James Madison University
JMU Scholarly Commons

Masters Theses The Graduate School

Spring 2011

The impact of cheating on IRT equating under the non-equivalent anchor test design

Daniel Pacheco Jurich
James Madison University

The Impact of Cheating on IRT Equating under the Non-equivalent Anchor Test Design

Daniel Jurich

A thesis submitted to the Graduate Faculty of

JAMES MADISON UNIVERSITY

In

Partial Fulfillment of the Requirements

for the degree of

Master of Arts

Department of Graduate Psychology

May 2011

Acknowledgements

I am indebted to numerous individuals who have contributed to my graduate experience and made the completion of this manuscript possible. First, I must thank my advisor, Dr. Christine DeMars, for her invaluable guidance, patience, and eagerness to assist me in all aspects of my development. Her mentorship was a catalyst for the actualization of this project. I also must express my gratitude for the continued support and, more importantly, friendship I have received from Dr. Joshua Goodman. Without the influence of these two wonderful individuals, I would have never developed into the student and individual I am now.

I would like to thank the other members of my committee, Dr. Dena Pastor and Dr. Sara Finney, for the time they invested into ensuring the quality of this manuscript. Their expertise, instruction, and dedication were essential to the completion of this project. They have also played an instrumental role in my growth as a student. Along these lines, I must also acknowledge all the faculty and staff at the Center for Assessment and Research Studies for instilling a culture of support that has enriched my education.

It is necessary for me to recognize the friends who supported me throughout this journey. I would like thank Megan France, James Koepfler, Becca Marsh and Anna Zilberberg for their friendship and guidance in my transition into graduate school. I must also acknowledge Jason Kopp, Jerusha Gerstner, and Megan Rodgers for keeping my spirits lifted throughout the trials and tribulations of this process.

Last, but certainly not least, I would like to express my appreciation for my parents, John and Sonia Jurich, for their unrelenting encouragement and understanding of my graduate experience.

Table of Contents

List of Tables

List of Figures

Abstract

The prevalence of high stakes test scores as a basis for significant decisions necessitates the dissemination of accurate and fair scores. However, the magnitude of these decisions has created an environment prone to examinees resorting to cheating. To reduce the risk of cheating, multiple test forms are commonly administered. When multiple forms are employed, the forms must be equated to account for potential differences in form difficulty. If cheating occurs on one of the forms, the equating procedure may produce inaccurate results. A simulation study was conducted to examine the impact of cheating on IRT true score equating. Recovery of equated scores and scaling constants were assessed for five IRT scaling methods under various conditions. Results indicated that cheating artificially increased the equated scores of the entire examinee group administered the compromised form and no scaling methods adequately mitigated this effect. Future research should focus on the identification and removal of compromised items.

CHAPTER 1

Introduction

The magnitude of the consequences associated with high-stakes tests has created an environment where some examinees resort to cheating (Cizek, 1999). A single score on a high-stakes test is often used to make potentially life-altering decisions about an individual. For example, test scores may determine whether an individual is certified to practice in a chosen profession or a student is accepted into an educational institution. With the increased prevalence of high-stakes testing, specifically in education through the No Child Left Behind legislation (NCLB, 2001), the threat of examinee impropriety has increased dramatically (Cohen & Wollack, 2006). When stakes are high, the testing organization must be certain that test scores accurately reflect each individual’s true ability. If any form of cheating, including prior knowledge of the item, influences the responses on a test, scores will inaccurately represent the individual’s ability and decisions based on the scores will be dubious at best (Haladyna & Downing, 2004).

The susceptibility of high-stakes tests to cheating behaviors is increased through the re-use of test items. To prevent the spread of test items, testing companies frequently use multiple forms when tests are administered on different occasions. Despite attempts to create parallel test forms, the multiple forms that testing programs develop often vary in terms of difficulty. As a result, slight differences in form difficulty may unfairly disadvantage some test-takers. To assess test-takers accurately, equating designs are utilized in an effort to produce comparable scores across forms of varying difficulty. The most common equating design employed in large-scale testing is the non-equivalent anchor test design (NEAT; Holland, 2007). Under the NEAT design, common items, also called anchor items, are placed on the multiple forms of the test and used to estimate the relationship between form difficulties. Consequently, as the number of forms increases, the exposure rate of anchor items also increases, rendering them prone to contamination. Because the anchor items are integral to the equating process, the contamination of anchor items will not only distort the meaning of an individual's test score, but also inaccurately represent the differences between the test forms, potentially changing the scores for many examinees.

Given the important decisions made from and the expanding use of high-stakes testing, it is necessary to investigate the potential effects of cheating on the equating process. Although much attention has been devoted to test security (Finkelman, Nering & Roussos, 2009; van der Linden & Veldkamp, 2004) and cheating detection (McLeod, Lewis & Thissen 2003; Wollack, 2006), little emphasis has been placed on understanding the interaction between cheating and the equating process. The purpose of this study is to examine how test scores are affected when cheating influences the equating process.

Cheating in High-Stakes Testing

The use of test scores to make high-stakes decisions has been frequently cited as motivation for test-takers to cheat (Nichols & Berliner, 2007). This motivation to engage in cheating behaviors undoubtedly increases as the stakes of the test increase. As previously stated, scores on educational tests, such as the SAT, GRE, LSAT and MCAT, often factor into selection. Low scores may impede access to higher education or to a desired career path. For certification and licensure tests, a single cut-off point may determine whether an individual receives a desired credential or is certified to practice in a certain occupation. In these situations, even a slight increase in performance may push an individual past the critically important cut-off point.

With the expansion of accountability testing, cheating has become disturbingly ubiquitous in society. Although it is difficult to ascertain the degree of cheating, particularly for large-scale exams due to the intense security surrounding testing procedures and possible negative publicity in revealing that cheating occurs (Cizek,1999), there is evidence that cheating behavior has seen a dramatic rise over time. A study conducted over the course of thirty years, from 1963 to 1993, on a sample of students from 99 United States based colleges revealed that test related cheating behaviors (copying from another examinee, providing answers to another student, and using prohibited notes) have each increased considerably (McCabe & Trevino, 1996). As a whole, the proportion of students from these colleges engaging in serious test related cheating has risen from 39% in 1963 to 64% in 1993, where serious test cheating was defined as copying answers, using prohibited notes, or having a confederate take the exam (McCabe, Trevino & Butterfield, 2001).

Although no studies report the prevalence of cheating in high-stakes testing, there is no shortage of individual examples of cheating incidents. The Educational Testing Service (ETS) investigated a case of cheating on the Test of English as a Foreign Language (TOEFL) which led to the arrest and the potential deportation of 61 students (Li, 2003). Cisco Systems, a software security corporation, and Pearson VUE, a leading certification testing company, implemented new anti-cheating software to detect fraudulent examinees taking the test for another individual (Baron & Wirzbicki, 2008). An eight-month trial run of this software on a fraction of the company's certification exams caught 200 of these proxy test takers. Moreover, the president of the exam security vendor CertGaurd, Robert Williams, suspects that entire test forms may be compromised (all items obtained and released) in as few as 24 hours after the form's initial use (Brodkin, 2008). It is clear that cheating still poses a serious threat despite the immense precautions and highly controlled environment employed in high-stakes testing.

The methods employed by cheaters have become more ingenious in reaction to the increased security and pressure of testing (Cizek, 1999). There are many small scale methods exploited to cheat on tests. These methods are local to the test-taking site and include looking at other answer sheets, relaying information about the test through codes, bringing in cheat sheets hidden in discrete places, and having a substitute take the test for an individual. Localized methods will impact a small number of scores and are difficult to carry out in the strictly controlled and proctored testing environment that many large-scale tests utilize.

Large-scale cheating methods pose a much greater dilemma to testing programs, those interpreting the test scores, and the public. Large-scale cheating refers to severe breaches of test security that provide future test-takers access to items. One example of a large-scale cheating method is item harvesting. This method involves individuals or groups who take examinations with the intent to memorize and distribute the items. Item harvesting may severely compromise the item bank used to create test forms. Moreover, distribution of these items has become effortless with the proliferation of the internet. Cizek (1999) describes one such instance where item harvesters on the east coast would distribute items to west coast clients before they took the same tests later that day.

Research on statistical methods for preventing and detecting cheating has become increasingly popular with the rise of testing. A large portion of this research focuses on controlling the exposure rate of items in computer adaptive tests. Exposure rate refers to the proportion of examinees receiving an item in comparison to the total number of examinees administered the test. Put simply, items administered to the most examinees have the highest exposure rates. Therefore, items with high exposure rates are more vulnerable to contamination. In addition, the potential dissemination of these items poses a serious threat to the validity of test scores as the items are frequently encountered by examinees. It is necessary to control the exposure rate of items to prevent widespread dissemination of specific items, as well as reduce the negative impact resulting from a security breach.

Several mathematical techniques have been developed in computer adaptive tests to limit the appearance of items (e.g., Stocking & Lewis, 1998), thus minimizing the opportunities for item harvesting. In non-adaptive testing, where forms are generated prior to administration of the test, item exposure is controlled through periodic testing administrations using different forms. Within each administration, the items that appear on a given form can be strictly controlled. In this case, testing companies can ensure that no item is repeated from one administration period to the next. However, the anchor items used in the NEAT equating design have necessarily appeared to a subset, if not all, of the examinees who have previously taken the exam. Therefore, the NEAT equating design introduces both positive and negative consequences for test-security. On one hand, the design controls the exposure of unique items on each form. In contrast, the threat of anchor set contamination is higher. Thus, exposure is minimized for a majority of the items; however, exposure rates are raised for the items vital to establishing score equity across forms.

Test Equating

Equating is a process that adjusts scores from multiple forms to account for differences in form difficulty (Kolen & Brennan, 2004). When the equating function is accurate, the resulting scores on either form can be used interchangeably. Without equating, the form administered may influence scores intended to measure the same construct on a specified scale. Thus, equating is necessary to be fair to examinees receiving different forms and to help ensure that accurate decisions are made based on the scores.

The need for equating can be demonstrated in the following situation. Consider two candidates applying for a position that uses scores on a certification exam as a measure of qualification. The candidates take the exam on different dates and thus are administered different forms. If the forms differ in difficulty, one candidate would be unfairly advantaged. The resulting scores would not necessarily be attributable to the ability of the candidates as they are confounded with form difficulty. This may create a situation where one candidate scores lower than the other solely due to the relative difficulty of their form. Moreover, failure to account for unequal difficulties in this situation may lead to the hiring of a less qualified candidate.

Equating plays an especially prominent role in high-stakes testing. The importance of test security requires that testing companies develop multiple forms, as repeated administration of the same items would assuredly result in items becoming compromised. In addition, the significant decisions rendered based on high-stakes testing require the equating function to be as precise as possible. Slight errors in the equating process can alter the scores for a large portion of the examinees (Holland & Dorans, 2006). The consequences of an equating error may lead to students incorrectly being denied entry to the college of their choice or qualified professionals failing to achieve certification. In addition to the consequences for examinees, legal action can be taken against the testing companies when errors are found or suspected in the equating process (Allalouf, 2007).

Prior to any statistical calculations, equating requires specific data collection procedures that allow for comparisons of difficulty across multiple forms. Three major equating designs exist: the single group design, random groups design, and the previously discussed non-equivalent anchor test design (Cook & Eignor, 1991; Kolen, 1988; Kolen & Brennan, 2004). Discussion on the single groups and random groups design are provided in chapter 2. In the NEAT design, anchor items are administered on both forms to be equated. Thus, equating scores on the two forms is facilitated through using the anchor items to account for existing ability differences between the groups taking the distinct forms. The logistic advantages of the NEAT design make it the most typically used equating design in practice.

The NEAT design is particularly useful when a unique test form is given at each testing occasion, and results from latter administrations are linked back to the base form. Such is the situation for many high-stakes testing companies, which commonly administer tests on multiple, successive occasions throughout the year. For example, College Board administers the SAT on approximately seven occasions over the course of one year and introduces several new forms at each testing date (Cook & Eignor, 1991).

The use of multiple forms over time requires that items selected as part of an anchor block be used in previous administrations of the test (Kolen & Brennan, 2004). Consequently, the anchor items pose a threat to test security, as these items have been exposed to previous test-takers. The exposure of anchor items leaves them vulnerable to potential theft or harvesting during a previous testing occasion. As the NEAT design relies heavily on the anchor items to link the results from the otherwise distinct forms, even the slightest contamination of the anchor items will damage the integrity of the equating function.

Item Response Theory in Equating

A common method for conducting the equating of test scores is through item response theory (IRT). The advantages of IRT, specifically the invariance of item and ability parameters, make its use in test equating particularly appealing (Hambleton, Swaminathan, & Rodgers, 1991; Skaggs & Lissitz, 1986). Under IRT, each test calibration results in item parameters which are on a different metric. If the assumptions of IRT hold, the resulting item parameter metrics differ by a linear transformation. A critical step in the equating process is to obtain the appropriate slope (A) and intercept (B) constants to place the item parameters from different tests on the same scale through a process termed scaling. Accurate scaling is essential for the equating process to reflect the true difference between forms. Under the NEAT design, scaling constants are estimated entirely through the anchor items. Thus, any source of error in the estimation of the anchor item parameters could result in inaccurate scaling constants.

Purpose

The proliferation of test scores as a basis for significant decisions necessitates the dissemination of accurate and fair scores. The validity of these scores has been threatened by the rise of cheating in high-stakes testing to epidemic status. Even though strict measures are taken to ensure test security, cheating has affected even the most secure tests. Moreover, equating requires test scores to accurately reflect examinee ability. If cheating causes the responses on a test form to inaccurately represent the ability distribution of a group of examinees, the equating function will incorrectly adjust the difficulty of the form. The prevalence of the NEAT design in large scale high-stakes testing further complicates this issue. The consistent exposure of anchor items may compromise the items specifically used to estimate the scaling constants and generate the equating function. When the equating process is compromised, the validity of the test scores for both cheaters and honest test takers comes into question. Moreover, honest test takers may be penalized if the equating inaccurately adjusts for the difficulty of the test.

The purpose of the current study is to examine the impact of cheating on the equating process. Specifically, the study investigates how alterations in the equating process caused by cheating translate into errors in equated observed scores. In particular, the scaling constants and equated scores will be evaluated. Scaling constants provide a measure for evaluating how cheating specifically affects IRT equating. The evaluation of equated scores serves as a global measure to analyze the impact that the interaction between cheating and equating has on the scores reported to examinees.

The current study will utilize a computer simulation to create a NEAT design equating procedure in which the anchor items have been exposed to a subset of the examinees. Multiple IRT scaling methods will be evaluated under a variety of realistic conditions to determine if a specific scaling method consistently produces more accurate scaling constants in the presence of compromised items. Equating of number correct raw test scores will be carried out for each condition included in the study.

The present study addresses a significant void in the literature. Previous studies have primarily focused on improving test-security through methods such as detection of cheaters and control of item exposure. Few studies have attempted to understand the effects of cheating on ability estimation in IRT (Guo, Tay, & Drasgow, 2009; Yi, Zhang, & Chang, 2008). Furthermore, no study has examined the interaction of cheating and the equating process. This study specifically addresses how cheating impacts the most commonly used equating design in high-stakes testing, which may have practical implications for both high-stakes testing companies applying equating procedures and examinees taking these tests.

CHAPTER 2

Literature Review

The purpose of this chapter is to discuss the relevant literature regarding Item Response Theory (IRT) equating. The chapter begins by outlining the basic tenets of item response theory in order to establish context for IRT equating. The following section presents the principles of equating, the prominent equating designs used in testing, and a brief overview of equating methods. Next, the five methods used in this study to place IRT parameters on a common scale are described and literature comparing the methods is evaluated. A comparison of IRT and conventional methods of equating is then presented, with a focus on the benefits IRT provides over conventional equating. The final section of the literature review assesses studies that have investigated the impact that compromised items can have on IRT estimation. At the conclusion of the chapter, the research questions addressed by the current study are presented and justified through the theoretical framework outlined in the literature review.

Item Response Theory

The use of item response theory for analyzing test data expanded in reaction to the shortcomings of classical test theory (Hambleton, Swaminathan, & Rogers, 1991). In common unidimensional IRT models, the probability of responding correctly to an item is dependent on the characteristics of that item and the ability of the examinee. Though IRT requires stringent assumptions to be met, IRT models allow for stronger inferences about the data than classical test theory. For example, the IRT standard error of measurement varies as a function of the latent ability, allowing practitioners to gauge the measurement quality of the instrument at different ability levels. The advantages of IRT have made the use of these models appealing for equating purposes (Cook & Eignor, 1991).

