants

```
%macro sheetz;
 %let group=R4;
 %do rep=1 %to 100;
 data newE;
  infile "&path.EQ30R&rep..ITM" firstobs=5 lrecl=200 truncover;
  input  item 2-3 a 37-46 aerr 47-56 b 57-66 berr 67-76 c 97-106 cerr 107-116;
  if item in (2  9  15  22  28);
 run;
 data new&group.;
  infile "&path.&group.30R&rep..ITM" firstobs=5 lrecl=200 truncover;
  input  item 2-3 a 37-46 aerr 47-56 b 57-66 berr 67-76 c 97-106 cerr 107-116;
    if item in (2 9 15 22 28);
 run;

filename myfileEQ dde  "excel|rep&rep!r4c1:r11c4";
 data newEQ; set newEQ;
  file myfileEQ;
   if _n_= 1 then do;
    put "$ Item Parameters";
    put "Item_Number";
   end;
   put item a b c;
   if _n_=5 then put "-1";
 run;

 filename myfile&group. dde  "excel|rep&rep!r12c1:r17c4";
 data new&group.; set new&group.;
  file myfile&group.;
   if _n_= 1 then do;
    put "$_Link_Item_Parameters";
   end;
   put item a b c;
 run;

* read in quadrature densities from phase 2 for equating group;
 data _null_;
  infile "&path.EQ30R&rep..PH2" missover;
  length phrase $28;
  input phrase 2-29;
  if phrase="QUADRATURE POINTS, POSTERIOR" then call symput ('quadstart', _n_);
 run;
%put &quadstart;
%let quadstart=%eval(&quadstart+4);
%let quadend=%eval(&quadstart+11);
*this works for 15 quadpoints;
```

```
data Eq_quad;
  infile "&path.RF30R&rep..PH2" missover firstobs=&quadstart obs=&quadend;
  input midpoint1 14-24 midpoint2 25-36 midpoint3 37-48 midpoint4 49-60 midpoint5 61-72;
  input density1 14-24 density2 25-36 density3 37-48 density4 49-60 density5 61-72;
  input;
  input;
  input;
  input midpoint6 14-24 midpoint7 25-36 midpoint8 37-48 midpoint9 49-60 midpoint10 61-72;
  input density6 14-24 density7 25-36 density8 37-48 density9 49-60 density10 61-72;
  input;
  input;
  input;
  input midpoint11 14-24 midpoint12 25-36 midpoint13 37-48 midpoint14 49-60 midpoint15 61-72;
  input density11 14-24 density12 25-36 density13 37-48 density14 49-60 density15 61-72;
run;

filename rRefquad dde  "excel|rep&rep!r18c1:r34c2";
data Refquad; set Refquad;
  file rRefquad;
    array midpoint[15];
    array density[15];
    put "$ Theta_Distribution Equating Form";
      do i=1 to 15;
        put midpoint[i] density[i];
      end;
put "-100";
run;

*read in Reference Group quads;
data _null_;
  infile "&path.&group.30R&rep..PH2" missover;
  length phrase $28;
  input phrase 2-29;
  if phrase="QUADRATURE POINTS, POSTERIOR" then call symput ('quadstart', _n_);
run;

%put &quadstart;
%let quadstart=%eval(&quadstart+4);
%let quadend=%eval(&quadstart+11);
%put &quadend;
```

```
data &group.quad;
  infile "&path.&group.30R&rep..PH2" missover firstobs=&quadstart obs=&quadend;
  input midpoint1 14-24 midpoint2 25-36 midpoint3 37-48 midpoint4 49-60 midpoint5 61-72;
  input density1 14-24 density2 25-36 density3 37-48 density4 49-60 density5 61-72;
  input;
  input;
  input;
  input midpoint6 14-24 midpoint7 25-36 midpoint8 37-48 midpoint9 49-60 midpoint10 61-72;
  input density6 14-24 density7 25-36 density8 37-48 density9 49-60 density10 61-72;
  input;
  input;
  input;
  input midpoint11 14-24 midpoint12 25-36 midpoint13 37-48 midpoint14 49-60 midpoint15 61-72;
  input density11 14-24 density12 25-36 density13 37-48 density14 49-60 density15 61-72;
run;

filename r&group.quad dde  "excel|rep&rep!r35c1:r50c2";
data &group.quad; set Refquad;
  file r&group.quad;
  array midpoint[15];
  array density[15];
  put "$ Theta_Distribution Reference Form";
    do i=1 to 15;
      put midpoint[i] density[i];
    end;
%end;
%mend sheetz;
%sheetz;
```

References

Angoff, W. H. (1971) Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508-600). Washington, DC: American Council on Education.

