EFFECTS OF OUTLIER ITEM PARAMETERS ON IRT CHARACTERISTIC CURVE LINKING METHODS UNDER THE COMMON-ITEM NONEQUIVALENT GROUPS DESIGN

By

FRANCISCO ANDRES JIMENEZ

A THESIS PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF ARTS IN EDUCATION

UNIVERSITY OF FLORIDA

2011

To my father, from whom I learned that persistency is one of the greatest qualities that a human being can possess. To my mother, who has always encouraged me to pursue my dreams and goals in life. To my sisters, who since the day they were born, represent a continuous source of joy and admiration to me.

ACKNOWLEDGMENTS

I wish to thank all who have helped me make progress in academia throughout these years at the University of Florida. I would specially like to show my gratitude to my advisor, Dr. David Miller, for his guidance, support, and patience during this process. I would also like to thank my committee member Dr. Walter Leite for introducing me to the world of simulation work using the R software. I am also grateful to Dr. James Algina for sharing his experience and knowledge on the application of statistical methods in the social sciences and education.

I owe my deepest gratitude to my family for their endless encouragement and support during all these years that I have been away from home. This task would have been impossible without their unconditional love and understanding.

Last, my thanks to Yasemin Kaya, Fernando Pagliai, and David Home for the unforgettable times and their invaluable friendship.

TABLE OF CONTENTS

page

LIST OF TABLES

LIST OF FIGURES

Figure page

Abstract of Thesis Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Master of Arts in Education

EFFECTS OF OUTLIER ITEM PARAMETERS ON IRT CHARACTERISTIC CURVE LINKING METHODS UNDER THE COMMON-ITEM NONEQUIVALENT GROUPS DESIGN

By

Francisco Andres Jimenez

August 2011

Chair: M. David Miller
Major: Research, Evaluation, and Measurement Methodology

The comparability of test scores on alternate forms of a test has become a matter of considerable importance for test developers and users. This study used a Monte Carlo simulation to investigate the effects of common items with outlier *b*- and *a*-parameter estimates on the Stocking-Lord (Stocking & Lord, 1983) and Haebara (Haebara, 1980) IRT characteristic curve linking methods under the common-item nonequivalent groups design. The simulation conditions examined in this study included different levels of ability distribution, types of outliers, and number of outliers. Results indicated that neither method seemed to be robust to the presence of common items with outlier *b*-parameter or items with an interaction of outlier *b*- and *a*-parameters, although the Haebara method tended to perform better than the Stocking-Lord method when there were outlier common items with extreme *a*-parameter.

CHAPTER 1
INTRODUCTION

In educational testing, there are many situations in which different groups of examinees are measured with different forms of a test that are supposed to measure the same latent construct (Crocker & Algina, 1986). Since it is very difficult to develop multiple forms of a test in such a way that they are entirely comparable in content, difficulty, or reliability, there is always a possibility that some examinees may be unfairly advantaged because of differences in the difficulty or reliability of the forms rather than real differences in their achievement levels (Dorans & Holland, 2000; Kolen & Brennan, 2004). Therefore, in these situations the comparability of test scores on alternate forms of a test becomes a matter of considerable importance for test developers and users.

Test equating refers to a statistical and psychometric process for adjusting test scores on different forms of a test so that the scores on these forms can be used interchangeably (Dorans & Holland, 2000; Kolen, 2004; Kolen & Brennan, 2004). The main purpose of test equating is to make scores on alternate forms of a test comparable by eliminating the effect of difficulty differences between such test forms (Cook & Eignor, 1991; Dorans & Holland, 2000; Hambleton & Swaminathan, 1985; Kolen & Brennan, 2004). Thus, if the equating process is successful, then in principle the administration of either form of the test should not make any substantial difference to the individuals' performance or comparability (Lord, 1980).

Equating can be accomplished using either classical test theory (CTT) or item response theory (IRT) methods. Since one of the major contributions of IRT to the measurement field has been its ability to place tests and groups of examinees on a common scale of measurement, IRT-based equating procedures are widely used

(Baker & Al-Karni, 1991; Hambleton & Swaminathan, 1985; Kolen & Brennan, 2004). According to Cook and Eignor (1991), IRT equating methods offer two main theoretical advantages over CTT methods: a) IRT equating be the most appropriate method when nonrandom groups of examinees with different ability levels are administered tests of differing difficulties, and b) because IRT provides invariant item parameter estimations, any resulting transformations should be the same regardless of the sample of individuals used to estimate them.

Another statistical IRT method similar to equating is linking (Kolen & Brennan, 2004). However, linking refers to the process of estimating a linear or nonlinear relationship between scores from separate tests that are not necessarily built to the same content or statistical specifications (Kim & Kolen, 2006; Kolen, 2004). In practical terms, linking assumes independent test administrations of multiple test forms to two or more samples of examinees (Kim & Kolen, 2006). Although the statistical procedures used for equating and linking are quite similar, their purposes are different. While equating is used to adjust for difficulty differences between forms of a test to be equated, linking refers to finding a correspondence between parameter estimates on a same scale (Dorans & Holland, 2000; Kolen & Brennan, 2004; Lee & Ban, 2010).

Equating and linking of scores of examinees on various tests can be conducted only under certain testing circumstances. For instance, two different tests administered to two different groups of examinees cannot be equated (Hambleton & Swaminathan, 1985). Therefore, there is a variety of designs that can be used for collecting data for performing linking and equating (Hambleton & Swaminathan, 1985; Kolen & Brennan, 2004). Of these designs, the common-item nonequivalent groups design is utilized

when different forms of a test cannot be administered at once because of test security issues or other practical apprehensions (Kolen & Brennan, 2004). In the common-item nonequivalent groups design, two forms of a test that share a common set of items are administered to different groups of examinees, which allows to adjust for differences on the difficulty of both test forms while preserving any ability differences on the separate groups (Kolen & Brennan, 2004). According to Kolen & Brennan (2004), the set of common items should be a representative sample of the total test forms in terms of content and statistical characteristics and their difficulty and discrimination parameters (*a*s and *b*s) should also be similar once they are transformed into the same proficiency scale.

However, some outlier item parameters can appear in the calibration of the common items due to reasons such as estimation errors, previous experience with the common items, context effects, or differential curriculum emphasis (Stocking & Lord, 1983; Hu, Rogers, & Vukmirovic, 2008; Kaskowitz & de Ayala, 2001; Kolen & Brennan, 2004; Michaelides & Haertel, 2004). While previous research has examined the effects of items with outlier difficulty parameter estimates (*b*-parameter) on IRT-based equating and linking methods (Bejar & Wingersky, 1981; Cohen & Kim, 1998; Hanson & Feinstein, 1997; Hu et al., 2008; Linn, Levine, Hastings, & Wardrop, 1981; Michaelides, 2003; 2010; Michaelides & Haertel, 2004; Sukin & Keller, 2008; Stocking & Lord, 1983), there are no studies addressing the effects of common items with outlier discrimination parameter (*a*-parameter), or with both outlier difficulty and discrimination parameters, on two popular IRT characteristic curve transformation linking methods: the Stocking-Lord method (SL; Stocking & Lord, 1983) and the Haebara (HA; Haebara, 1980) method.

Therefore, the purpose of this study is to examine the effects of having common-item with outlier *b*- and *a*-parameter estimates on the Stocking-Lord and Haebara methods under the common-item nonequivalent groups design using simulated data. The main advantage of using these two IRT characteristic curve methods to determine the effect of outlier common items is that both represent IRT linking methods that consider all of the item parameter estimates simultaneously and generate more stable results than other IRT linking methods (Kim & Cohen, 1998; Hanson & Béguin, 2002; Kolen & Brennan, 2004; Kim & Kolen, 2006; Lee & Ban, 2010). In addition, the following research questions are addressed:

1. Do the Stocking-Lord and Haebara methods differ with respect to bias in the linking process when there are common items with outlier *b*-parameter?
2. Do the Stocking-Lord and Haebara methods differ with respect to bias in the linking process when there are common items with outlier *a*-parameter?
3. Do the Stocking-Lord and Haebara methods differ with respect to bias in the linking process when there are common items with an interaction of outlier *b*- and *a*-parameters?
4. Is the performance of the Stocking-Lord and Haebara linking methods affected by ability distribution or number of outliers?

CHAPTER 2
LITERATURE REVIEW

The present literature review chapter begins with an overview of the IRT models, specifically of the three-parameter logistic model (3PL). Second, a description of the different IRT-based linking methods under the common-item nonequivalent group designs is given, with a focus on the Stocking-Lord and Haebara methods. Then, a series of studies comparing different IRT linking methods is presented. Finally, a discussion on the research studies examining the effects of estimation errors and outlier common items on IRT linking and equating is given.

2.1 Item Response Theory Parameter Estimation (IRT)

Item response theory (IRT) methods are widely used in many large-scale testing applications, including test linking and equating. One of the main advantages of the use of IRT equating methods over classical methods of equating (Angoff, 1982; Kolen & Brennan, 2004) relies on the possibility to obtain invariant item parameter estimations from different test forms that can be placed on a common scale of measurement (Baker & Al-Karin, 1991; Cook & Eignor, 1991; Kolen & Brennan, 2004; Lee & Ban, 2010).

Most IRT models establish that such common ability scale can be represented by a single latent trait or unidimensional construct, referred to as theta, θ (de Ayala, 2009; Kolen & Brennan, 2004). According to IRT, the ability parameter of an examinee, as well as the item parameters, is invariant across subsets of items or across different groups of examinees. This implies that, when the IRT model fits the dataset being analyzed, the same item characteristic curve (ICC) will be obtained for the test items regardless of the groups' ability distributions or the subsets of items used to estimate the item parameters (de Ayala, 2009; Hambleton, Swaminathan, & Rogers, 1991).