IRT equating specifically benefits from the invariance property. The invariance property implies that, when the IRT model strictly fits the data, item parameters are independent of the ability distribution of examinees (Hambleton, Swaminathan, & Rogers, 1991). Thus, after a linear scaling transformation the same item parameters would be obtained regardless of the sample's ability. Item parameter invariance alleviates much of the difficulty in disentangling group ability and form differences under the NEAT design. These advantages are discussed in more depth in subsequent sections.

Numerous IRT models exist, each of which describe the relationship between items and ability in different manners. The following section focuses solely on the three-parameter logistic (3-PL) IRT model given its widespread use in large-scale testing. The 3-PL model (Birnbaum, 1968) specifies that the probability of correct response to an item is a function of examinee ability and three item parameters. Mathematically the 3-PL model is defined as,

$$P(\theta_i) = c_j + (1 - c_j)\frac{e^{1.7a_j(\theta_i - b_j)}}{1 + e^{1.7a_j(\theta_i - b_j)}}, \tag{2.1}$$

where $P(\theta)$ is the probability of correct response for an examinee of ability θ_i on item j. The scale of θ is arbitrary, ranging from negative infinity to positive infinity, and is often constrained to be normally distributed with a mean of 0 with variance of 1 in a particular sample of examinees. The parameters $a_{j,}$ $b_{j,}$ and c_j refer to specific item parameters for item j. The a parameter denotes how well the item discriminates examinees of varying abilities. Higher values of a indicate that the item is useful in discriminating examinees.

The *b*-parameter indicates the difficulty of the item, and is on the same metric as θ. Difficult items will result in higher values of *b* in comparison to easier items. The *c*-parameter was developed to account for guessing behavior when multiple-choice items are administered. The *c*-parameter sets the absolute minimum probability of a correct response for a given item. Since the *c*-parameter often results in values below random chance due to quality distractors, this parameter is commonly referred to as the “pseudo-guessing parameter” (Hambleton, Swaminathan, & Rogers, 1991).

The relationship between the probability of a correct response and the underlying latent ability can be visually represented through an item characteristic curve (ICC). For any particular item with estimated item parameters, the probability of correct response can be calculated for a given ability. This is accomplished through substituting the item’s parameters and the specified ability into Equation 2.1. The relationship between probability of correct response given the ability and item parameters is commonly represented as,

$$p_{ij}(\theta_i; a_j; b_j; c_j) \,, \tag{2.2}$$

where p_{ij} is the probability of correct response for an examinee of ability θ_i on item *j*, given the item parameters a_j, b_j and c_j. Item response theory assumes a monotonic relationship between ability and the probability of correct response. That is, examinees of higher ability will always have a greater probability of responding to the item correctly than examinees with lower ability values. Figure 1 displays an example ICC. Figure 1 represents an item with the parameters: $a = 1$, $b = .75$ and $c = .2$.

A test characteristic curve (TCC) can be calculated by summing the individual ICCs on a test. The TCC will range from the sum of the *c*-parameters for all items to the total number of items and provides an expected number correct-score for a given ability (DeMars, 2010), referred to as the “true score”. Accordingly, the “true score” for an examinee is found by evaluating the TCC at the examinee’s estimated ability value. In true-score IRT equating, the TCC plays an integral role in converting a score on one form to an equivalent score on another form.

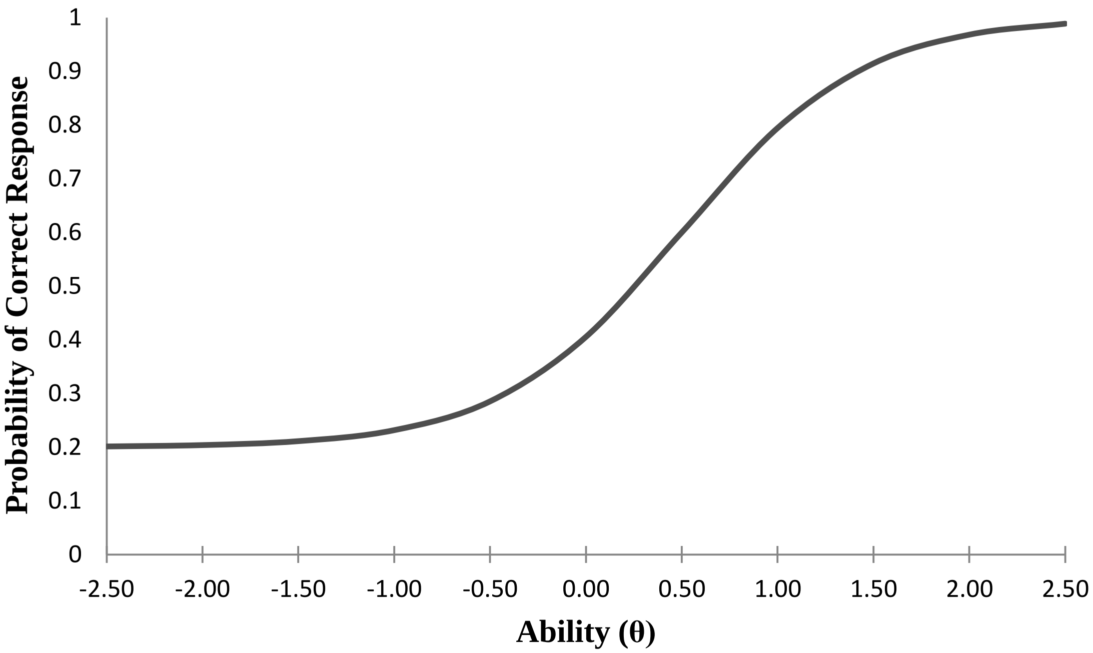

Figure 1. Sample Item Characteristic Curve *Note.* Item Parameters: a = 1, b = .75, c = .20.

Equating

As stated in the introduction, equating procedures were developed in order to ensure score comparability across multiple forms of a test by adjusting for minor differences in difficulty (Cook & Eignor, 1991; Kolen, 1988; Kolen & Brennan, 2004). Equating offers a method for disentangling the differences in test forms and examinee ability. Thus, equating is integral to preserve the accurate evaluation of an examinee’s ability. Several guidelines which define an equating procedure have been put forth.

Principles of equating. Several researchers have proposed requirements for a successful test equating (Angoff, 1984; Dorans & Holland, 2000; Kolen & Brennan, 2004; Lord, 1980). Although differing slightly, the underlying themes of these conceptualizations are in accord. Five properties consistently arise from the literature: (1) the equating function must be symmetric, (2) the different forms to be equated must measure the same construct, (3) the statistical reliabilities of the two test forms should be equivalent, (4) after equating, the form administered to the examinee should be inconsequential, and (5) the equating function should be invariant of the group used to develop the function. The following section explicates these properties.

The symmetry property requires that the transformation of a score on form A to the scale of form B be the inverse of form B to form A (Kolen & Brennan, 2004). To demonstrate this property, envision an equating function in which a form A score of 50 translates to a form B score of 55. The symmetry property requires that a score of 55 on form B convert to a form A score of 50. This property precludes the use of regression for equating, as the regression of X on Y is unequal to the inverse of the regression of Y on X, unless X and Y are perfectly correlated.

The same construct property states that truly equated scores can only exist if the forms are designed to measure the same cognitive ability or skill (Angoff, 1984). Strict care must be taken by the test developer to construct forms of the same specifications. No equating can produce interchangeable scores between forms that evaluate different constructs (Cook & Eignor, 1991).

Equal reliabilities across test forms has been argued by Lord (1980) as a necessity to equate forms. Strict adherence to this requirement would leave little hope for successful equating. Even forms built to be strictly parallel will often result in disparate reliabilities. However, Dorans and Holland (2000) maintain that experience has shown adequate equating functions can result from two forms that differ in reliability and recommend that high, as opposed to equal, reliabilities be required for equating. The authors contend that concerns about equal reliability should be secondary and minor violations of the requirement tolerated.

Equity, proposed by Lord (1980), requires the form given to an examinee to be a matter of indifference. Lord defined equity to hold only if the conditional distributions of equated scores for all examinees at a given true score are identical for each form. The corollary of this definition requires examinees at a given true score to have equivalent observed score means, standard deviations, and distributional shapes for converted scores (Kolen & Brennan, 2004). As Lord himself notes, equating under the equity property is possible only if the two forms are identical, thus absolving the need for equating. Livingston (2004) further argues the attainment of equity is impossible in practice, whereas Dorans and Holland (2000) deem this property "poorly specified" and claim the assumptions made by equity need not hold for successful equating. Issues with the practicality of the equity property arose quickly and a less restrictive form of equity, termed weak equity or first-order equity, was proposed by Morris (1982, also see Yen, 1983; Harris & Crouse, 1993). Weak equity requires only the expected value of the equated score from Form A to Form B to be identical for all examinees with a given true score. Therefore, examinees of the same ability are expected to obtain the same equated score on form A and form B on average, a much less stringent condition than Lord's equity. Statistics based on first order equity are used to evaluate the accuracy of equating in this study.

The final principle of equating commonly cited by the literature is population invariance. The equating relationship should be equivalent irrespective of the group used to develop the function. For example, the equating function found when using a group consisting of males should be identical to the function derived from a group of females (Kolen & Brennan, 2004). Lord and Wingersky (1984) describe how population invariance is theoretically inherent for true score methods of equating, however because observed scores are used in place of true scores in the equating function this assumption does not necessarily hold in practical situations. Dorans and Holland (2000) provide ample evidence that successful equating can be conducted when population invariance is approximated, yet equating should not be carried out under large violations of this assumption.

Equating designs. In order to equate forms of a test, data must be collected in a fashion that allows the ability of the examinees and difficulty of form to be evaluated independently. As introduced previously, three prominent designs have emerged: single-group, random equivalent groups, and non-equivalent anchor test design (NEAT). The choice of design is dependent on the practical constraints of the testing environment, with each design offering benefits and limitations. This section will describe the properties of each design with specific focus on the NEAT design because of its prevalence in large-scale testing and its importance in the current research.

Single-group design. In the single-group design, the same examinees are administered both test forms to be equated. Single-group designs provide the benefit of strict control over examinee ability as the same examinees take both forms (von Davier, Holland, & Thayer, 2004). Often administrators counter-balance the order of forms to protect against fatigue effects (Kolen & Brennan, 2004). Because the same examinees take both forms, the difficulty of the forms manifests in the mean score of the examinee group on each form. If examinees are clearly performing better on form A, form B can be considered more difficult and adjusted accordingly.

The assumptions and practical considerations required from the single-group design frequently render the design impractical. The administration of two full-length forms to a group of examinees doubles testing time and exposes two forms of a test to a subset of examinees. Given the length of most large-scale tests, requiring the completion of two forms is unfeasible under most circumstances (Cook & Eignor, 1991). Furthermore, the probability of fatigue influencing examinee scores in this situation is greatly increased. The group taking both forms of the test must also be a representative sample from the entire examinee population, an assumption deemed “rarely more than a convenient fiction” by von Davier, Holland, and Thayer (2004, p. 23). As Kolen (1988) expresses, the limitations of single-group designs obviates its use in large-scale assessments.

Random groups design. The random groups design, also denoted as the equivalent groups design (Holland & Dorans, 2006), involves administering the different forms to independent groups of equivalent ability. Obtaining a randomly equivalent sample can be accomplished through strict control of the testing environment (e.g., spiraling the test forms) or groups of test-takers may be assumed to have equivalent ability. If the groups can be considered comparable, the differences in form difficulty can be inferred through the mean score on each form, as in the single-group design.

The application of random-groups design circumvents limitations of the single-group design, such as order and fatigue effects, as examinees complete only one form. However, this equating design introduces other practical limitations. In most circumstances, the assumption that two groups of examinees are inherently equivalent cannot be justified. This assumption becomes more problematic in large-scale testing, when the test forms commonly are administered to examinees at different times and locations. If a method such as spiraling test forms is used, both the old and new test forms must be administered in the same session (Cook & Eignor, 1991; von Davier, Holland & Thayler, 2004). Thus, test security may be compromised if the old form items have been exposed and examinees receiving this form would be unfairly advantaged.

The non-equivalent anchor test design. The NEAT design disentangles ability of the examinees from difficulty of the form through the use of anchor items. Under the NEAT design, independent examinee groups take separate forms that contain a set of identical (anchor) items. The NEAT design is frequently employed in large-scale testing because it resolves the dependency on equivalent groups (Holland, 2007). The anchor items are used to assess potential differences in ability among examinees administered the different forms. Because the anchor items are identical, differences in mean scores on the anchor items for examinee groups administered separate forms reflect non-equivalent ability levels between these groups. After controlling for possible non-equivalent group ability, adjustments can be made for differences in form difficulty emerging from the unique items. Operational constraints arising from single-group and random groups designs are resolved as all examinees within a session receive a single form and examinees receiving the different forms can be sampled from non-equivalent groups (Dorans & Holland, 2006). Therefore, the NEAT design can be employed easily when multiple forms of a test are administered at different testing periods. The logistic advantages of the NEAT design make it the most typically used equating design in practice (Cook & Eignor, 1991).

There are two methods of incorporating the anchor set on the test (Kolen, 2007). When using internal anchors, examinee responses to these items count directly to their total score. Internal anchors are frequently employed in practice, with anchor items interspersed throughout the test. External anchors refer to when the anchor items do not contribute to an examinee's total score. This anchor design often requires an additional testing section and thus is used less frequently in practice due to increased testing time and the need to restructure the test (Petersen, 2007).

The gain in practical feasibility through use of the NEAT design comes at a price of statistical complexity (Kolen, 1988). The vital role assigned to anchor items under the NEAT design requires these items to undergo stringent evaluation. Two essential characteristics of anchor items have been examined in the literature: the number of anchor items and content representation.

As Angoff (1984) noted, the anchor tests must be long and reliable enough to accurately capture the differences in ability. Angoff recommended the use of 20 items or 20% of the test, depending on which number was greater. Since Angoff proposed this rule of thumb, several studies have empirically examined the effect of anchor set length on the accuracy of equating results. The focus of the following section will be on studies that investigated anchor sets in the context of item response theory. Many of these studies focused on a process specific to IRT termed scaling. Scaling, discussed in detail in a following section, is the process of creating a common scale for item parameter estimates of two separate forms. Scaling is a prerequisite to equating when using IRT methods.

McKinley and Reckase (1981) examined the number of anchor items necessary when scaling using a three-parameter logistic IRT model. The authors evaluated anchor sets at lengths of 5, 15, and 25 items on a test consisting of 50 items. Results suggested that the number of anchor items had little effect on the scaling procedures after a certain point. Therefore, McKinley and Reckase recommended the use of 15-item anchor sets, while advising against the use of 5 items. Regrettably, the results of this study are compromised by the use of unidimensional item-response theory models when the test evaluated was created to measure multiple constructs.

Vale, Maurelli, Gialluca, Weiss, and Ree (1981) conducted a simulation study to investigate the effect of anchor length on scaling accuracy. The researchers determined the transformation required to scale the unique items through obtaining two ability estimates, one from the anchor set and the other from unique items. Consistent with the McKinley and Reckase (1981) study, the results indicated scaling with 15 anchor items is comparable to 25 anchor items but scaling with 5 items was inadequate. Increasing the length of the anchor set resulted in moderately more accurate discrimination parameters, but had little effect on the difficulty parameter. The scaling method applied by Vale et. al. (1981) is extremely uncommon in modern IRT equating. As such, the effects of using a contemporary scaling transformation cannot be explicitly inferred from this study.

However, the accuracy of scaling should improve given the stronger mathematical theory and empirical research behind current methods.

The investigation of anchor test length was taken to an extreme by Wingersky and Lord (1984) who examined anchor lengths of 50, 25, and 2. Real data from two forms of the SAT mathematics test, 60 unique items per form, was used to evaluate scaling accuracy. Wingersky and Lord discovered the standard errors of unique items after scaling were nearly identical between the 50 and 25 item anchor sets. To examine an extreme case, the authors evaluated scaling using only 2 anchor items. Standard errors of unique items under scaling with 2 anchor items were comparable to scaling with 25 items, if the standard errors of the 2 item anchor were low to begin with. In a condition containing 2 anchor items with high standard errors, scaling suffered from more pronounced bias. Though the authors suggested accurate scaling can be obtained with five common items, it should be noted that increasing the anchor length always provided more accurate results.

Length of the anchor set is not the only variable test developers must consider when selecting anchor items. It is essential that the anchor items represent the content of the full-length test as shown by Klein and Jarijoura (1985). The authors specifically investigated whether increasing the length of the anchor set could compensate for poor content representation. The test used in this study contained 250 items. Under the content representative anchor condition, 60 anchor items were included. The non-representative anchor conditions used lengths of 101 and 105 anchor items. Results indicated that the use of representative anchors is of utmost importance to accurate equating. Simply adding items cannot compensate for a lack of content coverage. Klein and Jarijoura explained that when group differences exist on specific areas of the test, the differences would fail to manifest if the anchor did not include items representative of these areas. Similar sentiments are expressed by Cook and Petersen (1987).

