UNIVERSITY OF ALBERTA

THE CONSEQUENCES OF MULTIDIMENSIONALITY TO IRT EQUATING OUTCOMES USING A COMMON-ITEMS NONEQUIVALENT GROUPS DESIGN

by

Kathryn Louise Ricker

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for
the degree of Doctor of Philosophy
in
Psychological Studies in Education

Department of Educational Psychology
Edmonton, Alberta
Spring, 2007

Your file Votre référence
ISBN: 978-0-494-29728-5
Our file Notre référence
ISBN: 978-0-494-29728-5

Abstract

Two studies were conducted to evaluate the consequences of multidimensionality on equating outcomes when IRT true-score equating is employed under a common-items nonequivalent groups (CI-NEG) design. The Stocking-Lord (1983) scale transformation was employed. The first study was a simulation using realistic item parameters with two 68-item test forms, X and Y. A 2 (form parallelism) x 4 (correlation between dimensions) x 3 (group equivalence) x 3 (location of items measuring the second dimension) factorial design was employed, giving a total of 72 conditions. Each condition was replicated 100 times, and measures of Mean Absolute Difference and proportion of examinees with a Score Difference that Matters (Dorans & Feigenbaum, 1994) were examined. The second study used real data taken from two separate licensure tests. Each test was split to create two forms. The scores from each form were equated to each other using equivalent and nonequivalent groups through the common-items nonequivalent groups design. The equating procedures were conducted in a manner identical to those used in the simulation study.

In the simulation study, equating with parallel forms tended to be robust under most conditions, but equating with nonparallel forms tended not to be robust even under unidimensional conditions. In both simulated and real data studies, when the second dimension items were among both the unique and common items, equating tended to be more robust than when the items were in the other two locations. Equating was least robust when the items measuring the second dimension were only represented in the common items. The simulation study results support the previously established relationship between the correlation between test dimensions and IRT equating robustness. In both studies, IRT equating using a CI-NEG design tended to be more robust when the groups were equivalent. Equating benefits were limited with nonparallel forms. In many nonparallel conditions, error associated with scores was larger when equating was performed than when unequated. Collectively, these results provide evidence that factors beyond just the degree of multidimensionality present on tests mediate the robustness of IRT equating, including form parallelism, the location of the items measuring the second dimension, and group equivalence.

Acknowledgements

There are many people whom I need to thank for helping me complete this dissertation. First, I would like to thank my supervisor, Dr. Mark Gierl, for his help and support along the way. Without his guidance, I never would have gotten through. Second, thanks to Dr. Todd Rogers, who read more drafts of a dissertation than any committee member should have to. Third, thank you to my remaining committee members, Dr. Jacqueline Leighton, Dr. Stephen Hunka, Dr. Peter Hurd, and Dr. Daniel Bolt for your helpful feedback and comments.

Several other people also provided technical help along the way. They include Greg Sadesky, vBASIC whiz, Dr. Terry Ackerman, who modified the program M2GEN2 for me to accommodate my needs, and Dr. Michael Jodoin, who provided periodic technical and moral support.

A big thank you to my family and friends who saw me through the process. This was the hardest thing I've ever done in my life. Without you all, I would be even crazier than I am now!

Dedication

This dissertation is dedicated to the memory of my grandmothers, Olive Johnston and Marjory Ricker. These women taught me what it means to persevere over adversity.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Overview

Test score equating is a practice often used in large-scale testing programs in which multiple forms of the same test must be administered to ensure test security and equity in the testing process. Test equating can be used to adjust for small differences in test form difficulty, so that test scores are equivalent. However, for equating to be useful, it must be conducted in such a way that the equated test scores are reliable and can be interpreted validly.

It should be emphasized that test equating can only be conducted on test forms that are intended to be parallel in test content, specifications, and difficulty (Kolen & Brennan, 2004, p. 10). For example, while it would be possible to equate scores from two forms of the same Grade 9 math test, it is not possible to equate scores from two Grade 9 math tests that cover different concepts, nor would it be possible to equate scores from a Grade 9 math test to scores from a Grade 6 math test. Linking and scaling, respectively, are the methods that could be used to compare test scores in these two cases. Only equating allows for test scores to be used interchangeably (Kolen & Brennan, 2004, p. 10).

There are two main aspects of equating that must be considered. The first consideration in equating test forms is the data collection design. In testing programs, the common-items nonequivalent groups design is used commonly. This design has a subset of items that is identical on both tests. Performance on these items is used during equating to separate differences in item characteristics (difficulty, discrimination) of the unique items on each test form from ability differences between the groups assigned to each test form. This design is used when groups cannot be assumed to be randomly equivalent, such as in cases where it is not possible to administer the test forms at the same time or to randomly assign examinees to test forms.