Among the various IRT unidimensional models, the three-parameter logistic model (3PL) is specifically used in cases when the chance of success on an item needs to be addressed, as in large-scale assessments incorporating multiple-choice item tests (de Ayala, 2009; Lord, 1980). In this model, the probability that an examinee with ability equal to θ correctly responds to item j is defined as:

$$p(x_j = 1 \mid \theta, a_j, b_j, c_j) = c_j + (1 - c_j)\frac{e^{a_j(\theta - b_j)}}{1 + e^{a_j(\theta - b_j)}} \tag{2-1}$$

where the item parameters a_j, b_j, and c_j are associated with item j (Hambleton & Swaminathan, 1985; Kolen & Brennan, 2004). The item parameter b_j refers to as the difficulty or location parameter for item j. This parameter corresponds to the ability level where the probability of correctly responding to an item is equal to 0.5. On the other hand, the item parameter a_j represents the discrimination parameter. This parameter is proportional to the slope of the ICC at its inflexion point, b_j (de Ayala, 2009; Kolen & Brennan, 2004). Finally, the item parameter c_j is the lower asymptote or pseudo-guessing level parameter for item j. This parameter represents the probability that an examinee with very low ability level (i.e., $\theta = -\infty$) correctly responds to item j.

An example of the 3PL model is given in Figure 2-1. There are two items that have the same location (b_1 and b_2 equal to 0.00) and discrimination (a_1 and a_2 equal to 1.50) parameters, although they have different pseudo-guessing parameters. For item 1, c_1 = 0.10, and for item 2, c_2 = 0.05. Both ICCs have nonzero lower asymptotes and are asymptotic with respect to their corresponding a_j value (de Ayala, 2009). As said before, the lower asymptote corresponds to the lower possible value for the ICC and represents the smallest probability for a correct response (de Ayala, 2009).

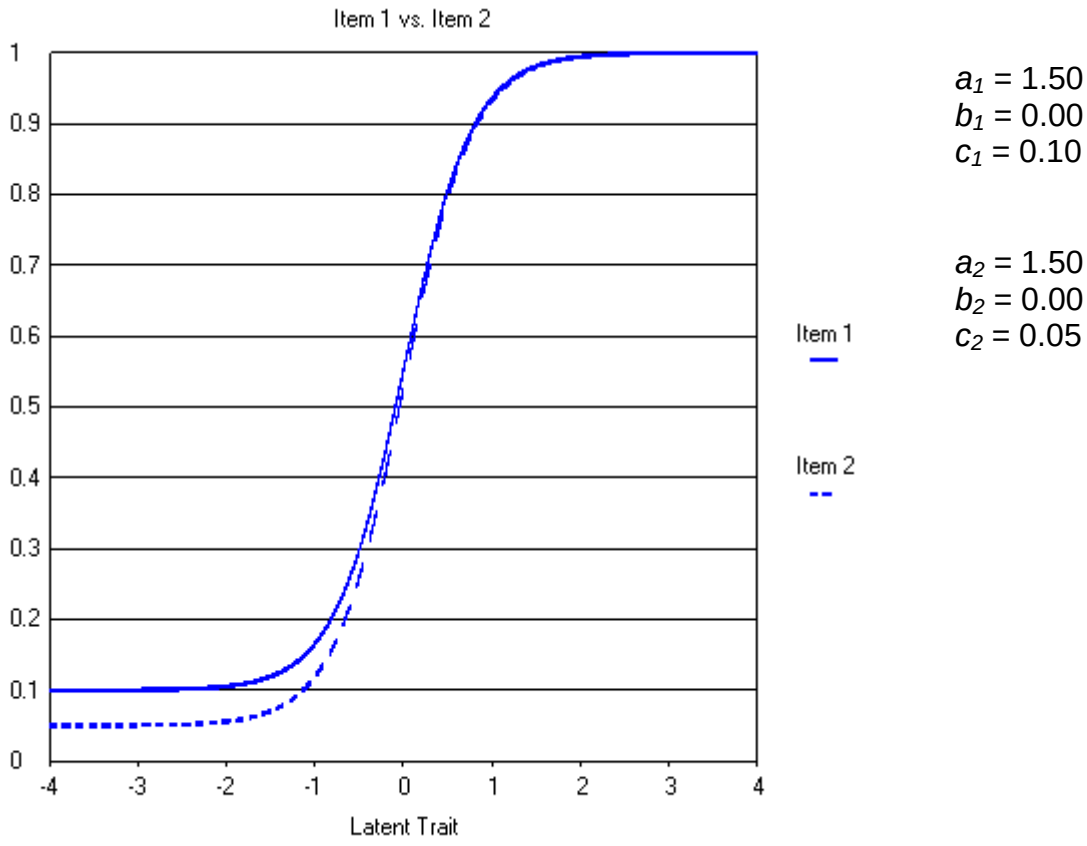

Figure 2-1. 3PL model ICCs for two items with different lower-asymptote values.

As can be seen in Figure 2-1, the ICC for item 1 is located higher than the ICC for item 2. This is due to the fact that as c_j increases, the probability of correct response also increases, all other things being equal (de Ayala, 2009). Thus, items with larger c_js become easier for examinees than those items with smaller c_js.

2.2 IRT Linking Methods Under the Common-Item Nonequivalent Groups Design

2.2.1 Common-item Nonequivalent Groups Design

There are several types of equating designs. Some approaches involve the administration of two test forms to a single group of examinees (single group with counterbalancing, random groups design, or common-item random groups design) and

others use two samples of examinees with each sample being administered one of the test forms, as the common-item nonequivalent groups design (de Ayala, 2009; Kolen & Brennan, 2004).

The common-item nonequivalent groups design is often used when more than one test form cannot be administered at once due to test security issues or other practical apprehensions (Kolen & Brennan, 2004). In this design, different groups of examinees take two test forms, Form *X* and Form *Y*, which have a set of items in common. When the score on the chosen set of common items does not contribute to the examinee's overall score on the total test, the set of common items is said to be external (Kolen & Brennan, 2004). On the other hand, when the performance of the examinees on the set of common items is taken into account in the examinee's overall score on the total test, the set of common items is said to be internal (Kolen & Brennan, 2004). According to Kolen and Brennan (2004) and de Ayala (2009), the set of common items should be proportionally representative of the total test forms in both content and statistical properties. In other words, the set of common items should represent a smaller version of the original test.

2.2.2 Scale Transformation Process

Under the common-item nonequivalent groups design, both groups of examinees are not considered to be equivalent and thus the item parameters obtained from the two different test forms (*X* and *Y*) need to be placed on the same IRT scale in order to conduct the IRT linking process (Kolen & Brennan, 2004). This scale transformation process is possible because the test forms have been created to have a set of items in common, which is generally known as the anchor test (de Ayala, 2009; Kolen & Brennan, 2004).

According to de Ayala (2009) and Kolen and Brennan (2004), the anchor test should measure the same construct, the same content specifications, and the same contextual effects as the noncommon items on the test. There are two subtypes of common-item nonequivalent groups designs: 1) the internal anchor design, which takes into account the score on the set of common items as part of the observed score of the examinees, and 2) the external anchor design, which does not consider the score on the common items as part of the examinees' observed score (de Ayala, 2009).

If an IRT model fits the responses on the test forms to be linked, then any linear transformation of the ability scale, based on the set of common items, will also fit the datasets without modifying the probability of correct response given that the item parameters have also been transformed (Kolen & Brennan, 2004). Therefore, one strategy to perform IRT linking is by using a linear transformation to convert the IRT parameters estimates into the same scale (de Ayala, 2009, Kolen & Brennan, 2004). The linear transformation from one metric to the other in terms of the IRT parameters can be represented by the following set of equations:

$$a_{Y_j} = \frac{a_{X_j}}{A}, \tag{2-1}$$

$$b_{Y_j} = Ab_{X_j} + B, \tag{2-2}$$

$$c_{Y_j} = c_{X_j}, \tag{2-3}$$

$$\theta_{Y_i} = A\theta_{X_i} + B, \tag{2-4}$$

where a_{Yj}, b_{Yj}, and c_{Yj} represent the item parameters for item j on Form Y and a_{Xj}, b_{Xj}, and c_{Xj} represent the item parameters for item j on Form X. Equation 2-3 shows that

the lower asymptote values for both forms remain the same since they are expressed on the probability metric. In addition to how the item parameters on both forms are related, Equation 2-4 shows the relationship between the theta values for examinees on both test forms (Kolen & Brennan, 2004). On the other hand, *A* and *B* are referred as the linking coefficients or metric transformation coefficients (de Ayala, 2009). These metric transformation coefficients in terms of groups of items can be expressed as follows:

$$A = \frac{\sigma(b_Y)}{\sigma(b_X)}, \tag{2-5}$$

$$= \frac{\mu(a_X)}{\mu(a_Y)}, \tag{2-6}$$

and

$$B = \mu(b_Y) - A\mu(b_X). \tag{2-7}$$

Several methods for computing the linking coefficients have been proposed. The most direct methods consist in substituting the means and standard deviations of the item parameter estimates of the common items for the parameters in Equations 2-1 and 2-2 (Kolen & Brennan, 2004). One of these methods is the Mean/Sigma procedure and was developed by Marco (as cited in Kolen & Brennan, 2004). Such method uses the means and standard deviations of the *b*-parameter estimates from the common items in Equations 2-5 and 2-7. In addition, Loyd and Hoover (1980) proposed the Mean/Mean transformation procedure, which consists in using the mean of the *a*-parameter estimates from the common items in Equation 2-6 to estimate the *A* constant. Following, the mean of the *b*-parameter estimates of the common items is used in place of the

parameters in Equation 2-7 to calculate the *B* constant (Kolen & Brennan, 2004). According to Kolen and Brennan (2004), the Mean/Sigma method is preferable over the Mean/Mean method because the estimates of the *b*-parameters are more stable than the estimates of the *a*-parameters.