In summary, initial studies have illuminated the variables important in selecting a quality anchor set. The number of anchor items must yield reliable data which allow for comparisons of ability across groups (Angoff, 1984). Although several studies have shown as few as five anchor item can result in adequate equating (Wingersky & Lord 1984), increasing the number of anchor items led to reduced error in the equating process (McKinley & Reckase, 1981; Wingersky & Lord 1984; Vale et al., 1981). The importance of length is mitigated by adequate content representation (Cook & Petersen, 1987; Klein & Jarjoura, 1985). Anchor sets should essentially take the form of a smaller version, “mini-test” (Kolen & Brennan, 2004, p. 19), of the total form. Thus, content representativeness is necessary to identify ability differences between groups accurately.

Equating methods. In general, two categories of equating exist: conventional and item response theory methods. Conventional methods equate test scores using observed score methods such as linear and equipercentile equating (Braun & Holland, 1982; Kolen & Brennan, 2004). Item response theory equating derives the equating relationship between two forms of a test using an IRT model. The following section describes IRT scaling, which places the item parameter estimates for the new form onto the base form metric. The next section describes the process of IRT equating, which uses these re-scaled item parameter estimates to equate the number-correct scores, with particular focus on true-score equating. Afterwards a brief review of literature comparing conventional and IRT equating is presented in order to explicate and justify the use of

IRT equating in this study. Although no definitively superior method of equating has been established, several theoretical and practical advantages of IRT based methods have emerged. Specifically, IRT equating has been shown to produce better results than conventional equating methods when examinee groups differ in ability (Lord & Wingersky, 1984; Petersen, Cook, & Stocking, 1983; Skaggs & Lissitz, 1986). This property is particularly beneficial when equating under the NEAT design, as the design accommodates differing level of abilities between groups.

Item Response Theory Scaling

IRT methods require the parameters of different forms to be on a common scale prior to equating number correct scores. As a result of the arbitrary nature of the latent ability scale however, IRT parameters from separate calibrations of the same items for different samples are commonly on different scales (Vale, 1986). As discussed previously, the ability metric for any single IRT calibration is typically constrained to be standard normal. Therefore, the group of examinees administered the test determines the metric for the ability distribution and thus the metric of the item parameters. Consequently, when the same items are administered to non-equivalent groups separately, the resulting parameters will not be on the same scale. For example, assume an IRT calibration was conducted for examinees scoring above the median of a test and another calibration for examinees below the median. The resulting ability distribution for each group will have an origin of 0 and variance of 1. Thus, it is impossible to identify the differences in ability across the two groups. A transformation of the item and ability parameters is required to establish a common metric.

When the IRT model fits the data for the calibrations to be scaled, a linear transformation can convert IRT parameters from different samples to a common scale. As the transformation is linear, the relationship between each pair of IRT parameters can be expressed through a multiplicative (A) and an additive (B) constant. This process is analogous to transforming any distribution in a linear fashion, such as when distributions are transformed to the z-score metric by dividing the variance and subtracting the mean to standardize the distribution. The ability values for examinee *i* on a calibration of a new form (*NF*) can be converted to the scale of the base form (*BF*) through the following,

$$\theta_{BFi} = A\theta_{NFi} + B. \tag{2.3}$$

The linear transformations for individual item parameters are calculated as,

$$a_{BFj} = a_{NFj} / A, \tag{2.4}$$

$$b_{BFj} = Ab_{NFj} + B, \tag{2.5}$$

$$c_{BFj} = c_{NFj}, \tag{2.6}$$

where *j* refers to item *j* on the specified test. The *c* parameter is independent of the scale transformation.

The A and B constants required for the transformation can be derived using several methods. The following section will focus on the properties of the predominant scaling methods in use. In general, these methods can be placed into two categories, moment and characteristic curve methods. Moment methods incorporate the mean and standard deviations of the IRT item parameters for the anchor items to form a common scale. Characteristic curve methods take into account all the item parameters simultaneously through each individual anchor item’s ICC or the TCC comprised of all anchor items. Another scaling method will be presented in which the anchor item parameters are fixed to be equivalent to the original calibration.

Moment methods.

Mean-sigma. Marco (1977) described a simple procedure in which the scaling constants are derived from the means and standard deviations of the *b*-parameter estimates of the anchor items. Specifically, the *A* constant is found through dividing the standard deviation of the *b*-parameters on the scale setting the metric, the base form, by the standard deviation of the new form b-parameters. This function can be expressed mathematically as,

$$A = \frac{\sigma(b_{BF})}{\sigma(b_{NF})} \tag{2.7}$$

where $\sigma(b_{BF})$ and $\sigma(b_{NF})$ represent the standard deviation of *b*-parameters for anchor items on the BF and NF scale respectively. In this example, the base form sets the metric of the parameters.

Derivation for the *B* constant is as follows,

$$B = \mu\left(b_{BF}\right) - A\mu(b_{NF}) \tag{2.8}$$

where $\mu\left(b_{BF}\right)$ and $\mu\left(b_{NF}\right)$ represents the mean of *b*-parameter for anchor items on NF and BF scale and *A* is the scaling constant defined previously.

Mean-Mean. Loyd and Hoover (1980) proposed a method similar to that of mean-sigma. The mean-mean procedure expresses the scaling constants through the mean of the anchor item *a*-parameters in place of the standard deviation of *b*-parameters used in mean-sigma. Mathematical notation for mean-mean is as follows:

$$A = \frac{\mu(a_{NF})}{\mu(a_{BF})} \tag{2.9}$$

and

$$B = \mu\left(b_{BF}\right) - A\mu(b_{NF}) \tag{2.10}$$

Theoretically, the use of the average *a*-parameter values may be appealing as means are typically more stable than standard deviations (Baker & Al-Karni, 1991). However, IRT values of the *b*-parameter are better estimated in comparison to *a*-parameters (Kolen & Brennan, 2004). Thus, each method provides theoretical advantages and limitations.

Characteristic curve methods.

Haebara method. Referring to Equations 2.1, 2.3, 2.4, and 2.5, it logically follows that the probability of correct response for an item on a new form scale can be placed on the scale of the base form through the following:

$$p_{ij}(\theta_{BFi}; a_{BFj}; b_{BFj}; c_{BFj}) = p_{ij}(A\theta_{NFi} + B;\ a_{NFj} / A; Ab_{NFj} + B; c_{NFj}). \tag{2.11}$$

The result of Equation 2.11 allows for all parameters for each individual anchor item to be considered simultaneously in producing a common metric. However, because item parameter values are estimates, there is no guarantee that one set of scaling constants will produce perfect concordance for the probability of correct response across all examinees and items (Kolen & Brennan, 2004). Haebara (1980) suggested that a function minimizing the difference between probabilities of correct response across samples would produce the correct scaling constants. Haebara defined the error resulting from scaling as the squared difference between ICCs for each anchor item conditional on ability across samples. As shown in Equation 2.12, the Haebara method locates the *A* and *B* constants that minimize this error across all examines, *i*, and anchor items, *j*:

$$F_h = \frac{1}{N}\sum_{i}^{N}\sum_{j}^{n}[\,p_{ij}\;(\theta_{BFi};a_{BFj};b_{BFj};c_{BFj})\;-p_{ij}(A\theta_{NFi}+B;a_{NFi}\,/\,A;Ab_{NFi}+B;c_{NFi})]^2. \quad (2.12)$$

Given the non-linear nature of the minimization function (F_H), a multivariate search technique is required to solve for *A* and *B* (Kim & Kolen, 2007).

Stocking-Lord Method. Stocking and Lord (1983) introduced a similar approach that applies the TCC for anchor items in place of individual ICCs. As a result, the summation across ICCs is performed prior to squaring the difference between parameter estimates,

$$F_{sl} = \frac{1}{N}\sum_{i}^{N}[\sum_{j}^{n} p_{ij}\;(\theta_{BFi};a_{BFj};b_{BFj};c_{BFj})\;\text{-}\;\sum_{j}^{n} p_{ij}(A\theta_{NFi}+B;a_{NFj}\,/\,A;Ab_{NFj}+B;c_{NFi})]^2. \, (2.13)$$

Thus, the Stocking and Lord approach evaluates the anchor test as a whole. As with the Haebara method, a multivariate search technique is used to solve for the scaling constants.

Fixed Anchor.

When a separate calibration of IRT item parameters is used for each form, a common metric can be obtained by fixing the anchor item parameter values to a previous estimation while estimating the unique item parameters (Mislevy & Bock, 1990). This procedure, termed fixed anchor, requires an initial IRT calibration to set the scale. The parameters for the anchor items from the initial calibration are held constant across any subsequent calibration. Thus, all future calibrations using the fixed parameters achieve a common metric.

Comparison of scaling methods. Despite the importance of placing IRT parameters on a common metric, the research on this topic is sparse. The majority of research on scaling compared characteristic curve methods against moment methods. Theoretically, the characteristic curve methods may produce more stable results as they incorporate all item parameters in calculating the scaling constants (Stocking & Lord, 1983). Research generally supports this conclusion. When developing the method, Stocking and Lord (1983) compared the results of their new method to a weighted mean-sigma procedure. The authors used model fit of the item parameters after scaling to assess the transformation quality. The Stocking-Lord method provided comparable or better fitting estimates for all comparisons.

Baker and Al-Karni (1991) explored the differences between the Stocking-Lord and mean-mean methods. Scaling constants resulting from the two methods were compared through a simulation study and using real test data. In slight contrast to Stocking and Lord (1983), the authors obtained similar scaling constants under both methods. However, Baker and Al-Karni noted that the largest differences between methods occurred when the item parameters posed calibration problems, such as when a low ability examinee group is administered a difficult set of items with low discrimination. The Stocking-Lord method proved more robust to these atypical combinations of IRT parameters.

Kaskowitz and de Ayala (2001) conducted a simulation study that investigated the robustness of Stocking and Lord’s method to parameter estimation error. Estimation error was created by varying the standard errors associated with the item parameters. Results from the study supported the notion that characteristic curve methods are robust to moderate estimation error. The authors concluded that Stocking and Lord’s method can provide accurate scaling even under the presence of moderate error.

Both characteristic curve methods were shown to recover equated scores with less error than the moment methods under a variety of conditions in a simulation study conducted by Hanson and Beguin (2002). The researchers compared the four separate estimation scaling methods described above when the ability distribution of examinees was both equivalent and non-equivalent. Different levels of sample size and number of common items were also investigated in the study. Characteristic curve methods provided significantly less biased and variable true score estimates under nearly all conditions. Differences between the characteristic curve methods themselves were negligible. When comparing bias and variability of true score estimates across the range of raw scores, the characteristic curve methods displayed nearly identical values at all raw score points. In comparison, the mean-sigma and mean-mean methods showed significantly larger errors that fluctuated at different levels of raw scores.

Hanson and Beguin's study also provided one of the few comparisons between the mean-mean and mean-sigma methods. The mean-mean method resulted in smaller error variance when recovering equated scores. When the distribution of ability was unequal however, the mean-sigma method yielded less biased results. The error variance of mean-sigma methods varied drastically across the range of raw scores, while the mean-mean variance remained relatively stable.

The fixed anchor method of equating has received little attention in the literature. Equating results from the fixed anchor method were compared to mean-sigma in a study conducted by Jodoin, Keller, and Swaminathan (2003). The authors found that occasionally equating results differed depending on the scaling method applied. Because

Jodoin et. al. (2003) evaluated real data with unknown true values, the accuracy of methods cannot be determined in this study.

Kim (2006) addressed the accuracy of various fixed anchor estimation methods in recovering the scaled item parameters of the new form through a simulation study. In general, the fixed anchor procedure was robust to unstable parameter estimates consistently producing relatively accurate item parameters. However, fixed anchor methods that do not update the prior ability distribution during estimation suffered increased bias when groups differed in ability distributions. Overall, Kim showed fixed anchor to be a viable method for establishing a common scale. Given the scarcity of research evaluating equating accuracy under various scaling methods, further investigation in this area would be beneficial.

Item Response Theory Equating

After a common metric is developed for the IRT parameters, examinee IRT ability estimates are on the same metric and can be compared directly. However, testing companies often use equated raw scores because they are more intuitive for practitioners to comprehend. In this situation, equating procedures must be applied to convert raw number correct scores for one form to another. Equating is necessary to convert raw scores because the unique items on different forms will cause equivalent ability estimates to result in different raw scores.

Two IRT based equating methods exist: true score equating and observed score equating (Lord, 1980). Although fundamentally different, results from the two methods have been shown to produce similar equated scores (Lord & Wingersky, 1984). IRT true-score equating was conducted in this study due to its theoretical and computational advantages.

In IRT true score equating, the true scores on two forms at a given ability are used to derive the equating relationship (Kolen & Brennan, 2004; von Davier & Wilson, 2007). The true score for an examinee of ability θ_i is equal to the probability of a correct response to all items at θ_i as represented by the TCC. Conceptually, this probability represents the expected score of an examine at θ_i for the entire test form. Thus, the true score for the base form, denoted τ_{BF}, and new form, $\tau_{NF,}$ are defined as:

$$\tau_{BF}(\theta_i) = \sum_{j}^{n} p_{ij}(\theta_{BFi}; a_{BFj}; b_{BFj}; c_{BFj};) \tag{2.14}$$

and

$$\tau_{NF}(\theta_i) = \sum_{j}^{n} p_{ij}(\theta_{NFi}; a_{NFj}; b_{NFj}; c_{NFj};) \tag{2.15}$$

IRT assumptions imply that for a given θ_i, true scores $\tau_{\mathrm{BF}}(\theta_i)$ and $\tau_{\mathrm{NF}}(\theta_i)$ are equivalent (Kolen & Brennan, 2004). Consequently, an equating relationship between the forms can be derived by finding the true score on the base form that corresponds to the true score on the new form X given θ_i. Kolen and Brennan (2004) outline this three-step process. First, a true score on the new form, τ_{NF}, is specified. Second, the ability, θ_i, corresponding to the true score selected in step 1 is calculated using a Newton-Raphson procedure. Third, the ability calculated in step two is used to compute the true score equivalent on the base form, τ_{BF}. In practice, this true score relationship is derived using number correct observed scores in place of the true scores (Kolen & Brennan, 2004).

Item response theory observed score equating uses the IRT model to generate a distribution of number-correct scores conditional on ability (Kolen & Brennan, 2004; Lord & Wingersky, 1984). The conditional score distribution is integrated over the ability distribution for a specified population. These calculations are repeated for the other form, integrating over the same ability distribution. The resulting score distributions are those that would have been observed if both forms had been given to a single group. The score distributions for both forms are then equated using traditional equipercentile techniques. A more thorough presentation of IRT observed score equating can be found in Kolen and Brennan (2004).

Comparison of equating methods. Lord (1980) noted one caveat in true score equating. The equating function is generated based on the true score relationship between two forms. However, the equating relationship based on true score relationships is used to convert observed scores on the new form to the equivalent observed score on the base form metric (Kolen & Brennan, 2004; Lord & Wingersky, 1984). Although theoretically unjustified, several studies have shown IRT true and observed methods to produce similar conversions (Lord & Wingersky, 1984). In addition, Han, Kolen, and Pohlmann (1997) found IRT true score equating to produce more stable results than IRT observed score and traditional equipercentile techniques. However, the mean differences in equating stability between the two IRT methods were not statistically significant, further suggesting the comparability between methods. These results have been used to justify the applying the true score equating function to observed scores.

Lord and Wingersky (1984) discussed the theoretical benefits and limitations of both observed-score equating methods and true-score IRT equating. As mentioned previously, successful equating requires invariance of the equating function across samples. Under equipercentile methods of equating, when the test forms differ in difficulty or the examinee groups sampled from a population differ in ability, invariance will not hold strictly. Lord and Wingersky (1984) noted that under observed-score IRT, invariance holds only for the group used to derive the equating function and not for subgroups with different ability distributions. If the assumptions of IRT hold, true-score IRT equating necessarily meets invariance. This occurs because IRT parameters are invariant across the examinee population. Pragmatically, this assumption is not guaranteed by true-score methods, however, because the invariance of the item parameters may not be met. Additionally, instead of estimating the true score from the ability estimate, sometimes the observed score is substituted in the equating function derived for true scores. Treating observed scores as if they are true scores cannot be theoretically justified, and thus invariance is not assured in practice (see also, Kolen and Brennan, 2004). Given that observed-score methods are population dependent and theoretical justification for true-score IRT is unclear, theory alone cannot determine which equating method is appropriate

Several empirical studies have compared the stability of equating under conventional and IRT methods. Kolen (1981) performed one of the initial studies comparing conventional equating to the then-burgeoning IRT methods. Equipercentile and linear equating comprised the conventional equating methods, while various IRT procedures were employed (including 3-PL true and observed score methods). Kolen used a cross validation statistic to evaluate the stability of equating procedures. 3-PL IRT observed and true-score equating methods were found to produce more stable cross-validation results. Petersen, Cook, and Stocking (1983) found similar results when investigating scale stability of IRT and conventional methods. Stability was evaluated through equating a form to three others then back to itself, i.e., form A to form B to form C to form D then back to form A (see also, Harris & Crouse, 1993). IRT methods generally performed comparably to or better than conventional methods. IRT equating proved particularly useful when the forms slightly differed in content and length. IRT methods also produced more stable results than traditional equipercentile when equating a test to itself (Han, Kolen, & Pohlmann, 1997).