With the data collection design selected, the second consideration in equating is the statistical method to be used to equate the test forms. Item response theory equating is a popular method for test equating. Item response theory, or IRT[1], is used often in test development and it is therefore advantageous to continue to use IRT during the equating process. IRT equating is based on the same theoretical underpinnings as IRT item and examinee calibration (Embretson & Reise, 2000) and so requires the same assumptions to be held: local independence, nonspeededness, and unidimensionality.

Unidimensionality is difficult to achieve under many operational testing conditions because most large scale tests measure complex constructs[2] and/or require multiple skills. For the purposes of this research, a test is defined as unidimensional if it contains only one dimension or if it meets the criterion of essential unidimensionality (Stout, 1987). By extension, a test is defined as multidimensional if it contains two or more dominant dimensions and fails to meet the criterion of essential unidimensionality.

Most tests contain more than one dimension, even if one dimension is dominant. As a result, multidimensional IRT (MIRT) models are necessary to account for additional dimensions. Multidimensional equating models are still under development (e.g.,

[1] It should be noted that IRT can be used in two contexts, 1) to refer to Item Response Theory as a general concept, including both unidimensional and multidimensional models, and 2) to refer specifically to unidimensional IRT. For the purposes of this dissertation, the acronym IRT refers only to unidimensional models and equating methods, while Multidimensional IRT (or MIRT) will be referred to specifically, if required.

[2] Because of their natural connections, there is often confusion and/or vagueness when discussing constructs and dimensions (Reckase, 2006). It should be noted that while substantive constructs and statistical dimensions might align closely, it is not necessarily always the case and therefore it becomes important to treat these two concepts as distinct. See p. 5 for definitions.

Oshima, Davey, & Lee, 2000) and are not widely used in practice. In the absence of a MIRT equating methodology, practitioners often choose to use unidimensional IRT equating even when the underlying data structure is multidimensional. But what are the consequences of failing to meet this assumption underlying the use of IRT on equating outcomes?

Purpose

The purpose of this research was to explore systematically four variables that are expected to affect equating outcomes using unidimensional IRT score equating with a common-items nonequivalent groups design when two dominant dimensions are present. There are several properties that can be used to choose criteria to evaluate the adequacy of the equated outcomes. This research focused on the equity property, which stipulates that the scores of tests, once equated, should be equivalent such that it does not matter which test form an examinee is administered. First, in order to have control over the variables, a simulation study was conducted. Four research questions were addressed:

1) What is the baseline error that is associated with equating? That is, how much error would be present if scores on a test form were equated to scores on the same test form?

2) How does the correlation between dimensions affect both the magnitude of equating error and the proportion of examinees with error in their equated scores that is large enough to matter?

3) Is the magnitude of equating error or the proportion of examinees with equating error that is large enough to matter in their equated scores different if the groups are randomly equivalent versus nonequivalent?

4) Does the location of the items (unique, common, both unique and common locations) related to the second dimension have an effect on the magnitude of equating error and the proportion of examinees who are affected by equating error?

Once these questions were addressed using a simulation design, data from two real tests were analyzed. The results of the real data study were compared to the simulation study results to see if they demonstrated the same patterns for the factors that were artificially manipulated during the simulation. If the results were found to align, then there is increased confidence that the simulation results are reasonable and might be generalizable to other real testing situations.

Rationale

This research is important because while, strictly speaking, IRT equating techniques should only be used under unidimensional conditions, there is some evidence to suggest that IRT equating is robust to violations of this assumption *under some conditions* (Camilli, Wang & Fesq, 1995; De Champlain, 1996; Dorans & Kingston, 1985). This study systematically explored key factors that might mediate the robustness of unidimensional IRT equating, including specific properties of the common-items nonequivalent groups design that have not yet been evaluated in the psychometric literature, including the location of the items measuring the second dimension, the degree of parallelism between the forms, and the degree of group ability differences.

Glossary of Technical Terms

Common-items Nonequivalent Groups Design: A data collection design used for equating in which the groups of examinees administered each test form are assumed to be nonequivalent in terms of ability. A correction for ability differences between groups is possible by comparing performance on a set of items that are common to both test forms. These common items are matched to either the remaining items on the test in terms of difficulty and content specifications for an internal anchor (as used in the present study) or to all of the items on the forms being equated for an external anchor.

Construct: a substantive ability or trait that that is measured by an assessment or test (Reckase, 2006).

Dimension: a statistically determined aspect of a data set. A dimension is a latent trait that is measured in the examinees by the some or all of the items on a test form. Dimension(s) may or may not align with the anticipated pattern of results based on the construct(s) being measured. Therefore, differing substantive and statistical interpretations the performance of a test are possible.

Equating: A method to correct for small differences in difficulty between alternate forms of the same test (built to the same specifications and difficulty) that allows test scores to be interchangeable.

Equity: "If an equating of tests X and Y is to be equitable to each applicant, it must be a matter of indifference to applicants at every given ability level θ whether they are to take test X or test Y." (Lord, 1980, p.195)

Essential Unidimensionality: A test that contains only one dominant dimension along with one or more weaker dimensions that do not interfere with the measurement of the dominant dimension meets the requirements of essential unidimensionality. Often used as an operational definition of unidimensionality in practical settings.

