# THE IMPACT OF ANCHOR ITEM EXPOSURE ON MEAN/SIGMA LINKING AND IRT TRUE SCORE EQUATING UNDER THE NEAT DESIGN

By

Moatasim A. Barri

B.S., King Abdul Aziz University
M.S.Ed., The University of Kansas
Ph.D., The University of Kansas

Submitted to the Department of Psychology and Research in Education, School of Education, and the Faculty of the Graduate School of the University of Kansas in partial fulfillment of the requirements for the degree of Master of Science in Education.

Bruce Frey, Ph.D., Chairperson,
Department of Psychology and Research in Education

William Skorupski, Ph. D.
Department of Psychology and Research in Education

Vicki Peyton, Ph.D.
Department of Psychology and Research in Education

Date defended:_____

The Thesis Committee for Moatasim A. Barri
Certifies that this is the approved version of the following thesis:

THE IMPACT OF ANCHOR ITEM EXPOSURE ON MEAN/SIGMA LINKING AND IRT TRUE SCORE EQUATING UNDER THE NEAT DESIGN

Chairperson Bruce Frey, Ph.D.

Date approved: _____

ABSTRACT

To compare examinees' true ability and their actual competence on the content being measured across different test administrations, test scores must be equated. One of the most common equating designs is called the nonequivalent anchor test (NEAT) design. This equating design requires two forms of a test, each of which is given to a group of examinees one year apart. The two forms have a set of items in common, usually called the anchor set, in order to control for differences in examinee ability. The anchor set can be treated as internal or external according to whether or not examinees' responses contribute to their total score. However, the anchor set is subject to exposure when it is used repeatedly, which most likely becomes a serious threat to test fairness and validity. Therefore, from time to time, the items in the anchor set must be evaluated for exposure. This study employed a Monte Carlo investigation to evaluate the impact of internal anchor item exposure on the equating process under the NEAT design. The study addressed a general scenario in which two forms of a small-scale dichotomously scored test were given to two small groups of examinees. Since mean/sigma linking and true score equating are main components of the equating process in the item response theory (IRT), the recovery of equating true scores and linking coefficients, slope and intercept, were assessed under various combinations of testing conditions using bias and mean squared error (MSE). Three testing conditions were manipulated in this study: (a) the number of exposed anchor items, (b) the percentage of examinees with preknowledge of the exposed anchor items, and (c) the difference in the means of ability distributions of groups taking the original form and new form. In each combination of testing conditions, the simulation process was replicated 100 times. The study results indicated that anchor

item exposure caused all examinees to receive inflated equating true scores. When anchor items were subject to low levels of exposure, the accuracy of equating true scores was still perturbing, while high levels of exposure distorted the test scores completely. The anchor item exposure became a serious threat to the test fairness to the extent that unqualified examinees might receive an unfair benefit over qualified examinees who completed an unexposed test form.

ACKNOWLEDGEMENTS

This thesis could not have been written without the support of God, my beloved immediate and extended families, and many helpful people.

O Allah, I thank you for everything you have provided me with to finish my academic work. To my darling parents, mom and dad, thank you for your love, support, patience, and tolerance. Thank you for persistent prayers asking Allah to help me get successful in my academics and to guide me toward the right path. Thank you for trying to understand why I did not always phone you. I will try to make it up to you.

To my darling and charming wife, Nahed, thank you for your love, support, and patience. Thank you for being always there for me. Thank you for all the times I did not join you for enjoyment. Thank you for all sacrifices you made for me to accomplish my academic goals. A very special thanks to you for making it possible to manage between my family duties and academics.

To my cute daughters, Hatoon, Hadwa, Hadeel, and Roaa, and my handsome son, Asaad, of whom I am so proud, thank you all.

To my beloved sisters, brothers, mother-in-law, and brothers-in-law, thank you all for your support, encouragement, and persistent prayers.

To my father-in-law, Ibrahim Barri, I wish you could have stayed around long enough to attend the completion of my thesis. I miss you a lot, my darling. May Allah forgive you, have mercy on you, and give you the best place in Paradise. Amen.

Many helpful people have contributed to the success of this research study. Deep appreciation and heartfelt thanks go to my thesis chair, Dr. Bruce Frey. Words cannot

express my appreciation and gratitude to him. I wish to deliver my special thanks to him for the time he devoted to reading and refining my thesis.

A deep appreciation goes to my thesis committee members, Dr. William Skorupski, and Dr. Vicki Peyton. Your support and expertise through this project were very much appreciated.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF APPENDICES

# CHAPTER 1

# INTRODUCTION

When a test is administered on different occasions, a test developer usually creates multiple forms of the test in order to rule out the release of test items to examinees. Even though the test developer cautiously attempts to create parallel forms of the test, the forms still vary in terms of difficulty to the extent that they might disadvantage some examinees in an unfair manner. Therefore, the test developer employs an equating design in order to assess the examinees fairly. One of the most common equating designs is called the nonequivalent anchor test design (NEAT; Holland, 2007; von Davier, Holland, & Thayer, 2004) or the common-item nonequivalent groups design (Kolen & Brennan, 1995), where anchor items or common items are placed on each form of the test for linking, a critical stage in the Item Response Theory (IRT) equating process. These anchor items do not remain invariant in terms of difficulty over different testing occasions (Wells, Subkoviak, & Serlin, 2002) because some of these items are exposed to examinees before the day of test administration. For example, a group of examinees intentionally memorize items during the test administration and then share these items with another group who will be administered the test subsequently. As a result, the possibility of anchor item compromise is greatly enhanced and becomes a major threat to the test fairness to the extent that it might possibly alter the test scores for some examinees who have already experienced the items (Han & Guo, 2011). For this reason, the anchor items must be evaluated frequently for compromise. Since the anchor items are part of the equating process under the NEAT design, it is necessary to evaluate

the potential impact of anchor item exposure on the equating process in terms of the accuracy of recovered scaling coefficients and IRT equating true scores.

To compare examinees' true ability and their actual competence on the content being measured across different test administrations, test equating must be conducted to produce comparable test scores. However, the equating process cannot be conducted unless item parameters across different test administrations are placed on a common scale, which can be accomplished using the linking method. There are a number of linking methods that can be used to obtain the scaling coefficients for placing ability and item parameters from different calibrations on a common metric. Among the most popular of these methods are mean/sigma method (Marco, 1977), mean/mean method (Loyd & Hoover, 1980), and test characteristic curve method (TCC; Haebara, 1980; Stocking & Lord, 1983). In this study, the mean/sigma linking method was considered. A brief description of the mean/sigma linking method is provided in the literature review section.

Once the item parameters are placed on the common scale, the IRT true score equating process can be conducted to relate number-correct scores on different forms. In this process, the true score on one form that corresponds to a given ability value is considered to be equivalent to the true score on another form that corresponds to that ability value (Kolen & Brennan, 1995). The IRT true score equating procedure is described in detail in the literature review section.

Few studies have evaluated the impact of exposed anchor items on the IRT equating process using Monte Carlo investigations (Jurich, DeMars, & Goodman, 2012; Jurich, Goodman, & Becker, 2010). However, studies of item exposure up to this date

have not placed a focused attention on the impact of exposed anchor items on the recovery of differences in the mean of ability distributions across two test administrations, which is the main goal of the current study. The impact of exposed anchor items was assessed under various testing conditions through a Monte Carlo investigation.

## Study Significance

This simulation study filled in a gap in the literature by examining the impact of common item exposure on the accuracy of scaling coefficients and equating true scores obtained through the IRT equating process according to the NEAT design. Due to crucial decisions made from a high-stakes test, this simulation study informs test developers who wish to create parallel forms of tests which assess examinees fairly.

## Study Hypotheses

The following hypotheses were investigated:

1. The difference in the means of examinee ability distributions across test administrations has an influence on the accuracy of scaling coefficients.
2. The difference in the means of examinee ability distributions across test administrations has an influence on the accuracy of the IRT equating true scores.

## Study Limitations

As with any Monte Carlo investigation, care must be taken when any generalization is made. The following limitations are applicable to this simulation study:

1. The *ltm* package in R was utilized to generate dichotomous item response data under the Rasch model. This model considers only the difficulty property of items and ignores other properties such as discrimination and guessing. The

Rasch model only deals with an assessment containing dichotomously scored items and cannot handle tests that are composed of polytomously scored items.

2. The study findings were limited to utilization of one equating design, known as the NEAT Design, and paid no heed to other data collection designs such as the Single-Group (SG) Design, the Equivalent-Groups (EG) Design, and the Counterbalanced (CB) Design (see Kolen & Brennan, 1995 or von Davier et al., 2004 for further details).

3. Research on transformations of IRT scales shows a variety of procedures that can be used for placing item and ability parameters from different tests on a single scale. The current simulation study limited its application to one procedure, described by Marco (1977) and called here as the mean/sigma method by Kolen & Brennan (1995), which uses the mean and standard deviations of the difficulty parameters from the common items to obtain the linking coefficients (intercept and slope) required for the linear transformation.

4. The literature review on equating designs showed that the NEAT equating design has two variations according to two kinds of anchor set: internal and external (Kolen & Brennan, 1995; von Davier et al., 2004). The current simulation study limited its application to the internal anchor set where examinees' responses to anchor items contribute to their total scores on the test forms to be equated.

5. Research on item parameter estimation through equating indicated that item parameters for two test forms can be estimated concurrently or separately according to whether the estimation is done through one run or two separate runs of the software being used for the IRT analysis (Hanson & Beguin, 2002; Kim &

Cohen, 1998; Petersen, Cook, & Stocking, 1983; Wingersky, Cook, & Eignor, 1987). The study limited its application to settings where the item-parameter estimates for two forms of a test were produced using two separate runs of the IRT analysis.

# CHAPTER 2

# LITERATURE REVIEW

## Equating

Equating plays an important role in the field of education. Most testing programs create multiple forms of a test in order to suppress any potential source of item exposure so that the test scores accurately reflect each test taker's true ability. Unfortunately, these multiple forms are subject to diversity in terms of difficulty, which unfairly disadvantage some examinees. To assess examinees accurately, equating designs are utilized in an effort to assure that examinees can be expected to earn the same score irrespective of the test form being administered. Equating occurs when test scores from different forms of the same test are compared. When test forms are built to be the same in content and difficulty, the equating process is utilized to adjust for differences in difficulty so that the forms can be used "interchangeably." After the equating process, a test score could accurately reflect examinee's true ability and could determine if s/he is fairly authorized to be engaged in an occupation or is honestly accepted into an educational institution.

## IRT Equating Process

A typical simulated equating process can be conducted through IRT in four main steps. First, ability and item parameters are generated according to the equating design or the data collection design being chosen. Second, the ability and item parameters being generated are estimated (calibrated) either concurrently across forms of a test or separately for each form. Third, the ability and item parameters being estimated are placed on a common scale, which can be accomplished using a linear transformation procedure, usually called a linking method. Fourth, once the item parameters are placed

on the common scale, the IRT true score equating can be conducted to link an equating true score on form 1 corresponding to a given ability value with an equating true score on form 2 associated with that ability value. The four steps of the equating process are described in some details next.

**IRT Equating Design**

The current simulation study focuses on the non-equivalent groups with anchor test design, which is often referred to as the NEAT design, for equating the test scores on multiple forms of a test. The NEAT design often is used "when more than one form per test date cannot be administered because of test security or other practical concerns" (Kolen & Brennan, 1995, p. 18). This design requires two different groups of examinees and two different forms of a test that have a set of items in common. For instance, a sample of examinees from population $P_1$ takes the original form (OF) and another sample of examinees from population $P_2$ takes the new form (NF). The two samples are administered a set of common items, A, usually called the anchor set, which controls for differences in examinee ability across different test administrations. The data collection for the NEAT design is described in Table 1.

Table 1

*Non-equivalent Groups with Anchor Test Design*

| Population | Sample | Original Form (OF) | New Form (NF) | Anchor A |
|---|---|---|---|---|
| $P_1$ | 1 | ✓ | | ✓ |
| $P_2$ | 2 | | ✓ | ✓ |

*Note. Modified from Table 2.4 of von Davier et al. (2004, p. 33).*

The anchor set in the NEAT design can be either internal or external according to the contribution of the anchor items to the examinee's total score on the test. When the anchor set serves as internal, examinees' responses to the anchor items contribute to their total score on the test. When the anchor set is treated as external, the score on the external anchor items does not count on the examinee's total score on the test. The overall content and difficulty of the anchor items should relatively match the other items of the test (Kolen & Brennan, 1995; von Davier et al., 2004). The anchor items should have similar locations in the original and new forms to help reach adequate equating.

**IRT Item Calibration Design**

The literature on IRT equating procedures shows that item parameters can be calibrated concurrently or separately according to whether the item calibration is performed through one run or two runs of the computer program being used for the IRT analysis (Hanson & Beguin, 2002; Kim & Cohen, 1998; Petersen et al., 1983; Wingersky et al., 1987). Concurrent calibration occurs when item parameters for two forms of a test are calibrated through a single run of the calibration software, while separate calibration occurs when item parameters for each form of a test are calibrated in a separate run of the calibration software. Wingersky et al. (1987) and Petersen et al. (1983) both compared IRT calibration methods for dichotomously scored tests and found that the concurrent calibration provided more accurate equating results than did the separate calibration. In their study, Hanson and Beguin (2002) found that concurrent calibration generally performed better than separate calibration; however, their study results were still insufficient to favor one method over another. In comparison with Kim and Cohen's (1998) study results, the separate calibration was found to be better or similar to the

concurrent calibration. In the current simulation, the ability and item parameters were estimated separately for each form of a test administered in the NEAT equating design.