As noted by Lord and Wingersky (1984), stability, though desirable, does not guarantee the accuracy of an equating method. In fact, a method may exhibit stability because it is producing results that are consistently inaccurate. Thus, studies demonstrating the accuracy of IRT methods are reviewed as well.

Cook and Eignor (1983) compared IRT methods of equating to linear and equipercentile methods across multiple tests using fit statistics for evaluation. In general, equating results between IRT and linear methods showed little difference. Equipercentile methods performed particularly poorly in comparison to other methods investigated. Based on the results, the authors promoted the use of IRT equating when applicable. Echoing a similar sentiment as Petersen, Cook, and Stocking (1983), the authors postulated that IRT methods outperform conventional methods when forms vary more in difficulty.

In reaction to the propagation of studies on IRT equating, Skaggs and Lissitz (1986) synthesized the literature and present a comprehensive review. Several common themes were identified in relation to the comparison of IRT and conventional methods.

First, Skaggs and Lissitz concluded the research to this point had not indicated a clearly superior method of equating in all situations. Second, when the test forms were reliable and similar in difficulty, most equating methods will yield reasonable results. Third, when the sample of examinees differed in ability, IRT methods consistently produced more accurate equating results. The capability of IRT methods to handle groups of examinees who differ in ability is also put forth by Cook and Eignor (1991) in a theoretical manner. Because differences in form difficulty or group ability will manifest in non-linear fashion, due to floor and ceiling effects, IRT offers a viable solution that conventional methods cannot provide.

Several practical advantages of IRT equating have also been discussed in the literature. As IRT scaling methods allow for alternative test forms to be placed on the same scale, any previously equated form can be used as the basis for scaling and equating. This allows a testing program great flexibility in regard to which items are used and when (Cook & Eignor, 1991). In addition, because the ability derived for each examinee is invariant of the items, when items are dropped or tests shortened, no rescoring, and thus no repeat equating, is necessary. Livingston (2004) also noted the flexibility of IRT, specifically in adaptive testing where items can be targeted to the examinee’s ability. However, these advantages come with drawbacks (Livingston, 2004). IRT is statistically and conceptually complex. The IRT model has stringent assumptions and large sample size requirements that may not be met in practice. Thus, the context for equating should always be considered in selecting equating methods.

In summary, item response theory provides the theoretical advantage of being population invariant and the ability to deal with non-linear differences in examinee ability. Empirically, IRT methods have performed comparably to conventional methods in terms of accuracy and stability. In cases where the examinee ability differs, IRT methods often yield superior equating results. These advantages prompted Lord and Wingersky (1984) to make the claim that, "conventional [non-IRT] equipercentile equating of observed scores is not recommended in situations when anchor tests are required". Though this claim may be overstating the superiority of IRT methods, the flexibility and statistical advantages offered by IRT equating has made the use of these methods popular in large-scale testing (Kolen & Brennan, 2004).

Effect of Compromised Items in Item Response Theory

Inherently, IRT equating is dependent on accurate parameter estimation. Compromised items introduce construct irrelevant variance that assuredly distorts the estimation of parameters. For example, the ability parameter will reflect both ability and prior knowledge of the item when compromised items are present. As stated in the introduction, few studies have examined the effect of compromised items on IRT parameters. No articles were found evaluating cheating under IRT true-score equating. Of the studies investigating cheating in IRT, Yi, Zhang, and Chang (2008) compared the error in ability estimates resulting from compromised items under various CAT selection criteria. The researchers simulated a 40-item adaptive test, generated from an item pool containing 480 items, for each of the 10,000 examinees. Thirty of these examinees, labeled thieves, could memorize 10 of the 40 items they received. Once a thief had memorized an item, any subsequent examinee would answer the compromised item correctly. Thus, the probability of correct response was independent of examinee ability or the item's parameters once the item was compromised. The resulting ability estimates displayed severe positive bias for all item selection methods included in the study. The mean difference between estimated and true abilities for low ability examinees, $-3.880 \leq \theta < -.890$, increased by an average of over one standard deviation in 5 of the 6 conditions included. The influence of compromised items on examinee ability estimates decreased as the initial true ability was higher.

Guo, Tay, and Drasgow (2009) compared the robustness of CAT and conventional tests to compromised items. Cheating was implemented through having simulated examinees randomly steal and share items with all successive test-takers. If one examinee compromised an item, the probability of correct response on that item for all examinees was set to .85. If the same item was stolen by two or more examinees, a correct response was guaranteed. The authors found IRT ability estimates for conventional tests to be highly sensitive to comprised items. As in Yi et. al., ability estimates were positively biased in all conditions. In addition, examinees with low abilities ($\theta = -3$) benefitted the most, with some obtaining estimates as high as $\theta = 2.23$. Although, the addition of items and use of multiple forms lowered bias, the impact of compromised items remained substantial.

Jurich, Goodman, and Becker (2010) investigated the effects of compromised items on the passing status of examinees. The researchers applied simulation techniques to create a situation in which cheating occurred on a new form of a previously calibrated test. Using IRT observed-score equating, the study compared how mean-sigma, Stocking-Lord, and fixed anchor methods recovered the correct status of examinees. Under all scaling methods, examinees passed at drastically higher rates than expected when cheating existed. Unexpectedly, both cheaters and honest test-takers completing the new form benefited from the compromised items. Jurich, Goodman, and Becker hypothesized that the scaling methods incorrectly adjust for differences in ability when anchor items become compromised. This occurs because the cheaters score higher on the anchor items then their true ability would suggest, thus inflating the ability distribution of the new form group as a whole. Consequently, when scaling is performed to place the NF abilities on the BF scale, the augmentation to the estimated NF ability distribution on the anchor items will cause the unique items on the NF to appear more difficult.

The distortion in difficulty arises from a constraint specific to the NEAT design. As anchor items are used to estimate the relationship between examinee group abilities, when cheating is present for one group, this group will appear more able than their true ability, as discussed above. The unique items to both forms are considered in the equating function. Thus, as the ability of the cheating group overestimates the proficiency of examinees, the unique items will appear more difficult because no cheating has occurred on these items. The equating function will reflect the misrepresented difficulty of the new form, incorrectly adjusting examinee's equated scores to be higher. Thus, the increased equated scores for new form examinees reflects the augmentation to the estimated NF ability distribution in the presence of cheating, benefitting both cheaters and honest test-takers.

Research Questions

No study has investigated the impact of cheating on IRT true-score equating. However, the result of Yi et al. (2008) and Guo et al. (2009) imply serious consequences for the equating function when cheating occurs. Given the reliance on ability estimates to generate the equating function, minor errors in the estimation of ability could seriously distort the relationship between two forms. Furthermore, estimation of item parameters will also be affected, resulting in inappropriate scaling constants. Studies on this topic are necessary to understand how the equating process is influenced by cheating.

Thus, this study will be conducted to address the gap in the literature and evaluate the accuracy of equating in the presence of cheating. Several specific hypotheses will be investigated:

Research question 1. To what extent do the proportions of cheaters and compromised anchor items affect the recovery of equated scores and scaling constants?

The results from Yi et al. (2008) and Guo et al. (2009) indicate ability estimates will be severely inflated when compromised items exist. The parameters of these items are also assuredly distorted. Jurich et al. (2010) showed that equated scores for examinees taking a compromised form were significantly over-estimated. Further investigation is required to ascertain the degree to which cheating affects scaling constants. In addition, understanding the magnitude of equating error under different levels of cheating will provide practical information for testing companies concerned about the security of their tests.

Research question 2. Which scaling method performs best in recovering true scaling constants under the presence of cheating?

Research on scaling constants suggests that characteristic curve methods may produce more stable results. Although characteristic curve methods may be less sensitive to error in item parameters (Kaskowitz & De Ayala 2001), it is necessary to evaluate this robustness when item parameters are biased by cheating. For example, if cheating significantly biases the *a*-parameter, while only moderately altering *b*, the mean-sigma method may produce more accurate results. Jurich et al. (2010) found the Stocking-Lord method to best classify examinees in the presence of cheating, however classification was poor for all methods investigated in the study. If one scaling method is found to perform better in recovering the true scaling constants, the use of this method would be desirable when cheating is suspected.

Research question 3. Do differences in the ability distributions of the NF and BF examinees affect the recovery of equated scores and scaling constants in the presence of cheating?

Given the purpose of utilizing the NEAT equating design is to account for possible difference in ability, it is necessary to evaluate the interaction of non-equivalent groups and cheating on the equating process. In addition, Hanson and Beguin (2002) found that differences in ability distributions cause the scaling constants to become more variable and produce more bias. Considering these results in conjunction with research question 2, it may be that one scaling method recovers parameters more efficiently when group abilities are non-equivalent and cheating has transpired.

Research question 4. Does the design of the anchor set, internal or external, impact the manner in which cheating affects equated scores?

The design of the anchor set will almost certainly alter the way cheating affects equated scores. As noted, an internal anchor is comprised of items that are accounted for in an examinee’s total score. In contrast, external anchor sets include items that are used for equating purposes only, and are not considered in the total test score. When the anchor is internal to the test, compromised items will count directly toward total score.

Accordingly, the internal anchor should considerably overestimate equated scores in comparison to the external anchor condition.

Research question 5. How are the equated scores for honest test-takers affected when anchor items have become compromised?

It is imperative to testing companies that test compromises do not impair scores for honest examinees. However, the benefit to ability estimates obtained by cheaters may come at a cost to honest test-takers. If the inflation of cheaters' scores leads to the underestimation of item difficulty, it is possible that estimates of honest test-takers ability will suffer. In contrast, Jurich et al. (2010) found that both cheaters and non-cheaters were incorrectly passed at a higher rate when cheating occurred. In either case, the validity of all scores on the test is suspect. Thus, it is necessary to investigate the degree the equating process affects honest test-takers when cheating is present. Furthermore, the impact on honest examinees may depend on the anchor design (see research question 4). Under the internal anchor, examinees with access to compromised items will benefit immensely in comparisons to non-cheaters as the items directly influence the total score. Given an external anchor, the compromised items only come into consideration when scaling the two test forms. Investigating the degree of change between these two designs should provide important information on how alterations to equated scores manifest under each design

CHAPTER 3

Method

A simulation was conducted to investigate the research questions presented in chapter 2. Simulation studies allow for systematic control over experimental conditions that are not manipulated easily in an actual testing environment. For example, systematically manipulating the degree of cheating in a testing situation would be unfeasible. Furthermore, true values of the parameters under investigation are known in simulations. These values provide an absolute level of comparison for the evaluation of parameter recovery. The results of a simulation are limited by the degree to which the conditions mimic real situations. Although it is impossible to capture the complexity of an authentic testing situation, the current study minimized this potential drawback through selecting conditions that reflect a typical large-scale examination.

Data were generated to simulate a non-equivalent anchor test (NEAT) equating design. The current study attempted to simulate a possible scenario in which anchor items have been exposed to the population through repeated test administrations. Specifically, the simulation addressed the common situation in which two administrations of a test are given at different, successive, testing occasions requiring the use of two forms. The first administration of the test was created as if the items were unique, thus none of the items on the original form were compromised. The second form required the use of anchor items for the NEAT equating, exposing these items to potential cheating. Hence, only anchor items on the new form were subject to possible cheating. In addition, various conditions were systemically altered to fully investigate the effects of cheating on the equating process. The following section provides a detailed description concerning the methods used to investigate the research hypotheses of this study. First, a discussion is provided detailing the data generation process and conditions included in the study. Next, the statistics compiled to evaluate the results are described. Last, the expected results pertaining to each research question are presented.

Test Generation

Two forms of a test, both containing 100 items, were generated for each replication of this study. One test represented a base form (BF), the form that sets the scale of the scores. The other generated test form represented a new form (NF) of the test that was equated to the scale of the BF. As this study employed a NEAT design to equate scores from the NF to the BF scale, a certain number of items were common across the two forms. The current study followed Angoff's (1984) original guidelines in using 20 items as the anchor set. This value should ensure reliable estimation of ability differences. The remaining 80 items were unique to each form. To assess the impact of cheating when the compromised items do not influence an examinee's total score, a facet was included that examined internal and external anchor sets.

Item parameters from the 1996 administration of the National Assessment of Educational Progress (NAEP) mathematics test (Allen, Donoghue, & Schoeps, 2001) were used as the basis for creating both the BF and the NF. Items with extreme location (*b*) or discrimination (*a*) parameters were excluded from selection to minimize the potential for estimation issues confounding the results. Specifically, items with *b*-parameters with absolute values greater than 2.5 or *a*-parameters above 1.7 or less than .5 were removed from the item set. This process left a total of 216 items available for selection.

For each replication, item parameters for both forms were randomly selected from the pool of 216. Anchor items were created by holding 20 randomly selected items constant across the two forms. That is, the item parameters were equivalent for these items across both forms. The unique items were then sampled without replacement from the remaining 196 item parameters.

Examinee Population

For each simulated test, the probability of correct response on each item was generated for 3,000 simulated test-takers using a 3 parameter logistic (3PL) IRT model. This sample size was chosen to reflect a typical sample for a large scale test. In addition, a large sample size reduces the risk of estimation problems confounding the results of the study. The probability of correct response was calculated using a randomly generated latent ability for each examinee in conjunction with the item parameters of the form administered to that examinee.

The IRT latent ability of examinees responding to the BF was generated from a standard normal distribution, denoted as N(0,1). The latent ability distribution of examinees administered the new form was systematically varied to compare situations in which the NF examinees have an ability distribution that differs from the BF group in mean and/or variance. Four levels of this condition were investigated including a condition in which the two groups of examinees were of equivalent ability. The specific distributions which NF examinees' abilities were generated from are: N(-.5,1), N(0,1.25), N(-.5,1.25), or N(0,1). The mean NF ability was selected to be lower than the mean BF ability to prevent a situation in which the majority of examinees score near the maximum possible value. A potential ceiling effect in the data would make capturing the benefits of cheating nearly impossible.

Cheating

Cheating conditions. Two conditions were manipulated to simulate the degree of cheating: 1) the proportion of compromised items, and 2) the number of examinees with access to these items (referred to as cheaters). The proportion of anchor items compromised was varied at 25% and 100%. In addition, two proportions of cheaters were examined, 10% and 50%. These levels were chosen to reflect low and high amounts of cheating in a real testing situation. To assign cheaters, the first *X* amount of generated examinees, where *X* corresponds to the proportion of cheaters in the current replication, were designated as cheaters. Delegation of compromised anchor items was carried out in the same manner.

As cheating often goes undetected, determining what constitutes a low and high degree of cheating is difficult. Although the condition including 50% cheaters and 100% compromised anchor items may seem extreme, situations have occurred in which the vast majority of examinees had access to a high number of test questions, if not the entire form. Therefore, results from the higher degree of cheating conditions still provides practical information. Furthermore, this condition provides a picture of how equating is influenced by an extreme case of cheating.

The simulation incorporated an additional condition that included no compromised items. This condition allowed for investigation of the quality of the equating process in the context of this simulation without the influence of cheating, and thus served as a baseline for relative comparisons.

Cheating Implementation. Cheating was implemented by adding .5 to the probability of answering a compromised item correctly for any examinee designated as a cheater. For example, if a cheating examinee's original probability of correct response on a compromised item was .3, the adjusted probability would increase to .8. If the cheating examinee's original probability of responding correctly was above .5, the examinee would necessarily get the item correct as the adjusted probability would exceed 1.

The cheating adjustment value of .5 was selected as it greatly improved the probability of correct response for a cheating examinee. However, the increase still allowed for incorrect responses to compromised items by cheaters with low abilities. Admittedly, the designated increase was arbitrary. However, there is a deficiency in research detailing benefits to having prior knowledge of an item. Further research should consider different methods to conceptualize and implement cheating behavior.