Item Response Theory (IRT): A test theory that models examinee performance based on the interaction between examinee characteristics (ability or abilities) and item characteristics (e.g., difficulty, discrimination).

Local Independence: An assumption of IRT that requires that examinee responses to items be independent of each other after accounting for ability.

Multdimensional: A test that contains more than one dimension (or fails to meet the criterion for essential unidimensionality).

Parallel Test Forms: Two or more test forms that are built to the same content specifications and having matching statistical characteristics (for the purposes of this dissertation, closely matching means and variances of *a*- and *b*-parameters).

Parameter: A property or characteristic of an item or person that is expressed numerically.

Robust: Relating to a result (or results) of a statistical procedure (in this case equating) that have an acceptable level of error (an acceptable level can have multiple definitions, depending on the scenario), despite violation(s) to the assumption(s) of the theory underpinning the procedure.

Scale Transformation: A statistical process by which test items that are not on the same scale are adjusted to be on a common scale.

Score Difference that Matters: Any score difference between the equated score and criterion score that is larger than one-half of a reporting scale unit (i.e., a score difference that would be detected after scores were rounded for reporting).

Target Score: The actual score an examinee achieved on Form X, which acts as a standard for comparison for the equated Form Y scores (equated to Form X).

Unidimensional: A test form that contains one dimension. A theoretical idea, as most real test forms will contain more than one dimension, but will meet the definition of essential unidimensionality (see p. 5).

Organization of Dissertation

The next two chapters will cover the relevant literature and methods used to conduct this simulation study. Chapter 2 is divided into two parts. The first part will introduce the relevant concepts of equating, and more specifically, IRT equating, discussing the significance of these concepts to the research question of interest. The second part reviews and critiques the body of literature dedicated to examining the effects of violating unidimensionality on IRT equating, highlighting the current understanding of the issue and what is still unknown. Chapter 3 contains the methods, procedures, dependent measures and software used to complete the research, which was approached using both simulated and real data analyses.

With the rationale, background, and methods outlined in Chapters 1, 2, and 3, Chapter 4 presents the results of the simulation study, while Chapter 5 presents the results of the real data study. In Chapter 6, the limitations of both studies are outlined, followed by a discussion of the results. The discussion includes the possible causes of the observed results, the relation of the simulated results to the real data results, the relation of these results to past studies, and conclusions about the implications of this research for IRT equating applications. Finally, recommendations for future research are proposed.

Chapter 2: Relevant Concepts and Literature Review

Overview

Large-scale standardized testing is an important part of education in North America. Students are administered many standardized tests over their lifetime. The decisions made from these test results can be as important as graduating from high school, being admitted to university, or being licensed to work in a chosen profession. Given the high-stakes nature of these tests and the potentially adverse effects of incorrect decisions based on these test scores, it is important that the entire testing process, from test development to score interpretation, be equitable. To protect equity, new test forms are often used for each administration in order to ensure that test items are not known to some examinees prior to the examination date. Ideally, alternate test forms would be parallel to previous test forms in both substantive content and statistical difficulty (Kolen & Brennan, 2004, p. 3). However, this task is nearly impossible to achieve during test development, especially if a new test form is required for every test administration, or if items are not piloted to gather information on their statistical characteristics. Additionally, pilot testing of new test forms may not always align with actual performance during a real test administration because pilot results are based on small samples using incomplete or experimental test sections, and pilot samples are often not representative of the population. As a result, differences between the test forms are a threat to equity that must be addressed.

Test score equating is a method used to adjust for score differences that occur between test forms. It provides a method for comparing and interpreting test scores from different forms. However, test equating is not infallible. Threats to the statistical assumptions that underpin the theory upon which equating methods are based have the potential to create faulty equating functions, resulting in systematic equating errors. Faulty equating potentially poses as much threat to test equity as the original differences in form difficulty (Harris & Crouse, 1993).

With the stakes and implications of testing being so high, it is important for all stakeholders to feel confident about the procedures used to come to decisions based on test scores. Because equating plays an important role in the testing process, it is important to investigate the conditions which might pose a threat to equating, and to document when and to what extent equating errors occur under specific conditions, so that testing practitioners can make more informed choices about which equating methods and procedures to use.

The following chapter is divided into two main sections. First, the relevant concepts of IRT equating are reviewed. The organization of this section is intended to mirror the steps that are taken in conducting an equating study. Second, a critical review of literature related to using unidimensional IRT under multidimensional conditions will provide a summary of what is known about this topic to date, as well as demonstrate where there are gaps in the current body of knowledge. A brief summary follows both sections.