**IRT-Based Linking Method**

Once ability and item parameters are estimated for both forms of the test based on the NEAT equating design, a linear transformation is required to place the estimated parameters on a single scale. The literature shows a number of linking or transformation methods that can be used to place ability and item parameters from different calibrations on a common scale. These methods are as follows: mean/sigma method (Marco, 1977), mean/mean method (Loyd & Hoover, 1980), and test characteristic curve method (TCC; Haebara, 1980; Stocking & Lord, 1983). The current study used the mean/sigma method for developing this common scale. This method uses the anchor items across the OF and NF to determine the liner transformation needed to convert the estimated parameters from the NF scale to the OF scale. In the mean/sigma method, the linear transformation matches the mean and standard deviation of anchor item b-values across the OF and NF to obtain the appropriate scaling coefficients, slope (X) and intercept (Y), as follows:

$$X = \frac{\sigma_{b_{OF}}}{\sigma_{b_{NF}}}, \text{ and}$$

$$Y = \mu_{b_{OF}} - X * \mu_{b_{NF}},$$

where $\sigma_{b_{OF}}$ and $\mu_{b_{OF}}$ are, respectively, the standard deviation and mean of the item difficulty parameters of the anchor items for the OF, and $\sigma_{b_{NF}}$ and $\mu_{b_{NF}}$ are, respectively, the standard deviation and mean of item difficulty parameters of the anchor items for the new form. Once the appropriate slope, X, and intercept, Y, have been computed; the ability and item difficulty parameters derived from the NF are transformed as follows:

$$b_j^* = X * b_j + Y, \text{ and}$$

$$\theta_i^* = X * \theta_i + Y,$$

where $b_j^*$ and $b_j$ are, respectively, the rescaled and initial item difficulty parameters for item $j$ and $\theta_i^*$ and $\theta_i$ are, respectively, the rescaled and initial ability parameters for examinee $i$.

**IRT True Score Equating**

Once the item difficulty parameters are on the common scale, IRT true score equating is used to develop a relationship between number-correct scores, sometimes called true scores, on the OF and NF. In this process, the true score on the NF corresponding to a given $\theta$ is considered to be equivalent to the true score on the OF corresponding to that $\theta$ (Kolen & Brennan, 1995). The IRT true score equating function depends on the test characteristic curves for the OF and NF, which give number-correct true score as a function of the ability variable (Hanson & Beguin, 2002). The true scores on the NF and OF, which correspond to $\theta$, are defined as the test characteristic curve values for the NF and OF at a particular ability value. Mathematically, the IRT true scores on the NF and OF can be respectively defined as:

$$\tau_{NF}(\theta) = \sum_{j=1}^{n} p_j$$

$$\tau_{OF}(\theta) = \sum_{j=1}^{n} p_j$$

where n is the number of items, $p_j$ is the item characteristic curve value at a particular ability value, $\theta$, and is mathematically defined as

$$p_j = \frac{e^{(\theta - b_j)}}{1 + e^{(\theta - b_j)}},$$

and $\theta$ is the ability value corresponding to $\tau_{NF}$ and $\tau_{OF}$.

An IRT true score equating process is conducted through IRT in three main steps (Kolen & Brennan, 1995):

1. Specifying a true score on the NF ($\tau_{NF}$), which is an integer greater than or equal to 0 and less than or equal to n,
2. Finding a $\theta$ value that corresponds to the $\tau_{NF}$ specified in step 1, and
3. Calculating a true score on the OF ($\tau_{OF}$) that corresponds to that $\theta$ value.

To find the $\theta$ value that corresponds to the $\tau_{NF}$ requires solving a nonlinear equation through an iterative process using the Newton-Raphson method. The $\theta$ value is calculated as:

$$\theta = \theta^{\circ} - \frac{f(\theta^{\circ})}{f'(\theta^{\circ})},$$

where $\theta^{\circ}$ is an initial value for $\theta$ and is chosen by guessing, $f(\theta^{\circ})$ is a function of the variable $\theta^{\circ}$ and is defined as

$$f(\theta^{\circ}) = \tau_{NF} - \sum_{j=1}^{n} p_j,$$

and $f'(\theta^{\circ})$ is the first derivative of the function with respect to $\theta^{\circ}$ and is expressed as

$$f'(\theta^{\circ}) = -\sum_{j=1}^{n} p_j',$$

where

$$p_j' = 1.7\,(1 - p_j)p_j.$$

The iterative process is repeated until the value of $f(\theta^{*})$ is close to 0 or until $\tau_{NF}$ nearly equals $\sum_{j=1}^{n} p_j$ at a specified level of precision. If $f(\theta^{*})$ is not close to 0, for example, in the first iteration, then the $\theta$ value calculated in the first iteration is redefined as $\theta^{*}$ for the second iteration and so on. Once $\theta$ that corresponds to $\tau_{NF}$ is determined, $\tau_{OF}$ is found by calculating the sum of the item characteristic curve values for the OF items at the $\theta$ level corresponding to $\tau_{NF}$.

## The Influence of Test Exposure in IRT

Test exposure occurs when examinees gain preknowledge of answers to some of the test items prior to testing day. The scores obtained on the exposed test might be improperly inflated and do not reflect the examinees' true ability and their actual competence on the content being tested (McLeod & Lewis, 1999; Stocking, Ward, & Potenza, 1998; Zara, 2006). Test exposure can come from various different sources: (a) when one examinee shares test items with a future examinee, (b) when a group of examinees construct a bank of operational test items and spread them to others, or (c) when a set of disclosed test items, which are reused in subsequent test administrations, are intentionally reviewed by a test developer (Segall, 2002). Some recent research has investigated the impact of exposed items on the recovery of examinee ability parameters using simulation studies. Guo, Tay, and Drasgow (2009) investigated the effect of exposed items on ability estimates at different test systems and found that exposed items led to large score gains for low-ability examinees, bringing about scores that inaccurately reflect the examinee's true ability. Yi, Zhang, and Chang (2008) employed the impact of exposed items on ability estimates as an evaluation criterion in determining the detrimental effects of organized item theft in Computerized Adaptive Testing (CAT).

The results of this study showed that the examinees' ability estimates were considerably overestimated in the presence of exposed items, especially the testing condition that included examinees with a low level of ability or the condition with the organized item theft group.

Few studies have investigated the impact of exposed anchor items on the IRT equating process using Monte Carlo investigations. Jurich, DeMars, and Goodman (2012) conducted a simulation study examining the impact of cheating on test characteristic curve linking (TCC; Stocking & Lord, 1983) and IRT true score equating under the NEAT design. The results of this study indicated that the recovery of the estimated TCC linking constants and equated scores became less accurate as a result of an increase in either the proportion of exposed anchor items or proportion of cheaters. However, studies of item exposure to date have not examined the impact of exposed anchor items on the recovery of differences in the mean of ability distributions across two testing occasions, which is the ultimate goal of the current study. Jurich et al.'s (2012) study was limited to using test length, which reflects a larger standardized test. Since test length had less of an effect on the accuracy of equating results (Wu, 2012) and the accuracy of recovered ability and item parameters, the current study considered a different scenario in which a small sample of examinees are administered dichotomously scored items which, relatively, represent a typical-length classroom examination.

# CHAPTER 3

# METHOD

A simulation study was conducted to evaluate the potential impact of exposed common items on the scaling coefficients and number-correct true scores obtained from the equating process under the NEAT design. This simulation deals with a common scenario in which two parallel forms of a small-scale test, the original form (OF) and new form (NF), are administered on two testing occasions and are offered to different groups of examinees. The two forms have a set of items in common in order to meet the NEAT equating design assumptions. The items on the OF were unique, meaning that they were not subject to any source of item exposure, while common items on the NF were under the influence of possible item exposure.

## Test Generation

The latent trait model (*ltm*) package (Rizopoulos, 2006) in R was utilized to generate dichotomous item response data under the Rasch model (Rasch, 1980), usually called the one-parameter logistic (1-PL) model. In the 1-PL model, the probability of a correct answer is defined as a logistic function of the difference between the examinee's ability and item difficulty parameter. The Rasch model can be expressed as:

$$P(X_{ij} = 1) = \frac{e^{(\theta_i - b_j)}}{1 + e^{(\theta_i - b_j)}}$$

$$i = 1, 2, \ldots \ldots \ldots, N$$

$$j = 1, 2, \ldots \ldots \ldots, n$$

where $P(X_{ij} = 1)$ is the probability that an examinee $i$ with ability $\theta$ responds to item $j$ correctly, $b_j$ is the difficulty trait for item *j*, *N* is a number of examinees, and *n* is a

number of dichotomous items. Item response data were generated separately to create two forms of a dichotomously scored test. The number of items administered was held constant. Two forms of a test, each of which included 50 items, were generated for each replication of this simulation study. The 50-item test reflects a typical-length classroom examination. The first form of the test represented the original form (OF) from which the scale of scores was set, while the second form represented the new form (NF), which was equated to the OF. Since this study utilized the NEAT design to equate scores from the NF to the OF, a set of test items functioned as an anchor set across the two forms. As demonstrated in the literature (Budescu, 1985; Kolen & Brennan, 1995; Li, 2012; Ricker & von Davier, 2007; Yang & Houang, 1996), the longer the anchor length is, the smaller the linking and equating errors are. This study acted in accordance with Angoff's (1984) recommendation to use at least 20% of the length of the operational items for the anchor items. Thus, 10 items were anchored across the two forms, while the remaining 40 test items were distinctive in each form. The last 10 items (items 41 through 50) on the OF and NF were considered common items, and were in the same location in each form to ensure adequate equating. In each testing condition, the anchor set was treated as internal where examinees' responses to anchor items contribute to their total scores. It was assumed that the 10-item anchor set produced should proportionally match the content and average difficulty parameter of the distinctive item set; otherwise, systematic errors are introduced in recovering scaling coefficients during the equating process (Haberman & Dorans, 2009; Wei, 2010). The item difficulty parameter for both the anchor and distinctive item sets were sampled separately from a normal distribution, $b \sim N(0.00, 1.00)$, to create the OF and the NF of the test.

## Ability Generation

In the current study, the sample size was held constant to rule out its influence on recovering true ability. The literature suggested that the 1-PL model requires a small sample size to estimate stable and accurate item parameters in comparison with the other IRT models, such as the 3-PL model (Kim, Barton & Choi, 2010; Yen & Fitzpatrick, 2006). The *ltm* package in R was used to generate a sample size of 500 simulated examinees. This small sample size represents a typical size of class section. The latent ability of examinees taking the OF was sampled from a standard normal distribution, $\Theta_{OF} \sim N(0.00, 1.00)$, while the ability distribution of examinees taking the NF was varied according to the mean of examinee ability. Five levels of the NF ability distribution were considered in the current simulation study as follows: (a) $\Theta_{NF} \sim N(-0.50, 1.00)$, (b) $\Theta_{NF} \sim N(-0.25, 1.00)$, (c) $\Theta_{NF} \sim N(0.00, 1.00)$, (d) $\Theta_{NF} \sim N(0.25, 1.00)$, and (e) $\Theta_{NF} \sim N(0.50, 1.00)$.

## Simulation of Anchor Item Exposure

Based on the NEAT equating design, the anchor items are placed on the OF and NF, and different groups of examinees are administered the two forms. When a group of examinees intentionally memorize some anchor items during the OF administration and then share them with other examinees who will be administered the NF subsequently, these anchor items become exposed. To simulate the magnitude of anchor item exposure, two testing conditions were considered: (a) the number of exposed items and (b) the proportion of examinees with preknowledge of these items (Guo et al., 2009; Jurich et al., 2012; Yi, Zhang, & Chang, 2008). First, the number of exposed anchor items was set at two or 10 items. The former reflects a scenario where examinees have access to a small

amount of test questions, while the latter reflects a worst-case scenario where examinees may have preknowledge of a large amount of test questions or may have access to the entire forms administered previously. Second, the percentage of examinees with preknowledge of the exposed anchor items was investigated at four different levels, 5%, 10%, 15% and 20%.

The probability of answering the exposed anchor item correctly for any examinee who has preknowledge of that item was set to 1. This reflects a real scenario where the simulee can remember the right response to the exposed common item regardless of their low ability and the high level of item difficulty.

A condition which includes examinees with no preknowledge of the answers to the exposed anchor items worked as a baseline for relative comparisons. The baseline condition reflected a scenario where two courses of testing happen simultaneously, so there is no time for examinees to share the anchor items with each other. This condition also served as an evaluative criterion for assessing the accuracy of recovered scaling coefficients and equating true scores within the equating process.

## Calibration

In the current simulation study, the *ltm* package in R was used to generate item and ability parameters for each form of a test administered in the NEAT equating design. Once the probability of correct answers for anchor items was modified for exposure, dichotomous responses were generated by comparing the probability of correct answer to a random draw from a uniform distribution with a mean of 0 and standard deviation of 1. Each simulated item response dataset was generated based on the Rasch model. Item parameters were estimated separately for each form of a test using approximate Marginal

Maximum Likelihood Estimation (MMLE) as implemented in the *ltm* package in R. The maximum number of Gauss-Newton iterations was increased to 500. Besides the modification described earlier, default *ltm* package options were used for estimation. The item parameter estimates for the OF and NF were then equated using the mean/sigma method via the *ltm* package in R in order to place the item parameter estimates from the NF onto the metric of the OF.

## Summary of Conditions

This simulation study manipulated three factors to examine the influences of anchor item exposure on the equating process. One factor manipulated was the mean in latent trait ability. For all testing conditions, latent trait ability for the OF administration was drawn from a normal distribution with a mean of 0 and standard deviation of 1. For the NF administration, latent ability was varied among five levels as follows: (a) $\theta \sim N(-0.50, 1.00)$, (b) $\theta \sim N(-0.25, 1.00)$, (c) $\theta \sim N(0.00, 1.00)$, (d) $\theta \sim N(0.25, 1.00)$, and (e) $\theta \sim N(0.50, 1.00)$. Two levels of exposed anchor items (2 and 10 items) and four levels of examinees with preknowledge (5%, 10%, 15%, and 20%) were considered to simulate the degree of item exposure. The interaction of the three conditions resulted in a $5 \times 2 \times 4$ simulation for a total of 40 distinctive conditions. A baseline condition with anchor items displaying no exposure was added for each NF ability distribution, resulting in a total of 45 conditions overall. The simulation process was replicated 100 times for each condition. Generally, 100 replications are considered enough for Monte Carlo investigations, as demonstrated in the literature (Harwell, Stone, Hsu, & Kirisci, 1996).