Scoring

After the probability of a correct response was determined for each examinee on each item, a dichotomous score was created for each simulated response by comparing the probability of correct response to a random number generated from a uniform distribution ranging from 0 to 1. If the probability of correct response was greater than the randomly generated value, the response was scored as correct. When the probability of correct response was less than the random value the response was scored as incorrect. This is a standard procedure for scoring generated IRT data (Macdonald & Paunonen, 2002) and allows for error in the data, thus better approximating reality.

Item parameters for each of the simulated tests were estimated separately using BILOG-MG (Zimowski et al., 2003). The FLOAT command was applied to remove the influence of incorrectly specified prior-distributions on the item parameter estimates (Hendrickson & Kolen, 1999). When the FLOAT command is applied in BILOG, *b* parameter's prior distribution standard deviation is set equal to the mean of the distribution. As the mean of the *b*-parameters can be negative, the prior is inappropriate for a standard deviation. Thus, the standard deviation of the prior-distribution for the *b* parameter was set to 1. The maximum number of Gauss-Newton iterations for the expectation-maximization algorithm was increased to 100. Ability distributions were estimated using an empirical distribution over 40 quadrature points. Aside from the modifications described above, default BILOG-MG options were used for estimation.

Scaling

Once item parameters for both forms were estimated, the NF item parameters were placed on the BF scale. Scaling was applied using five common IRT scaling methods. The five methods under investigation in this study include: (a) the mean-mean method (MM: Loyd & Hoover, 1983) (b) the mean-sigma method (MS: Marco, 1977), (c) the Stocking-Lord (SL: Stocking & Lord, 1983) approach, (d) the Haebara method and (e) the fixed anchor method (FA: Lord, 1980; Kolen & Brennan, 2004). The five scaling methods utilized in this study allowed for a comparison of moment, characteristic curve, and concurrent methods of scaling. As discussed in chapter 2, each of these methods calculate the slope (A) and intercept (B) scaling constants necessary to transform the parameter estimates from the NF to the BF scale. Both scaling constants were retained after the completion of a replication.

Equating

In the final step of the simulation process, true score IRT equating was applied to establish equivalence between NF and BF scores. As discussed in chapter 2, true score equating converts an examinee's raw score on the NF to an equivalent score on the BF. This conversion was accomplished by calculating the ability value corresponding to a number correct total score on the NF. The resulting ability was then used to derive the expected true score on the BF. Because the minimum true score for the three-parameter IRT model will equal the sum of the *c* parameters, no true score exists for observed scores below this sum. Kolen (1981) proposed a method for equating scores that fall outside the range of possible true scores that can be used to address this issue. First, a score of 0 on the NF was set to a score of 0 on the BF. Second, a score of the sum of the *c*-parameter on the NF was set to equal the sum of the *c*-parameter on the BF. Linear interpolation was then conducted to find the equivalent BF score for NF scores that fall outside the range of possible true scores. This procedure was applied when an observed score fell below the lowest possible true score.

When the anchor items were internal to test scoring, all 100 items were used to estimate the true score relationship between the forms. When the anchor was external to the test, the true score relationship was determined through the 80 items unique to each form. The external anchor condition presented a situation in which the compromised items do not count towards examinees' total scores, yet are used in the scaling process. Note that the maximum possible number correct score differed between the two anchor conditions, 80 for the external anchor and 100 for the internal. Estimated scaling constants and equated scores were retained at the completion of each replication to examine the effects of cheating on the equating process.

Summary of Conditions

The current study varied five conditions to investigate the effects of cheating on the equating process across a range of factors. Two proportions of cheaters, 10% and 50%, and two proportions of compromised items, 25% and 100%, were included to explore how the degree of cheating affects the equating process. True ability for examinees administered the new form was generated from four normal distributions: N(0,1), N(-.5,1), N(0,1.25), N(-.5, 1.25). Five scaling methods were used to place item parameters on the same scale. These methods included mean-mean, mean-sigma, Stocking-Lord, Haebara, and fixed anchor. The anchor items were manipulated to be either internal or external to the scoring of the test. To allow for exploration of complex interactions, the five conditions were fully crossed. The interaction of conditions resulted in a 2 x 2 x 4 x 5 x 2 design, for a total of 160 unique conditions. The simulation process was replicated 500 times for each combination of conditions. New tests and examinees were generated for each replication. Appendix A displays a summary of the study's simulation design.

Comparison Criteria

The accuracy of recovered scaling constants and equated scores was used to evaluate the effects of cheating on the equating process. Comparisons across all 160 conditions in the study were made to assess how the experimental conditions alter the effects of cheating on parameter recovery. Within each replication, two scaling constants (A and B) and 100 equated scores were produced.

Scaling Constants. In evaluating the accuracy of recovering scaling constants, it is necessary to calculate the true scaling constants without the influence of cheating or error in estimation. True scaling parameters were derived by calculating the A and B constants necessary to set the estimated NF examinee ability distribution, which was constrained to be N(0,1), equal to the true NF distribution. These constants can be derived by solving for the linear transformation that shifts the constrained, N(0,1), new form distribution to the true distribution. As seen in Equation 2.3, the A scaling constant corresponds to the slope of a linear transformation and the B constant corresponds to the intercept. For example, to shift the estimated N(0,1) distribution to a true distribution of N(-.5,1.25) the scores would first need to be multiplied by 1.25 to expand the variance, then .5 would be subtracted from each score to shift the mean. Therefore, the true scaling constants in this situation would be 1.25 and -.5 for A and B respectively. Recall that the true NF ability distribution was a condition under investigation in the study, thus requiring the computation of different true scaling constants depending on the condition used to generate the data. The true scaling constant values were compared to the estimated values retained after each replication across each condition.

To quantify errors in recovery, bias and root mean squared error (RMSE) were calculated for both A and B scaling constants. Bias identifies systematic deviation of the estimated scaling constant from the true parameter. Bias is mathematically defined as the average difference of the estimated parameter from its true value across all replications,

$$bias_{\lambda} = \frac{\sum_{j=1}^{m}\left(\hat{\lambda}_j - \lambda_j\right)}{m} \tag{3.1}$$

where λ is the true value of the scaling constant, $\hat{\lambda}$ is the estimated value of the scaling constant and *m* represents the total number of replications.

RMSE provides a measure of absolute accuracy in parameter recovery. The RMSE statistic incorporates both the bias and the variability of the sampled parameter. RMSE is computed by taking the square root of the average squared deviation between the estimated parameter and the true value. Mathematically, RMSE can be expressed by:

$$RMSE_{\lambda} = \sqrt{\frac{\sum_{j=1}^{m}\left(\hat{\lambda}_j - \lambda_j\right)^2}{m}} \tag{3.2}$$

The mathematical terms in RMSE are equivalent to those in the measure of bias. RMSE differs from bias in that it captures both bias and sampling variability. Thus, RMSE assesses the overall variability of the estimated scaling constant around the true scaling constant. As with bias, RMSE was calculated for each A and B scaling constant across replications.

Equated Scores. Quantifying the impact of cheating on equated scores involved comparing an examinee’s equated score derived from the observed responses to the examinee’s equated score derived from that examinee’s true ability (θ) and the true item parameters on the BF (Hanson & Beguin, 2002). In other words, this process compared $\hat{\tau}$, the estimated true score, to τ, where τ is the true-score of the examinee given perfect conditions of measurement. Before calculating the actual true score for an examinee, the new form ability estimate was scaled to the base form metric using the true equating constants. The BF true score equivalent (τ) was obtained by evaluating the TCC at the true latent ability associated with an examinee administered the NF using the true item parameters on the BF. To clarify, true item parameters refer to the generated item parameters from the test creation process and not the estimated item parameters. Thus, these item parameters reflect the true properties of the item and are unaffected by conditions under investigation or error associated with parameter estimation. The true score obtained when using true parameters represents the equated score an examinee would receive under perfect estimation and no extraneous influences. As the current study examined an equating design that sets a clear base form, only recovery of true scores for examinees administered the new form was assessed.

As with scaling constants, bias and RMSE were used to evaluate the recovery of equated scores. As equated scores were produced for each examinee in a replication, it was necessary to sum across all examinees within a replication prior to averaging across replications when calculating bias and RMSE. Accordingly, bias in equated scores is mathematically defined as:

$$bias_{\lambda} = \frac{\sum_{j=1}^{m}\sum_{i=1}^{n}\left(\frac{\hat{\tau}_{ji} - \tau_{ji}}{n}\right)}{m} \qquad (3.3)$$

where *i* is the individual examinee and *n* denotes the sample size. $\hat{\tau}_{ji}$ and τ_{ji} refer to the estimated and population true score for examinee *i* and replication *j* respectively. The mathematical notation is equivalent to that of Equation 3.1 for all other terms. RMSE for equated scores is calculated as,

$$RMSE_{\lambda} = \sqrt{\frac{\sum_{j=1}^{m}\sum_{i=1}^{n}\left(\frac{\hat{\tau}_{ji} - \tau_{ji}}{n}\right)^{2}}{m}} \qquad (3.4)$$

where the terms are equivalent to Equation 3.3. In addition to the overall bias and RMSE, these indices were plotted as a function of the new form raw score to examine how deviations from truth vary across the observed score.

Evaluating Honest Test Takers. To address research question 5 and examine how cheating influences ability estimates for honest test takers, a separate calculation of bias and RMSE for equated scores was conducted including only honest examinees. That is, Equations 3.3 and 3.4 were applied using only the honest test takers in the given replication. The resulting statistics captured the degree of deviation an honest examinee's ability estimate, on the true score metric, was from truth.

Expected Results

Conditions without cheating. The accuracy of the equating and scaling processes under conditions in which cheating was not present should produce consistent results with negligible deviations from truth. Given that the data were generated in a method conducive to adequate model fit and item parameters were constrained to reasonable values, estimation of scaling coefficients and equated scores in conditions free of cheating should accurately recover the true values, aside from sampling error. The conditions with no cheating can serve as a baseline to compare against cheating conditions. Recovery of true parameters under conditions without cheating would suggest that any deviation from truth in cheating conditions, beyond the deviations in the baseline condition, arises due to a main effect of cheating or an interaction with cheating and another manipulation.

Effects of cheating on scaling constants. Scaling constants were expected to deviate from true values within cheating conditions. As discussed previously, cheating will mask the true ability differences between examinee groups by making new form examinees appear more able. Hence, estimation of the scaling constants will be affected by this shift in ability. Furthermore, estimation of the discrimination parameter may be severely altered when cheating occurs as cheating can become the primary factor influencing whether an anchor item was responded to correctly. Thus, scaling methods that utilize both location and difficulty parameters (Stocking-Lord, Haebara, and mean-mean) may produce scaling constants further from truth in comparison to the mean-sigma and fixed anchor methods. It is difficult to predict the direction of bias in either the A or B scaling constants under any of the scaling methods. The multiple interactions of conditions in this study were expected to produce different and unpredictable results in regards to the scaling coefficients.

Effects of Cheating on Equated Scores. The effect of increasing the proportion of compromised items and increasing the proportion of cheaters was expected to create positive bias in the estimated equated score. Positive bias was expected to occur because as the degree of cheating increases, the cheating examinees in the new form group will score higher on the anchor items than their true ability predicts. As a result, the ability distribution of NF examinees on the anchor items will be overestimated. Following from the results of Jurich et. al. (2010), the augmented NF ability distribution will cause the unique items on the NF to be estimated as more difficult. Thus, the positive bias in equated scores will arise as examinees are correctly answering unique items above what their true ability would predict due to the inflated difficulty. This process was expected to benefit both cheaters and honest test-takers.

In conditions where the anchor items, and thus compromised items, were internal to test scoring, cheaters were expected to benefit at a higher degree than non-cheaters in terms of equated scores. The equated scores were expected to reflect the fact that the compromised items were used in the estimation of cheaters' ability levels. In external anchor conditions, cheaters and non-cheaters should achieve similar benefits toward their equated scores. As external anchor items were excluded in the calculation of a total score, the compromised items in this situation were not used in the estimation of ability. Therefore, any impact on the equated scores was expected to arise from the use of compromised items in the scaling process.

CHAPTER 4

Results

The results of this study are reported and interpreted in reference to the five research questions posited in chapter two.

Research question 1. *To what extent do the proportions of cheaters and compromised anchor items affect the recovery of equated scores and scaling constants?*

Research question 1 addressed how the amount of compromised anchor items and proportion of cheaters influenced IRT equating. Specifically, the research question explored how the cheating conditions affect equated scores and the scaling constants produced by the four scaling methods investigated in the study. To explore the extent cheating influenced the recovery of these parameters, bias and RMSE for equated scores and the scaling constants were aggregated over the various ability distribution conditions examined in the study.

Table 1 presents the bias for equated scores under the different levels of cheating by the five scaling methods. A positive value for bias indicates the average equated score is overestimated whereas a negative bias reflects the average equated score is underestimated. Table 1 reports the results for the internal anchor; results for the external anchor followed the same pattern. As expected, bias across the five scaling methods for the condition including no cheating approached zero except for the fixed anchor condition. The fixed anchor method showed a slight positive bias when no cheating occurred[1]. For the cheating condition including the 25% compromised anchor items with

[1] We believe this result occurred because, when the NOAdujust option is specified, BILOG does not adjust the quadrature density to reflect that the new form ability distribution may not be distributed N(0,1) relative to the scale of the fixed item parameters (Kim, 2006).

10% cheaters, bias increased slightly for each scaling method. Within this cheating condition, both moment methods produced less biased scores than the characteristic curve and fixed anchor methods. The fixed anchor method yielded the largest bias in equated scores, with the individual scores on average being 1.253 raw score points above the expected scores.

Table 1

Internal Anchor Equated Score Bias Aggregated across Ability Distribution

% Compromised Anchor Items	% Cheaters	Scaling Method				
		SL	HB	MM	MS	FA
0	0	-0.036	-0.039	0.008	0.007	0.676
25	10	0.812	0.603	0.419	0.409	1.253
	50	4.296	2.924	2.812	2.842	3.696
100	10	3.621	2.992	2.829	2.767	3.227
	50	18.972	17.515	16.078	19.461	14.758

Note. MM = Mean-mean, MS = Mean-sigma, SL = Stocking-Lord, HB = Haebara, FA = Fixed anchor

When the proportion of cheaters was increased to 50%, bias in the equated scores increased considerably. The condition including 100% compromised items and 10% cheaters resulted in a similar trend, with a smaller magnitude of bias across the scaling methods. The most extreme cheating condition including 100% compromised items and 50% cheating resulted in drastically large positively biased equated scores. The average equated score across all scaling methods was 17.36 above the true equated score.

Figure 2 depicts the equated score bias as a function of the new form raw score by the different cheating conditions. The figure presents the SL method in the internal anchor condition with N(0,1) new form ability distribution. Trends seen in Figure 2 replicate in the various scaling methods and new form ability distribution conditions. The baseline condition trend noticeably differed from the cheating conditions. In the baseline condition, scores were well recovered near the middle of the new form score distribution. Examinees obtaining low raw scores received underestimated equated scores on average whereas equated scores were overestimated for examinees scoring high. When cheating occurred, equated scores were consistently overestimated. Specifically, equated scores for examinees with low raw scores were drastically overestimated. Overestimation decreased at higher raw scores values. The disparity among cheating conditions is readily apparent in the figure. In the extreme cheating condition, the positive bias greatly exceeded even the moderate cheating conditions explored in this study. At the peak bias for the extreme cheating condition, equated scores were inflated by approximately 23 points above the raw score. This value is considerably larger than in either of the two moderate cheating conditions.

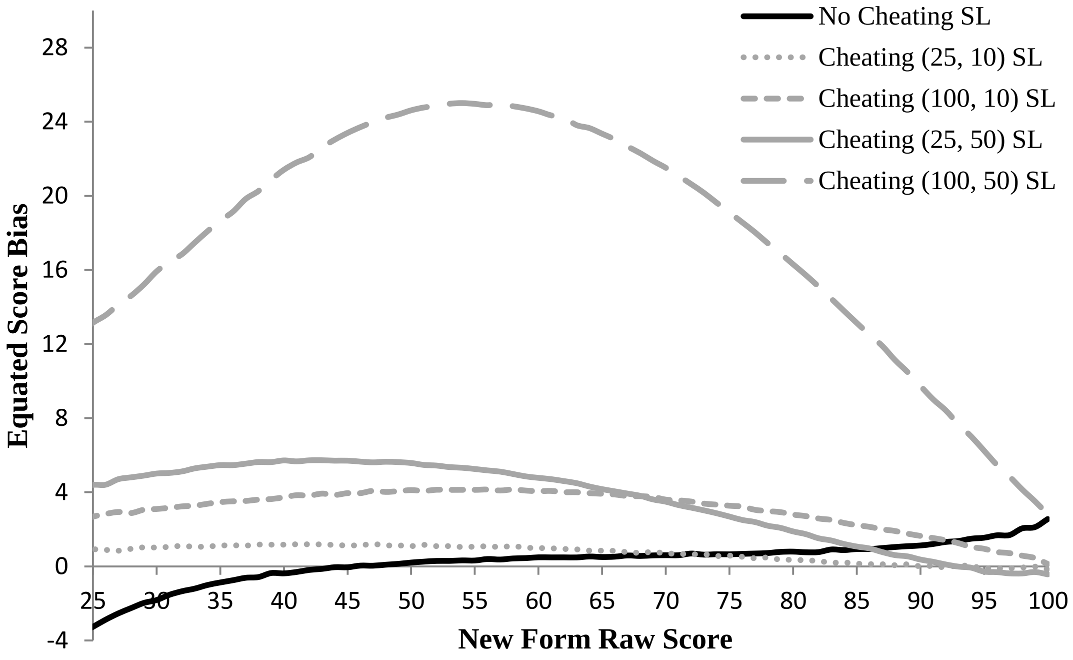

FIGURE 2 Equated Score Bias for the Stocking-Lord Scaling Method at Varying Degrees of Cheating as a Function of the New Form Raw Score. *Note*. Biases are presented for the N(0,1) new form ability distribution and the internal anchor. In the legend, the first number in parenthesis represents the percentage of compromised anchor items, whereas the second number indicates the proportion of cheaters.