Relevant IRT Equating Concepts

This section reviews the concepts that are relevant to the research conducted to examine the effects of multidimensionality on unidimensional IRT equating, specifically its effects on equating equity. The review includes 1) general equating concepts including: a) properties of equating, b) data collection designs used for equating, and c) equating methods; 2) concepts specific to IRT equating, including: a) assumptions of IRT equating, b) dimensionality and data structure, c) two-parameter logistic IRT models, d) IRT parameter estimation, e) Stocking and Lord scale transformation, and f) IRT true score equating; and, finally; 3) criteria for assessing the adequacy of equating, including a) simulated data, b) indices, c) circular equating, and d) practical criteria.

General Equating Concepts

Properties of equating. The relationship between the scores on test forms should possess three properties once equated: symmetry, invariance, and equity (Kolen & Brennan, 2004, pp. 10-13). These properties are desirable because they contribute to both the validity and reliability of the equating function. These properties ensure that equating is contributing to the fairness of the testing process.

The first property, symmetry, describes the relationship between the scores of two equated test forms. Symmetry requires that the equation that converts scores from form X onto the scale of form Y should be the inverse of the equation to convert scores from form Y onto the scale of form X (Kolen & Brennan, 2004, p. 10). This property dictates that no advantage or disadvantage in final equated score exists because of the direction (Y to X or X to Y) of the equating. The second property, invariance, requires that the equating function for any given subgroup in the population of examinees should be the same for all other subgroups within the population (Kolen & Brennan, 2004, p. 13). This property implies that the tests must be free of bias that would affect test scores. The third property, equity, also relates to avoiding bias in equating. This research will focus on the equity property as the criterion to evaluate equating outcomes.

Lord (1980) stated, "If an equating of tests X and Y is to be equitable to each applicant, it must be a matter of indifference to applicants at every given ability level θ whether they are to take test X or test Y" (p.195). This definition requires that both test forms measure the same psychological construct, and that the construct aligns closely with the dimension. From a statistical standpoint, this property requires that the conditional frequency distribution for the new form (after transformation) must be identical to the conditional frequency distribution for the old form at every θ. To meet this requirement, the forms must have, at a minimum, equal means and variances.

Lord's (1980) definition of equity is now referred to as strong equity, and is considered a theoretical requirement rather than a practical rule. In practice, a less restrictive definition of equity, known as weak equity, is used commonly (Morris, 1982). Morris' definition of equity requires that examinees with the same true score have the same score on form X as on form Y, once the Y scores have been placed on the equated score scale. Weak equity also requires that the test forms measure the same psychological properties or area of achievement. Weak equity is also known as first-order equity.

First-order equity (originally proposed by Divgi, 1981) is achieved when the means of the X and equated Y conditional distributions are equal across the θ distribution. In other words, the expected true score on X is equal to the expected equated score on Y, for all θ. An additional equity criterion, second-order equity, requires the variances of the conditional distributions be equal (Morris, 1982).

Equity is critical to research designed to examine the robustness of equating using multidimensional data because it is the property that is most likely to be sensitive to equating error. According to Lord (1980), equity in unidimensional equating is potentially compromised under multidimensional conditions. The result of poor equity in equating is bias. Bias occurs as a result of equating when, for a given θ, the expected performance on test X is systematically better (or worse) than on test Y after the equating function has been applied (Bolt, 1999).

Data collection designs used for equating. There are two important components to test equating: 1) the equating design that is used to administer the tests, and 2) the statistical method used to equate the tests. The equating design dictates how subsequent equating steps will be conducted. The most technically simple equating design is the random groups design. The random groups design is desirable because it makes equating calculations very straightforward, and only requires that each group take one test form. Any differences in test scores observed between test forms are attributed to differences in the forms themselves because the groups of examinees are assumed to be randomly equivalent. However, this assumption is very difficult to meet in practice because alternate test forms are most often used at different times and in different locations. As a result, there is a need for an equating method that is able to account for differences between test form difficulties separate from group differences.

Another equating design is the single groups design. This design is similar to a repeated measures data collection design, where each person in the group acts as their own control for ability. Because both groups are identical in ability, any differences in ability are attributed to differences in form difficulty. This design is very strong in theory because the groups are identical, but is not usually practical in an operational setting because test-takers would be required to take both test forms, ideally at the same time.

The additional testing burden can lead to fatigue and reduced motivation among examinees (not to mention being undesirable from a customer service point of view).

A more practical design for equating is the common-items nonequivalent groups design. This equating design utilizes a set of items that are common to all test forms. These common (sometimes called anchor) test items are used as a bridge between forms, where performance on the common items is used to control for the effects of differences in ability between the groups so that equating across the forms can occur. This process requires an additional step in the equating process, adding complexity to equating calculations. However, this design is more feasible in practical testing situations because examinees only need to take one form and groups are not required to be randomly equivalent over time.

The importance of the characteristics of the common items set (or anchor test) in this equating design cannot be overemphasized. The items in the common items set must be a surrogate of the remaining items on both test forms, both in statistical specifications and substantive content (Kolen & Brennan, 2004, p. 19). When the common items do not meet these requirements, the validity and reliability of equating is threatened.