## Evaluation Criteria

The accuracy of scaling coefficients was used as dependent variables to explore how the extent of anchor item exposure impacts the equating process. The scaling coefficients were obtained by calculating the slope and intercept values necessary to place ability and item parameters from the NF administration onto the scale of the OF administration.

To determine errors in the recovered scaling coefficients, bias was calculated to evaluate the deviation between the estimated and true scaling coefficients across all replications. Bias for recovered scaling coefficients is mathematically defined as a measure of the average difference between the estimated and true scaling coefficient across all replications:

$$bias_{\delta} = \frac{\sum_{j=1}^{k}(\hat{\delta}_j - \delta_j)}{k},$$

where $\delta$ is the true value of the scaling coefficient, $\hat{\delta}$ is the estimated value of the scaling coefficient, and $k$ is the total number of replications.

To determine the precision of recovered scaling coefficients, Mean Squared Error (MSE) was used by taking the average squared deviation of the estimated scaling coefficient from the true parameter. Mathematically, MSE can be defined as follows:

$$MSE_{\delta} = \frac{\sum_{j=1}^{k}(\hat{\delta}_j - \delta_j)^2}{k}.$$

The mathematical terms in the MSE formula are similar to those mentioned previously in Bias formula.

The accuracy of IRT equating true scores was also used as dependent variables to explore the influence of exposed anchor items on the equating process. In the IRT true

score equating, the equating true scores on the NF and OF are considered to be equivalent if both true scores correspond to one ability value.

To determine the impact of anchor item exposure on the IRT true score equating, bias and MSE were employed to quantify errors in the recovery of IRT equating true scores as comes next:

$$Bias_{\tau} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n+1}\left(\hat{\tau}_{ij} - \tau_{ij}\right)}{k(n+1)}$$

and

$$MSE_{\tau} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n+1}\left(\hat{\tau}_{ij} - \tau_{ij}\right)^{2}}{k(n+1)}$$

where k is the total number of replications, n is the number of items, $\tau(\theta_j)$ is the true number-correct true score, and $\hat{\tau}(\theta_{kj})$ is the estimated number-correct true score.

# CHAPTER 4

# RESULTS

This simulation study was conducted to evaluate the impact of anchor item exposure on the accuracy of recovered linking coefficients and IRT equating true scores for a test containing dichotomously scored items. The number of exposed items in the anchor set, the proportion of examinees with preknowledge of the anchor items, and the means of the ability distributions of groups taking the NF of the test were manipulated to evaluate their relative effects on the results. The number of exposed dichotomous items in the anchor set was two or 10 items. The proportion of examinees with preknowledge was varied at 5%, 10%, 15%, and 20%. The ability distribution of the group taking the OF was N(0,1), while the ability distributions of the group taking the NF were (a) N(-0.5,1), (b) N(-0.25,1), (c) N(0,1), (d) N(0.25,1), or (e) N(0.5,1). The interaction of these three factors yielded a total of 40 unique conditions. A non-item-exposure condition was added for each NF ability distribution, yielding a total of 45 conditions overall.

Table 2 exhibits bias and MSE results for the linking coefficient X under different levels of anchor item exposure in the five ability distributions of groups taking the NF. The linking coefficient X was recovered very well in the nonexposure conditions, bias almost approached zero. The linking coefficient X was consistently overestimated with the proportion of examinees with preknowledge of either a small or large amount of anchor items on a test. The exposure conditions, including 20% examinees with preknowledge of the anchor items, produced the most positively biased estimate of the linking coefficient X.

Table 2.

*Bias and MSE for Linking Coefficient X (Standard Deviation)*

| Anchor Item exposure (n) | Examinees with Preknowledge (%) | NF Ability Distribution | | | | |
|---|---|---|---|---|---|---|
| | | N(-0.50, 1) | N(-0.25, 1) | N(0.00, 1) | N(0.25, 1) | N(0.50, 1) |
| Bias | | | | | | |
| 0 | 0 | -0.001 | 0.002 | 0.001 | -0.001 | -0.002 |
| 2 | 5 | 0.081 (0.061) | 0.065 (0.058) | 0.058 (0.059) | 0.053 (0.062) | 0.046 (0.059) |
| | 10 | 0.160 (0.055) | 0.132 (0.056) | 0.121 (0.057) | 0.116 (0.061) | 0.104 (0.063) |
| | 15 | 0.237 (0.064) | 0.198 (0.061) | 0.184 (0.063) | 0.177 (0.065) | 0.162 (0.070) |
| | 20 | 0.320 (0.069) | 0.273 (0.065) | 0.258 (0.060) | 0.251 (0.065) | 0.231 (0.066) |
| 10 | 5 | 0.093 (0.051) | 0.075 (0.052) | 0.063 (0.051) | 0.054 (0.053) | 0.050 (0.055) |
| | 10 | 0.166 (0.054) | 0.136 (0.057) | 0.115 (0.060) | 0.102 (0.063) | 0.084 (0.064) |
| | 15 | 0.249 (0.067) | 0.198 (0.064) | 0.172 (0.060) | 0.149 (0.063) | 0.125 (0.063) |
| | 20 | 0.307 (0.080) | 0.249 (0.072) | 0.216 (0.069) | 0.186 (0.070) | 0.160 (0.072) |
| MSE | | | | | | |
| 0 | 0 | 0.002 | 0.002 | 0.003 | 0.003 | 0.003 |
| 2 | 5 | 0.010 | 0.007 | 0.007 | 0.007 | 0.006 |
| | 10 | 0.029 | 0.021 | 0.018 | 0.017 | 0.015 |
| | 15 | 0.060 | 0.042 | 0.038 | 0.036 | 0.031 |
| | 20 | 0.107 | 0.079 | 0.070 | 0.067 | 0.058 |
| 10 | 5 | 0.011 | 0.008 | 0.007 | 0.006 | 0.005 |
| | 10 | 0.030 | 0.022 | 0.017 | 0.014 | 0.011 |
| | 15 | 0.067 | 0.043 | 0.033 | 0.026 | 0.020 |
| | 20 | 0.101 | 0.067 | 0.052 | 0.039 | 0.031 |

It appears that the number of exposed items in the anchor set affected the linking coefficient X differentially according to the proportion of examinees with preknowledge and the mean of the NF ability distribution. For example, in the conditions including the NF ability distributions with lower means (-0.50 and -0.25), the 10-exposed-item condition resulted in a more biased estimate of the linking coefficient X when the proportions of examinees with preknowledge were 5%, 10%, or 15%. However, this influence was reversed in the condition including the highest proportion of examinees with preknowledge (20%), where the conditions including 10 exposed items led to a less biased estimate of the linking coefficient X. In general, differences between the two amounts of exposed anchor items tended to be trivial.

Results indicated that the linking coefficient X was influenced by the mean of the ability distribution of the group taking the NF. Across all exposure conditions, bias for the linking coefficient X decreased as the mean of the NF ability distribution increased. The item exposure conditions including the NF ability distribution with a mean of -0.50 consistently yielded the most positively biased estimate of the linking coefficient X, while the exposure conditions including the NF ability distribution with a mean of 0.50 consistently produced the least positively biased estimate of the linking coefficient X.

MSE seems to be a function of bias in the evaluation criterion of interest. This means that, when a bias value is small for a given condition, the corresponding MSE value for that condition must also be small. MSE results indicated that the recovery of the estimated linking coefficient X became less accurate as the proportion of examinees with preknowledge increased. In addition, MSE indicated that the linking coefficient X was more accurately estimated as the mean of the NF ability distribution increased.

Table 3 exhibits the bias and MSE results for the linking coefficient Y for each combination of the three testing factors: (a) the number of exposed anchor items, (b) the percentage of examinees with preknowledge of the anchor items, and (c) the mean of the ability distribution of the group taking the NF. For the testing conditions including no item exposure, the linking coefficient Y was recovered very well, with nearly no bias in the item parameter estimation. Overall, the results indicated that the estimated linking coefficient Y was positively biased in all testing conditions. This suggests that anchor items on the NF became easier than their pairs on the OF due to exposure. In addition, overestimation of the linking coefficient Y indicates that the group taking the NF benefited more from the exposed anchor items than did the group taking the OF.

Results indicated that the linking coefficient Y was influenced by the number of exposed anchor items. The overestimation of the linking coefficient Y was more intense in the testing conditions including 10 exposed anchor items than did the conditions including 2 exposed anchor items. This indicates that increasing the number of exposed items in the anchor set greatly benefited the group taking the NF. This result is applied to all testing conditions regardless of the mean of the ability distribution of the group taking the NF or even the proportion of examinees with access to the anchor items.
Results suggest that the anchor item exposure had an influence on the accuracy of the linking coefficient Y as a result of a change in the proportion of examinees with preknowledge of the anchor items. Bias in the linking coefficient Y systematically increased as the proportion of examinees with preknowledge increased.

Variations in the mean of the ability distribution of the group taking the NF seemed to influence the linking coefficient Y bias differentially according to the degree

Table 3.

*Bias and MSE for Linking Coefficient Y (Standard Deviation)*

| Anchor Item exposure (n) | Examinees with Preknowledge (%) | NF Ability Distribution | | | | |
|---|---|---|---|---|---|---|
| | | N(-0.50, 1) | N(-0.25, 1) | N(0.00, 1) | N(0.25, 1) | N(0.50, 1) |
| Bias | | | | | | |
| 0 | 0 | -0.008 | 0.005 | -0.001 | -0.006 | 0.006 |
| 2 | 5 | 0.016 (0.049) | 0.042 (0.050) | 0.041 (0.051) | 0.048 (0.054) | 0.064 (0.057) |
| | 10 | 0.044 (0.057) | 0.083 (0.056) | 0.091 (0.049) | 0.102 (0.052) | 0.127 (0.055) |
| | 15 | 0.068 (0.060) | 0.116 (0.055) | 0.138 (0.049) | 0.164 (0.054) | 0.202 (0.059) |
| | 20 | 0.103 (0.057) | 0.164 (0.054) | 0.199 (0.050) | 0.236 (0.048) | 0.280 (0.053) |
| 10 | 5 | 0.148 (0.052) | 0.157 (0.050) | 0.146 (0.047) | 0.142 (0.048) | 0.154 (0.052) |
| | 10 | 0.323 (0.058) | 0.322 (0.057) | 0.303 (0.058) | 0.303 (0.059) | 0.305 (0.064) |
| | 15 | 0.501 (0.060) | 0.493 (0.056) | 0.470 (0.057) | 0.460 (0.063) | 0.453 (0.073) |
| | 20 | 0.726 (0.055) | 0.699 (0.060) | 0.661 (0.062) | 0.643 (0.072) | 0.632 (0.077) |
| MSE | | | | | | |
| 0 | 0 | 0.004 | 0.003 | 0.003 | 0.004 | 0.004 |
| 2 | 5 | 0.003 | 0.004 | 0.004 | 0.005 | 0.007 |
| | 10 | 0.005 | 0.010 | 0.011 | 0.013 | 0.019 |
| | 15 | 0.008 | 0.016 | 0.022 | 0.030 | 0.044 |
| | 20 | 0.014 | 0.030 | 0.042 | 0.058 | 0.081 |
| 10 | 5 | 0.025 | 0.027 | 0.024 | 0.023 | 0.026 |
| | 10 | 0.108 | 0.107 | 0.095 | 0.095 | 0.097 |
| | 15 | 0.254 | 0.246 | 0.224 | 0.216 | 0.211 |
| | 20 | 0.530 | 0.492 | 0.441 | 0.418 | 0.405 |

of exposure (the number of exposed anchor items and the proportion of examinees with preknowledge of the anchor items). In general, the difference in the linking coefficient bias among the NF ability distributions was relatively moderate.

As corresponded to the linking coefficient X, MSE results indicated that the recovery of the estimated linking coefficient Y became less accurate as the proportion of examinees with preknowledge increased. As opposed to the linking coefficient X, the MSE results indicated that the estimated linking coefficient Y was less accurately recovered as the number of exposed anchor items increased. As with bias for the linking coefficient Y, results indicated that the NF ability distribution was found to influence MSE for the linking coefficient Y differentially according to the extent of exposure. In general, however, MSE results indicated that the recovery of the estimated linking coefficient Y was the least accurate in the conditions including 10 exposed anchor items and the NF ability distribution with a mean of - 0.50, and the most accurate in the conditions including the two exposed anchor items and the NF ability distribution with a mean of - 0.50.

Table 4 displays the bias and MSE of the equating true scores for each of the exposure conditions and NF ability distributions. The estimated equating true scores were recovered well in the no-exposure conditions, with almost no errors in estimation. An examination of the bias for equating true scores overall indicated that the estimated equating true scores were positively biased in all exposure conditions. This suggests that prior knowledge of anchor items causes examinees to receive inflated test scores on content being tested.

Table 4.