A potential limitation of both the Mean/Sigma and Mean/Mean methods relates to the use of only means and standard deviations of the item location parameter estimates in the transformation equations. When combinations of *a*-, *b*-, and *c*-parameter estimates generate almost identical ICCs over the range of ability at which most examinees score, both methods can be excessively influenced by marginal differences in the *a*- or *b*-parameter estimates (Kolen & Brennan, 2004).

In contrast to the methods presented above, a second type of scale transformation procedure considers all of the item parameters simultaneously in order to determine the *A* and *B* constants (de Ayala, 2009). These methods are commonly known as the IRT characteristic curve transformation linking methods and their purpose is to relate as closely as possible the initial total characteristic function (TCF) with that of the new metric (de Ayala, 2009). The most relevant characteristic curve methods were developed by Haebara (HA; 1980) and Stocking and Lord (SL; 1983) in the early 80s to overcome the flaws of the Mean/Sigma and Mean/Mean methods.

2.2.3 Haebara Method

The HA procedure consists in a weighted least squares method involving the use of a loss function and an optimization process (Haebara, 1980). The function used by Haebara (1980) to denote the difference between the ICCs is the sum of the squared difference between the ICCs for each of the test items for examinees located at a

particular ability level (Kolen & Brennan, 2004). Then, for a given ability level θ_i, the sum, over items, of the squared difference can be expressed as:

$$Hdiff(\theta_i) = \sum_{j:V} \left[p_{ij}(\theta_{Y_i}; \hat{a}_{Y_j}, \hat{b}_{Y_j}, \hat{c}_{Y_j}) - p_{ij}(\theta_{Y_i}; \frac{\hat{a}_{X_j}}{A}, A\hat{b}_{X_j} + B, \hat{c}_{X_j}) \right]^2 \quad (2\text{-}8)$$

(Kolen & Brennan, 2004).

Equation 2-8 shows that the summation is over the set of common items (*j:V*). As seen, the difference between each ICC on the two scales is squared and summed. This difference is then added up over examinees. The Haebara function aims to find values for *A* and *B* such that they minimize the following condition over examinees:

$$Hcrit = \sum_i Hdiff(\theta_i). \quad (2\text{-}9)$$

(Kolen & Brennan, 2004).

2.2.4 Stocking-Lord Method

Stocking and Lord (1983) developed a robust procedure that gives small weights to those item parameter estimates whose perpendicular distance from the weighted estimate of the transformation is large. Because of its iterative nature, the weighting procedure is repeated until changes on the perpendicular distances become marginally small. The SL method can be described by the following equation:

$$SLdiff(\theta_i) = \left[\sum_{j:V} p_{ij}(\theta_{Y_i}; \hat{a}_{Y_j}, \hat{b}_{Y_j}, \hat{c}_{Y_j}) - \sum_{j:V} p_{ij}(\theta_{Y_i}, \frac{\hat{a}_{X_j}}{A}, A\hat{b}_{X_j} + B, \hat{c}_{X_j} \right]^2 \quad (2\text{-}10)$$

(Kolen & Brennan, 2004).

In contrast to the HA method, the SL method summates across items for each set of parameter estimates before it is squared. In this method, the term $SLdiff(\theta_i)$

represents the squared difference between the test characteristic curves for a given theta level of an individual, whereas the term $Hdiff(\theta_i)$ in the HA method is the sum of the squared difference between the ICCs for a given ability level. Similar to the HA method, the SL method aims to find values for *A* and *B* such that they minimize the following condition over examinees:

$$SLcrit(\theta_i) = \sum_i SLdiff(\theta_i). \tag{2.11}$$

(Kolen & Brennan, 2004).

2.3 Studies Comparing Different IRT Linking Methods

Several empirical studies have compared both characteristic curve methods against one another or against other methods of estimation. Baker and Al-Karni (1991) were one of the first authors to compare two IRT linking methods: the SL and Mean/Mean methods. The authors simulated data for four 60-item tests under the 3PL model and examined three basic types of equating designs: an IRT parameter recovery study, horizontal equating, and vertical equating. In the recovery section of the study, three different normal distributions of ability were used: low (M = -0.5, SD = 0.25), medium (M = 0, SD = 1) and high (M = 0.5, SD = 2.25). Additionally, two ranges for the discrimination parameters were used: low (0.6 to 0.94) and high (1.53 to 1.87), as well as two normal distributions for the difficulty parameters: low (-0.5 to 0.35) and high (0.5 to 0.25). In the horizontal equating section of the study, results from the recovery study were analyzed in order to select two datasets (one with the largest overall root mean square values and another with the lowest values) within each of the three ability levels. In the vertical equating design section of the study, three different normal distributions of ability were specified: low (M = -0.5, SD = 1), medium (M = 0, SD = 1), and high (M =

0.5, *SD* = 1); and for each of the 60-item test, three different distributions of *a*, *b*, and *c* parameters were defined for a set of 15 common items and for the remaining 45 unique items.

Baker and Al-Karni's (1991) simulation results indicated that, in the recovery study, the values for the *A* and *B* transformation coefficients were very similar under the SL and Mean/Mean methods across conditions and overall the root mean square error values for both methods were small except for the low-ability, low-discrimination, and high-difficulty dataset. When comparing the SL method to the Mean/Mean method in terms of performance in this particular dataset, the SL coefficients produced TCFs closer to the underlying TCFs than the Mean/Mean method did (Baker & Al-Karni, 1991). In the vertical equating design, the values of the loss function and root mean square errors were quite small for both IRT linking methods, although the SL method showed smaller values than the Mean/Mean method did (Baker and Al-Karni, 1991).

Kim and Cohen (1998) examined the performance of three different methods for estimating a common metric using simulated data based on the 2PL model: 1) the SL method, 2) concurrent calibration using marginal maximum a posteriori estimation, and 3) concurrent calibration using marginal maximum likelihood estimation. They simulated data for 50 items and 500 examinees and manipulated factors such as the mean of the ability distribution (*M* = 0 and 1), number of common items (*j* = 5, 10, 25, and 50), and IRT estimation program (BILOG and MULTILOG; Kim & Cohen, 1998). Results indicated that, for the discrimination parameters, the SL method generally produced smaller root mean square differences between the equating coefficients and expected values than the other two methods did (Kim & Cohen, 1998). In terms of the item

difficulty parameters, the SL method likewise produced better root mean square differences than both concurrent calibration methods, especially under the 5- and 10-common-item conditions (Kim & Cohen, 1998). Finally, Kim and Cohen (1998) also reported that all three linking methods performed equally well with larger numbers of common items (i.e., more than 5 common items).

In another similar study, Hanson and Béguin (2002) investigated the performance of concurrent versus separate item parameter estimation under the common-item equating design using simulated data based on the 3PL model. For a 60-item dichotomous test, they simulated conditions under different sample sizes (n = 1,000 and 3,000), number of common-items (j = 10 and 20), equivalent versus non-equivalent groups, and IRT estimation program (BILOG-MG and MULTILOG). When comparing the SL and HA methods to the Mean/Mean and Mean/Sigma methods, they concluded that both IRT characteristic curve methods outperformed the Mean/Mean and Mean/Sigma methods in terms of the total error in the estimated true score equating function at a given score (Hanson & Béguin, 2002). On the other hand, the SL and HA methods performed equally well and neither method had consistently lower errors across conditions than the other (Hanson & Béguin, 2002). However, when comparing both IRT characteristic curve methods to the concurrent calibration method, the latter resulted in lower errors across conditions except in conditions under a sample size of 3,000, equivalent groups, and 10 common items.

In a different set of studies, Béguin, Hanson, and Glas (2000), Béguin and Hanson (2001), and Kim and Kolen (2006) evaluated the performance of unidimensional IRT linking methods on multidimensional data. In order to compare the SL method against

the concurrent calibration method, Béguin et al. (2000) simulated multidimensional compensatory data for 1) equivalent and nonequivalent groups, 2) different levels of covariance between two dimensions, and 3) unidimensional and multidimensional parameter estimation. Results showed that, under the equivalent groups conditions, the SL method resulted in generally larger mean square errors than the concurrent estimation method, whereas in the nonequivalent groups conditions, both SL and concurrent methods performed poorly. Therefore, Béguin et al. (2000) concluded that multidimensionality could affect the performance of unidimensional IRT linking methods.

Similarly, Béguin and Hanson (2001) carried out a simulation study to compare the performance of the SL and concurrent calibration methods for unidimensional IRT models when applied to multidimensional noncompensatory data. They simulated data based on real item parameters taken from three pairs of test forms of a standardized measure of language comprehension and examined the following conditions: 1) equivalent and nonequivalent groups, 2) three levels of covariance between the test forms, and 3) multidimensional and unidimensional estimation. Using the same evaluation criteria than Béguin et al. (2000) did, Béguin and Hanson (2001) also concluded that the concurrent estimation method led to smaller mean square errors than the SL estimation method under the equivalent groups condition, although for Béguin et al. (2000) the effect of the type of multidimensionality seemed to vary more as a function of the size of the covariance between dimensions.

Kim and Kolen (2006) conducted a simulation study to examine the degree to which four IRT linking methods (Mean/Mean, Mean/Sigma, HA, and SL) and the concurrent calibration method were robust to format effects in mixed-format tests under

the common-item nonequivalent groups design. The authors developed a series of simulations using two types of mixed-format tests as part of the linking design: wide-range and narrow-range, each of which had a set of 12 multiple-choice (MC) items and five constructed-response (CR) items in common, had different target information functions, and was based on the item parameters of the science assessment from the 1996 National Assessment of Educational Progress report (as cited in Kim & Kolen, 2006). While the wide-range test aimed to replicate common standardized achievement and aptitude tests, the narrow-range test aimed to characterize screening tests used other testing situations (Kim & Kolen, 2006).