Table 2 displays the RMSE of equated scores produced by the five scaling methods for the internal anchor set. RMSE is a function of both the bias and sampling variability of the parameter of interest. Thus, for conditions in which the bias is large, the RMSE must also be large. A RMSE unusually larger than the corresponding bias suggests estimation of the parameter is highly variable. For the current conditions, the increase in RMSE follows the increase in cheating proportions closely. Examining the bias and RMSE of the largest cheating conditions concurrently reveals that as cheating increases, sampling variability of the equated score estimates do not increase. The rise in RMSE in larger cheating conditions seems to be entirely a function of the increased bias in equated scores.

Table 2

Equated Score Root Mean Squared Error (RMSE) Aggregated across Ability Distribution

% Compromised Anchor Items	% Cheaters	Scaling Method				
		SL	HB	MM	MS	FA
Internal Anchor Set						
0	0	4.848	4.849	4.891	4.899	4.927
25	10	4.963	4.943	4.957	4.971	5.088
	50	6.660	5.848	5.832	5.901	6.255
100	10	6.564	6.269	6.161	6.248	6.396
	50	21.112	19.748	18.042	22.178	17.091

Table 3 displays the bias and RMSE for the A scaling constant across the cheating conditions and scaling methods. The A scaling constant is recovered well in the baseline condition, with nearly no bias in estimation. For the moderate cheating conditions, the A constant was consistently underestimated. In the extreme cheating condition, however, the A scaling constant was highly overestimated by the mean-sigma scaling method. In contrast, for the same cheating condition mean-mean severely underestimated the constant. RMSE indicates that the recovery of the A scaling constant became less accurate as the proportion of cheating increases.

Bias and RMSE for the B scaling constant are presented in Table 4. When no cheating is present, the B constant was recovered with virtually no bias. Bias in the B scaling constant systematically increased as cheating increased. Bias severely increased from the moderate cheating conditions to the extreme condition. As with equated scores, the RMSE for the B constant indicated that the parameter was less accurately recovered as cheating increases. However, the small discrepancy between the RMSE and bias again suggests that the inaccuracy in recovering the B constant was largely attributable to the bias in estimation.

Table 3

Bias and Root Mean Squared Error for Scaling Constant A

% Compromised Anchor Items	% Cheaters	Scaling Method			
		SL	HB	MM	MS
		Bias			
0	0	-0.001	-0.001	-0.005	-0.003
25	10	-0.029	-0.014	-0.014	-0.008
	50	-0.145	-0.097	-0.039	-0.091
100	10	-0.050	-0.012	-0.068	0.002
	50	0.000	0.075	-0.221	0.257
		RMSE			
0	0	0.030	0.029	0.042	0.046
25	10	0.042	0.034	0.045	0.049
	50	0.152	0.117	0.058	0.133
100	10	0.067	0.038	0.084	0.053
	50	0.335	0.127	0.245	0.321

Table 4

Bias and Root Mean Squared Error for Scaling Constant B

% Compromised Anchor Items	% Cheaters	Scaling Method			
		SL	HB	MM	MS
		Bias			
0	0	0.001	0.001	0.005	0.004
25	10	0.060	0.044	0.033	0.031
	50	0.288	0.200	0.185	0.193
100	10	0.237	0.194	0.192	0.178
	50	1.182	1.088	0.963	1.258
		RMSE			
0	0	0.032	0.032	0.047	0.048
25	10	0.069	0.055	0.058	0.069
	50	0.293	0.206	0.196	0.293
100	10	0.240	0.197	0.199	0.240
	50	1.344	1.091	0.968	1.344

Research question 2. Which scaling method performs best in recovering true scaling constants under the presence of cheating?

Results corresponding to the second research question attempted to determine whether certain scaling methods performed more accurately in recovering the true scaling constants when cheating occurred. To address this question, scaling methods were compared on their bias and accuracy in recovering the true A and B scaling constants.

In reference to the A scaling constant, reexamining the RMSE from Table 3 shows that the Haebara method provided the most accurate recovery on average. As noted, there was a systematic underestimation of the A constant when cheating was present across the scaling methods, except in the most extreme condition of cheating. For the extreme condition, the mean-mean method highly underestimated the true A scaling constant, whereas the mean-sigma method overestimated the scaling constant. Although the bias of the Stocking-Lord method for the largest degree of cheating approximated zero, as seen in Table 3, this occurred solely due to the aggregation of the data across the differing ability distributions. Additional detail on this effect is presented under research question 3.

The mean-mean scaling method performed most accurately in recovering the B scaling constant when cheating was present. Despite this fact, the mean-mean method still displayed large amounts of bias when a moderate degree of cheating was present. Results regarding the scaling constants indicate that no scaling method would adequately combat the effects of cheating to justify employing these methods when cheating has been known to occur.

The accuracy of recovering scaling constants corresponded with accuracy in recovering equated scores, displayed in Table 1. At low to moderate conditions, the two moment methods along with the Haebara method yielded lower biases in comparison to the Stocking-Lord method. Under the extreme cheating condition mean-sigma and Stocking-Lord produced larger biases than the other methods. The fixed anchor method of scaling yielded large biases at low and moderate degrees of cheating. However, fixing the anchor item parameters displayed the least amount of bias within the extreme cheating condition.

Research question 3. Do differences in the ability distributions of the NF and BF examinees affect the recovery of equated scores and scaling constants in the presence of cheating?

Question 3 addresses the influence of the different new form ability distributions on equated scores and scaling constants. To investigate this question, recovery of the equated scores and two scaling constants were examined for each of the cheating conditions and scaling methods by the four ability distributions conditions. Tables reporting the RMSE for subsequent research questions are included in the appendix as the RMSE values consistently indicated that bias accounted for the majority of inaccurate recovery.

Table 5 contains the equated score bias by the various ability distributions. As noted, RMSE for equated scores was largely a function of bias. This trend held across subsequent analysis. Thus, RMSE for the current and subsequent analyses are presented in the Appendix. Generating the new form examinees from an ability distribution with a mean of -.5 increased the bias in equated scores for a majority of the cheating conditions. This increase in bias was more pronounced in conditions with a larger degree of cheating. The fixed anchor method of scaling was particularly affected when the mean of the new form ability distribution was below the old form.

Table 5

Internal Anchor Equated Score Bias by Ability Distribution

% Compromised Anchor Items	% Cheaters	Ability Distribution	Scaling Method				
			SL	HB	MM	MS	FA
0	0	N(0.0,1.00)	0.020	0.016	-0.001	-0.001	0.055
		N(-0.5,1.00)	-0.156	-0.165	-0.082	-0.082	1.197
		N(0.0,1.25)	0.068	0.067	0.120	0.120	0.143
		N(-0.5,1.25)	-0.075	-0.071	-0.007	-0.007	1.306
25	10	N(0.0,1.00)	0.873	0.666	0.423	0.422	0.616
		N(-0.5,1.00)	0.754	0.511	0.380	0.356	1.826
		N(0.0,1.25)	0.835	0.666	0.472	0.476	0.669
		N(-0.5,1.25)	0.784	0.568	0.402	0.384	1.902
	50	N(0.0,1.00)	4.313	3.020	3.031	2.936	3.006
		N(-0.5,1.00)	4.598	3.063	2.832	3.036	4.468
		N(0.0,1.25)	3.969	2.764	2.767	2.649	2.930
		N(-0.5,1.25)	4.302	2.850	2.619	2.745	4.380
100	10	N(0.0,1.00)	3.516	2.987	2.718	2.825	2.524
		N(-0.5,1.00)	3.811	3.068	3.160	2.872	3.993
		N(0.0,1.25)	3.409	2.896	2.575	2.670	2.481
		N(-0.5,1.25)	3.746	3.015	2.862	2.700	3.912
	50	N(0.0,1.00)	18.006	16.783	14.938	18.157	13.654
		N(-0.5,1.00)	20.485	18.921	18.133	22.128	16.537
		N(0.0,1.25)	17.340	15.963	13.982	16.423	13.070
		N(-0.5,1.25)	20.056	18.395	17.260	21.134	15.773

Changes to the variance of the new form examinees ability distribution appeared to affect the equated score bias dependent on the degree of cheating. In the baseline and lowest cheating condition, a larger new form examinee variance led to more biased estimates of the equated scores. This effect was reversed within higher cheating conditions, where a larger new form variance resulted in less biased estimates on average. This trend was consistent across the five scaling methods.

Tables 6 and 7 show the bias for the A and B scaling constants by the varying ability distributions investigated in this study, respectively. Results suggest that the A scaling constant is affected by both the mean and variance of the new form ability distribution. On average, the bias in estimating the A constant was larger when the mean of the ability distribution was -.5. This effect was more pronounced in the two moment methods of scaling. It is interesting to note that the -.5 mean ability condition increased the absolute value of bias even when the scaling constants yielded biases in opposite directions. For instance, in the cheating condition including 100% compromised items, 50% cheaters, and a variance of 1, the underestimation of the A scaling constant in the mean-mean method increased when the ability distributions differed. For the same condition, the other three scaling methods overestimated the value of the A scaling constant. This overestimation was more severe at the -.5 mean ability value. Increasing the variance of the ability distribution also resulted in less accurate recovery of the A constant. The conditions with a larger variance consistently produced higher biases except in the extreme cheating condition. In the extreme condition, a more variable new form ability distribution led to less biased, or more negatively biased, estimation of the A scaling constant. Also within the extreme cheating condition, the Stocking-Lord method showed both positive and negative bias at different levels of the new form ability distribution. When aggregating across these ability conditions, the Stocking-Lord bias averaged to zero as seen in Table 3.

Table 6

Internal Anchor Bias for Scaling Constant A by Ability Distribution

% Compromised Anchor Items	% Cheaters	Ability Distribution	Scaling Method			
			SL	HB	MM	MS
0	0	N(0.0,1.00)	0.001	0.002	0.000	0.001
		N(-0.5,1.00)	-0.002	-0.002	-0.007	-0.008
		N(0.0,1.25)	0.000	0.000	-0.003	-0.001
		N(-0.5,1.25)	-0.003	-0.003	-0.011	-0.003
25	10	N(0.0,1.00)	-0.022	-0.008	-0.006	-0.004
		N(-0.5,1.00)	-0.026	-0.013	-0.014	-0.010
		N(0.0,1.25)	-0.032	-0.016	-0.012	-0.007
		N(-0.5,1.25)	-0.036	-0.021	-0.022	-0.013
	50	N(0.0,1.00)	-0.121	-0.081	-0.031	-0.069
		N(-0.5,1.00)	-0.131	-0.088	-0.039	-0.090
		N(0.0,1.25)	-0.160	-0.107	-0.039	-0.099
		N(-0.5,1.25)	-0.169	-0.112	-0.047	-0.104
100	10	N(0.0,1.00)	-0.024	0.006	-0.043	0.019
		N(-0.5,1.00)	-0.036	-0.005	-0.074	0.005
		N(0.0,1.25)	-0.061	-0.016	-0.058	-0.003
		N(-0.5,1.25)	-0.080	-0.032	-0.096	-0.014
	50	N(0.0,1.00)	0.040	0.113	-0.128	0.223
		N(-0.5,1.00)	0.107	0.147	-0.194	0.404
		N(0.0,1.25)	-0.092	0.010	-0.214	0.103
		N(-0.5,1.25)	-0.053	0.029	-0.348	0.299

Table 7

Internal Anchor Bias for Scaling Constant B across Ability Distribution

% Compromised Anchor Items	% Cheaters	Ability Distribution	Scaling Method			
			SL	HB	MM	MS
0	0	N(0.0,1.00)	0.001	0.001	0.001	0.000
		N(-0.5,1.00)	0.002	0.001	0.007	0.008
		N(0.0,1.25)	0.001	0.000	0.004	0.004
		N(-0.5,1.25)	0.002	0.002	0.010	0.006
25	10	N(0.0,1.00)	0.054	0.040	0.026	0.026
		N(-0.5,1.00)	0.066	0.047	0.039	0.036
		N(0.0,1.25)	0.052	0.040	0.027	0.027
		N(-0.5,1.25)	0.069	0.050	0.039	0.035
	50	N(0.0,1.00)	0.256	0.181	0.179	0.175
		N(-0.5,1.00)	0.321	0.221	0.196	0.219
		N(0.0,1.25)	0.251	0.176	0.173	0.168
		N(-0.5,1.25)	0.323	0.220	0.191	0.210
100	10	N(0.0,1.00)	0.207	0.174	0.162	0.164
		N(-0.5,1.00)	0.256	0.205	0.224	0.190
		N(0.0,1.25)	0.215	0.181	0.163	0.166
		N(-0.5,1.25)	0.272	0.215	0.218	0.191
	50	N(0.0,1.00)	1.075	1.013	0.851	1.142
		N(-0.5,1.00)	1.232	1.139	1.073	1.387
		N(0.0,1.25)	1.145	1.021	0.845	1.088
		N(-0.5,1.25)	1.277	1.179	1.083	1.416

Results suggest that the mean of the ability distribution was the main distributional factor influencing estimation of the B scaling constant. The conditions including a mean NF ability distribution of -.5 consistently produced a more positively biased estimate of the B constant. The additional overestimation of the B scaling constant was more severe in conditions including more cheating. For example, in the condition with the highest degree of cheating, conditions including a mean of -.5 consistently overestimated the B constant by approximately .15 to .30 more than the condition with a mean new form distribution of 0. Results showed that the variance of the new form distribution had little effect on recovery of the B constant.

Research question 4. Does the design of the anchor set, internal or external, impact the manner in which cheating affects equated scores?

Question four explores potential differences that may arise in equated score bias when implementing an external anchor set. Bias and RMSE for the equated scores were compared between the two anchor conditions to identify possible differential effects. Recall, in an external anchor set, the anchor items are used to compute the scaling constants only. Anchor items are not included in calculations of the examinees' abilities or raw scores. Thus, the maximum possible score on the simulated test with an external anchor set was 80. The difference between total possible scores of the internal and external tests precludes direct comparisons between these conditions. To alleviate these issues, bias and RMSE for the external anchor conditions were converted to a proportion correct. The proportion correct scores for the external anchor were then multiplied by 100 to make the values directly comparable to the internal anchor test. Note that the anchor type employed does not affect the calculation of scaling constants. Thus, no comparisons were made across internal and external conditions for recovery of the scaling constants.

Bias for equated scores for both the internal and external anchor conditions are displayed in Table 8. Results indicate that cheating on the anchor items has less of an effect on the accuracy of equated scores under the external anchor condition. Across the majority of conditions and scaling methods, an internal anchor resulted in more positively biased scores than an external anchor set. The difference in bias between the two anchor sets was exacerbated at higher degrees of cheating. Results of research question 5 expand on the differential effects of the anchor sets.

Table 8

Equated Score Bias by Anchor Type across Ability Distribution

% Compromised Anchor Items	% Cheaters	Scaling Method				
		SL	HB	MM	MS	FA
Internal Anchor Set						
0	0	-0.036	-0.039	0.008	0.007	0.676
25	10	0.812	0.603	0.419	0.409	1.253
	50	4.296	2.924	2.812	2.842	3.696
100	10	3.621	2.992	2.829	2.767	3.227
	50	18.972	17.515	16.078	19.461	14.758
External Anchor Set						
0	0	-0.050	-0.053	-0.006	0.009	0.844
25	10	0.781	0.460	0.258	0.219	1.495
	50	4.311	2.478	1.963	2.364	3.484
100	10	3.304	2.363	2.569	2.031	2.938
	50	15.769	13.365	14.189	14.545	11.081

Research question 5. How are the equated scores for honest test-takers affected when anchor items have become compromised?