*Bias and MSE for IRT Equating True Score (Standard Deviation)*

| Anchor Item exposure (n) | Examinees with Preknowledge (%) | NF Ability Distribution | | | | |
|---|---|---|---|---|---|---|
| | | N(-0.50, 1) | N(-0.25, 1) | N(0.00, 1) | N(0.25, 1) | N(0.50, 1) |
| Bias | | | | | | |
| 0 | 0 | 0.009 | 0.009 | 0.015 | 0.004 | -0.005 |
| 2 | 5 | 0.393 (0.306) | 0.334 (0.309) | 0.298 (0.306) | 0.269 (0.318) | 0.232 (0.332) |
| | 10 | 0.818 (0.314) | 0.708 (0.310) | 0.617 (0.299) | 0.545 (0.323) | 0.451 (0.337) |
| | 15 | 1.203 (0.324) | 1.020 (0.305) | 0.917 (0.301) | 0.840 (0.318) | 0.741 (0.323) |
| | 20 | 1.627 (0.295) | 1.397 (0.297) | 1.273 (0.290) | 1.162 (0.291) | 1.007 (0.302) |
| 10 | 5 | 1.174 (0.290) | 1.021 (0.308) | 0.886 (0.309) | 0.809 (0.318) | 0.714 (0.316) |
| | 10 | 2.391 (0.351) | 2.063 (0.352) | 1.816 (0.358) | 1.679 (0.352) | 1.492 (0.340) |
| | 15 | 3.630 (0.352) | 3.146 (0.357) | 2.771 (0.340) | 2.506 (0.345) | 2.222 (0.368) |
| | 20 | 4.963 (0.385) | 4.304 (0.383) | 3.810 (0.391) | 3.474 (0.384) | 3.117 (0.358) |
| MSE | | | | | | |
| 0 | 0 | 0.143 | 0.147 | 0.158 | 0.173 | 0.181 |
| 2 | 5 | 0.330 | 0.273 | 0.240 | 0.229 | 0.217 |
| | 10 | 1.006 | 0.778 | 0.616 | 0.529 | 0.423 |
| | 15 | 2.035 | 1.474 | 1.216 | 1.059 | 0.868 |
| | 20 | 3.572 | 2.658 | 2.229 | 1.888 | 1.478 |
| 10 | 5 | 1.809 | 1.403 | 1.088 | 0.935 | 0.756 |
| | 10 | 7.191 | 5.382 | 4.208 | 3.617 | 2.880 |
| | 15 | 16.386 | 12.317 | 9.580 | 7.862 | 6.239 |
| | 20 | 30.513 | 22.929 | 18.002 | 14.990 | 12.089 |

Results indicated that the proportion of examinees with preknowledge of the anchor items had an influence on the recovery of estimated equating true scores as a result of anchor item exposure. Bias in the estimated equating true score consistently increased as the proportion of examinees with preknowledge increased. This suggests that increasing the number of examinees who have access to test items used to place the OF and NF of a test on a single scale causes the entire group of examinees to receive overestimated equating true scores.

Bias results shown in Table 4 indicated that the amount of exposed items in the anchor set had more of an influence on the accuracy of estimated equating true scores. A corollary of this influence is that the estimated equating true scores were more overestimated under the conditions including the large amount of exposed anchor items than under the conditions including the small amount of exposed anchor items. This suggests that increasing the number of exposed items in the anchor set greatly overestimates true abilities of the entire group of examinees taking the exposed form of a test.

Changes to the mean of the ability distribution of examinees taking the NF of a test seemed to influence the accuracy of estimated equating true scores. The exposure conditions including the NF ability distribution with a mean of -0.50 consistently produced the most positively biased estimate of the equating true score, while the bias in estimating the equating true score was the least under conditions including the NF ability distribution with a mean of 0.50. However, the overestimation of the equating true score was more intense in the conditions including a large amount of exposed anchor items. Since examinees with lower ability benefited the most from the exposed anchor items,

this explains a possible scenario where unqualified examinees might appear qualified as a result of exposure.

As with both linking coefficients X and Y, MSE for the equating true score showed that the recovery of the estimated equating true score became less accurate as the proportion of examinees with preknowledge of the anchor items. As with the linking coefficient Y, MSE for the equating true score indicated that the recovered equating true score became less accurate as the number of exposed items in the anchor set increased. As with the linking coefficient X and contrasted to the linking coefficient Y, the conditions with a smaller mean of the NF ability distribution consistently produced the most overestimated MSE values, while the conditions including a larger mean of the NF ability distribution consistently produced the least overestimated MSE values.

# CHAPTER 5

# DISCUSSION, IMPLICATIONS, CONCLUSIONS, FUTURE DIRECTIONS

## Discussion

The goal of the current simulation study was to evaluate the impact of the anchor item exposure on the linking coefficients and equating results under the NEAT design. This simulation dealt with a common scenario in which two parallel forms of a small-scale dichotomously scored test, the OF and NF, were administered on two testing occasions and offered to different small groups of examinees. The OF was the form that set the scale of scores and the NF was the form that was equated to the scale of the OF. The impact of anchor item exposure on the equating process was assessed through manipulating three main factors: (a) the number of exposed anchor items, (b) the proportion of examinees with preknowledge of the anchor items, and (c) the mean of the ability distribution of the group taking the NF of a test.

Of the three factors, the number of exposed anchor items and the proportion of examinees with preknowledge of the anchor items were found to have an influence on the accuracy of recovered equating true scores. These two factors were assessed at different levels in order to explore how the degree of anchor exposure impacts the equating process. Results revealed that an increase in either the number of exposed anchor items or the proportion of examinees with preknowledge overestimated the equating true scores. In spite of the fact that overestimation of the equating true scores was expected, the magnitude of bias at the lowest degree of anchor exposure was still perturbing. Therefore, the results suggested that the inclusion of slightly exposed items in the anchor set causes the entire group of examinees to receive inflated scores on a test form being

equated. These results are compatible with previous research study findings (Jurich et al., 2012).

Results indicated that variations in the mean of the ability distribution of the group taking the NF of a test had an influence on the recovery of estimated equating true scores as a consequence of anchor item exposure. The conditions including examinees with low mean ability consistently produced the most overestimated equating true scores, while the conditions including examinees with high mean ability produced the least overestimated equating true scores. The results suggested that examinees with low mean ability benefited more from gaining prior knowledge of the exposed anchor items than did examinees with high mean ability. This means that the true abilities of examinees coming from a lower ability distribution were more overestimated than the abilities of examinees being derived from a higher ability distribution. This leads to a possible scenario where unqualified examinees might appear qualified as a result of anchor item exposure, while qualified examinees have less to gain from exposure.

Evaluating the influence of anchor item exposure on the linking coefficient X provided more insight about the positively biased equating true scores. Overall, results indicated that exposed anchor items introduced positive bias in the linking coefficient X in all conditions. As with the equating true scores, the linking coefficient X was also overestimated with the degree of exposure. The overestimation of the linking coefficient X arose because exposure occurs on the anchor items used to equate the NF to the OF. As examinees with preknowledge respond correctly to the anchor items on the NF, the difficulty values of these items will be underestimated. Thus, the exposure implementation will underestimate the difficulty parameters of the anchor items, thus

getting these parameters closer to each other and then reducing the variability among them. The linking procedure must overestimate the linking coefficient X to account for this diminished variability (see the equation shown below that calculates the linking coefficient X under the mean/sigma linking method). In addition to the underestimated difficulty values of the

$$X = \frac{\sigma_{b_{OF}}}{\sigma_{b_{NF}}}$$

exposed anchor items and their diminished variability, the exposure implementation will inflate the true abilities of examinees and have them obtain overestimated equating true scores.

As with the equating true scores, results indicated that changes to the mean in the ability distribution of the group taking the NF of a test had an influence on the recovery of the linking coefficient X as a consequence of anchor item exposure. The overestimation in the linking coefficient X was found to be greater for the group with a low mean ability and less for the group with a high mean ability. This result leads to an actual scenario where anchor item exposure is more advantageous for low-ability examinees and less advantageous for high-ability examinees. Since low-ability examinees benefit the most from the exposed anchor items, this leads to a potential scenario where unskilled individuals might seem to be proficient as a result of exposure.

Evaluating the influence of anchor item exposure on the linking coefficient Y provided more insight about the positively biased equating true scores. Overall, results indicated that the linking coefficient Y was positively biased in all testing conditions. The overestimation of the linking coefficient Y revealed that anchor items on the NF of a test became easier than their pairs on the OF because the anchor items on the NF were

under the influence of exposure. In addition, the overestimated linking coefficient Y revealed that the group taking the NF always had higher ability than the group taking the OF regardless of the difference in the ability distributions between the two groups. The artificial correct response to each exposed anchor item on the NF is the usual reason why the group taking the NF had higher ability than the group taking the OF in all conditions. That is also the usual reason why all exposure conditions produced a positively biased estimate of the equating true score.

As with the equating true scores, results indicated that the linking coefficient Y bias increased positively with the number of exposed items in the anchor set and/or the proportion of examinees with preknowledge of the anchor items. As the degree of exposure increases, the entire group of examinees responding to the NF can receive benefit from the exposure after the equating process is conducted. Logically, the positively biased estimate of the linking coefficient Y indicates that the anchor items on the NF become easier than their pairs on the OF, so the difficulty values for these NF anchor items will be underestimated. Since the mean/sigma method exposure uses the anchor items to obtain the linking coefficients, this leads to that the mean of difficulty parameters for the anchor items on the NF will be underestimated, which, in turn overestimates the linking coefficient Y (refer to the equation below).

$$Y = \mu_{b_{OF}} - X * \mu_{b_{NF}}$$

In this simulation study, both linking coefficients X and Y were used to rescale the difficulty parameters for all items (unique and anchor) on the NF (the test form including exposed anchor items) for the purpose of IRT equating. Since both linking coefficients were overestimated in all exposure conditions, this artificial dual

overestimation would cause the rescaled difficulty parameters ($b^*s$) for all items on the NF to be increased as well (refer to the equation below).

$$b_j^* = X * b_j + Y$$

All items on the NF seem to be more difficult than they are in reality, so answering these items correctly will inflate the true abilities of examinees. This artificial inflation in the ability estimates causes the equating true scores to be inflated as well.

**Implications for Test Developers**

The study results suggest that the inclusion of slightly exposed items in the anchor set causes the entire group of examinees to receive inflated scores on a test form being equated. These inflated scores, in turn, could not accurately reflect examinees' true abilities and their actual competence on the content being tested, so any decision made on the basis of these artificial scores will be doubtful. Based on these findings, the researcher recommends that test developers should remove exposed items from the anchor set.

As demonstrated in this simulation study, the overestimation in equating true scores was more severe in the exposure conditions including a mean ability distribution of -0.50. This finding indicated that unqualified examinees administered a test form where exposure occurred might be more proficient in the content being tested than is the case in the real life. The worst-case scenario occurs when, as a result of exposure, unqualified examinees take an unfair advantage over qualified examinees who completed an unexposed test form. As noted above, the anchor item exposure can cause detrimental influences on the equating true scores and become a serious threat to the test's fairness and validity. Based on these findings, the researcher recommends that focused attention

must be given to methods that detect exposed anchor items on the test to ensure that scores accurately reflect the true abilities of examinees.

## Conclusions

This simulation study attempted to close a gap in the literature by examining the impact of anchor item exposure on the accuracy of linking coefficients and equating true scores obtained through the IRT equating process according to the NEAT design. Due to crucial decisions made on the basis of high-stakes testing, it is incumbent on test developers to create parallel forms of a test in order to lessen the risk of cheating and assess examinees fairly. It is also incumbent on test developers to be confident that a test score produced by an examinee is an accurate reflection of their true abilities and actual competence on the content of interest being measured. Based on the study findings, however, anchor item exposure can have detrimental influences on equating true scores. If some examinees have prior knowledge of anchor items on a test, equating true scores for all examinees taking the test might be inflated. Even when items in the anchor set are subject to low levels of exposure, equating true scores still challenge the accuracy, while high levels of exposure will completely twist the test scores even under the most favorable circumstances. In this simulation study, the anchor item exposure became a serious threat to the test fairness and validity to the extent that unqualified examinees might receive an unfair benefit over qualified examinees who completed an unexposed test form. The conclusions of this study suggest that the anchor items on a test must be evaluated frequently for exposure, so individuals dealing with test scores can be confident that each score represents the true ability of the examinee on the construct of interest being measured.

## Future Directions

As with any Monte Carlo investigation, only a restricted number of factors could be investigated, so this must be taken into consideration and care exercised when any generalization is made to other testing settings. Further research is needed to evaluate the effect of anchor item exposure on the equating process under factors different than the ones considered in the current study.

The findings of this study will only be applicable to a test containing dichotomously scored items that measure a common dimension. Further research is required to evaluate the impact of anchor item exposure on linking and equating results for a test containing dichotomous items that measure multiple dimensions, polytomous items that measure one or multiple dimensions, or both types of items that measure one or multiple dimensions.

In the current simulation study, the item exposure was implemented in a way that added a magnitude of 1 to the probability of correct response for an anchor item and applied this modified probability to any examinee who has preknowledge of that item. However, it is important to adhere to a real scenario in which an examinee forgets the right response to the exposed anchor item due to his or her low ability and high level of item difficulty. Thus, future studies may consider different methods for item exposure implementation.

The current simulation study used the Rasch model to calibrate item and ability parameters. This model considered only the difficulty property of items and ignored other properties, such as discrimination and guessing. The two- or three-parameter

logistic IRT models can be employed as alternatives for item parameter calibration in future studies.

The study findings were limited to utilization of one equating design, known as the NEAT Design. Other data collection designs such as the Single-Group (SG) Design, the Equivalent-Groups (EG) Design, and the Counterbalanced (CB) Design can be studied to allow for a comparison with the NEAT design under the same testing conditions considered in this study. The study also restricted its application to the internal anchor design and did not consider the external one, where examinees' responses to anchor items do not contribute to their total scores on the test forms to be equated. Additional research on the external anchor design might provide further insight into the effect of anchor item exposure on the equating process.

Finally, the study findings were limited to using one linear transformation procedure, called here the mean/sigma method, for placing the ability and item parameters from different calibrations on a common scale. Other methods, such as the mean/mean method and test characteristic curve method can be studied to allow for a comparison with the mean/sigma method under the same conditions considered in the current study.

## REFERENCES


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

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. doi: 10.1111/j.1745-3984.1985.tb01045.x

Guo, J., Tay, L., & Drasgow, F. (2009). Conspiracies and test compromise: An evaluation of the resistance of test systems to small-scale cheating. *International Journal of Testing, 9*(4), 283-309.