Kim and Kolen (2006) simulated data under three factors: 1) three levels of format-effects as correlations between θ_1 and θ_2 (0.5, 0.8, and 1), 2) two types of mixed-format tests (wide-range and narrow-range), and 3) three levels of nonequivalence in linking as values for bivariate distributions ($\mu_1 = 0$, $\sigma_1 = 1$, $\mu_2 = 0$, $\sigma_1 = 1$; $\mu_1 = 0.5$, $\sigma_1 = 1$, $\mu_2 = 0.5$, $\sigma_1 = 1$; and $\mu_1 = 1$, $\sigma_1 = 1$, $\mu_2 = 1$, $\sigma_1 = 1$). For the narrow-range, mixed-format test, results indicated that under a combination of nonequivalence and format effects, the Mean/Mean and Mean/Sigma methods had consistency larger mean square errors across point scores than the SL and HA methods, yet there were no substantial differences between these last two. On the other hand, the concurrent calibration method showed consistently lower mean square errors across all score points than the four IRT linking methods (Kim & Kolen, 2006). Nevertheless, the difference between the concurrent calibration method and the SL and HA methods was only slight. For the wide-range, mixed-format test, the pattern of results was very similar to what was observed in the narrow-range test. However, the wide-range showed the maximum

values for the root mean square errors in the middle of the observed score range, whereas the narrow-range showed the maximum values in the lower part.

In a recent simulation study, Lee and Ban (2010) compared the performance of four different IRT linking methods in the random groups equating design: 1) concurrent calibration method, 2) SL method, 3) HA method, and 4) proficiency transformation. In order to simulate data, the authors used real item parameters from two 75-item ACT English forms, each of which was taken for approximately 3,000 examinees and was analyzed using a 3PL model. The authors manipulated four different factors in their study: 1) combinations of sampling designs (normal distributions with proficiency means of 0, 0, and 0; 0.5, and 0.5; and 0, 1, and 1), sample size (n = 500 and 3,000), and number of items (j = 25 and 75).

Lee and Ban (2010) stated that, on average, the SL and HA methods outperformed the concurrent calibration and proficiency transformation methods under all conditions, except when the samples were obtained from the same population. On the other hand, mean square errors were substantially lower for n = 3,000 and linking errors were prompt to increase as the difference in the proficiency level between the sampling designs also did (Lee & Ban, 2010). With respect to the SL and HA methods compared to one another, the HA method generally tended to perform better than the SL method, producing lower linking errors in average across conditions (Lee & Ban, 2010).

2.4 Effects of Estimation Errors and Outlier Common Items on IRT Linking and Equating

Relatively few empirical studies have addressed the implications of estimation errors and outlier common items on IRT linking and equating methods. Kaskowitz and

de Ayala (2001) investigated the effect of errors in item parameter estimates on both the estimation of linking coefficients with the SL method and the subsequent estimation of the examinees' true score. Using a horizontal internal anchor equation design, the authors simulated data by generating item parameter estimates for initial and target metrics from known sampling distributions (Kaskowitz & de Ayala, 2001). Four different factors were manipulated: number of common items (5, 15, and 25), IRT model (2PL and 3PL), relationship of the standard error to the parameter estimate (nonrelated error and related error), and error level based on the sample size (low, moderate, and high).

Results showed that the SL method was overall robust in relation to the amount of error in the item parameter estimates (Kaskowitz & de Ayala, 2001). In the nonrelated error condition, the bias and root mean square errors for *A* and *B* improved as the level of error decreased, and the poorest estimation of *A* took place in the conditions under the 5-common items factor. In the related error condition, the bias and root mean square errors were higher than for the nonrelated error condition, yet a similar pattern of improvement was seen as the level of error decreased. Interestingly, with the error level being constant, *A* and *B* always showed greater levels of estimation error and bias for the 3PL than for the 2PL model (Kaskowitz & de Ayala, 2001).

As part of his dissertation work, Michaelides (2003) conducted a three-phase study that examined the effects of misbehaving common items in IRT-based equating processes. Of interest is the first applied stage of his study where the author examined the consequences of either retaining or eliminating items identified as outliers by using the delta-plot method proposed by Angoff in 1972 (as cited in Michaelides, 2003). The author analyzed two test administrations (Year 1 and 2) from four statewide

assessments (8th-grade Mathematics, 11th-grade Science, 6th-grade Social Studies, and 6th-grade Science) by fitting the 1PL and 3PL model for dichotomous items, and the graded response model for polytomous items. Michaelides (2003) compared the SL, HA, Mean/Sigma, and Mean/Mean IRT-based linking methods. Results indicated that when outliers (three items in this case) were included in the linking process, the Year 2 mean was higher and the 1PL-model analyses gave positive gains in the standard error of equating for Year 2 that were as twice as large than for Year 1. However, when the 3PL was fitted, there was a slight decline with the SL method, yet a larger drop with the HA and Mean/Mean and a similar drop in magnitude gain with the Mean/Mean method.

In another study examining the effects of outlier common-items on IRT linking, Hu et al. (2008) conducted a simulation study to compare four IRT-based linking methods when items with outlier *b*-parameters were either ignored or considered under the common-item nonequivalent groups design. The authors simulated data under the 2PL (short-answer, SA, and open-ended response, OR, items) and 3PL (multiple-choice, MC, items) models by using real item parameters from the Massachusetts Comprehensive Assessment System mathematics tests (as cited in Hu et al., 2008). The following conditions were manipulated in their linking design: 1) group ability distribution (M = 0, SD = 1 and M = 1, SD = 1), 2) number/score points and types of outliers (no outliers, three MC items with three score points from one content area, three MC items randomly distributed across content areas, three MC items with extreme *b*-parameters, five MC and one OR items with nine score points from one content area, five MC and one OR items with nine score points that were randomly distributed across content areas), and 3) IRT-based equating designs: concurrent calibration, separate

calibration with the SL and HA methods, separate calibration with the Mean/Sigma method, and calibration with fixed common item parameters (FCIP).

Study results showed that when no outlier common items were present and groups were equivalent, all IRT-based equating methods performed well (Hu et al., 2008). For items with three score points, all IRT-based linking methods showed similar mean square error results no matter if outliers were present in one content area or across the five content areas. However, for equivalent groups, the concurrent calibration with outliers included and the fixed common item parameter calibration had the smallest mean square errors, followed by the characteristic curve methods, and finally, by the Mean/Sigma method. For nonequivalent groups, the characteristic curve, Mean/Sigma and FCIP methods had moderate mean square errors, whereas the concurrent calibration method had large mean square errors (Hu et al., 2008). Similar to the results obtained with three score points, there were no differences between the IRT-based linking methods under the nine score points conditions when outliers were present in one content area or across the five content areas. For equivalent groups, mean square errors tended to increase as the number/score points of outliers also increased (Hu et al., 2008). For nonequivalent groups when outliers were included in the design, the mean square errors for both characteristic curve and Mean/Sigma methods also tended to increase as the number/score points of outliers increased, although for the concurrent calibration method tended to remain the same and for the FCIP method they remained small (Hu et al., 2008).

Sukin and Keller (2008) also conducted a simulation study to explore the effects of removing or maintaining items with misbehaving *b*-parameters under the common-item

nonequivalent groups design. Using real item parameters from an operational statewide testing program, two test forms were simulated, each of which had a twenty-item internal anchor and was analyzed based on the 3PL model. Sukin and Keller (2008) manipulated conditions under two normal ability distributions (M = 0, SD = 1 and M = 0.2, SD = 1), four different IRT-based linking methods (Mean/Sigma, Mean/Mean, SL, and HA), three cut scores (-0.75, 0, and 0.75 on the θ metric), and two types of outlier items (a shift of 0.3 and 0.8 in the b-parameter scale). Sukin and Keller (2008) simulation results showed that the overall classification accuracy was not affected by the presence of outlier common items, although the proportion of examinees over- and under-classified was. When the outlier item was kept in the estimation, about 1% fewer examinees were being identified as under-classified in comparison to when the outlier item was excluded from the IRT-based linking methods. A similar pattern of results and no real observed differences were obtained across both types of ability distribution, IRT-linking methods, and cut scores (Sukin & Keller, 2008).

CHAPTER 3
METHODS

In the present study, a Monte Carlo simulation was conducted to compare the performances of the Stocking-Lord (SL; Stocking & Lord, 1983) and Haebara (HA; Haebara, 1980) IRT linking methods in the presence of outlier common items. Data were simulated for a common-item nonequivalent groups matrix design. In addition, several factors were manipulated to reflect different levels of group ability differences, types of outliers, and number of outliers. Following there is a detailed description of the equating design, factors manipulated, and factors held constant in this study. Finally, a description of the data generation process, evaluation criteria, and data analysis process is given.

3.1 Equating Design

A common-item nonequivalent groups design using an external anchor test was employed in the current study. Since the results of this study are in part intended to be generalized to large-scale achievement tests using IRT-based linking methods, the common-item nonequivalent groups design was chosen to closest replicate the horizontal equating process utilized in the 2005 and 2006 administrations of the Florida Comprehensive Assessment Test (FCAT) mathematics test for 4th-Grade (Florida Department of Education (FDOE), 2006, 2007). In order to maintain the comparability of the FCAT scale year after year, FCAT developers indeed use the SL procedure to establish a statistical correspondence between the performance of current-year students on the external anchor items and the performance of students on the same set of items in the previous years of testing (FDOE, 2006, 2007).