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

Baker, F. (1984). Ability metric transformations involved in vertical equating under item response theory. *Applied Psychological Measurement, 8*(3), 261-271.

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

Bené, N. H. (2008, March).*The effect of common item test characteristics on equating scores of a credentialing examination.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Bejar, I., & Wingersky, M. (1981). An Application of Item Response Theory to Equating the Test of Standard Written English. (Report No. ETS-RR-81-35) Princeton, NJ's College Entrance Examination Board. (ERIC Document Reproduction Service No. ED211595)

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

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

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

Camilli, G., Yamamoto, K., & Wang, M. (1993). Scale shrinkage in vertical equating. *Applied Psychological Measurement*, *17*(4), 379-388.

Cook, L.L., Dunbar, & Eignor (1981, April). *IRT Equating: A flexible alternative to conventional methods for solving practical testing problems.* Paper presented at the annual meeting of the American Educational Research Association, Los Angeles, CA.

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

Cook L. L., & Douglass, J. B. (1982). Analysis of fit and vertical equating with the three-parameter model. (ERIC Document Reproduction Service No. ED226007)

Crocker, L.M. & Algina, J. (1986). *Introduction to Classical & Modern Test Theory*. Belmont, CA: Wadsworth Group.

Cui, Z., & Kolen, M. (2008). Comparison of parametric and nonparametric bootstrap methods for estimating random error in equipercentile equating. *Applied Psychological Measurement*, *32*(4), 334-347.

Davey, T., Nering, M., & Thompson, T. (1997). ACT research report series: Realistic simulation of item response data.

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

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

Dorans, N. J. & Kingston, N. M. (1985). The effects of violations of unidimensionality on the estimation of item and ability parameters and on item response theory equating of the GRE verbal scale. *Journal of Educational Measurement*, 32(4), 249-262.

Dorans, N. J., Kubiak, A., Melican, G. J. (1998). *Guidelines for selection of embedded common items for score equating.* (Report No. SR-98-02). Princeton, NJ: Educational Testing Service.

Duong, M.Q. & Reckase, M.D. (2008, March). *Effects of Anchor Characteristics on Linking Multidimensional Item Calibrations in the Non-Equivalent groups with Anchor-Test Design.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Efron, B., & Tibshirani, R. (1985). The bootstrap method for assessing statistical accuracy. *Behaviormetrika*, *17*, 1-35.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58*(3), 357-381.

Felan, G. D. (2002, February). *Test equating: Mean, linear, equipercentile, and item response theory.* Paper presented at the annual meeting of the American Educational Research Association, Austin, TX

Fitzpatrick, A. (2008). NCME 2008 presidential address: The impact of anchor test configuration on student proficiency rates. *Educational Measurement: Issues and Practice*, *27*(4), 34-40.

Gao, X, Hanson, B. A., & Harris, D. J. (1999). *Effect of using different common item sets under the common item non-equivalent groups design.* Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Gao, X., Zhu, R., Chen, H. & Harris, D.J. (2008, March). *Impact of Anchor-Item Selections on IRT Scale Transformation and Equating.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Gialluca, K.A., Crichton, L.I., Vale, C.D., and Ree, M.J. (1984). *Methods for Equating Mental Tests.* (Interim Report for Period March 1982-October 1984.) Brooks Air Force Base, TX: Air Force Systems Command (ERIC Document Reproduction Service No. ED251 512)

Gustafsson, J. (1979). The Rasch model in vertical equating of tests: A critique of Slinde and Linn. *Journal of Educational Measurement, 16*(3), 153-158.

Haebara, T. (1980). Equated logistic ability scales by a weighted least squares method. *Japanese Psychological Research, 22* (3), 144-149.

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

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

Hanson, B. (1996). Testing for differences in test score distributions using loglinear models. *Applied Measurement in Education*, *9*(4), 305-321.

Hanson, B. A., & Béguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement, 26* (1), 3-24.

Harris, D.J. (1991a) A comparison of Angoff's design I and design II for vertical equating using traditional and IRT methodology. *Journal of Educational Measurement, 28(* 3), 221-235.

Harris, D. (1991b). Effects of passage and item scrambling on equating relationships. *Applied Psychological Measurement, 15*(3), 247-256.

Harris, D. J. & Crouse, J. D. (1993). A study of criteria used in equating. *Applied Measurement in Education, 6* (3), 195-240.

Harris, D., & Hoover, H. (1987). An application of the three-parameter IRT model to vertical equating. *Applied Psychological Measurement, 11*(2), 151-159.