Haberman, S., & Dorans, N. J. (2009, April). *Scale consistency, drift, stability: Definitions, distinctions and principles*. Paper presented at the American Educational Research Association (AERA) and the National Council on Measurement in Education, San Diego, CA. Retrieved from http://www.ets.org/Media/Conferences_and_Events/AERA_2009_pdfs/AERA_NCME_2009_Haberman1.pdf

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

Han, K. T., & Guo, F. (2011, January). *Potential impact of item parameter drift due to practice and curriculum change on item calibration in computerized adaptive testing* (GMAC RR-11-02). Reston, Virginia: Graduate Management Admission Council. Retrieved from http://www.gmac.com/~/media/Files/gmac/Research/research-report-series/rr1102_itemcalibration.pdf

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

Harwell, M., Stone, C. A., Hsu, T.-C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement, 20*(2), 101–125. doi: 10.1177/014662169602000201

Holland, P. W. (2007). A framework and history for score linking. In N. J. Dorans, M. Pommerich, & P.W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 5-30). New York, NY: Springer.

Jurich, D. P., DeMars, C. E., & Goodman, J. T. (2012). Investigating the impact of cheating on IRT equating under the nonequivalent anchor test design. *Applied Psychological Measurement, 36*(4), 291- 308.

Jurich, D. P., Goodman, J. T., & Becker, K. A. (2010, May). *Assessment of various equating methods: Impact on the pass-fail status of cheaters and non-cheaters*. Poster presented at National Council on Measurement in Education, Denver, CO.

Kim, D., Barton, K., & Choi, S. (2010, May). *Sample size impact on screening methods in the Rasch model*. Paper presented at the American Educational Research Association, Denver, CO.

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. doi: 10.1177/01466216980222003

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

Li, Y. (2012). *Examining the impact of drifted polytomous anchor items on test characteristic curve (TCC) linking and IRT true score equating* (ETS RR–12-09). Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/Media/Research/pdf/RR-12-09.pdf

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement*, *17*(3), 179-193. doi: 10.1111/j.1745-3984.1980.tb00825.x

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement*, *14*(2), 139-160. doi: 10.1111/j.1745-3984.1977.tb00033.x

McLeod, L. D., & Lewis, C. (1999). Detecting item memorization in the CAT environment. *Applied Psychological Measurement*, *23*(2), 147–160. doi: 10.1177/01466219922031275

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

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests* (expanded edition). Chicago: The University of Chicago Press.

Ricker, K. L., & von Davier, A. A. (2007, December). *The impact of anchor test length on equating results in a nonequivalent groups design* (RR-07-44). Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/Media/Research/pdf/RR-07-44.pdf

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software, 17*(5), 1-25. Retrieved from http://www.jstatsoft.org/v17/i05/

Segall, D. O. (2002). An item response model for characterizing test compromise. *Journal of Educational and Behavioral Statistics, 27*(2), 163-79. doi: 10.3102/10769986027002163

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*(2), 201-210. doi: 10.1177/014662168300700208

Stocking, M. L., Ward, W. C., & Potenza, M. T. (1998). Simulating the use of disclosed items in computerized adaptive testing. *Journal of Educational Measurement, 35*(1), 48–68. doi: 10.1111/j.1745-3984.1998.tb00527.x

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

Wei, H. (2010, May). *Impact of non-representative anchor items on scale stability*. Paper presented at the annual meeting of the National Council on Measurement in Education, Denver, CO. Retrieved from http://www.pearsonassessments.com/NR/rdonlyres/C6DC49A6-2479-4C01-9BF1-FD3B14EA048C/0/12_NCME_Wei_42210.pdf

Wells, C. S., Subkoviak, M. J., & Serlin, R. C. (2002). The effects of item parameter drift on examinee ability estimates. *Applied Psychological Measurement, 26*(1), 77-87. doi: 10.1177/0146621602261005

Wingersky, M. S., Cook, L. L., & Eignor, D. R. (1987, June). *Specifying the characteristics of linking items used for item response theory item calibration*. ETS Research Report 87-24. Princeton, NJ: Educational Testing Service.

Wu, Y.-F. (2012). *How test length and sample size have an impact on the standard errors for IRT true score equating: Integrating SAS® and other software*. Poster session presented at proceedings of the SAS® Global Forum 2012 Conference, Cary, NC: SAS Institute Inc.

Yang, W., & Houang, R. T. (1996, April). *The effect of anchor length and equating method on the accuracy of test equating: comparisons of linear and IRT-based equating using an anchor-item design*. Paper presented at the 1996 annual meeting of the American Educational Research Association, New York, NY. Retrieved from http://www.eric.ed.gov/PDFS/ED401308.pdf

Yen, W., & Fitzpatrick, A. R. (2006). Item response theory. In Robert L. Brennan (Ed). *Education Measurement* (4th ed., pp. 111-153).Westport, CT: Praeger.

Yi, Q., Zhang, J., & Chang, H. (2008). Severity of organized item theft in computerized adaptive testing: A simulation study. *Applied Psychological Measurement,32*(7), 543-558. doi: 10.1177/0146621607311336

Zara, A. (2006). *Defining item compromise*. Unpublished paper presented at the 2006 annual meeting of the National Council on Measurement in Education, San Francisco, CA. Retrieved from https://www.ncsbn.org/2006.04_Zara_-_AERA_-_Defining_Item_Compromise.pdf

# APPENDICES

## Appendix A

### R Script for Evaluating the Accuracy of Recovered Linking Coefficients (Slope and Intercept) and Recovered Equating True Scores in the Non-exposure Conditions

```
setwd("C:/Users/Moatasim/Desktop/Master's Thesis/R Scripts")

#Setting the test length
n <- 50

#Setting the number of distinctive items
n.d <- 40

#Generating b values for distinctive items on the original form (OF)
b.d1_seed <- 3490123
set.seed(b.d1_seed)
b.d1 <- rnorm(n.d,0,1)

#4738237
#Generating b values for distinctive items on the new form (NF)
b.d2_seed <- 3490123
set.seed(b.d2_seed)
b.d2 <- rnorm(n.d,0,1)

#Setting the number of common items
n.a <- 10

#Generating b-values for common items
b.a_seed <- 1728240
set.seed(b.a_seed)
b.a <- rnorm(n.a,0,1)
mean(b.a)

true.b1 <- c(b.d1,b.a)
true.b2 <- c(b.d2,b.a)
mean(true.b1)
#Setting the number of examinees
N <- 500

#Generating theta-values for group 1 taking the OF
true.theta1_seed <- 2199345
set.seed(true.theta1_seed)
true.theta1 <- rnorm(N,0,1)

#Generating theta-values for group 2 taking the NF
```

```
true.theta2_seed <- 5692104
set.seed(true.theta2_seed)
true.theta2 <- rnorm(N,0,1)

item_seed <- 2126354
set.seed(item_seed)
c1 <- vector(mode="numeric",n)
a1 <- matrix(1,n,1)
c2 <- vector(mode="numeric",n)
a2 <- matrix(1,n,1)
fix <- matrix(1,1,2)
fix[,1] <- n+1

library("ltm")

#Generating item response data set for the OF and NF
p1 <- matrix(0,N,n)
u1 <- matrix(0,N,n)

p2 <- matrix(0,N,n)
u2 <- matrix(0,N,n)

for (i in 1:N) {
    for (j in 1:n){
    p1[i,j] <- c1[j] + (1-c1[j])/(1+exp(-a1[j]*(true.theta1[i]-true.b1[j])))
    r1 <- runif(1,0,1)
        if (r1 <= p1[i,j]) {
        u1[i,j] <- 1
        }
    }
}

for (i in 1:N) {
    for (j in 1:n){
    p2[i,j] <- c2[j] + (1-c2[j])/(1+exp(-a2[j]*(true.theta2[i]-true.b2[j])))
    r2 <- runif(1,0,1)
        if (r2 <= p2[i,j]) {
        u2[i,j] <- 1
        }
    }
}

#Running IRT analysis
#Rasch
dichrasch1 <- rasch(data=u1,IRT.param=TRUE,constraint=fix,control=list(iter.qN=500))
dichrasch2 <- rasch(data=u2,IRT.param=TRUE,constraint=fix,control=list(iter.qN=500))
```

```
#Item Difficulty Parameters for the OF and NF items
b1 <- coef(dichrasch1)[,1]
b2 <- coef(dichrasch2)[,1]

#Ability parameters for group 1 taking the OF
t.est1 <- factor.scores(dichrasch1,method="EAP",resp.patterns=u1)
theta1 <- t.est1$score.dat[,n+3]

#Ability parameters for group 2 taking the NF
t.est2 <- factor.scores(dichrasch2,method="EAP",resp.patterns=u2)
theta2 <- t.est2$score.dat[,n+3]

#Equating Process for True Parameters

#The true Item Parameters for the Common Items across the Two Forms

F1.anchor.item.par <- c(b1[41],b1[42],b1[43],b1[44],b1[45],
                        b1[46],b1[47],b1[48],b1[49],b1[50])

F2.anchor.item.par <- c(b2[41],b2[42],b2[43],b2[44],b2[45],
                        b2[46],b2[47],b2[48],b2[49],b2[50])

mean.F1.anchor.item.par <- mean(F1.anchor.item.par)
mean.F2.anchor.item.par <- mean(F2.anchor.item.par)

SD.F1.anchor.item.par <- sd(F1.anchor.item.par)
SD.F2.anchor.item.par <- sd(F2.anchor.item.par)

#True Scaling Coefficients

true.x <- SD.F1.anchor.item.par/SD.F2.anchor.item.par
true.y <- mean.F1.anchor.item.par - (true.x * mean.F2.anchor.item.par)

#Rescaled True Item Parameters

rescaled_b2     <- true.x * b2 + true.y
rescaled_theta2 <- true.x * theta2 + true.y

#IRT True Score Equating Process

# tau1: True Score on the OF
# tau2: True Score on the NF
# tau_theta: Theta to which tau 1 and tau 2 correspond
# icc1: Item Characteristic Curve value for item j on the OF at a particular ability value
# icc2: Item Characteristic Curve value for item j on the NF at a particular ability value
```