Equating methods. Once the equating design is selected, an equating method must be chosen. Three main classes of equating methods are used with the common-items nonequivalent groups design: 1) linear, 2) equipercentile, and 3) IRT. Linear equating methods find a slope and *y*-intercept coefficient that is applied to the raw test scores to linearly transform them to the target score scale. As a result, the equating conversion line is linear, which implies the assumption that forms differ in difficulty uniformly across the entire score scale range. There are many different types of linear equating methods, which differ greatly in procedure and complexity (e.g., Mean; Linear; Tucker (Gulliksen, 1950, pp. 299-301); Levine observed and true score (Levine, 1955)), but as a group they are considered to be the simplest form of equating.

Equipercentile methods (e.g., chain (Dorans, 1990; Livingston, Dorans, & Wright, 1990), frequency estimation (Angoff, 1971; Braun & Holland, 1982)) use the cumulative score distributions from each form to equate the scores. Because they align the distributions, equipercentile conversions are typically non-linear. A score on one form is set to equal the score on the target scale that is of the same percent rank in the cumulative score distribution of the target test form. Equipercentile methods are more accurate than linear conversions because they do not assume a linear relationship between forms. However, the cost of a non-linear conversion is that larger sample sizes than those required for linear equating are required to ensure accuracy of the conversion across the entire score range, so that sample sizes are adequate across the entire score scale.

IRT equating is a commonly used method, especially in cases where IRT is used in other steps of the testing process (e.g., test development). IRT also produces a non-linear conversion, but requires even larger sample sizes than equipercentile methods because adequate sample sizes are also required for accurate item and examinee parameter estimation. As a result, IRT equating methods are generally only used with large-scale assessments where several thousand examinees would take a form at any given administration. As with any other equating method, the assumptions that underlie the IRT procedures must be met in order for the results of the equating to be accurate.

Concepts Specific to IRT Equating

Assumptions of IRT equating. IRT equating is based on three strong assumptions because of its connections to item response theory (IRT). These assumptions are unidimensionality, local independence, and non-speededness. Of these assumptions, two, unidimensionality and local independence, are critical to research examining the effects of multidimensionality on IRT equating.

First, unidimensionality requires that there is one singular factor or trait that the test is measuring. Second, the items are assumed to be locally independent, so the response to one item is not dependent on the responses to any other items once the ability θ is accounted for. Local independence can also be called conditional independence, meaning that test items are independent of each other once conditioned on ability θ. Mathematically, local independence is represented as:

$$P(U_1, U_2, \ldots U_i, \ldots U_n \mid \theta) = P(U_1 \mid \theta) P(U_2 \mid \theta) \ldots P(U_i \mid \theta) \ldots P(U_n \mid \theta),$$

where U_i is the response of an examinee with ability θ to item i of n items. Local independence is an essential assumption of maximum likelihood estimation, an important statistical procedure used commonly to estimate item and ability parameters in an IRT model. When local independence is compromised, so is the accuracy of the maximum likelihood estimates (Hambleton, Swaminathan, & Rogers, 1991).

The first assumption of IRT, unidimensionality, is a straightforward concept, at least in theory. A test is unidimensional if it measures only one trait. A test is multidimensional if it measures more than one trait. The second assumption of IRT, local independence, is related to the first assumption. While distinct concepts, unidimensionality and local independence are linked in unidimensional IRT equating. Local independence requires that responses to items on a test be uncorrelated, conditional on the dimension or dimensions accounted for in the IRT model. If a test is unidimensional, local independence is achieved automatically because only one trait related to performance exists. Unidimensionality is not a necessary condition for local independence; rather as long as all dimensions are fully explicated in the measurement model, local independence will be achieved (Hambleton, Swaminathan, & Rogers, 1991). However, it is important to note that when a unidimensional model is employed under conditions of multidimensionality, not only is the assumption of unidimensionality violated, but also the assumption of local independence is violated.

Dimensionality and data structure. Dimensionality refers to the number of distinct dimensions or factors that are present in a set of test items. These dimensions are unique constructs, which might be related to one another, but that are distinct from one another in some meaningful way. Dimensionality can be assessed both substantively and statistically. In an ideal situation, the results of both assessments closely align; however, this is not always the case.

As with many other areas of psychometrics, assessing the dimensionality of a test is not always straightforward. Pure unidimensionality is rarely observed in real tests with more than one item. Strictly speaking, unidimensionality is "the existence of one latent trait underlying the data" (Hattie, 1985, p. 139). However, unidimensional and multidimensional are not mutually exclusive categories. A test may have essential unidimensionality, where one factor dominates a factor structure that contains weaker or theoretically unimportant additional factors (Stout, 1987).