This study employed a linking design in which in Year 1 (i.e., 2005) a test Form *X* was administered to a sample of 4th-Grade students. This test form included a set of unique (or core) items (i.e., U_X) and a set of common items (i.e., C_Z). In Year 2 (i.e., 2006), another test Form *Y* was administered to another sample of 4th-Grade students as well. The unique items administered on Form *Y* (i.e., U_Y) were different from those items administered on Form *X* in Year 1. Nonetheless, the same set of common items was used on Form *Y* (i.e., C_Z). In this sense, examinees' scores on U_X and U_Y can be linked thorough such set of common items, C_Z.

The FCAT mathematics test composition and format vary depending upon the grade. According to the Sunshine State Standards (SSS; Florida Department of Education, 1996), the FCAT mathematics tests are meant to assess five different content areas: 1) number sense, concepts, and operations, 2) measurement, 3) geometry and spatial sense, 4) algebraic thinking, and 5) data analysis and probability (FDOE, 2006, 2007). Likewise, there are three types of FCAT mathematics item formats: multiple-choice (MC), gridded-response (GR), and performance task (PT; i.e., short- and extended-response items). The item format used in the FCAT 4th-Grade mathematics test only contains dichotomously scored, multiple-choice (MC) items. In 2005 and 2006, each FCAT 4th-Grade mathematics form had 39 unique items plus 6 to 8 common or field-test items that were distributed across 30 different forms. For technical reasons, the unique and common items in Forms 27 to 30 were finally selected for an early-return IRT calibration sample of 4th-grade students (FDOE, 2006, 2007)

Therefore, in the present study, the IRT characteristic curve linking process was conducted by using real item parameters from those 39 unique and 22 common items (61 items in total) used in such early-return IRT calibration sample (FDOE, 2006, 2007). As suggested by Kolen and Brennan (2004), a common item set should be at least 20% of the length of a total test containing 40 or more items. In this sense, the 22 common items thoroughly represented, in statistical, format, and content terms, the same characteristics of the unique items. Given this equating design, the unique items in the test Forms *Y* and *X* were linked to be expressed on the same scale.

3.2 Factors Manipulated

3.2.1 Group Ability Differences

Similar to Hu et al. (2008) and Lee and Ban (2010), samples of item responses for Form *Y* were generated by sampling the latent trait (θ) from a normal independent distribution with mean equal to 0 and standard deviation equal to 1 (*NID* (0, 1)). In addition, two sets of item responses were created for Form *X* by sampling θ from an *NID* (0, 1) distribution and an *NID* (1, 1) distribution. Only for comparison purposes, the samples with *NID* (0, 1) for both Forms *Y* and *X* examined the specific situation in which the two groups were equivalent. The samples with *NID* (0, 1) for Form *Y* and *NID* (1, 1) for Form *X* examined the case in which the groups are nonequivalent. According to Kolen and Brennan (2004), the larger the difference between groups of examinees, the more difficult it becomes for the statistical methods to identify differences due to the groups or test forms. Hence, the result of any linking method becomes critical when the groups differ, especially in terms of the representativeness of the content and statistical properties. In addition, the findings of the current study can be compared with similar studies that have also included this condition (Hanson & Béguin, 2002; Hu et al., 2008).

3.2.2 Types of Outliers

In the current study, an outlier common item was conceptually defined and operationalized primarily based on what Hu et al. (2008) graphically described as an outlier. In terms of its *b*-parameter, Hu et al. (2008) defined a common item as an outlier if in the scatterplot of difficulties the distance between the intersection point drawn from each item's *x*- and *y*-axis position and its predicted position on the straight line was equal to or more than two score points. Thereafter, in the current study all the *b*-parameters for the outliers in Year 2 (i.e., Form *Y*) were two score points lower than in Year 1 (i.e., Form *X*). This meant that only outliers located on the left side of the straight line were taken into consideration.

Hu et al. (2008) gave a theoretical and practical explanation for considering this operational definition of an outlier. In educational practice, the presence of outliers is often due to the misbehavior of some common items, a change in the instructional emphasis of a certain content area, and item parameter estimation errors (Kaskowitz & de Ayala, 2001; Kolen & Brennan, 2004; Michaelides, 2003). Similarly, the most likely result of the exposure of some common items, and subsequently of the exposure to the instructional emphasis such items measure, is that those items will become easier when they are administered in the second year (Hu et al., 2008).

Although Hu et al. (2008) did not examine common items with outlier *a*-parameter estimates, the current study used a somewhat similar operational definition to identify outlier items based on their *a*-parameter. Theoretically, the discrimination parameter can vary from $-\infty$ to $+\infty$. However, an item with a negative *a*-parameter may indicate that its performance is severely inconsistent with the IRT model or that it is behaving in a counterintuitive fashion (de Ayala, 2009). Therefore, in this study, an *a*-parameters

that was greater than 1.000 in Year (i.e., Form *X*) became an outlier by dropping one point score in Year 2 (i.e., Form *Y*). On the other hand, an *a*-parameter that was between 0.500 and 0.999 on Form *X* became an outlier by dropping 0.5 point scores on Form *Y*. Lastly, an *a*-parameter that was below 0.500 on Form *X* became an outlier by dropping to a fixed value of 0.100. Similar to the case of items with outlier *b*-parameter, only outliers with *a*-parameter located on the left side of the straight line were taken into consideration for this study.

3.2.3 Number of Outliers

Seven combinations of number of common items with outlier *b*- and *a*-parameters were examined: 1) There were no outliers in the common items, 2) the outliers were two common items with extreme *b*-parameter, 3) the outliers were two common items with extreme *a*-parameter, 4) the outliers were two common items with extreme *b*- and *a*-parameters (interaction of outlier parameters), 5) the outliers were four common items with extreme *b*-parameter, 6) the outliers were four common items with extreme *a*-parameter, 7) the outliers were four common items with extreme *b*- and *a*-parameters (interaction of outlier parameters).

The first condition was used as the baseline against which the other conditions were compared. Hu et al. (2008) suggested that the baseline condition should well control the number/score points of outliers and representativeness of the common items. Conditions 2, 3, 5, and 6 simulated the effect of the increase of outlier common items with either extreme *b*- or *a*-parameters due to the impact of differential curriculum emphasis on the common items (Hu et al., 2008). Conditions 4 and 7 reflected the effect of an increase in the presence of outlier common items with an interaction of extreme *b*- and *a*-parameters.

3.2.4 IRT Characteristic Curve Linking Methods

Two commonly used IRT characteristic curve linking methods, the Stocking-Lord and Haebara methods, were used to link the two test forms with and without outliers.

3.3 Factors Held Constant

3.3.1 Sample Size

The sample size for this study was held constant at 3,000 for each test form. According to several research studies on equating methods (Hanson & Béguin, 2002; Hu et al., 2008; Kolen & Brennan, 2004; Lee & Ban, 2010), this sample size seems to be adequate to obtain stable parameter estimates. Therefore, the total sample size considering both test forms was 6,000.

3.3.2 IRT Model

The three-parameter logistic model (3PL; Lord, 1980) was used to simulate the data for test Forms *X* and *Y*. A three-parameter logistic model was chosen to account for MC items response data from low-ability examinees for which guessing is always an important factor to consider in test performance (de Ayala; 2009; Hambleton & Swaminathan, 1985). Additionally, the 3PL model was indeed used by the FCAT specialists to calibrate and analyze MC items from the 2005 and 2006 FCAT 4th-Grade Mathematics tests (FDOE, 2006, 2007).

3.4 Data Generation

Each of the simulated datasets for Forms *X* and *Y* included responses to 61 items in total. Of these, 39 were unique dichotomous items and 22 were common dichotomous items. In order to simulate the item responses, item parameters a_j, b_j, and c_j were taken from the 2005 and 2006 FCAT 4th-Grade Mathematics test calibrations (FDOE, 2006, 2007). Item parameters corresponding to the 2005 FCAT 4th-Grade

Mathematics tests were used to simulate responses on Form *X*, whereas item parameters corresponding to the 2006 FCAT 4th-Grade Mathematics test were utilized to simulate responses on Form *Y*. Table 3-1 contains the item parameters used for the unique items on each form in the present simulation study. Table 3-2 shows the item parameters used for the common items on both forms.

On Form *X*, the a_j parameters for the unique items used in the simulation ranged from 0.513 to 1.429, the b_j parameters ranged from -2.004 to 1.011, and the c_j parameters ranged from 0.034 to 0.562. On the other hand, on Form *Y*, the a_j parameters for the unique items ranged from 0.530 to 1.491, the b_j parameters ranged from -2.640 to 0.802, and the c_j parameters ranged from 0.063 to 0.404. With respect to the set of common items on both Forms *X* and *Y*, the a_j parameters ranged from 0.549 to 1.556, the b_j parameters ranged from -1.911 to 0.427, and the c_j parameters ranged from 0.036 to 0.386.

In the first condition, all *a*- and *b*-parameters for the common items remained the same from Form *X* (Year 1) to Form *Y* (Year 2) because there were no outliers present in the common items. In Condition 2, the *b*-parameter for item 40 dropped from 0.132 on Form *X* to -1.868 on Form *Y,* and the *b*-parameter for item 49 dropped from -1.027 on Form *X* to -3.027 on Form *Y*. In Condition 3, the *a*-parameter for item 42 dropped from 1.156 on Form *X* to 0.156 on Form *Y*, and the *a*-parameter for item 52 dropped from 1.556 on Form *X* to 0.556 on Form *Y*. In Condition 4, the *b*-parameter for common item 44 dropped from -0.434 to -2.434 and the *a*-parameter did from 1.197 to 0.197 from Form *X* to Y, respectively. Similarly, the *b*-parameter for common item 54 dropped

from -0.001 to -2.001 and the *a*-parameter did from 0.816 to 0.316 from Form *X* to *Y*, respectively.