```
# tcc1: Test Characteristic Curve value over items on the OF at a particular ability value
# tcc2: Test Characteristic Curve value over items on the NF at a particular ability value
# icc2.d: The First Derivative value of the Item Characteristic Curve for item j on the NF
at a particular ability value
# s_icc2.d: Sum of the First Derivative values of the Item Characteristic Curve over items
on the NF at a particular ability value
# theta0: The Starting Value of Theta
# ts: Total Score (  0 < ts  < 50)

ts <- 51

icc1          <- vector(mode="numeric",n)
icc2          <- vector(mode="numeric",n)
icc2.d        <- vector(mode="numeric",n)
tau_theta     <- vector(mode="numeric",ts)
tau1          <- vector(mode="numeric",ts)
tau2          <- vector(mode="numeric",ts)

#Applying the Newton-Raphson Method to find theta (tau_theta)

theta0 <- -14.7

#specifying a true score on the new form (tau2)
for (k in 1:1){

tau2[k] <- k-1
h <- 0
repeat{
h <- h +1
#Calculating the ICC values for the new form at theta0
for (j in 1:n){
            icc2[j] <- c2[j] + (1-c2[j])/(1+exp(-a2[j]*(theta0 -rescaled_b2[j])))
           }

tcc2 <- sum(icc2)
tcc2 <- round(tcc2, digits = 4)

#Finding theta (tau_theta) corresponding to the true score on the new form (tau2)
if (tcc2 == tau2[k]) {
                tau_theta[k] <- theta0
                break
                }

if (tcc2 != tau2[k]) {
                icc2.d     <- (1.7)*(1-icc2)*(icc2)
                s_icc2.d   <- sum(icc2.d)
```

```
                new_theta <- theta0 - ((tau2[k] - tcc2)/(-1*s_icc2.d))
                theta0 <- new_theta
                next
               }
    }
cat("Done with the IRT True Score Equating Number",k,"\n")
}

#Finding the true score on the original form (tau1)
for (k in 1:1){

for (j in 1:n){
            icc1[j] <- c1[j] + (1-c1[j])/(1+exp(-a1[j]*(tau_theta[k]-b1[j])))
           }
tcc1 <- sum(icc1)
tau1[k] <- round(tcc1, digits = 4)
            }

#########################################################################
#Setting a starting value for theta
theta0 <- -5

#specifying a true score on the new form (tau2)
for (k in 2:ts){

tau2[k] <- (k-1)

repeat{

#Calculating the ICC values for the new form at theta0
for (j in 1:n){
            icc2[j] <- c2[j] + (1-c2[j])/(1+exp(-a2[j]*(theta0 -rescaled_b2[j])))
           }

tcc2 <- sum(icc2)
tcc2 <- round(tcc2, digits = 4)

#Finding theta (tau_theta) corresponding to the true score on the new form (tau2)
if (tcc2 == tau2[k]) {
                tau_theta[k] <- theta0
                break
                }

if (tcc2 != tau2[k]) {
                icc2.d     <- (1.7)*(1-icc2)*(icc2)
                s_icc2.d   <- sum(icc2.d)
```

```
                new_theta <- theta0 - ((tau2[k] - tcc2)/(-1*s_icc2.d))
                theta0 <- new_theta
                next
               }
    }
cat("Done with the IRT True Score Equating Number",k,"\n")
}

#Finding the true score on the original form (tau1)
for (k in 2:ts){

for (j in 1:n){
            icc1[j] <- c1[j] + (1-c1[j])/(1+exp(-a1[j]*(tau_theta[k]-b1[j])))
           }
tcc1 <- sum(icc1)
tau1[k] <- round(tcc1, digits = 4)
            }

#Evaluating the accuracy of recovered linking constants and equating true scores for the
baseline (non-exposure) conditions

NREPS <- 100

bias.slope     <- vector(mode="numeric",NREPS)
bias.intercept <- vector(mode="numeric",NREPS)

MSE.slope     <- vector(mode="numeric",NREPS)
MSE.intercept <- vector(mode="numeric",NREPS)

bias.IRT.true.score <- vector(mode="numeric",NREPS)
MSE.IRT.true.score <- vector(mode="numeric",NREPS)

for (nr in 1:NREPS) {

#Generate data set
p1 <- matrix(0,N,n)
u1 <- matrix(0,N,n)

p2 <- matrix(0,N,n)
u2 <- matrix(0,N,n)

for (i in 1:N) {
    for (j in 1:n){
    p1[i,j] <- c1[j] + (1-c1[j])/(1+exp(-a1[j]*(theta1[i]-b1[j])))
    r1 <- runif(1,0,1)
        if (r1 <= p1[i,j]) {
```

```
        u1[i,j] <- 1
        }
    }
}


for (i in 1:N) {
    for (j in 1:n){
    p2[i,j] <- c2[j] + (1-c2[j])/(1+exp(-a2[j]*(theta2[i]-b2[j])))
    r2 <- runif(1,0,1)
        if (r2 <= p2[i,j]) {
           u2[i,j] <- 1
                    }
            }
           }


#Run IRT analysis
#Rasch
dichrasch1 <- rasch(data=u1,IRT.param=TRUE,constraint=fix,control=list(iter.qN=500))
dichrasch2 <- rasch(data=u2,IRT.param=TRUE,constraint=fix,control=list(iter.qN=500))

#Item Difficulty Parameters Estimates
estimated.b1 <- coef(dichrasch1)[,1]
estimated.b2 <- coef(dichrasch2)[,1]

#Equating Process for Estimates

#The Estimated Item Parameters for the Common Items across the Two Forms

F1.estimated.anchor.items <- c(estimated.b1[41],estimated.b1[42],estimated.b1[43],
                    estimated.b1[44],estimated.b1[45],estimated.b1[46],
                    estimated.b1[47],estimated.b1[48],estimated.b1[49],
                    estimated.b1[50])

F2.estimated.anchor.items <- c(estimated.b2[41],estimated.b2[42],estimated.b2[43],
                    estimated.b2[44],estimated.b2[45],estimated.b2[46],
                    estimated.b2[47],estimated.b2[48],estimated.b2[49],
                    estimated.b2[50])

#The Mean and Standard Deviation of the Estimated Item Parameters
#for the Common Items across the Two Forms

mean.F1.estimated.anchor.items <- mean(F1.estimated.anchor.items)
mean.F2.estimated.anchor.items <- mean(F2.estimated.anchor.items)
```

```
SD.F1.estimated.anchor.items <- sd(F1.estimated.anchor.items)
SD.F2.estimated.anchor.items <- sd(F2.estimated.anchor.items)

#Estimated Scaling Coefficients

estimated.x <- SD.F1.estimated.anchor.items/SD.F2.estimated.anchor.items
estimated.y <- mean.F1.estimated.anchor.items -(estimated.x *
mean.F2.estimated.anchor.items)

#Rescaled Estimated Item and ability Parameters

rescaled.estimated.b2 <- estimated.x * estimated.b2 + estimated.y

t.est2 <- factor.scores(dichrasch2,method="EAP",resp.patterns=u2)
estimated.theta2 <- t.est2$score.dat[,n+3]
rescaled.estimated.theta2 <- estimated.x * estimated.theta2 + estimated.y

#Estimated IRT True Score Equating

est.icc1      <- vector(mode="numeric",n)
est.icc2      <- vector(mode="numeric",n)
est.icc2.d    <- vector(mode="numeric",n)
est.tau_theta <- vector(mode="numeric",ts)
est.tau1      <- vector(mode="numeric",ts)
est.tau2      <- vector(mode="numeric",ts)


theta0 <- -14.7

#Specifying an estimated true score on the new form (est.tau2)
for (k in 1:1){

est.tau2[k] <- k-1

#Finding estimated theta (est.tau_theta) corresponding to (est.tau2)

repeat{

#Calculating estimated ICCs for the new form items
for (j in 1:n){
      est.icc2[j] <- c2[j] + (1-c2[j])/(1+exp(-a2[j]*(theta0-rescaled.estimated.b2[j])))
           }
est.tcc2 <- sum(est.icc2)
est.tcc2 <- round(est.tcc2, digits = 4)
```

```
if (est.tcc2 == est.tau2[k]) {
                    est.tau_theta[k] <- theta0
                    break
                    }

if (est.tcc2 != est.tau2[k]) {
                est.icc2.d   <- (1.7)*(1-est.icc2)*(est.icc2)
                s_est.icc2.d <- sum(est.icc2.d)
                new_theta    <- theta0 - ((est.tau2[k] - est.tcc2)/(-1*s_est.icc2.d))
                theta0       <- new_theta
                next
                    }
    }
#cat("Done with the IRT True Score Equating Number",k,"\n")
}

###Finding the estimated true score on the original form (est.tau1)
###corresponding to (est.tau_theta)
for (k in 1:1){

for (j in 1:n){
     est.icc1[j] <- c1[j] + (1-c1[j])/(1+exp(-a1[j]*(est.tau_theta[k]-estimated.b1[j])))
          }
est.tcc1 <- sum(est.icc1)
est.tau1[k] <- round(est.tcc1, digits = 4)
}

#################################

theta0 <- -5

#Specifying an estimated true score on the new form (est.tau2)
for (k in 2:ts){

est.tau2[k] <- (k-1)

#Finding estimated theta (est.tau_theta) corresponding to (est.tau2)

repeat{

#Calculating estimated ICCs for the new form items
for (j in 1:n){
     est.icc2[j] <- c2[j] + (1-c2[j])/(1+exp(-a2[j]*(theta0-rescaled.estimated.b2[j])))
          }
est.tcc2 <- sum(est.icc2)
est.tcc2 <- round(est.tcc2, digits = 4)
```

```
if (est.tcc2 == est.tau2[k]) {
                        est.tau_theta[k] <- theta0
                        break
                        }

if (est.tcc2 != est.tau2[k]) {
                    est.icc2.d   <- (1.7)*(1-est.icc2)*(est.icc2)
                    s_est.icc2.d <- sum(est.icc2.d)
                    new_theta    <- theta0 - ((est.tau2[k] - est.tcc2)/(-1*s_est.icc2.d))
                    theta0       <- new_theta
                    next
                        }
     }
#cat("Done with the IRT True Score Equating Number",k,"\n")
}

###Finding the estimated true score on the original form (est.tau1)
###corresponding to (est.tau_theta)
for (k in 2:ts){

for (j in 1:n){
      est.icc1[j] <- c1[j] + (1-c1[j])/(1+exp(-a1[j]*(est.tau_theta[k]-estimated.b1[j])))
            }
est.tcc1 <- sum(est.icc1)
est.tau1[k] <- round(est.tcc1, digits = 4)
}

bias.IRT.true.score[nr] <- mean(est.tau1 - tau1)
MSE.IRT.true.score[nr] <- mean((est.tau1 - tau1)^2)

bias.slope[nr]     <- (estimated.x - true.x)
bias.intercept[nr] <- (estimated.y - true.y)

MSE.slope[nr]      <- ((estimated.x - true.x)^2)
MSE.intercept[nr]  <- ((estimated.y - true.y)^2)

cat("Done with replication number", nr,"\n")
}

mean(bias.slope)
mean(MSE.slope)

mean(bias.intercept)
mean(MSE.intercept)
```

```
mean(bias.IRT.true.score)
mean(MSE.IRT.true.score)
```