Hendrickson, A. B., & Kolen, M. J. (2003). IRT equating of the MCAT. Washington DC: American Association of Medical Colleges.

Holland, P.W. (2007). A framework and history for score linking. In Dorans, N.J., Pommerich, M, & Holland, P.W. (Eds.), *Linking and Aligning Scores and Scales* (pp. 5-29). New York: Springer.

Holmes, S. (1982). Unidimensionality and vertical equating with the Rasch model. *Journal of Educational Measurement, 19* (2), 139-147.

Hu, H., Rogers, W., & Vukmirovic, Z. (2008). Investigation of IRT-based equating methods in the presence of outlier common items. *Applied Psychological Measurement, 32* (4), 311-333.

Huddleson, E. M. (1957). *Equating.* Test Development Memorandum. Princeton, NJ: Educational Testing Service.

Jaeger, R. M., (1981). Some explanatory indices for selection of a test equating method. *Journal of Educational Measurement, 18*(1), 23-53.

Keller, R. R., Kim, W., Nering, M., Keller, L. A. (2006, July). *What breaks the equating: preliminary results of a preliminary investigation.* Paper presented at the Measured Progress internship results session, Dover, NH.

Kim, D., Choi, S., Lee, G., & Um, K. (2008). A comparison of the common-item and random-groups equating designs using empirical data. *International Journal of Selection and Assessment, 16*(2), 83-92.

Kim, S., & Lee, W. (2006). An extension of four IRT linking methods for mixed-format tests. *Journal of Educational Measurement, 43*(1), 53-76.

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

Klein, L., & Jarjoura, D. (1985). The importance of content representation for common-item equating with nonrandom groups. *Journal of Educational Measurement, 22*(3), 197-206.

Kolen, M. J. (2007). Data collection designs and linking procedures. In Dorans, N.J., Pommerich, M, & Holland, P.W. (Eds.), *Linking and Aligning Scores and Scales* (pp. 31-54). New York: Springer.

Kolen, M. J. (1988). An NCME instructional module on traditional equating methodology. *Educational Measurement: Issues and Practice, 7,* 29-36

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

Kolen, M. J. & Brennan, R.L. (2004). *Test equating, scaling, and linking methods and practices.* (2nd ed.). New York: Springer.

Kolen, M. J. & Whitney, D. R. (1981, April). C*omparison of four procedures for equating the tests of general educational development.* Paper presented at the annual meeting of the American Educational Research Association, Los Angeles, CA.

Leary, L. F., & Dorans, N. J. (1985) Implications for altering the context in which test items appear: a historical perspective on an immediate concern. *Review of Educational Research, 55*(3) 387-413.

Levine, R. (1955). *Equating the score scales of alternate forms administered to samples of different ability.* (Research Bulletin 55-23). Princeton, NJ: Educational Testing Service.

Liou, M., & Cheng, P. (1995). Asymptotic standard error of equipercentile equating. *Journal of Educational and Behavioral Statistics, 20*(3), 259-286.

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

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

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

Loyd, B., & Hoover, H. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17*(3), 179-193

Lu, Y. (2008, March). *The comparison of selecting different common item in MIRT Equating.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Marco, G. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement, 14*(2), 139-160.

McKinley, R., & Reckase, M. (1981). *A comparison of procedures for constructing large item pools (Research Report 81-3).* Columbia, MO: University of Missouri, Department of Educational Psychology.

Meyers, J., Miller, G., & Way, W. (2009). Item position and item difficulty change in an IRT-based common item equating design. *Applied Measurement in Education, 22*(1), 38-60.

Michaelides, M. (2006). *Effects of misbehaving common items on aggregate scores and an application of the Mantel-Haenszel statistic in test equating.* (CSE Report 688). Los Angeles, CA: Center for the Study of Evaluation.

Michaelides, M. & Haertel, E. H. (2004). *Sampling of Common Items: An unrecognized source of error in test equating*. (CSE Report 636). Los Angeles, CA: Center for the Study of Evaluation.

Morris, C.N. (1982). On the foundations of test equating. In P.W. Holland & D.B. Rubin (Eds.) , Test equating (pp. 169-191). New York: Academic.

Muraki, E., Hombo, C., & Lee, Y. (2000). Equating and linking of performance assessments. *Applied Psychological Measurement, 24*(4), 325-337.

Ogasawara, H. (2000). Asymptotic standard errors of IRT equating coefficients using moments. *Economic Review (Otaru University of Commerce), 51*(1), 1-23.