Table 3-1. Item parameters for unique items 1 to 39 on Forms *X* and *Y* (j = 39)

Item	Form *X*			Form *Y*		
number	a_j	b_j	c_j	a_j	b_j	c_j
1	1.173	-0.634	0.183	0.530	-2.022	0.214
2	0.822	-0.396	0.309	0.762	-2.640	0.149
3	0.513	-0.986	0.094	0.751	0.048	0.287
4	0.524	0.202	0.054	0.692	0.490	0.291
5	1.071	-0.437	0.122	0.559	-1.134	0.143
6	0.977	-2.004	0.110	0.608	-1.631	0.083
7	0.635	0.357	0.300	0.568	-0.772	0.063
8	0.989	-1.299	0.208	0.758	-0.466	0.068
9	0.779	-0.378	0.314	0.905	0.373	0.251
10	0.948	0.296	0.080	0.609	-0.210	0.092
11	0.902	-1.723	0.113	1.297	-0.396	0.145
12	0.804	-0.274	0.130	0.733	0.239	0.090
13	0.717	0.769	0.228	1.353	0.065	0.275
14	1.073	-0.041	0.124	1.200	0.662	0.115
15	1.134	0.078	0.208	0.580	0.638	0.208
16	0.872	0.116	0.060	0.933	0.276	0.275
17	1.071	0.074	0.095	0.714	-1.605	0.280
18	0.912	-0.580	0.266	0.532	-1.625	0.404
19	0.820	-1.350	0.048	0.825	0.194	0.159
20	0.848	0.072	0.060	1.491	-0.731	0.229
21	1.221	-0.303	0.149	0.715	-1.660	0.210
22	1.178	-0.082	0.115	1.207	0.643	0.172
23	1.367	1.011	0.348	0.740	-1.549	0.152
24	1.273	0.499	0.094	0.620	0.189	0.098
25	1.429	0.525	0.325	0.976	0.519	0.264
26	0.710	0.447	0.141	0.705	-1.073	0.209
27	0.559	-0.144	0.034	1.029	0.771	0.088
28	0.906	-1.173	0.562	0.939	0.802	0.245
29	1.177	-0.435	0.084	1.058	-0.348	0.161
30	0.936	-0.664	0.155	0.831	0.567	0.204
31	0.636	-1.036	0.035	0.864	0.130	0.214
32	0.801	-1.351	0.162	0.806	-1.163	0.190
33	1.066	-1.436	0.155	0.808	-0.761	0.166
34	0.946	-1.078	0.354	0.792	0.081	0.215
35	0.619	-1.527	0.093	0.671	0.045	0.296
36	0.892	-0.039	0.193	0.805	-0.342	0.065
37	1.002	0.208	0.335	0.728	-0.212	0.174
38	0.844	-0.682	0.185	1.352	-0.252	0.273
39	1.233	-0.543	0.253	0.984	-0.774	0.173

Table 3-2. Item parameters for common items 40 to 61 on Forms *X* and *Y* (*j* = 22)

Item number	Form *X* and Form *Y*		
	a_j	b_j	c_j
40	1.226	0.132	0.076
41	0.561	-0.209	0.036
42	1.156	-0.579	0.055
43	0.791	0.352	0.268
44	1.197	-0.434	0.206
45	0.549	-1.228	0.085
46	1.293	0.045	0.132
47	0.706	-0.705	0.198
48	1.006	-1.911	0.148
49	0.861	-1.027	0.289
50	0.603	-0.615	0.095
51	0.879	-0.846	0.092
52	1.556	0.008	0.232
53	0.511	-1.799	0.125
54	0.816	-0.001	0.231
55	0.747	-0.503	0.193
56	0.685	-0.082	0.386
57	1.036	-1.148	0.197
58	0.635	-1.691	0.158
59	0.654	0.427	0.200
60	0.595	-1.533	0.151
61	1.302	0.319	0.106

In Condition 5, the *b*-parameters for item 41 dropped from -2.209 to -2.209 from Form *X* to Form *Y*, for item 51 did from -0.846 to -2.846, for item 55 did from -0.503 to -2.503, and for item 61 did from 0.319 to -1.681. In Condition 6, the *a*-parameters for item 45 dropped from 0.549 to 0.049 from Form *X* to *Y*, for item 48 did from 1.006 to 0.006, for item 56 did from 0.685 to 0.185, and for item 60 did from 0.595 to 0.095. In Condition 7, the *b*-parameter for item 43 dropped from 0.352 to -1.648 and the *a*-parameter dropped from 0.791 to 0.291 from Form *X* to Form *Y*. For item 46, the *b*-parameter dropped from 0.045 to -1.995 and the *a*-parameter dropped from 1.293 to 0.293. For item 53, the *b*-parameter dropped from -1.799 to -3.799 and the *a*-parameter

dropped from 0.511 to 0.011. Finally, for item 59 the *b*-parameter dropped from 0.427 to -1.573 and the *a*-parameter did from 0.654 to 0.154.

Overall, this study involved a common-item nonequivalent groups simulation design composed of seven conditions based on different outlier common-item scenarios (either the *b*-parameter, *a*-parameter, or an interaction of *a*- and *b*-parameters), two levels of ability distribution (*NID* (0, 1) for Form Y and *NID* (0, 1) and (1, 1) for Form *X*), and two IRT characteristic curve linking methods (SL and HA). For each of the conditions, 1000 datasets were generated. Once the datasets were created, the response samples for the two test forms were analyzed with the 3PL model and then linked by using one of the characteristic curve methods. Following the linking process, test responses to Form *X* and Form *Y* were equated by using IRT true-score equating (Kolen & Brennan, 2004). The entire data simulation process was done with the statistical software R, version 2.12-2 (R Development Core Team, 2011).

3.5 Evaluation Criteria

In order to evaluate the performance of the SL and HA methods in the presence of common items, the unweighted mean square error for the *b*-parameters (MSE_b) and the unweighted mean square error for the *a*-parameters (MSE_a) were used. Equations 3-1 and 3-2 show each formula respectively:

$$MSE_b = \frac{\sum_{r=1}^{1000}\sum_{j=1}^{39}(b_{jr}^{*} - b_j)^2}{1000 \times 39} \qquad (3\text{-}1)$$

and

$$MSE_a = \frac{\sum_{r=1}^{1000}\sum_{j=1}^{39}(a_{jr}^{*} - a_j)^2}{1000 \times 39}, \quad (3\text{-}2)$$

(Hu et al., 2008)

where b^*_{jr} is the *b*-parameter for item *j* on the equated test (Form *Y*) for replication *r*, b_j is the true value for the *b*-parameter for item *j* on the equated test Form *Y*, a^*_{jr} is the *a*-parameter for item *j* on the equated test (Form *Y*) for replication *r*, and a_j is the true value for the *a*-parameter for item *j* on the equated test Form *Y*. In both formulas, such difference is summed across score points (39 points) and across replications (1000 replications) and then divided by the number of replications times the number of score points.

Hu et al. (2008) concluded that in order to claim that the size of the unweighted mean square error for a specific condition was considered small, moderate, or large, it was necessary to develop absolute rules for interpreting the sizes of the estimated error for each parameter. Such criterion was sustained on the magnitude of the square roots of the unweighted mean square errors (referred to as bias), which correspond to the difference between the observed *b*- and *a*-parameters and their corresponding true *b*- and *a*-parameters. In the case of the *b*-parameters, bias values of 0.2500 (one fourth of the standard deviation of the distribution of the *b*-parameters) and 0.5000 (one half of the standard deviation of the distribution of the *b*-parameters) were set as the cutoff scores (Hu et al., 2008). In the specific metric of mean square errors, such values correspond to 0.0625 and 0.2500. Therefore, in this study, the same criteria for interpreting the values of MSE_b were used: a) $MSE_b \leq 0.06$: small, b) $0.06 < MSE_b \leq 0.25$: moderate, and c) $MSE_b > 0.25$: large (Hu et al., 2008).

Similar criteria were developed to interpret the size of the MSE_a. According to de Ayala (2009), reasonably good values of the *a*-parameter range from approximately 0.8 to 2.5. Under a normal distribution of values for the *a*-parameters, the expected median for this range would then be 1.6500. Therefore, for the *a*-parameters, bias values of 0.4125 (one fourth of a possible median for the *a*-parameters) and 0.8250 (one half of the standard deviation of the true *a*-parameters) were established as the cutoff scores. In the metric of mean square errors, such values correspond to 0.1702 and 0.6806. Consequently, the criteria for interpreting the values of MSE_a were: a) $MSE_a \leq 0.17$: small, b) $0.17 < MSE_a \leq 0.68$: moderate, and c) $MSE_a > 0.68$: large.

3.6 Data Analysis

The R software, version 2.12-2 (R Development Core Team, 2011) was used to fit the 3PL model, conduct IRT linking, and IRT true-score equating, as well to estimate the mean square errors for the *b*- and *a*-parameters. The 3PL model was fitted using the LTM package (Rizopoulos, 2006) through the function *tpm*. The *tpm* function fitted the 3PL model on the simulated datasets based on marginal maximum likelihood by using the Gauss-Hermite rule for the approximation of the required integrals (Rizopoulos, 2006) and returned the item parameters as part of the function outputs.

Next, the *plink* function from the PLINK package (Weeks, 2010) was utilized to conduct linking with the SL and HA methods to place item parameters on both Forms *X* and *Y* onto a common scale. Following, IRT true-score equating was carried out with the *equate* function from the same statistical package in order to relate number-correct scores across Forms *X* and *Y* (Weeks, 2010). Finally, mean square errors *b*- and *a*-parameters were calculated in R for each replication by applying the formulas described in the Evaluation Criteria section (Section 3.5).