## Appendix B

## R Script for Evaluating the Accuracy of Recovered Linking Coefficients (Slope and Intercept) and Recovered Equating True Scores in the Exposure Conditions Including 2 Exposed Anchor Items

```
setwd("C:/Users/Moatasim/Desktop/Master's Thesis/R Scripts")

#Setting the test length
n <- 50

#Setting the number of distinctive items
n.d <- 40

#Generating b values for distinctive items on the original form (OF)
b.d1_seed <- 3490123
set.seed(b.d1_seed)
b.d1 <- rnorm(n.d,0,1)

#4738237
#Generating b values for distinctive items on the new form (NF)
b.d2_seed <- 3490123
set.seed(b.d2_seed)
b.d2 <- rnorm(n.d,0,1)

#Setting the number of common items
n.a <- 10

#Generating b-values for common items
b.a_seed <- 1728240
set.seed(b.a_seed)
b.a <- rnorm(n.a,0,1)
mean(b.a)

true.b1 <- c(b.d1,b.a)
true.b2 <- c(b.d2,b.a)
mean(true.b1)

#Setting the number of examinees
N <- 500

#Generating theta-values for group 1 taking the OF
true.theta1_seed <- 2199345
set.seed(true.theta1_seed)
true.theta1 <- rnorm(N,0,1)
```

```
#Generating theta-values for group 2 taking the NF
true.theta2_seed <- 5692104
set.seed(true.theta2_seed)
true.theta2 <- rnorm(N,-0.50,1)

item_seed <- 2126354
set.seed(item_seed)
c1 <- vector(mode="numeric",n)
a1 <- matrix(1,n,1)
c2 <- vector(mode="numeric",n)
a2 <- matrix(1,n,1)
fix <- matrix(1,1,2)
fix[,1] <- n+1

library("ltm")

#Generating item response data set for the OF and NF
p1 <- matrix(0,N,n)
u1 <- matrix(0,N,n)

p2 <- matrix(0,N,n)
u2 <- matrix(0,N,n)

for (i in 1:N) {
    for (j in 1:n){
    p1[i,j] <- c1[j] + (1-c1[j])/(1+exp(-a1[j]*(true.theta1[i]-true.b1[j])))
    r1 <- runif(1,0,1)
        if (r1 <= p1[i,j]) {
        u1[i,j] <- 1
        }
    }
}

for (i in 1:N) {
    for (j in 1:n){
    p2[i,j] <- c2[j] + (1-c2[j])/(1+exp(-a2[j]*(true.theta2[i]-true.b2[j])))
    r2 <- runif(1,0,1)
        if (r2 <= p2[i,j]) {
        u2[i,j] <- 1
        }
    }
}

#Running IRT analysis
#Rasch
```

```
dichrasch1 <- rasch(data=u1,IRT.param=TRUE,constraint=fix,control=list(iter.qN=500))
dichrasch2 <- rasch(data=u2,IRT.param=TRUE,constraint=fix,control=list(iter.qN=500))

#Item Difficulty Parameters for the OF and NF items
b1 <- coef(dichrasch1)[,1]
b2 <- coef(dichrasch2)[,1]

#Ability parameters for group 1 taking the OF
t.est1 <- factor.scores(dichrasch1,method="EAP",resp.patterns=u1)
theta1 <- t.est1$score.dat[,n+3]

#Ability parameters for group 2 taking the NF
t.est2 <- factor.scores(dichrasch2,method="EAP",resp.patterns=u2)
theta2 <- t.est2$score.dat[,n+3]

#Equating Process for True Parameters

#The true Item Parameters for the Common Items across the Two Forms

F1.anchor.item.par <- c(b1[41],b1[42],b1[43],b1[44],b1[45],
                        b1[46],b1[47],b1[48],b1[49],b1[50])

F2.anchor.item.par <- c(b2[41],b2[42],b2[43],b2[44],b2[45],
                        b2[46],b2[47],b2[48],b2[49],b2[50])

mean.F1.anchor.item.par <- mean(F1.anchor.item.par)
mean.F2.anchor.item.par <- mean(F2.anchor.item.par)

SD.F1.anchor.item.par <- sd(F1.anchor.item.par)
SD.F2.anchor.item.par <- sd(F2.anchor.item.par)

#True Scaling Coefficients

true.x <- SD.F1.anchor.item.par/SD.F2.anchor.item.par
true.y <- mean.F1.anchor.item.par - (true.x * mean.F2.anchor.item.par)

#Rescaled True Item Parameters

rescaled_b2     <- true.x * b2 + true.y
rescaled_theta2 <- true.x * theta2 + true.y

#IRT True Score Equating Process

# tau1: True Score on the OF
# tau2: True Score on the NF
# tau_theta: Theta to which tau 1 and tau 2 correspond
```

```
# icc1: Item Characteristic Curve value for item j on the OF at a particular ability value
# icc2: Item Characteristic Curve value for item j on the NF at a particular ability value
# tcc1: Test Characteristic Curve value over items on the OF at a particular ability value
# tcc2: Test Characteristic Curve value over items on the NF at a particular ability value
# icc2.d: The First Derivative value of the Item Characteristic Curve for item j on the NF
at a particular ability value
# s_icc2.d: Sum of the First Derivative values of the Item Characteristic Curve over items
on the NF at a particular ability value
# theta0: The Starting Value of Theta
# ts: Total Score (  0 < ts  < 50)

ts <- 51

icc1          <- vector(mode="numeric",n)
icc2          <- vector(mode="numeric",n)
icc2.d        <- vector(mode="numeric",n)
tau_theta     <- vector(mode="numeric",ts)
tau1          <- vector(mode="numeric",ts)
tau2          <- vector(mode="numeric",ts)

#Applying the Newton-Raphson Method to find theta (tau_theta)

theta0 <- -14.7

#specifying a true score on the new form (tau2)
for (k in 1:1){

tau2[k] <- k-1
h <- 0
repeat{
h <- h +1
#Calculating the ICC values for the new form at theta0
for (j in 1:n){
            icc2[j] <- c2[j] + (1-c2[j])/(1+exp(-a2[j]*(theta0 -rescaled_b2[j])))
           }

tcc2 <- sum(icc2)
tcc2 <- round(tcc2, digits = 4)

#Finding theta (tau_theta) corresponding to the true score on the new form (tau2)
if (tcc2 == tau2[k]) {
                tau_theta[k] <- theta0
                break
                }

if (tcc2 != tau2[k]) {
```

```
                icc2.d     <- (1.7)*(1-icc2)*(icc2)
                s_icc2.d   <- sum(icc2.d)
                new_theta <- theta0 - ((tau2[k] - tcc2)/(-1*s_icc2.d))
                theta0 <- new_theta
                next
                }
    }
cat("Done with the IRT True Score Equating Number",k,"\n")
}

#Finding the true score on the original form (tau1)
for (k in 1:1){

for (j in 1:n){
            icc1[j] <- c1[j] + (1-c1[j])/(1+exp(-a1[j]*(tau_theta[k]-b1[j])))
           }
tcc1 <- sum(icc1)
tau1[k] <- round(tcc1, digits = 4)
            }

#############################################################################
#Setting a starting value for theta
theta0 <- -5

#specifying a true score on the new form (tau2)
for (k in 2:ts){

tau2[k] <- (k-1)

repeat{

#Calculating the ICC values for the new form at theta0
for (j in 1:n){
            icc2[j] <- c2[j] + (1-c2[j])/(1+exp(-a2[j]*(theta0 -rescaled_b2[j])))
           }

tcc2 <- sum(icc2)
tcc2 <- round(tcc2, digits = 4)

#Finding theta (tau_theta) corresponding to the true score on the new form (tau2)
if (tcc2 == tau2[k]) {
                tau_theta[k] <- theta0
                break
                }

if (tcc2 != tau2[k]) {
```

```
                icc2.d     <- (1.7)*(1-icc2)*(icc2)
                s_icc2.d   <- sum(icc2.d)
                new_theta <- theta0 - ((tau2[k] - tcc2)/(-1*s_icc2.d))
                theta0 <- new_theta
                next
                }
    }
cat("Done with the IRT True Score Equating Number",k,"\n")
}

#Finding the true score on the original form (tau1)
for (k in 2:ts){

for (j in 1:n){
            icc1[j] <- c1[j] + (1-c1[j])/(1+exp(-a1[j]*(tau_theta[k]-b1[j])))
           }
tcc1 <- sum(icc1)
tau1[k] <- round(tcc1, digits = 4)
            }

#Evaluating the accuracy of recovered linking constants and equating true scores for
exposure conditions

NREPS <- 100

bias.slope     <- vector(mode="numeric",NREPS)
bias.intercept <- vector(mode="numeric",NREPS)

MSE.slope     <- vector(mode="numeric",NREPS)
MSE.intercept <- vector(mode="numeric",NREPS)

bias.IRT.true.score <- vector(mode="numeric",NREPS)
MSE.IRT.true.score <- vector(mode="numeric",NREPS)

for (nr in 1:NREPS) {

#Generate data set
p1 <- matrix(0,N,n)
u1 <- matrix(0,N,n)

p2 <- matrix(0,N,n)
u2 <- matrix(0,N,n)

for (i in 1:N) {
    for (j in 1:n){
    p1[i,j] <- c1[j] + (1-c1[j])/(1+exp(-a1[j]*(theta1[i]-b1[j])))
```

```
    r1 <- runif(1,0,1)
        if (r1 <= p1[i,j]) {
        u1[i,j] <- 1
        }
    }
}

#specifying the percentage of examinees with pre-knowledge

percent <- 0.05

M <- percent*N

for (i in 1:(N-M)) {
    for (j in 1:n){
    p2[i,j] <- c2[j] + (1-c2[j])/(1+exp(-a2[j]*(theta2[i]-b2[j])))
    r2 <- runif(1,0,1)
        if (r2 <= p2[i,j]) {
           u2[i,j] <- 1
                    }
            }
           }

for (i in (N-M+1):N){
      for (j in 1:n){
       if (j==43|j==45){
       p2[i,j] <- 1 + c2[j] + (1-c2[j])/(1+exp(-a2[j]*(theta2[i]-b2[j])))
       r2 <- runif(1,0,1)
          if (r2 <= p2[i,j]){
                    u2[i,j] <- 1
                    }
                }
       if (j!=43|j!=45) {
       p2[i,j] <- c2[j] + (1-c2[j])/(1+exp(-a2[j]*(theta2[i]-b2[j])))
       r2 <- runif(1,0,1)
          if (r2 <= p2[i,j]){
                    u2[i,j] <- 1
                    }
                  }

                }
              }


#Run IRT analysis
#Rasch
```

```
dichrasch1 <- rasch(data=u1,IRT.param=TRUE,constraint=fix,control=list(iter.qN=500))
dichrasch2 <- rasch(data=u2,IRT.param=TRUE,constraint=fix,control=list(iter.qN=500))

#Item Difficulty Parameters Estimates
estimated.b1 <- coef(dichrasch1)[,1]
estimated.b2 <- coef(dichrasch2)[,1]

#Equating Process for Estimates

#The Estimated Item Parameters for the Common Items across the Two Forms

F1.estimated.anchor.items <- c(estimated.b1[41],estimated.b1[42],estimated.b1[43],
                    estimated.b1[44],estimated.b1[45],estimated.b1[46],
                    estimated.b1[47],estimated.b1[48],estimated.b1[49],
                    estimated.b1[50])

F2.estimated.anchor.items <- c(estimated.b2[41],estimated.b2[42],estimated.b2[43],
                    estimated.b2[44],estimated.b2[45],estimated.b2[46],
                    estimated.b2[47],estimated.b2[48],estimated.b2[49],
                    estimated.b2[50])

#The Mean and Standard Deviation of the Estimated Item Parameters
#for the Common Items across the Two Forms

mean.F1.estimated.anchor.items <- mean(F1.estimated.anchor.items)
mean.F2.estimated.anchor.items <- mean(F2.estimated.anchor.items)

SD.F1.estimated.anchor.items <- sd(F1.estimated.anchor.items)
SD.F2.estimated.anchor.items <- sd(F2.estimated.anchor.items)

#Estimated Scaling Coefficients

estimated.x <- SD.F1.estimated.anchor.items/SD.F2.estimated.anchor.items
estimated.y <- mean.F1.estimated.anchor.items -(estimated.x *
mean.F2.estimated.anchor.items)

#Rescaled Estimated Item and ability Parameters

rescaled.estimated.b2 <- estimated.x * estimated.b2 + estimated.y

t.est2 <- factor.scores(dichrasch2,method="EAP",resp.patterns=u2)
estimated.theta2 <- t.est2$score.dat[,n+3]
rescaled.estimated.theta2 <- estimated.x * estimated.theta2 + estimated.y

#Estimated IRT True Score Equating
```

```
est.icc1      <- vector(mode="numeric",n)
est.icc2      <- vector(mode="numeric",n)
est.icc2.d    <- vector(mode="numeric",n)
est.tau_theta <- vector(mode="numeric",ts)
est.tau1      <- vector(mode="numeric",ts)
est.tau2      <- vector(mode="numeric",ts)


theta0 <- -14.7

#Specifying an estimated true score on the new form (est.tau2)
for (k in 1:1){

est.tau2[k] <- k-1

#Finding estimated theta (est.tau_theta) corresponding to (est.tau2)

repeat{

#Calculating estimated ICCs for the new form items
for (j in 1:n){
      est.icc2[j] <- c2[j] + (1-c2[j])/(1+exp(-a2[j]*(theta0-rescaled.estimated.b2[j])))
           }
est.tcc2 <- sum(est.icc2)
est.tcc2 <- round(est.tcc2, digits = 4)


if (est.tcc2 == est.tau2[k]) {
                      est.tau_theta[k] <- theta0
                      break
                      }

if (est.tcc2 != est.tau2[k]) {
                  est.icc2.d   <- (1.7)*(1-est.icc2)*(est.icc2)
                  s_est.icc2.d <- sum(est.icc2.d)
                  new_theta    <- theta0 - ((est.tau2[k] - est.tcc2)/(-1*s_est.icc2.d))
                  theta0       <- new_theta
                  next
                       }
    }
#cat("Done with the IRT True Score Equating Number",k,"\n")
}

###Finding the estimated true score on the original form (est.tau1)
###corresponding to (est.tau_theta)
for (k in 1:1){
```

```
for (j in 1:n){
      est.icc1[j] <- c1[j] + (1-c1[j])/(1+exp(-a1[j]*(est.tau_theta[k]-estimated.b1[j])))
            }
est.tcc1 <- sum(est.icc1)
est.tau1[k] <- round(est.tcc1, digits = 4)
}

#################################

theta0 <- -5

#Specifying an estimated true score on the new form (est.tau2)
for (k in 2:ts){

est.tau2[k] <- (k-1)

#Finding estimated theta (est.tau_theta) corresponding to (est.tau2)

repeat{

#Calculating estimated ICCs for the new form items
for (j in 1:n){
       est.icc2[j] <- c2[j] + (1-c2[j])/(1+exp(-a2[j]*(theta0-rescaled.estimated.b2[j])))
            }
est.tcc2 <- sum(est.icc2)
est.tcc2 <- round(est.tcc2, digits = 4)


if (est.tcc2 == est.tau2[k]) {
                        est.tau_theta[k] <- theta0
                        break
                        }

if (est.tcc2 != est.tau2[k]) {
                    est.icc2.d   <- (1.7)*(1-est.icc2)*(est.icc2)
                    s_est.icc2.d <- sum(est.icc2.d)
                    new_theta    <- theta0 - ((est.tau2[k] - est.tcc2)/(-1*s_est.icc2.d))
                    theta0       <- new_theta
                    next
                        }
     }
#cat("Done with the IRT True Score Equating Number",k,"\n")
}

###Finding the estimated true score on the original form (est.tau1)
```

```
###corresponding to (est.tau_theta)
for (k in 2:ts){

for (j in 1:n){
      est.icc1[j] <- c1[j] + (1-c1[j])/(1+exp(-a1[j]*(est.tau_theta[k]-estimated.b1[j])))
           }
est.tcc1 <- sum(est.icc1)
est.tau1[k] <- round(est.tcc1, digits = 4)
}

bias.IRT.true.score[nr] <- mean(est.tau1 - tau1)
MSE.IRT.true.score[nr] <- mean((est.tau1 - tau1)^2)

bias.slope[nr]     <- (estimated.x - true.x)
bias.intercept[nr] <- (estimated.y - true.y)

MSE.slope[nr]      <- ((estimated.x - true.x)^2)
MSE.intercept[nr]  <- ((estimated.y - true.y)^2)

cat("Done with replication number", nr,"\n")
}

mean(bias.slope)
mean(MSE.slope)

mean(bias.intercept)
mean(MSE.intercept)

mean(bias.IRT.true.score)
mean(MSE.IRT.true.score)

sd(bias.slope)
sd(bias.intercept)
sd(bias.IRT.true.score)
```