This weaker, less restrictive definition allows tests to meet the unidimensionality and local independence requirements of IRT in practice. Stout (1987) defined essential unidimensionality statistically as a condition that is met such that there is one trait θ where the conditions of essential local independence are met (covariance among items approaches 0). In simple terms, essential unidimensionality occurs when there is one strong or dominant dimension and one or more weak or inconsequential dimensions that do not interfere with the measurement of the primary dimension (Stout, 1987, 1990). It could be argued, then, that multiple dimensions on a test, however well constructed, might be a natural consequence of a test containing complex subject matter or even simple subject matter in an applied setting. For example, the context used in passages on a reading comprehension test (e.g., sports, history, popular culture) might lead to a secondary dimension on a test if more than one test item is assigned to each passage. However, the test would be essentially unidimensional if the passage context (the secondary dimension) had a very small impact on the probability of examinees answering the question correctly.

Essential unidimensionality might also be met when the correlations between dimensions is very high, such that the lesser dimensions do not interfere with the measurement of the dominant dimension (Stout, 1990). For example, reading comprehension and vocabulary skills might be very highly related, so measurement of one dimension (assuming that the constructs align with the dimensions) does not hinder measurement of the other.

Stout's T is the test statistic that is employed to test for essential unidimensionality (Nandakumar & Stout, 1993). To use this statistic, the test items are split into three subtests. A short assessment test of homogenous items (i.e., assumed to be unidimensional) is designated AT1. These items are selected using either factor analysis or expert opinion, but does not include *all* of the items that measure that dimension. A second assessment set with an equal number of items is chosen to match the item difficulty distribution (but not necessarily dimension) with AT1, and is designated AT2. A partitioning test (PT) is also created that contains all remaining test items. Examinee scores on PT are used to partition the examinees into *K* groups based on the raw score on the PT items. A *T*-statistic is calculated for both AT1 and AT2, based on the difference between estimated examinee and item *p*-value variance for each group *k*. Stout's *T* is based on the assumption that if the AT1 and PT are measuring the same dimension, then the *T*-statistic for AT1 will be small because the groups formed based on PT test scores will have small score variability within groups on the AT1 test (the AT2 test statistic is used to correct for test length and difficulty differences between AT1 and the total test). Thus, Stout's *T* is only significant (i.e., the null hypothesis, number of dimensions = 1 is rejected) exceeds a critical value. The final *T* value calculated:

$$T = (T_1 - T_2 / \sqrt{2}),$$

where T_1 is the *T* statistic for AT1 and T_2 is the *T* statistic for AT2. A significant T-value does not specify how many dimensions are present on a test, only that more than one dimension is present. The program DIMTEST (Nandakumar & Stout, 1993) operationalizes Stout's procedure.

While Stout's definition has been influential in dimensionality research (e.g., Bolt, 1999; DeChamplain, 1996; Meara, Robin, & Sireci, 2000; Nandakumar, 1991, 1993), other methods of assessing dimensionality exist. Another method is based on the analysis of the residual covariance matrix after fitting a non-linear factor analysis model to an item response matrix. The number of factors for the analysis is set to the number of factors hypothesized to underlie the item response matrix. This procedure is employed in the computer program NOHARM (Fraser & McDonald, 1988).

If most tests display at least some degree of multidimensionality and there are potentially detrimental consequences to violating the assumptions of IRT models, then it is important to examine test dimensionality before choosing an IRT model. However, assessing the dimensionality of a test or set of items is one of the most researched but least understood areas of IRT. Traditionally, the focus of dimensionality assessment has been on determining if a test is unidimensional (and thus whether unidimensional models and techniques can be used), rather than specifically trying to identify how many dimensions are present.

One of the areas of uncertainty in assessing dimensionality is the judgment that must be made in the process. Using more than one method of assessing dimensionality might yield different dimensional structures. For example, in a simulation study Finch (2002) found that NOHARM tended to be more conservative (i.e., retaining a null hypothesis of unidimensionality) than DIMTEST when the two-parameter logistic IRT model was fit to multidimensional data.

It is not always clear how many dimensions underlie a set of response vectors in a data matrix, nor is it necessarily any clearer once the dimensionality of the data has been assessed through statistical means. Therefore, more than one IRT model can be fit separately to any given data set. Once parameters have been estimated, it is useful to examine the goodness-of-model fit to the data. One approach to comparing IRT models is to assess their accuracy in reproducing the item response patterns of the original data (Embretson & Reise, 2000, p. 243). However, this type of assessment of model-data fit is still somewhat subjective, particularly in a case where it is not so important to determine which model is the best fit to the data (i.e., unidimensional versus multidimensional), but rather to determine if a particular model fits the data *adequately*.

If it is difficult to determine whether a test is unidimensional or multidimensional (Hattie, 1985), it is understandable why it might seem reasonable to apply unidimensional IRT techniques to data that may be deemed multidimensional. Unidimensional techniques are much simpler and easier to implement than multidimensional techniques because they require fewer subjects to achieve stable parameter estimates.

The correlation between dimensions in the data also affects dimensionality. If the correlation between dimensions is high, a dimensionality assessment might determine only one dimension is present because the multiple dimensions present are so closely related. If multiple dimensions are perfectly correlated (i.e., 1.0), then dimensions cannot be distinguished from each other, and the test is unidimensional. On the other hand, if the correlation between dimensions is low, the dimensions are easier to distinguish from one another, and the test may be judged to be multidimensional.