CHAPTER 4
RESULTS

Convergence rates of the 3PL model calibrations and of the characteristic curve linking methods were near 99.8% in all conditions with the exception of the last outlier common-item condition (four common items with an interaction of extreme *b*- and *a*-parameters) for which the IRT characteristic curve linking results converged for very few replications, yet generating some aberrant results in terms of IRT estimation. Study results are presented in Tables 4-1 to 4-6.

Table 4-1 illustrates the MSE_b and MSE_a values under the first outlier condition, that is, when there were no outliers present in the common items. The MSE_b and MSE_a values for both methods were small when the two groups were equivalent. On the other hand, when the two groups differed in ability, both methods also had small MSE_a values but large MSE_b values. These results indicate that the SL and HA methods performed equally well in terms of estimating the *a*-parameters under equivalent and nonequivalent groups, yet they seemed to be more sensitive to the presence of nonequivalent groups when estimating the *b*-parameters.

Table 4-1. Mean square errors under the no outliers condition

	Stocking-Lord		Haebara	
Criteria	Equivalent Groups	Nonequivalent Groups	Equivalent Groups	Nonequivalent Groups
MSE_b	0.0111	1.6074	0.0156	1.4139
MSE_a	0.0071	0.0381	0.0079	0.0083

Table 4-2 shows the values for the condition in which the outliers were two common items with extreme *b*-parameter. Similar to the condition with no outliers, the MSE_b and MSE_a values for both SL and HA methods were small under the equivalent groups condition. As expected, the MSE_b values for both characteristic curve methods

were very large under the nonequivalent groups condition but their MSE_a values remained within a small range. Moreover, the MSE_b values for the two methods under the nonequivalent groups condition were larger than they were when there were no outliers in the common items (Table 4-1). Therefore, these results suggest that the presence of two outlier common items with outlier *b*-parameter had a specific negative effect on the calibration of the *b*- and *a*-parameters when the groups were not equivalent.

Table 4-2. Mean square errors under the second outlier condition

	Stocking-Lord		Haebara	
Criteria	Equivalent Groups	Nonequivalent Groups	Equivalent Groups	Nonequivalent Groups
MSE_b	0.0675	2.1942	0.0528	1.8502
MSE_a	0.0213	0.1039	0.0306	0.0376

The MSE_b and MSE_a values for the outlier condition in which there were two outlier common items with extreme *a*-parameter are shown in Table 4-3. For equivalent groups, the MSE_a value for the SL method was moderate and for the HA method was small. For nonequivalent groups, the MSE_a value for the SL method was large and for the HA method was small. Thus, the HA method seemed to perform equally well in either group ability condition under the presence of two outlier common items with extreme *a*-parameter. On the other hand, the MSE_b values for the SL method were moderate under the equivalent groups condition and very large under the nonequivalent condition. For the HA method, the MSE_b values were large in both the equivalent and nonequivalent groups conditions. Not only the characteristic curve methods were again sensitive to group equivalence, but they were also affected by the presence of two common items with outlier *a*-parameter towards the estimation of the discrimination parameters and towards the calibration of the difficulty parameters as well. The MSE_b

and MSE_a values significantly increased for all combinations of factors in relation to the previous two conditions.

Table 4-3. Mean square errors under the third outlier condition

	Stocking-Lord		Haebara	
Criteria	Equivalent Groups	Nonequivalent Groups	Equivalent Groups	Nonequivalent Groups
MSE_b	0.1827	1.8205	0.3136	1.7806
MSE_a	0.3778	0.7192	0.1230	0.1294

The MSE_b and MSE_a values for the outlier condition in which there were two outlier common items with an interaction of extreme *b*- and *a*-parameters are shown in Table 4-4. The MSE_a values for the SL method were moderate under both the equivalent and nonequivalent groups conditions. The MSE_a values for the HA method were small under the equivalent groups condition and moderate under the nonequivalent groups condition. The MSE_b values for the SL and HA methods were very large under the nonequivalent groups condition but moderate under the equivalent groups condition. In comparison to the effects caused by having no outliers at all, the presence of two common items with an interaction of outlier *a*- and *b*-parameters did have an overall impact on the errors of both parameter estimations. However, the interesting finding is that the MSE_b and MSE_a values were less sensitive to the presence of two common items with an interaction of outlier *a*- and *b*-parameters than they had been when two common items with outlier *b*- or *a*-parameter were introduced separately.

Table 4-4. Mean square errors under the fourth outlier condition

	Stocking-Lord		Haebara	
Criteria	Equivalent Groups	Nonequivalent Groups	Equivalent Groups	Nonequivalent Groups
MSE_b	0.1412	1.4644	0.2117	1.2808
MSE_a	0.2335	0.4216	0.0839	0.0765

Table 4-5 shows the MSE_b and MSE_a values for the outlier condition in which there were four outlier common items with extreme *b*-parameter. Under the equivalent groups condition, the MSE_a value for the SL method was small and for the HA method was moderate. Under the nonequivalent groups condition, the MSE_a value for both the SL and HA methods was small. With respect to the MSE_b values, under the equivalent groups condition such values were moderate for both SL and HA methods. However, the MSE_b values under the nonequivalent groups condition were very large for the two methods. As expected, the SL and HA methods performed worse under four outlier items with extreme *b*-parameter than under two outlier items. Interestingly, varying from two outlier *b*-parameters to four outlier *b*-parameters did not have a clear impact on the MSE_a values.

Table 4-5. Mean square errors under the fifth outlier condition

	Stocking-Lord		Haebara	
Criteria	Equivalent Groups	Nonequivalent Groups	Equivalent Groups	Nonequivalent Groups
MSE_b	0.2206	2.8661	0.1702	2.3255
MSE_a	0.0395	0.1704	0.0898	0.1029

Table 4-6 shows the MSE_b and MSE_a values for the outlier condition in which there were four outlier common items with extreme *a*-parameter. For both equivalent and nonequivalent groups, the SL method had large MSE_a values. In contrast, the HA method had moderate MSE_a values under both the equivalent and nonequivalent groups conditions. Similar to the outlier condition in which there were two outlier common items with extreme *a*-parameters, the HA method performed relatively well in either group ability condition under the presence of four outlier common items. With respect to the MSE_b values, the SL and HA methods exhibited large values under the two group ability conditions. These results may indicate that when there were more

common items with outlier *a*-parameter, they not only affected the discrimination parameter estimation but also tended to drag the difficulty parameter calibration into estimation errors.

Table 4-6. Mean square errors under the sixth outlier condition

	Stocking-Lord		Haebara	
Criteria	Equivalent Groups	Nonequivalent Groups	Equivalent Groups	Nonequivalent Groups
MSE_b	0.4182	1.3100	0.5594	1.7256
MSE_a	1.6095	1.9773	0.2081	0.2040

CHAPTER 5
DISCUSSION

The purpose of this study was to examine the effects of having common item with outlier *b*- and *a*-parameter estimates on two widely used IRT characteristic curve linking methods: the Stocking-Lord and Haebara methods. By conducting a series of Monte Carlo simulations, this study compared the performance of both methods under different outlier conditions based on the types of outliers, the number of outliers, and group ability distributions.

Several studies had previously compared the Stocking-Lord and Haebara methods under the common-item nonequivalent groups design, although they did not show consistent conclusions about which method is more preferable across different testing conditions (Baker & Al-Karni, 1991; Béguin & Hanson, 2001; Béguin et al., 2000; Hanson & Béguin, 2002; Kim & Cohen, 1998; Kim & Kolen, 2006; Lee & Ban). On the other hand, there have been very few studies examining the effects of estimation errors and outlier common items on the Stocking-Lord and Haebara methods (Hu et al., 2008; Kaskowitz & de Ayala, 2001; Michaelides, 2003; Sukin & Keller, 2008). Moreover, no previous studies have addressed the specific effects of common items with outlier *a*-parameter or with an interaction of outlier *b*- and *a*-parameters.

Overall, this study accomplished its main purpose. The first research question addressed whether the Stocking-Lord and Haebara linking methods differed with respect to bias in the linking process when there were common items with outlier *b*-parameter. Mean square errors for the *b*-parameters were very large for both characteristic curve methods under the nonequivalent groups condition. Although the Stocking-Lord and Haebara methods were already sensitive to group nonequivalence

when there were no outliers in the common items, both methods became more biased in the calibration of the *b*-parameters when there were two common-item outliers with extreme *b*-parameters. When the number of common items with outlier *b*-parameter increased from two to four, mean square errors for the *b*-parameters slightly increased under the equivalent groups condition, but significantly increased under the nonequivalent groups condition. When comparing the Stocking-Lord and Haebara linking methods to one another in the presence of extreme common-item outliers with extreme *b*-parameter, the Haebara method performed slightly better, yet still poorly, than the Stocking-Lord method when there were two and four outlier common items. These results tend to confirm Lee and Ban's (2010) and Michaelides' (2003) findings on the Haebara method over the Stocking-Lord method.

The second research questions addressed whether the Stocking-Lord and Haebara linking methods differed with respect to bias in the linking process when there were common items with outlier *a*-parameter. Under the equivalent and nonequivalent groups conditions, the Haebara method consistently outperformed the Stocking-Lord method by generating smaller mean square errors for the *a*-parameters. Interestingly, both methods performed poorly in terms of the mean square errors for the *b*-parameters, which indicates that linking estimation errors in the *a*-parameters also affect the calibration of the *b*-parameters to a great extent. When the number of common items with outlier *a*-parameter increased from two to four, the Haebara method again outperformed the Stocking-Lord method by showing low to moderate mean square errors for the *a*-parameters. Similar to when there were two outlier common items with extreme *a*-parameter, the mean square errors for the *b*-parameters were very large for

both methods. Kolen and Brennan (2004) pointed out that the characteristic curve methods do not explicitly account for the error in estimating item parameters and ignoring the error in parameter estimates might lead to problems in the estimation of the *A* and *B* coefficients, and therefore, in the estimation of the linking relationship between different test forms.