## Appendix C

### R Script for Evaluating the Accuracy of Recovered Linking Coefficients (Slope and Intercept) and Recovered Equating True Scores in the Exposure Conditions Including 10 Exposed Anchor Items

```
setwd("C:/Users/Moatasim/Desktop/Master's Thesis/R Scripts")

#Setting the test length
n <- 50

#Setting the number of distinctive items
n.d <- 40

#Generating b values for distinctive items on the original form (OF)
b.d1_seed <- 3490123
set.seed(b.d1_seed)
b.d1 <- rnorm(n.d,0,1)

#4738237
#Generating b values for distinctive items on the new form (NF)
b.d2_seed <- 3490123
set.seed(b.d2_seed)
b.d2 <- rnorm(n.d,0,1)

#Setting the number of common items
n.a <- 10

#Generating b-values for common items
b.a_seed <- 1728240
set.seed(b.a_seed)
b.a <- rnorm(n.a,0,1)
mean(b.a)

true.b1 <- c(b.d1,b.a)
true.b2 <- c(b.d2,b.a)
mean(true.b1)
#Setting the number of examinees
N <- 500

#Generating theta-values for group 1 taking the OF
true.theta1_seed <- 2199345
set.seed(true.theta1_seed)
true.theta1 <- rnorm(N,0,1)

#Generating theta-values for group 2 taking the NF
```

```
true.theta2_seed <- 5692104
set.seed(true.theta2_seed)
true.theta2 <- rnorm(N,0.50,1)

item_seed <- 2126354
set.seed(item_seed)
c1 <- vector(mode="numeric",n)
a1 <- matrix(1,n,1)
c2 <- vector(mode="numeric",n)
a2 <- matrix(1,n,1)
fix <- matrix(1,1,2)
fix[,1] <- n+1

library("ltm")

#Generating item response data set for the OF and NF
p1 <- matrix(0,N,n)
u1 <- matrix(0,N,n)

p2 <- matrix(0,N,n)
u2 <- matrix(0,N,n)

for (i in 1:N) {
    for (j in 1:n){
    p1[i,j] <- c1[j] + (1-c1[j])/(1+exp(-a1[j]*(true.theta1[i]-true.b1[j])))
    r1 <- runif(1,0,1)
        if (r1 <= p1[i,j]) {
        u1[i,j] <- 1
        }
    }
}

for (i in 1:N) {
    for (j in 1:n){
    p2[i,j] <- c2[j] + (1-c2[j])/(1+exp(-a2[j]*(true.theta2[i]-true.b2[j])))
    r2 <- runif(1,0,1)
        if (r2 <= p2[i,j]) {
        u2[i,j] <- 1
        }
    }
}

#Running IRT analysis
#Rasch
dichrasch1 <- rasch(data=u1,IRT.param=TRUE,constraint=fix,control=list(iter.qN=500))
dichrasch2 <- rasch(data=u2,IRT.param=TRUE,constraint=fix,control=list(iter.qN=500))
```

```
#Item Difficulty Parameters for the OF and NF items
b1 <- coef(dichrasch1)[,1]
b2 <- coef(dichrasch2)[,1]

#Ability parameters for group 1 taking the OF
t.est1 <- factor.scores(dichrasch1,method="EAP",resp.patterns=u1)
theta1 <- t.est1$score.dat[,n+3]

#Ability parameters for group 2 taking the NF
t.est2 <- factor.scores(dichrasch2,method="EAP",resp.patterns=u2)
theta2 <- t.est2$score.dat[,n+3]

#Equating Process for True Parameters

#The true Item Parameters for the Common Items across the Two Forms

F1.anchor.item.par <- c(b1[41],b1[42],b1[43],b1[44],b1[45],
                        b1[46],b1[47],b1[48],b1[49],b1[50])

F2.anchor.item.par <- c(b2[41],b2[42],b2[43],b2[44],b2[45],
                        b2[46],b2[47],b2[48],b2[49],b2[50])

mean.F1.anchor.item.par <- mean(F1.anchor.item.par)
mean.F2.anchor.item.par <- mean(F2.anchor.item.par)

SD.F1.anchor.item.par <- sd(F1.anchor.item.par)
SD.F2.anchor.item.par <- sd(F2.anchor.item.par)

#True Scaling Coefficients

true.x <- SD.F1.anchor.item.par/SD.F2.anchor.item.par
true.y <- mean.F1.anchor.item.par - (true.x * mean.F2.anchor.item.par)

#Rescaled True Item Parameters

rescaled_b2     <- true.x * b2 + true.y
rescaled_theta2 <- true.x * theta2 + true.y

#IRT True Score Equating Process

# tau1: True Score on the OF
# tau2: True Score on the NF
# tau_theta: Theta to which tau 1 and tau 2 correspond
# icc1: Item Characteristic Curve value for item j on the OF at a particular ability value
# icc2: Item Characteristic Curve value for item j on the NF at a particular ability value
```

```
# tcc1: Test Characteristic Curve value over items on the OF at a particular ability value
# tcc2: Test Characteristic Curve value over items on the NF at a particular ability value
# icc2.d: The First Derivative value of the Item Characteristic Curve for item j on the NF
at a particular ability value
# s_icc2.d: Sum of the First Derivative values of the Item Characteristic Curve over items
on the NF at a particular ability value
# theta0: The Starting Value of Theta
# ts: Total Score (  0 < ts  < 50)

ts <- 51

icc1          <- vector(mode="numeric",n)
icc2          <- vector(mode="numeric",n)
icc2.d        <- vector(mode="numeric",n)
tau_theta     <- vector(mode="numeric",ts)
tau1          <- vector(mode="numeric",ts)
tau2          <- vector(mode="numeric",ts)

#Applying the Newton-Raphson Method to find theta (tau_theta)

theta0 <- -14.7

#specifying a true score on the new form (tau2)
for (k in 1:1){

tau2[k] <- k-1
h <- 0
repeat{
h <- h +1
#Calculating the ICC values for the new form at theta0
for (j in 1:n){
            icc2[j] <- c2[j] + (1-c2[j])/(1+exp(-a2[j]*(theta0 -rescaled_b2[j])))
           }

tcc2 <- sum(icc2)
tcc2 <- round(tcc2, digits = 4)

#Finding theta (tau_theta) corresponding to the true score on the new form (tau2)
if (tcc2 == tau2[k]) {
                tau_theta[k] <- theta0
                break
                }

if (tcc2 != tau2[k]) {
                icc2.d     <- (1.7)*(1-icc2)*(icc2)
                s_icc2.d   <- sum(icc2.d)
```

```
                new_theta <- theta0 - ((tau2[k] - tcc2)/(-1*s_icc2.d))
                theta0 <- new_theta
                next
               }
    }
cat("Done with the IRT True Score Equating Number",k,"\n")
}

#Finding the true score on the original form (tau1)
for (k in 1:1){

for (j in 1:n){
            icc1[j] <- c1[j] + (1-c1[j])/(1+exp(-a1[j]*(tau_theta[k]-b1[j])))
           }
tcc1 <- sum(icc1)
tau1[k] <- round(tcc1, digits = 4)
            }

#########################################################################
#Setting a starting value for theta
theta0 <- -5

#specifying a true score on the new form (tau2)
for (k in 2:ts){

tau2[k] <- (k-1)

repeat{

#Calculating the ICC values for the new form at theta0
for (j in 1:n){
            icc2[j] <- c2[j] + (1-c2[j])/(1+exp(-a2[j]*(theta0 -rescaled_b2[j])))
           }

tcc2 <- sum(icc2)
tcc2 <- round(tcc2, digits = 4)

#Finding theta (tau_theta) corresponding to the true score on the new form (tau2)
if (tcc2 == tau2[k]) {
                tau_theta[k] <- theta0
                break
                }

if (tcc2 != tau2[k]) {
                icc2.d     <- (1.7)*(1-icc2)*(icc2)
                s_icc2.d   <- sum(icc2.d)
```

```
                new_theta <- theta0 - ((tau2[k] - tcc2)/(-1*s_icc2.d))
                theta0 <- new_theta
                next
               }
    }
cat("Done with the IRT True Score Equating Number",k,"\n")
}

#Finding the true score on the original form (tau1)
for (k in 2:ts){

for (j in 1:n){
            icc1[j] <- c1[j] + (1-c1[j])/(1+exp(-a1[j]*(tau_theta[k]-b1[j])))
           }
tcc1 <- sum(icc1)
tau1[k] <- round(tcc1, digits = 4)
            }

#Evaluating the accuracy of recovered linking constants and equating true scores for
exposure conditions

NREPS <- 100

bias.slope     <- vector(mode="numeric",NREPS)
bias.intercept <- vector(mode="numeric",NREPS)

MSE.slope     <- vector(mode="numeric",NREPS)
MSE.intercept <- vector(mode="numeric",NREPS)

bias.IRT.true.score <- vector(mode="numeric",NREPS)
MSE.IRT.true.score <- vector(mode="numeric",NREPS)

for (nr in 1:NREPS) {

#Generate data set
p1 <- matrix(0,N,n)
u1 <- matrix(0,N,n)

p2 <- matrix(0,N,n)
u2 <- matrix(0,N,n)

for (i in 1:N) {
    for (j in 1:n){
    p1[i,j] <- c1[j] + (1-c1[j])/(1+exp(-a1[j]*(theta1[i]-b1[j])))
    r1 <- runif(1,0,1)
        if (r1 <= p1[i,j]) {
```

```
          u1[i,j] <- 1
          }
     }
}

#specifying the percentage of examinees with pre-knowledge

percent <- 0.20

M <- percent*N

for (i in 1:(N-M)) {
     for (j in 1:n){
     p2[i,j] <- c2[j] + (1-c2[j])/(1+exp(-a2[j]*(theta2[i]-b2[j])))
     r2 <- runif(1,0,1)
          if (r2 <= p2[i,j]) {
             u2[i,j] <- 1
                        }
               }
              }

for (i in (N-M+1):N){
        for (j in 1:n){
         if (j==41|j==42|j==43|j==44|j==45|j==46|j==47|j==48|j==49|j==50){
         p2[i,j] <- 1 + c2[j] + (1-c2[j])/(1+exp(-a2[j]*(theta2[i]-b2[j])))
         r2 <- runif(1,0,1)
            if (r2 <= p2[i,j]){
                        u2[i,j] <- 1
                        }
                    }
        if (j!=41|j!=42|j!=43|j!=44|j!=45|j!=46|j!=47|j!=48|j!=49|j!=50){
        p2[i,j] <- c2[j] + (1-c2[j])/(1+exp(-a2[j]*(theta2[i]-b2[j])))
        r2 <- runif(1,0,1)
          if (r2 <= p2[i,j]){
                      u2[i,j] <- 1
                      }
                    }

                  }
               }


#Run IRT analysis
#Rasch
dichrasch1 <- rasch(data=u1,IRT.param=TRUE,constraint=fix,control=list(iter.qN=500))
dichrasch2 <- rasch(data=u2,IRT.param=TRUE,constraint=fix,control=list(iter.qN=500))
```

```
#Item Difficulty Parameters Estimates
estimated.b1 <- coef(dichrasch1)[,1]
estimated.b2 <- coef(dichrasch2)[,1]

#Equating Process for Estimates

#The Estimated Item Parameters for the Common Items across the Two Forms

F1.estimated.anchor.items <- c(estimated.b1[41],estimated.b1[42],estimated.b1[43],
                    estimated.b1[44],estimated.b1[45],estimated.b1[46],
                    estimated.b1[47],estimated.b1[48],estimated.b1[49],
                    estimated.b1[50])

F2.estimated.anchor.items <- c(estimated.b2[41],estimated.b2[42],estimated.b2[43],
                    estimated.b2[44],estimated.b2[45],estimated.b2[46],
                    estimated.b2[47],estimated.b2[48],estimated.b2[49],
                    estimated.b2[50])

#The Mean and Standard Deviation of the Estimated Item Parameters
#for the Common Items across the Two Forms

mean.F1.estimated.anchor.items <- mean(F1.estimated.anchor.items)
mean.F2.estimated.anchor.items <- mean(F2.estimated.anchor.items)

SD.F1.estimated.anchor.items <- sd(F1.estimated.anchor.items)
SD.F2.estimated.anchor.items <- sd(F2.estimated.anchor.items)

#Estimated Scaling Coefficients

estimated.x <- SD.F1.estimated.anchor.items/SD.F2.estimated.anchor.items
estimated.y <- mean.F1.estimated.anchor.items -(estimated.x *
mean.F2.estimated.anchor.items)

#Rescaled Estimated Item and ability Parameters

rescaled.estimated.b2 <- estimated.x * estimated.b2 + estimated.y

t.est2 <- factor.scores(dichrasch2,method="EAP",resp.patterns=u2)
estimated.theta2 <- t.est2$score.dat[,n+3]
rescaled.estimated.theta2 <- estimated.x * estimated.theta2 + estimated.y

#Estimated IRT True Score Equating

est.icc1     <- vector(mode="numeric",n)
est.icc2     <- vector(mode="numeric",n)
```

```
est.icc2.d    <- vector(mode="numeric",n)
est.tau_theta <- vector(mode="numeric",ts)
est.tau1      <- vector(mode="numeric",ts)
est.tau2      <- vector(mode="numeric",ts)


theta0 <- -14.7

#Specifying an estimated true score on the new form (est.tau2)
for (k in 1:1){

est.tau2[k] <- k-1

#Finding estimated theta (est.tau_theta) corresponding to (est.tau2)

repeat{

#Calculating estimated ICCs for the new form items
for (j in 1:n){
       est.icc2[j] <- c2[j] + (1-c2[j])/(1+exp(-a2[j]*(theta0-rescaled.estimated.b2[j])))
            }
est.tcc2 <- sum(est.icc2)
est.tcc2 <- round(est.tcc2, digits = 4)


if (est.tcc2 == est.tau2[k]) {
                        est.tau_theta[k] <- theta0
                        break
                        }

if (est.tcc2 != est.tau2[k]) {
                    est.icc2.d   <- (1.7)*(1-est.icc2)*(est.icc2)
                    s_est.icc2.d <- sum(est.icc2.d)
                    new_theta    <- theta0 - ((est.tau2[k] - est.tcc2)/(-1*s_est.icc2.d))
                    theta0       <- new_theta
                    next
                        }
     }
#cat("Done with the IRT True Score Equating Number",k,"\n")
}

###Finding the estimated true score on the original form (est.tau1)
###corresponding to (est.tau_theta)
for (k in 1:1){

for (j in 1:n){
```

```
      est.icc1[j] <- c1[j] + (1-c1[j])/(1+exp(-a1[j]*(est.tau_theta[k]-estimated.b1[j])))
           }
est.tcc1 <- sum(est.icc1)
est.tau1[k] <- round(est.tcc1, digits = 4)
}

#################################

theta0 <- -5

#Specifying an estimated true score on the new form (est.tau2)
for (k in 2:ts){

est.tau2[k] <- (k-1)

#Finding estimated theta (est.tau_theta) corresponding to (est.tau2)

repeat{

#Calculating estimated ICCs for the new form items
for (j in 1:n){
      est.icc2[j] <- c2[j] + (1-c2[j])/(1+exp(-a2[j]*(theta0-rescaled.estimated.b2[j])))
           }
est.tcc2 <- sum(est.icc2)
est.tcc2 <- round(est.tcc2, digits = 4)


if (est.tcc2 == est.tau2[k]) {
                    est.tau_theta[k] <- theta0
                    break
                    }

if (est.tcc2 != est.tau2[k]) {
                  est.icc2.d   <- (1.7)*(1-est.icc2)*(est.icc2)
                  s_est.icc2.d <- sum(est.icc2.d)
                  new_theta    <- theta0 - ((est.tau2[k] - est.tcc2)/(-1*s_est.icc2.d))
                  theta0       <- new_theta
                  next
                       }
    }
#cat("Done with the IRT True Score Equating Number",k,"\n")
}

###Finding the estimated true score on the original form (est.tau1)
###corresponding to (est.tau_theta)
for (k in 2:ts){
```

```
for (j in 1:n){
      est.icc1[j] <- c1[j] + (1-c1[j])/(1+exp(-a1[j]*(est.tau_theta[k]-estimated.b1[j])))
            }
est.tcc1 <- sum(est.icc1)
est.tau1[k] <- round(est.tcc1, digits = 4)
}

bias.IRT.true.score[nr] <- mean(est.tau1 - tau1)
MSE.IRT.true.score[nr] <- mean((est.tau1 - tau1)^2)

bias.slope[nr]     <- (estimated.x - true.x)
bias.intercept[nr] <- (estimated.y - true.y)

MSE.slope[nr]      <- ((estimated.x - true.x)^2)
MSE.intercept[nr]  <- ((estimated.y - true.y)^2)

cat("Done with replication number", nr,"\n")
}

mean(bias.slope)
mean(MSE.slope)

mean(bias.intercept)
mean(MSE.intercept)

mean(bias.IRT.true.score)
mean(MSE.IRT.true.score)

sd(bias.slope)
sd(bias.intercept)
sd(bias.IRT.true.score)
```