Research question five investigated the influence of cheating on the equated scores assigned to honest test-takers. To explore how cheating affects honest test-takers, equated score bias for honest examinees was compared to the bias in dishonest examinees scores. Table 9 contains the bias in equated scores for honest and dishonest test takers. Results were aggregated across the scaling methods to ease the interpretability of the findings.

Examining the bias for honest test takers indicates that equated scores were positively biased in all conditions. In the internal anchor condition, bias for dishonest examinees is consistently more positive across all conditions. Specifically, increasing the amount of compromised items greatly benefitted the cheating test takers in the internal anchor condition. In the most extreme cheating condition, dishonest test takers obtained equated scores 20 to 26 raw score points above their true scores on average across the various ability distributions and scaling methods. In comparison, honest examinees benefited by 14 to 16 score points with these conditions. In stark contrast, differences between dishonest and honest test takers do not arise in the external anchor condition. Across all cheating conditions and ability distributions within the external cheating condition, bias for equated scores was nearly identical between honest and dishonest test takers.

Table 9

Bias for Honest and Dishonest Test Takers by Anchor Type

% Compromised Anchor Items	% Cheaters	Ability Distribution	Internal		External	
			Honest	Dishonest	Honest	Dishonest
25	10	N(0.0,1.00)	0.421	2.211	0.492	0.512
		N(-0.5,1.00)	0.567	2.554	0.603	0.583
		N(0.0,1.25)	0.456	2.130	0.678	0.679
		N(-0.5,1.25)	0.616	2.538	0.797	0.803
	50	N(0.0,1.00)	2.362	4.160	3.027	3.037
		N(-0.5,1.00)	2.566	4.633	2.856	2.849
		N(0.0,1.25)	2.158	3.874	2.974	2.983
		N(-0.5,1.25)	2.375	4.384	2.819	2.814
100	10	N(0.0,1.00)	2.200	9.341	2.468	2.460
		N(-0.5,1.00)	2.551	10.846	2.627	2.643
		N(0.0,1.25)	2.121	8.971	2.675	2.675
		N(-0.5,1.25)	2.456	10.367	2.794	2.788
	50	N(0.0,1.00)	12.274	20.341	13.668	13.689
		N(-0.5,1.00)	13.788	24.693	13.055	13.060
		N(0.0,1.25)	11.503	19.209	14.025	14.045
		N(-0.5,1.25)	13.459	23.588	14.400	14.376

Figure 3 graphically depicts the equated score bias for the Stocking-Lord scaling method for both internal and external anchor conditions. The figure illustrates that cheaters in the internal anchor condition benefit the most from cheating across the scale of new form scores. However, the magnitude of bias for honest examinees in the internal anchor condition was nearly identical to the external anchor condition across the entire distribution. The figure also illustrates that cheaters and honest examinees in the external anchor condition benefited equivalently from cheating.

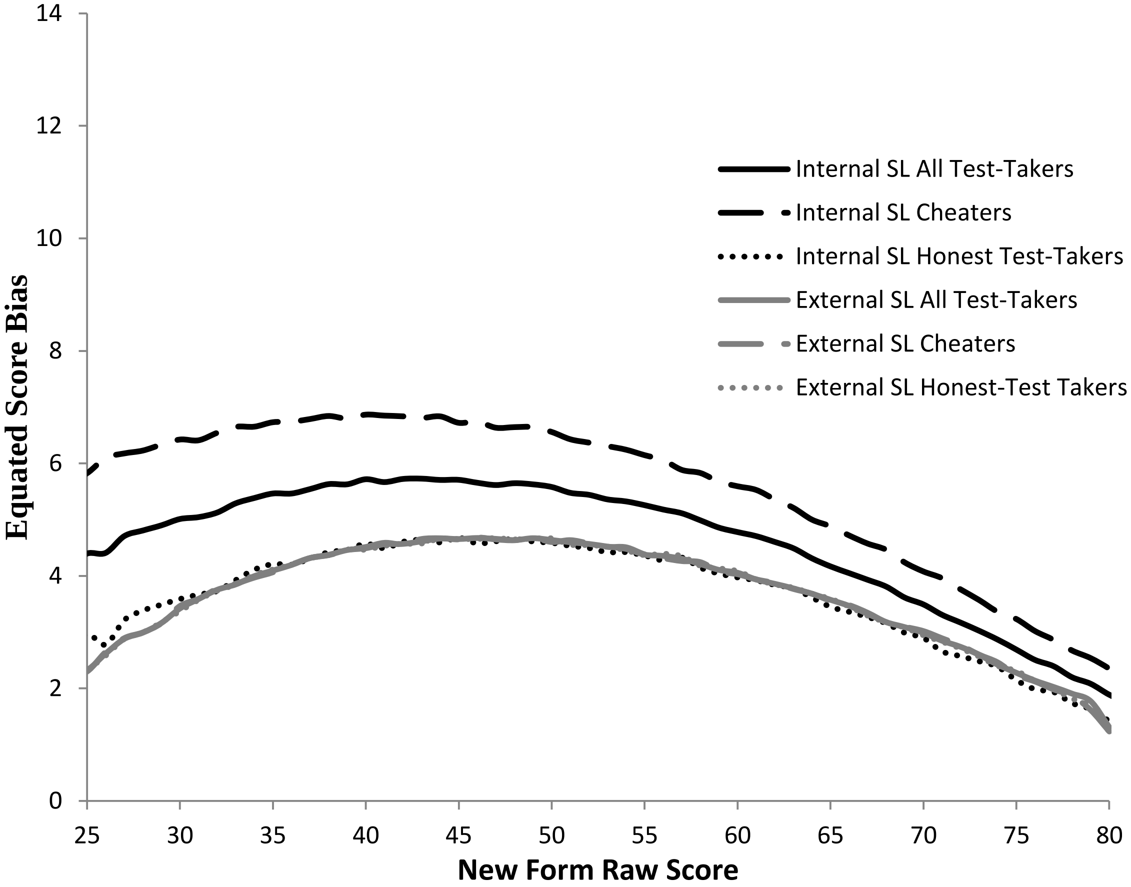

FIGURE 3 Equated Score Bias for the Stocking-Lord Method Scaling Method for Internal and External Anchor Tests. *Note.* Bias is presented for the condition including 25% compromised items, 50% cheaters, and a N(0,1) new form ability distribution.

CHAPTER 5

Discussion

This study employed simulation techniques to examine the effects of cheating on IRT equating under various realistic testing conditions. The study focused on identifying how cheating influences the recovery of equated scores and IRT scaling constants when anchor items are compromised in the non-equivalent anchor test equating design. Furthermore, the study attempted to determine whether a certain method of scaling performed best in recovering the correct IRT scaling constants. Recovery was assessed for various new form examinee ability distributions, levels of cheating, and internal and external anchor sets. The following discussion expands on the results presented in chapter four, followed by general implications of the study and recommendations for future research.

The prominent question of interest in this study concerned the impact of compromised items and proportion of cheaters on equated scores and scaling constants obtained from IRT true score equating. Results indicated that an increase in either compromised items or proportion of cheaters led to positively biased equated scores. Although overestimated equated scores were predictable, the extent of bias at even moderate degrees of cheating was disconcerting. Thus, the results suggest that scores obtained from even slightly compromised tests overestimate examinees' true abilities.

Compromised items introduce positive bias in the equating procedure in part because cheaters respond correctly to the compromised items above the level implied by their true ability. Therefore, cheaters obtain inflated raw scores that, in turn, inflate estimates of their true ability. Although this process helps explain why cheaters benefit directly from compromised items, it does not account for the extensive degree of bias in equated scores or why honest takers benefited from cheating as well.

Investigating the effects of cheating on the scaling constants provides further insight into biased equated scores. Specifically, the severe overestimation of the B scaling constants reveals how all the examinees in the new form group can benefit from cheating. When scaling the new form parameters to the metric of the base form, a large B constant will have two major effects. First, as seen in Equation 2.3, a large B constant will cause ability estimates for new form examinees to be increased. As a result, all examinees will receive an overestimated ability if the B constant is positively biased. Furthermore, Equation 2.5 indicates that the *b* item parameters for the new items will also be artificially increased when the B constant is overestimated. Thus, the new form unique items appear more difficult than truth. Because the inflated new form *b* parameters cause the unique new form items to appear more difficult, responding correctly to the these items will considerably increase an examinee’s ability estimate. Therefore, both cheaters and honest test-takers benefit from responding to these inaccurately difficult unique items.

In less technical terms, access to compromised anchor items distorts the ability differences between new form examinees and base form examinees. Because cheaters increase the average score on the anchor items, the ability of the new form group as a whole is inflated when scaled to the base form. Consequently, as the new form group appears more proficient, the unique new form items, where no cheating occurred, appear more difficult. This benefits both honest and dishonest examinees.

The large positive bias in the B scaling constant arises because cheating occurs on the items specifically used to scale the test form. As cheaters are artificially responding correctly to several compromised items, the difficulty of the anchor items will be underestimated for the new form group. When deriving the scaling constants, each scaling method must overestimate the B constant to account for decreased difficulty. Excessive degrees of cheating increase this bias as a large portion of the examinees have access to a majority of anchor items. Under these circumstances, the *b* parameter for the majority of anchor items will be severely underestimated, drastically inflating the B constant to overcorrect for this effect.

For example, the moment methods calculate the B scaling constant through subtracting the mean new form anchor item *b* parameters, multiplied by the A constant, from the mean of the base form parameters as seen in Equations 2.8 and 2.10. Therefore, the B constant estimate increases because the mean of the new form anchor item *b* parameter decreases when cheating occurs. The two moment methods produce unequal B constants because the methods calculate the A constant using different item parameters. These differences are discussed subsequently. Characteristic curve scaling methods utilize all three IRT item parameters in the calculation of the B constant. Thus, the underestimated *b* parameters assuredly play a role in the overestimation of B in characteristic curve methods.

The influence of cheating on estimating the A scaling constant depended on the scaling method used to derive A. For the majority of conditions and methods the A constant is slightly underestimated. This underestimation occurs because cheating introduces a factor irrelevant to the construct that influences responses, thus decreasing the discriminatory power, the *a* parameter, of new form anchor items. However, under the extreme cheating condition, the moment methods diverged in the direction of bias. Under extreme cheating conditions, the mean-mean method underestimated the A constant. As shown in Equation 2.9, the A constant for the mean-mean scaling method is derived through dividing the mean of the new form anchor item *a* parameters by the mean of the base form anchor item *a* parameters. As cheating decreases the anchor item *a* parameters for the new form, the mean of the new form anchor items will decrease in turn. Thus, the numerator decreases in Equation 2.9, while the denominator remains unaffected. This causes the A constant to be underestimated in the mean-mean scaling method.

Calculation for the A constant in the mean-sigma method applies the standard deviation of the anchor item *b* parameters, as shown in Equation 2.7. Specifically, the A constant is found by dividing the standard deviation of the base form anchor item *b* parameters by the standard deviation of the new form anchor item *b* parameters. As discussed previously, cheating will cause the *b* parameters for the anchor items to be underestimated. If this underestimation is severe, as in the extreme cheating condition, the fact that a majority of the examinees respond correctly to the anchor items causes the standard deviation of the b parameters to decrease considerably. Thus, the A scaling constant will be overestimated for the mean-sigma method. As discussed previously, the characteristic curve methods consider all item parameters when estimating the scaling constants. Thus, it is difficult to identify the specific factors influencing the bias in these scaling approaches.

Underestimation of the A scaling constant has several effects on scaling. First, an underestimated A constant will underestimate the true variability of ability and *b* parameters as indicated by Equations 2.3 and 2.5. Second, the new form unique item *a* parameters will be inflated, as shown in Equation 2.4. Inflating the new form *a* parameters will cause correct responses to increase an examinee's ability estimate at a higher rate and incorrect responses to decrease the estimated ability more drastically.

As noted, the moment methods considerably diverged in estimating the A constant at extreme degrees of cheating. This led to the methods producing different B scaling constants. Equations 2.8 and 2.10 show that the B constant is a function of the anchor item mean *b* parameters and the A scaling constant. Specifically, the mean of the old form anchor item *b* parameters is subtracted from the product of the new form anchor item *b* parameters and the A scaling constant. Because the new form anchor item *b* parameters are underestimated when compromised, and thus more negative on average, multiplying them by a larger A results in an even more negative value. Because this negative value is subtracted, B is more positively biased. For this reason, the positively biased A constant found for the mean-sigma method leads to larger overestimation of the B constant in comparison to the mean-mean method. This effect can be seen in Table 4, where the mean-sigma method produced a considerably larger bias than mean-mean method under extreme cheating.

This study attempted to determine if a particular scaling method was robust to the effects of cheating and could accurately recover scaling constants as well as equated scores. Results indicated that cheating severely influences the recovery of scaling constants for each method. No scaling method investigated in this study produced results that could justify applying these methods to counteract cheating given the consequences associated with high-stakes tests. In general, the mean-mean method performed best in recovering equated scores at low to moderate degrees of cheating. The fixed anchor method displayed the least bias under extreme cheating. Yet, all scaling methods produced consistently overestimated equated scores with little differences among the methods. Perhaps more important, the degree of bias, even at moderate cheating conditions for the best performing scaling methods, would not be acceptable for a high-stakes test.

Results regarding the influence on the new form examinee ability distribution suggest that the mean ability of the new form group can influence recovery of equated scores. Bias in the equated scores increased when the new form examinees had a lower mean ability than the base form examinees. The overestimated equated scores occur for lower and middle ability examinees because examinees with higher ability have less to gain from cheating. Figures 2 and 3 help illustrates this effect; the largest bias in equated scores occurs for examinees at true raw scores slightly at or below the middle of the score scale. Equated score bias decreases sharply at higher values of new form raw scores. When the new form examinee population is less able than the base form examinee population, there is a larger proportion of new form examinees falling within the area of the distribution where cheating is most beneficial. Thus, the overall bias resulting from cheating increases when the new form examinees are less able on the test in comparison to the examinees setting the metric. These results have negative implications for testing programs. As lower ability examinees benefit the most from cheating, unqualified examinees may appear qualified solely because cheating occurred.

Another condition investigated by this study was whether employing an external anchor would alter the influence of cheating on equated scores. Employing an external anchor negates the direct influence of compromised anchor items on total scores as anchor items are used for scaling purposes only. Thus, cheating examinees that have access to the compromised items should not receive a benefit from these items specifically. Results demonstrated that the overall bias for the external anchor was less positive than when the anchor employed was internal. However, the overall bias for the external condition was still considerable.

To explore the differential effects of anchor types further, bias in equated scores was compared between honest and dishonest test-takers within the two anchor conditions. As expected, within the internal condition cheaters benefited from the compromised items at a higher degree than honest test takers. This difference was exacerbated when the proportion of compromised anchor items was increased. In contrast, cheaters and honest test takers benefitted equally from cheating in the external anchor condition. This result supports the hypothesis that the direct benefit from the compromised items in the internal anchor introduces additional bias for cheating examinees. Bias in equated scores arises in the external anchor condition solely as a result of the impact of cheating on IRT scaling discussed previously. This result is demonstrated in Figure 3, where honest examinees in the internal anchor condition benefited at an equal rate as all examinees administered an external anchor test across the range of ability. The considerable bias obtained in the external condition displays the severe impact that cheating has on the scaling process. Employing an external anchor will prevent bias attributable to scores on compromised items directly, however, the substantial amount of bias related to inaccurate scaling will continue to plague the scores obtained from equating.

The overall trends found in this study demonstrate the detrimental effects of cheating on equated scores estimated under the non-equivalent anchor test design. When examinees have access to the items used to scale a new form of a test to a common metric, equated scores for the entire group of examinees will be overestimated. Although relatively few examinees may be cheating on the test, scores for the entire group of examinees administered the form will appear more proficient. Clearly, decisions made based on the test scores when any amount of cheating has occurred will be dubious at best. Specifically, these results suggest that if cheating occurs on their form, under-qualified examinees —whether they engage in cheating behaviors or not—may be unfairly given preference over qualified examinees that completed an uncompromised form.

Unfortunately, no scaling methods examined in this study were robust to the bias introduced by cheating. Given the detrimental implications of compromised anchor items, focus should be given to identifying these items and removing them from the equating process and scoring. As such, future studies must address the detection of cheating and compromised items. Several results of this study may be useful in developing models that identify compromised items under the NEAT design. For example, understanding *a* and *b* parameters for the compromised anchor items will be for underestimated for cheating examinees, a mixture IRT (von Davier & Carstensen, 2007) model can be specified that contains a two class mixture with a “cheater” class containing lower *a* and *b* parameters in reference to the other, “honest examinee”, class. Items that show large parameter differences across the two classes may be removed from scaling, equating and scoring to protect against the negative consequences of cheating.