A related concept to dimensionality is data structure, which refers to the way in which the dimensions are measured by the test items. There are two types of multidimensional structures. Simple structure occurs when multiple dimensions arise because items on the test do not all measure the same dimension, but no items measure more than one dimension. Complex structure occurs when some or all items measure more than one identifiable dimension.

*Two-parameter logistic IRT models.*Once a decision about the dimensionality of the data has been made, an appropriate model must be selected. The two-parameter logistic (2-PL) model is relevant to a discussion of equating because both the *a*- and *b*-parameters are affected by scale transformations of IRT equating. These models are used on dichotomously scored items (i.e., correct/incorrect). By convention, these items are scored as 1 = correct, 0 = incorrect. The 2-PL model is defined by the equation:

$$p_{ij}(x_{ij} = 1 \mid \theta_j; a_i, b_i) = \left[\frac{e^{Da_i(\theta_j - b_i)}}{1 + e^{Da_i(\theta_j - b_i)}} \right],$$

where $p_{ij}(\theta)$ is the probability that examinee j with ability θ_j will answer an item i correctly, x_{ij} is the response for item i by examinee j (1= correct, 0= incorrect), a_i is the discrimination parameter for item i, indicating where on the θ scale the item will be maximally discriminating, b_i is the difficulty parameter (expressed in the metric of θ) of item i, and D is the constant 1.7, which scales the logistic function close to the normal ogive function. $p_{ij}(\theta)$ is in the closed interval [0,1], given that it is a probabilistic value. A higher $p_{ij}(\theta)$ indicates the examinee is more likely to answer the item correctly, but does not guarantee a correct response (nor does a low *p*-value guarantee an incorrect response) . The ability or theta (θ) distribution is usually set, by convention, to the unit normal distribution $(\mu = 0, \sigma = 1)$ with a lower bound of -∞, and upper bound of + ∞. This convention is somewhat arbitrary (which will be of importance in the discussion of scale transformation later in this chapter). The *a*-parameter, as a measure of the items' discrimination is bound by -∞ and +∞, though items with negative discrimination are undesirable because the probability of answering the item correctly decreases with increasing examinee ability (Hambleton et al., 1991).

The 2-PL model is also relevant to research examining multidimensionality underlying data because it has a multidimensional analogue, the M2PL model (Reckase, 1979). The M2PL model (shown here for two dimensions) is specified by:

$$p_{ij}(x_{ij} = 1 | \boldsymbol{\theta}_j; \mathbf{a}_i, d_i) = \frac{e^{(\mathbf{a}_i'\boldsymbol{\theta}_j + d_i)}}{1 + e^{(\mathbf{a}_i'\boldsymbol{\theta}_j + d_i)}},$$

where x_{ij} is the response for examinee *j* on item *i*, $\boldsymbol{\theta}_j$ is the vector of order (1,2) containing the ability estimates for examinee *j* on both dimensions 1 and 2, $\mathbf{a}_i$ is the vector of order (1,2) containing discrimination parameters a_1 and a_2 for item *i*, and d_i is a composite difficulty parameter for item *i*. Note how the 2-PL model can also be considered a special case of the M2PL model where the vectors $\boldsymbol{\theta}_j$ and $\mathbf{a}_i$ contain only one θ and a value, respectively and are therefore scalars, and d_i is analogous to b_i.

IRT parameter estimation. Item and examinee parameter estimation, as mentioned earlier, can be conducted simultaneously using maximum marginal likelihood estimation (MMLE; Bock & Aitkin, 1981). For the purposes of the proposed research, it is relevant to discuss MMLE because this is the method of estimation used by the program BILOG (Mislevy & Bock, 1990), which was used in the present study.

To estimate parameters, maximum likelihood estimation methods must estimate either the item or ability parameters first, then use the estimated value(s) to find the other parameter(s). MMLE first estimates the item parameters, then the ability parameters. MMLE is based on the assumption that all examinee responses to a given item are independent of each other in addition to local independence of an individual examinee's responses to each item. Mathematically, this idea is represented by the formula:

$$p(\mathbf{x} | \theta) = \prod_{j=1}^{n} [p_j(\theta)]^{x_j} [1 - p_j(\theta)]^{1-x_j},$$

where **x** is a vector of order $(1, j)$ of dichotomous outcomes for each examinee on that item (correct = 1, incorrect = 0) (Mislevy & Bock, 1990, p1-6). By integrating over θ, the probability of a response vector becomes a function of only the item parameter estimates. Parameter estimates of a and b are found such that the likelihood of the examinee responses to a given item is maximized. In other words, MMLE jointly estimates parameters a and b that would explain the pattern of correct and incorrect examinee response to items that was observed. Once the item parameters are estimated, they are used to estimate examinee ability values. MMLE for estimating item and parameters requires large sample sizes in order to be able to produce stable parameter estimates. Large sample sizes are also important so that the prior approximate ability distribution fits the data (Hambleton et al., 1991).