Ogasawara, H. (2001a). Item response theory true score equatings and their standard errors. *Journal of Educational and Behavioral Statistics, 26*(1), 31-50.

Ogasawara, H. (2001b). Least squares estimation of item response theory linking coefficients. *Applied Psychological Measurement, 25,* 373-383.

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

Parshall, C. G., Houghton, P. D., and Kromrey, J. D. (1995). Equating error and statistical bias in small sample linear equating. *Journal of Educational Measurement,* 32(1), 37-54.

Patz, R., & Yao, L. (2007). Methods and models for vertical scaling. *Linking and aligning scores and scales* (pp. 253-272). New York, NY : Springer

Petersen, N. (2007). Equating: Best practices and challenges to best practices. In Dorans, N.J., Pommerich, M, & Holland, P.W. (Eds.), *Linking and Aligning Scores and Scales* (pp. 59-71). New York: Springer.

Petersen, N. Cook, L. L. & Stocking, M. L. (1983). IRT versus conventional equating methods: a comparative study of scale stability. *Journal of Educational Statistics, 8* (2), 137-156.

Petersen, N., Kolen, M., & Hoover, H. (1989). Scaling, norming, and equating. *Educational measurement (3rd ed.)* (pp. 221-262). New York, NY England: Macmillan Publishing.

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

Ragland, S. Pashley, P.J. & Armstong, R.L. (2009, April). *Deriving IRT Scale Transformation Constants: A predicted score distribution approach.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Schmitt, A. P., Cook, L. L., Dorans, N. L., and Eignor, D.R. (1990). Sensitivity of equating results to different sampling strategies. *Applied Measurement in Education, 3* (1) 53-71.

Sinharay S., & Holland, P.W. (2006). Is it necessary to make anchor tests mini-versions of the tests being equated or can some restrictions be relaxed? *Journal of Educational Measurement, 44,* (3), 249-275.

Skaggs, G. (2005). Accuracy of random groups equating with very small samples. *Journal of Educational Measurement, 42*(4), 309-330.

Skaggs, G. (1990). To Match or Not to Match Samples on Ability for Equating: A Discussion of Five Articles. *Applied Measurement in Education,* (ERIC Document Reproduction Service No. EJ407912)

Skaggs, G., & Lissitz, R. (1982, April). Effect of examinee ability on test equating invariance. Paper presented at the annual meeting of the National Council on Measurement in Education, Los Angeles, CA.

Skaggs, G. & Lissitz, R. W. (1988). An exploration of the robustness of four test equating methods. *Applied Psychological Measurement,12* (1), 69-82.

Slinde, J., & Linn, R. (1978). An exploration of the adequacy of the Rasch model for the problem of vertical equating. *Journal of Educational Measurement, 15*(1), 23-35.

Stocking, M. L. & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement,* 7(2), 201-210.

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

Tong, Y. & Kolen, M.J. (2005). Assessing Equating Results on Different Equating Criteria. *Applied Measurement in Education*, *29*(6), 418-432.

Vale, C., Maurelli, V., Gialluca, K. Weiss, D., Ree, M. (1981). *Methods for linking item parameters (AFHRL-TR-81-10)*. Brooks Air Force Base, TX: Air Force Human resources Laboratory.

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

von Davier, A., & Wilson, C. (2007). IRT true-score test equating: A guide through assumptions and applications. *Educational and Psychological Measurement*, *67*(6), 940-957.

von Davier, M., & von Davier, A. (2007). A unified approach to IRT scale linking and scale transformations. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *3*(3), 115-124.

Wolkowitz, Amanda A. (2008). A comparison of classical test theory and item response theory methods for equating number-right scored to formula scored assessments. Ph.D. dissertation, University of Kansas, United States -- Kansas. Retrieved March 21, 2009, from Dissertations & Theses: Full Text database. (Publication No. AAT 3297800).

Yang, W., & Houang, R. (1996). The Effect of Anchor Length and Equating Method on the Accuracy of Test Equating: Comparisons of Linear and IRT-Based Equating Using an Anchor-Item Design. (ERIC Document Reproduction Service No. ED401308)

Yi, Q., Harris, D., & Gao, X. (2008). Invariance of equating functions across different subgroups of examinees taking a science achievement test. *Applied Psychological Measurement*, *32*(1), 62-80.

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

Zeng, L. (1991). ACT research report series: Standard errors of linear equating for the single-group design. (ERIC Document Reproduction Service No. ED344944)

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1995) [Computer software and manual] BILOG-MG: multiple-group item analysis and test scoring. Chicago: Scientific Software Int'l.