The third research questions addressed whether the Stocking-Lord and Haebara linking methods differed with respect to bias in the linking process when there were common items with an interaction of outlier *b*- and *a*-parameters. Although two common-item scenarios for this condition were part of the original linking design, only the condition with two common items with outlier both *b*- and *a*-parameters was examined in this study. When there were four outlier common items with extreme *b*- and *a*-parameters, the Stocking-Lord and Haebara methods had serious convergence problems when estimating the *A* and *B* coefficients, as well as the posterior true-score equating process. Therefore, those results had to be discarded from any type of further analysis.

Instead, when there were two common items with outlier *b*- and *a*-parameters, both the Stocking-Lord and Haebara methods performed poorly in terms of mean square errors for the *b*- and *a*-parameters under the nonequivalent groups condition, though they were less sensitive than in the previous conditions where both item parameters were analyzed independently. This difference between the two scenarios is small and is likely due to sampling variability of the common items.

The last research question addressed whether the performance of the Stocking-Lord and Haebara linking methods was affected by changes in the ability distribution of

number of outliers. Based on the findings, the Stocking-Lord and Haebara methods tended to perform better under the equivalent groups condition than under the nonequivalent groups condition. These results confirm what had also been found in previous studies (Hanson & Béguin, 2002; Hu et al., 2008; Lee & Ban, 2010; Sukin & Keller, 2008). However, special attention must be given to the nonequivalent groups conditions because in real testing situations it is more common to find groups with large ability differences (i.e., retesters vs. first-time testers or private vs. public elementary school students). In addition, as Kolen and Brennan (2004), a large ability difference can better show which equating methods are most sensitive to group differences.

With respect to the number of outliers, both the Stocking-Lord and Haebara linking methods performed worse in terms of mean square errors for the *b*- and *a*-parameters when having four outlier common items than when having two outlier common items, except when there were four common items with outlier *a*-parameters because the Haebara method performed equally well under both ability distribution levels.

5.2 Conclusion

In conclusion, the Stocking-Lord method is not a robust IRT-based method to the effects of the presence of common items with outlier *b*- and *a*-parameters under the common-item nonequivalent design. On the other hand, the Haebara method is not a robust method to the effects of the presence of common-items with outlier *b*-parameter under the common-item nonequivalent design, although it is more robust method to the effects of common items with outlier *a*-parameters. Finally, neither the Stocking-Lord nor the Haebara method is a robust IRT-based linking method to the effects of the presence of common items with an interaction of outlier *b*- and *a*-parameters.

5.1 Limitations and Suggestions for Future Research

There are multiple limitations in this Monte Carlo simulation study involving the Stocking-Lord and Haebara linking methods. As with any simulation study, some restraints need to be taken before drawing conclusions because of the small number of conditions investigated. In this simulation study, the results are appropriate to only the use of two test forms, equivalent and nonequivalent groups, two and four outlier common items, two IRT characteristic curve linking methods, and one IRT model (the 3PL model). Future simulation research including more test forms (and more items) and different IRT models is necessary.

Another limitation of this simulation study has to do with how realistic the simulated data were. Although this study employed real item parameters from the 2005 and 2006 FCAT 4th-Grade Mathematics test administration and cautions were taken to ensure that the simulated responses would match real test responses, the simulation of data under the 3PL model is not an easy task because problems in estimating the *c*-parameter can influence the estimation of the items' other parameters (de Ayala, 2009). Furthermore, the *c*-parameter is considered to be more reflective of a person characteristic rather than a genuine item characteristic, which may have an impact when simulating data under the 3PL model (de Ayala, 2009; Hambleton & Swaminathan, 1985). Therefore, future work using actual student responses is needed in order to determine the real effects of outlier items on the Stocking-Lord and Haebara methods.

The current study was also limited to analyzing the effects of outliers located on the left side of the straight line of the *b*- and *a*-parameters, where the parameters from Year 2 (Form *Y*) were two points lower than in Year 1 (Form *X*). Future studies should

also focus on other types of outlier characteristics in order to establish more conclusive remarks about the performance of characteristic curve linking methods. Likewise, this study was only limited to examining the effects of two and four common items (out of 22 common items in total). Thus, it would be interesting to determine the effects of having a larger percent of common items with outlier *b*- and *a*-parameters.

On the other hand, this study did not investigate linking or equating methods that do consider the presence of outlier items as other research studies have done in the past (Hu et al., 2008; Michaelides, 2003). Therefore, it would be of interest to investigate the performance of other methods, such as the delta-plot method (Michaelides, 2010), under the presence of outlier common items with extreme *a*-parameters.

Another limitation of the present study is strictly related with the comparability of tests scores across test forms. This study did not examine the effects of common items with outlier *b*- or *a*-parameters on the true score of the examinees, and therefore, it was not possible to quantify the effect of such types of common items on the actual estimation of the ability levels of the examinees. Further studies ought to consider the estimation of mean square errors for the true scores in order to determine the magnitude of the bias on the estimation of the examinees' ability levels. Similarly, this study only considered the presence of outlier common items as part of an external anchor test, which does not consider the performance of the examinees on such items as part of their total observed score. Therefore, future research should also analyze the effects of having outlier common items in an internal anchor test on both the test calibration process and true score estimation.

REFERENCES

Angoff, W. H. (1982). Summary and derivation of equating methods used at ETS. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 55-70). New York: Academic Press.

Baker, F. B., & Al-Karni, A. (1991). A comparison of two procedures for computing IRT equating coefficients. *Journal of Educational Measurement*, *28*(2), 147-162.

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

Béguin, A. A., Hanson, B. A., & Glas, C. A. W. (2000, April). *Effect of multidimensionality on separate and concurrent estimation in IRT equating*. Paper presented at the annual meeting of the National Council of Measurement in Education, New Orleans.

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

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

Cook, L., & Eignor, D. (1991). An NCME instructional module on IRT equating methods. *Educational Measurement: Issues and Practice, 10*(3), 37-45.

Crocker, L. & Algina, J. (2008). *Introduction to classical and modern test theory.* Mason, OH: Cengage Learning.

de Ayala, R. J. (2009). *The theory and practice of item response theory.* New York: The Guilford Press.

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

Florida Department of Education (1996). *Sunshine State Standards.* Retrieved September 20, 2010, from the Florida Department of Education website: http://www.fldoe.org/bii/curriculum/sss/

Florida Department of Education (January, 2006). *Technical report for 2005 FCAT test administrations.* Unpublished. Tallahassee, FL: Author.

Florida Department of Education (January, 2007). *Technical report for 2006 FCAT test administrations.* Unpublished. Tallahassee, FL: Author.

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

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

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamental of item response theory (Measurement methods for the social sciences).* California: Sage Publications, Inc.

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.

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

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

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

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.

Kim, S. –H., & Kolen, M. J. (2006). Robustness to format effects of IRT linking methods for mixed-format tests. *Applied Measurement in Education, 19*(4), 357-381.

Kolen, M. J. (2004). Population invariance in equating and linking: Concept and history. *Journal of Educational Measurement, 41*(1), 3-14.

Kolen, M. J., & Brennan, R. L. (2004). Test equating, scaling, and linking. New York: Springer-Verlag.

Lee, W. –C., & Ban, J. –C. (2010). A comparison of IRT linking procedures. *Applied Measurement in Education*, *23*(1), 23-48.

Linn, R. L., Levine, M. V., Hastings, C. N., & Wardrop, J. L. (1981). An investigation of item bias in a test of reading comprehension. *Applied Psychological Measurement*, *5*(2), 159-173.

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

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

Michaelides, M. P. (2003). *Effects of common-item selection on the accuracy of item response theory test equating with nonequivalent groups* (Unpublished Doctoral dissertation). Stanford University, Stanford, CA.

Michaelides, M. P. (2010). Sensitivity of equated aggregate scores to the treatment of misbehaving common items. *Applied Psychological Measurement*, *34*(5), 365-369.

Michaelides, M. P., & Haertel, E. H. (2004). *Sampling of common items: An unrecognized source of error in test equating* (CSE Report 636). Los Angeles: The Regents of The University of California.

R Development Core Team. (2011). *R: A language and environment for statistical computing* [Computing Software]. Vienna, Austria: R Foundation for Statistical Computing. Program available at www.r-project.org

Rizopoulos, D. (2006). Ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software*, *17*(5), 1–2.

Sukin, T., & Keller, L. (2008). The effect of deleting anchor on the classification of examinees. NERA Conference Proceedings 2008. Downloaded on February 20, 2011 from http://digitalcommons.uconn.edu/nera_2008/19

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

Weeks, J. P. (2010). Plink: An R Package for Linking Mixed-Format Tests Using IRT-Based Methods. *Journal of Statistical Software*, *35*(12), 1-33.

BIOGRAPHICAL SKETCH

Francisco Andres Jimenez was born in Santiago, Chile in 1982. He received a Bachelor of Science degree in Psychology from the University of Chile in 2005, obtained his professional title of Psychologist from the University of Chile in 2007, and received his Master of Arts degree in Education from the University of Florida in 2011. He has been an Assistant Professor in the Department of Psychology at the University of Chile since 2008. He is currently a Fulbright doctoral student in the Research and Evaluation Methodology Program in the College of Education at the University of Florida.