Admittedly, simulation studies cannot capture the complexity of applied situations completely. Although this simulation was designed to address this issue by exploring realistic conditions and using real item parameters, several limitations remain. Perhaps most important, the cheating adjustment value selected in this study may not reflect the actual benefit of knowing the item prior to administration. Although prior item knowledge assuredly benefits examinees, future research should examine the functional relationship between probability of correct response and prior knowledge to gauge the strength of this effect. In addition, because details of cheating often go unreported, the cheating conditions selected in this study may not reflect the degree of cheating occurring in actual high stakes testing. However, the results of this study should generalize well to other degrees of cheating. That is, the positive bias in equated scores should be a function of the degree of cheating such that any increase in cheating will only further overestimate the equated scores.

In conclusion, this study addressed a gap in the literature by exploring the effects of cheating on equated scores obtained from IRT equating under the NEAT design. If examinees have prior access to the items used to scale a test form, equated scores obtained for the all examinees administered the form may be overestimated. Even small amounts of cheating call into question the results obtained from the test, with large degrees of cheating distorting the scores completely. Given the influential decisions made based on high-stakes tests, it is imperative that scores reflect the examinees' true abilities on the attribute measured. The impact that cheating has on the entire distribution of scores is a severe threat to the validity of inferences made from test scores. Scores for all examinees, both honest and dishonest, may indicate that the examinee is more proficient than reality. Thus, examinees administered forms where no cheating occurred may be unfairly disadvantaged. Research investigating methods to detect compromised items must take priority to ensure that scores, and thus decisions made, from high stakes tests accurately reflect the ability of the examinees.

Appendix A

Simulation Conditions Investigated

			New Form Examinee Ability Distributions																			
			N(0,1)					N(-.5,1)					N(0,1.25)					N(-.5,1.25)				
			Scaling Method																			
Anchor	Cheaters (%)	Compromised Items (%)	MM	MS	SL	HB	FA	MM	MS	SL	HB	FA	MM	MS	SL	HB	FA	MM	MS	SL	HB	FA
Internal	10	25	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500
		100	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500
	50	25	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500
		100	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500
	0	0	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500
External	10	25	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500
		100	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500
	50	25	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500
		100	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500
	0	0	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500

Note. MM = Mean-mean, MS = Mean-sigma, SL = Stocking-Lord, HB = Haebara, FA = Fixed anchor

Appendix B

Internal Anchor Equated Score RMSE by Ability Distribution

% Compromised Anchor Items	% Cheaters	Ability Distribution	Scaling Method				
			SL	HB	MM	MS	FA
0	0	N(0.0,1.00)	4.556	4.562	4.588	4.592	4.554
		N(-0.5,1.00)	4.542	4.542	4.558	4.563	4.712
		N(0.0,1.25)	4.330	4.330	4.364	4.375	4.318
		N(-0.5,1.25)	4.434	4.435	4.475	4.483	4.614
25	10	N(0.0,1.00)	4.629	4.602	4.609	4.615	4.582
		N(-0.5,1.00)	4.539	4.504	4.536	4.542	4.782
		N(0.0,1.25)	4.448	4.416	4.438	4.453	4.452
		N(-0.5,1.25)	4.467	4.426	4.457	4.464	4.850
	50	N(0.0,1.00)	6.002	5.232	5.208	5.290	5.093
		N(-0.5,1.00)	5.927	5.130	5.038	5.262	5.647
		N(0.0,1.25)	5.801	5.037	5.017	5.106	5.116
		N(-0.5,1.25)	5.828	5.013	4.910	5.096	5.773
100	10	N(0.0,1.00)	5.383	5.125	5.141	5.151	4.974
		N(-0.5,1.00)	5.397	5.040	5.220	5.033	5.473
		N(0.0,1.25)	5.297	4.985	4.959	4.970	4.911
		N(-0.5,1.25)	5.429	5.025	5.092	4.963	5.564
	50	N(0.0,1.00)	15.041	13.512	12.783	14.756	11.328
		N(-0.5,1.00)	15.392	13.911	14.520	16.629	12.647
		N(0.0,1.25)	15.762	13.863	12.552	14.335	11.873
		N(-0.5,1.25)	16.373	14.642	14.716	17.226	13.066

Appendix C

Internal Anchor RMSE for Scaling Constants A by Ability Distribution

% Compromised Anchor Items	% Cheaters	Ability Distribution	Scaling Method			
			SL	HB	MM	MS
0	0	N(0.0,1.00)	0.025	0.024	0.037	0.041
		N(-0.5,1.00)	0.027	0.026	0.039	0.044
		N(0.0,1.25)	0.035	0.032	0.043	0.051
		N(-0.5,1.25)	0.033	0.032	0.046	0.049
25	10	N(0.0,1.00)	0.034	0.027	0.039	0.042
		N(-0.5,1.00)	0.037	0.031	0.043	0.045
		N(0.0,1.25)	0.046	0.037	0.045	0.056
		N(-0.5,1.25)	0.048	0.040	0.051	0.053
	50	N(0.0,1.00)	0.126	0.100	0.048	0.109
		N(-0.5,1.00)	0.136	0.107	0.057	0.124
		N(0.0,1.25)	0.166	0.126	0.060	0.142
		N(-0.5,1.25)	0.175	0.133	0.067	0.152
100	10	N(0.0,1.00)	0.042	0.030	0.058	0.046
		N(-0.5,1.00)	0.050	0.030	0.086	0.052
		N(0.0,1.25)	0.075	0.039	0.074	0.056
		N(-0.5,1.25)	0.089	0.048	0.109	0.059
	50	N(0.0,1.00)	0.119	0.136	0.138	0.257
		N(-0.5,1.00)	0.140	0.167	0.207	0.427
		N(0.0,1.25)	0.628	0.090	0.226	0.189
		N(-0.5,1.25)	0.139	0.097	0.357	0.355

Appendix D

Internal Anchor RMSE for Scaling Constants B by Ability Distribution

% Compromised Anchor Items	% Cheaters	Ability Distribution	Scaling Method			
			SL	HB	MM	MS
0	0	N(0.0,1.00)	0.028	0.027	0.042	0.044
		N(-0.5,1.00)	0.033	0.032	0.052	0.053
		N(0.0,1.25)	0.034	0.033	0.043	0.044
		N(-0.5,1.25)	0.035	0.035	0.049	0.051
25	10	N(0.0,1.00)	0.061	0.049	0.051	0.052
		N(-0.5,1.00)	0.075	0.059	0.067	0.066
		N(0.0,1.25)	0.062	0.052	0.051	0.051
		N(-0.5,1.25)	0.077	0.062	0.065	0.064
	50	N(0.0,1.00)	0.261	0.186	0.189	0.187
		N(-0.5,1.00)	0.326	0.228	0.209	0.236
		N(0.0,1.25)	0.257	0.182	0.183	0.178
		N(-0.5,1.25)	0.327	0.227	0.202	0.229
100	10	N(0.0,1.00)	0.209	0.177	0.168	0.171
		N(-0.5,1.00)	0.260	0.208	0.231	0.197
		N(0.0,1.25)	0.218	0.184	0.170	0.173
		N(-0.5,1.25)	0.275	0.218	0.225	0.198
	50	N(0.0,1.00)	1.078	1.016	0.856	1.157
		N(-0.5,1.00)	1.236	1.142	1.078	1.394
		N(0.0,1.25)	1.781	1.025	0.850	1.110
		N(-0.5,1.25)	1.280	1.182	1.089	1.430

Appendix E

Equated Score RMSE by Anchor Type across Ability Distribution

% Compromised Anchor Items	% Cheaters	Scaling Method				
		SL	HB	MM	MS	FA
Internal Anchor Set						
0	0	4.847	4.849	4.890	4.899	4.925
25	10	4.962	4.942	4.957	4.970	5.084
	50	6.658	5.847	5.831	5.899	6.237
100	10	6.561	6.266	6.158	6.246	6.377
	50	21.058	19.698	17.962	22.015	17.019
External Anchor Set						
0	0	5.582	5.584	5.621	5.629	5.687
25	10	5.651	5.609	5.638	5.648	5.833
	50	7.362	6.379	6.304	6.485	6.759
100	10	6.721	6.305	6.379	6.286	6.538
	50	19.552	17.478	17.053	19.671	15.286

Appendix F

RMSE for Honest and Dishonest Test Takers for Internal and External Conditions

% Compromised Anchor Items	% Cheaters	Ability Distribution	Internal		External	
			Honest	Dishonest	Honest	Dishonest
25	10	N(0.0,1.00)	4.969	5.355	5.760	5.752
		N(-0.5,1.00)	5.053	5.558	5.727	5.726
		N(0.0,1.25)	4.764	5.189	5.551	5.552
		N(-0.5,1.25)	4.935	5.471	5.669	5.668
	50	N(0.0,1.00)	5.573	6.468	6.712	6.724
		N(-0.5,1.00)	5.800	6.841	6.764	6.763
		N(0.0,1.25)	5.371	6.257	6.527	6.531
		N(-0.5,1.25)	5.650	6.676	6.670	6.675
100	10	N(0.0,1.00)	5.505	10.558	6.445	6.447
		N(-0.5,1.00)	5.759	11.867	6.543	6.553
		N(0.0,1.25)	5.269	10.374	6.283	6.285
		N(-0.5,1.25)	5.598	11.550	6.524	6.530
	50	N(0.0,1.00)	14.174	21.497	16.927	16.952
		N(-0.5,1.00)	16.641	25.819	18.344	18.358
		N(0.0,1.25)	13.611	20.769	17.172	17.190
		N(-0.5,1.25)	16.216	24.982	19.098	19.086

References

Allalouf, A. (2007). Quality Control Procedures in the Scoring, Equating, and Reporting of Test Scores. *Educational Measurement: Issues and Practice, 26*, 36-43.

Allen, N.L., Donoghue, J.R. & Schoeps, T.L. (2001). *The NAEP 1998 Technical Report*, (NCES publication No. 2001-509). Washington, DC: National Center for Education Statistics.

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

Baron, K., Wirzbicki, A. (2008, July 22). *Study confirms widespread cheating on job exams*. Boston.com. Retrieved from http://www.boston.com/jobs/news/articles/2008/07/22/study_confirms_widespread_cheating_on_job_exams/

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

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In. Lord, F. M. & Novick, M. R., *Statistical theories of mental test scores* (pp. 397-472). Reading, MA: Addison-Wesley.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in microcomputer environment. *Applied Psychological Measurement, 6,* 431-444.

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

Brodkin, J. (2008, September 3). Don't be fooled by suspicious test preparation Web sites. *Network World*. Retrieved from http://www.networkworld.com/newsletters/edu/2008/090108ed1.html?hpg1=bn

Cizek, G.J. (1999). *Cheating on tests: How to do it, detect it, and prevent it.* Mahwah, NJ: Erlbaum.

Cohen, A. S., & Wollack, J. A. (2006). Test administration, security, scoring, and reporting. In R. L. Brennan (Ed.) *Educational Measurement* (4th ed., pp. 355-386). Westport, CT: American Council on Education/Praeger.

Cook, L. L., & Eignor, D. R. (1985). *An investigation of the feasibility of applying item response theory to equate achievement tests* (Research Report 85-31). Princeton, NJ: Educational Testing Service.

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

Cook, L. L., & Petersen, N. S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. *Applied Psychological Measurement, 11,* 225-244.

DeMars, C. (2010). *Item Response Theory*. Oxford University Press.

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

Finkelman, M., Nering, M. L., & Roussos, L. A. (2009). A conditional exposure control method for multidimensional adaptive testing. *Journal of Educational Measurement, 46*, 84-103.

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*, 283-309.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice 23*, 17–27.

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

Hambleton, R., & Swaminathan, H. & Rogers, J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage Publications.

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

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

Hanson, B. A., & Béguin, 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,* 3-24.

Hendrickson, A. B., & Kolen, M. J. (1999). IRT equating of the MCAT. *MCAT Monographs*. Washington, DC: Association of American Medical Colleges.

Holland, P.W. (2007) A framework and history for score linking. In: Dorans NJ, Pommerich M, Holland PW. (Eds), *Linking and Aligning Scores and Scales(* pp. 5–30). New York, NY: Springer, 2007.

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 187–220). Westport, CT: American Council on Education/Praeger.

Jodoin, M. G., Keller, L. A., & Swaminathan, H. (2003). A comparison of linear, fixed common item, and concurrent parameter estimation equating procedures in capturing academic growth. *The Journal of Experimental Education, 71,* 229–250.

Jurich, D. P., Goodman, J.T., & Becker, K. A. (May, 2010) *Assessment of Various Equating Methods: Impact on the Pass-Fail Status of Cheaters and Non-*

Cheaters. Poster presented at the annual meeting of the National Council on Measurement in Education, Denver, CO.

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

Kim, S. (2006). A comparative study of IRT fixed parameter calibration methods. *Journal of Educational Measurement, 43,* 355-381.

Kim, S., & Kolen, M. J. (2007). Effects on scale linking of different definitions of criterion functions for the IRT characteristic curve methods. *Journal of Educational and Behavioral Statistics*, 32, 371-397.

Klein, L. W, & Jarjoura, D. (1985) Effect of number of common items in common-item equating with nonrandom groups. *Journal of Educational Measurement, 22,* 197–206.

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

Kolen, M. J. (2007). Data collection designs and linking procedures. In N. J. Dorans, M. Pommerich, & P.W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 31–55). New York: Springer.

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

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

Li, K. (2003, March 5). Fraudulent TOEFL takers face possible deportation. *Daily Princetonian.* Avaliable at: http://www.dailyprincetonian.com/2003/03/05/7516/

Livingston, S.A. (2004). *Equating test scores (without IRT).* Educational Testing Service, Princeton, NJ.

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

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

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

Macdonald, P., & Paunonen, S. V. (2002). A monte carlo comparison of item and person statistics based on item response theory versus classical test theory. *Educational and Psychological Measurement, 62*, 921.

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

McCabe, D. L., & Trevino, L. K. (1996). What we know about cheating in college: Longitudinal trends and recent developments. *Change, 28*, 29-33.

McCabe, D. L., Trevino, L. K., & Butterfield, K. D. (2001). Cheating in academic institutions: A decade of research. *Ethics & Behavior*, *11*, 219–233.

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

McLeod, L., Lewis, C., & Thissen, D. (2003). A Bayesian method for the detection of item preknowledge in computerized adaptive testing. *Applied Psychological Measurement*, *27*, 121-137.

Mislevy, R. J.,& Bock, R. D. (1990). BILOG-3 (2nd ed.): *Item analysis and test scoring with binary logistic models.* [Computer software and manual] Mooresville, IN: Scientific Software.

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

Nichols, S. L., & Berliner, D. C. (2007). *Collateral damage: How high-stakes testing corrupts America's schools*. Cambridge, MA: Harvard Education Press.

No Child Left Behind Act of 2001, Pub. L. No. 107-110.

Petersen, N. S. (2007). Equating: Best practices and challenges to best practices. In N. J. Dorans, M. Pommerich, & P.W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 59–72). New York: Springer.

Petersen, N. S., Cook, L. L., & Stocking, M. L. (1983). IRT versus conventional equating methods: A comparative study in scale stability. *Journal of Educational Statistics,* 8, 136-156.

Skaggs, G., & Lissitz, R. W. (1986). IRT test equating: Relevant issues and a review of recent research. *Review of Educational Research,* 56, 495-529.

Stocking, M. L., & Lewis, C. (1998). Controlling item exposure conditional on ability in computerized adaptive testing. *Journal of Educational and Behavioral Statistics, 23*, 57–76.

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

Wollack, J. A. (2006). Simultaneous use of multiple answer copying indexes to improve detection rates. *Applied Measurement in Education*, 19, 265-288.

Wollack, J. A., Cohen, A. S., & Serlin, R. C. (2001). Defining error rates and power for detecting answer copying. *Applied Psychological Measurement*, 25(4), 385-404.

Vale, D. C. (1986). Linking item parameters onto a common scale. *Applied Psychological Measurement*, *10*, 133–144.

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

van der Linden, W.J., Veldkamp, B.P. (2004). Constraining item-exposure rates in computerized adaptive testing with shadow tests. *Journal of Educational and Behavioral Statistics*, *29*, 273–291.

von Davier, A. A., Holland, P. W., Thayer, D. T. (2004). *The Kernel Method of Test Equating*. New York: Springer-Verlag.

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

Yen, W.M. (1983). Tau-equivalence and equipercentile equating. *Psychometrika, 48,* 353- 369.

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

Zimowski, M., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). BILOG-MG 3: Item analysis and test scoring with binary logistic models. [Computer software]. Chicago, IL: Scientific Software