The central basis for likelihood estimation of a response vector is local independence. If local independence is violated, then errors in parameter estimation can occur. Research in this area has pointed to errors in over estimating both difficulty and discrimination parameters (Ackerman, 1987, 1991; Yen 1993). If estimation errors occur because of the presence of a second dimension, it seems reasonable that test forms with lower correlations between the first and second dimensions would have greater parameter estimation errors associated with each test form than test form with higher correlations. This parameter estimation error in turn would make equating these test forms more prone to equating error than test forms with higher correlations between dimensions. Ackerman (1991) described this error as "filtering" (p.23) of the multidimensional characteristics of the test items. It has been demonstrated previously that higher correlations between dimensions reduce the error associated with using a unidimensional model on multidimensional items (Ackerman, 1987, 1989). The direct link of parameter estimation to IRT equating also means that error that occurs in the parameter estimation step would be expected to also contribute to equating error.

Stocking and Lord scale transformation. In IRT equating, when item parameters are estimated for each test separately, the items are set on a θ scale that is relative to the ability levels of the examinees that took each test. If two groups are non-equivalent, the item parameters that are estimated for each test form will be on discordant scales. This artifact is a result of the general convention of setting the mean ability to zero and the standard deviation to one for each sample when estimating the IRT item parameters. For example, items will appear more difficult on a test form that was administered to a group of examinees with lower mean ability than if they were administered to a group of examinees with higher mean ability. Therefore, it is necessary to first scale item parameters from each test form on to a common metric before the test scores can be equated. In the common-items design, the common-items set is used as a bridge between scales. The common items make it possible to separate differences in ability between groups from differences in item difficulty between test forms. For the purposes of this proposal, the discussion of equating will be framed in equating form Y to form X.

In the common-items design, form X and form Y are associated with scales P and Q, respectively. Scales P and Q are assumed to differ by a linear transformation. The differences in common-item parameter estimates between test forms are used to calculate scaling coefficients that align the remaining items on both test forms. The A transformation coefficient adjusts the slope of the θ scale, while the B coefficient changes the y-intercept. The formula for A is:

$$A = \frac{a_{Qi}}{a_{Pi}},$$

while the formula for B is:

$$B = b_{P_i} - Ab_{Qi}$$

(from Kolen & Brennan, 2004, pp. 162-163). By rearrangement, the formulas are:

$$a_{P_i} = \frac{a_{Qi}}{A}, \text{ and } b_{P_i} = Ab_{Q_i} + B$$

are used to transform each a and b parameter for item i from scale Q (Form Y) to scale P *(Form X*).

The Stocking and Lord procedure (Stocking & Lord, 1983) uses an iterative approach to find A and B transformation constants that minimizes the function:

$$SLcrit = \sum_j SLdiff(\theta_j),$$

where:

$$SLdiff(\theta_j) = [\sum_{i:V} p_{ji}(\theta_{Pj}; \hat{a}_{P_i}, \hat{b}_{P_i}) - \sum_{i:V} p_{ji}(\theta_{Pj}; \frac{\hat{a}_{Qi}}{A}, A\hat{b}_{Qi} + B)]^2.$$

In this formula, *SLdiff* is the squared difference between the test characteristic curves of the common items set V in each test form, and $\hat{a}_{Q_i}$, $\hat{b}_{Qi}$, $\hat{a}_{P_i}$, and $\hat{b}_{P_i}$ are the a and b parameter estimates for scales Q and P, respectively. Once the calibration coefficients have been calculated, they are used to adjust all test items onto the common metric. Based on usage in the equating literature, the Stocking and Lord procedure is a commonly used scale transformation (e.g., Bolt, 1999; Camilli, Wang, & Fesq, 1995; De Champlain, 1996; Dorans & Kingston, 1985; Stocking & Eignor, 1986; Thomasson,

1993) and has been found to produce empirically the most accurate transformation results (Baker & Al-Karni, 1991).

Scale transformation methods are susceptible to error from using a model under conditions where the required assumptions are violated. As indicated earlier, local dependence caused by multidimensionality can lead to errors in estimating the *a*- and *b*-parameters (Ackerman, 1987; Yen 1993). These parameters are both used in the Stocking and Lord calculation of *A* and *B*, and are also the parameters that are transformed onto the common scale. It seems plausible that the location of the multidimensional items might be a factor in how much error is introduced. One might expect to see more error if the common items contain a second dimension because they are directly involved in the calculation of the *A* and *B* scaling coefficients.

IRT true score equating. With the test form parameter estimates on a common scale, it is now possible to equate the test scores. IRT true-score equating is based on the assumption that for every ability θ, there is a corresponding true score, τ, on test X and test Y (τ_x and τ_y, respectively). τ_x and τ_y are the summed probabilities of answering each item correctly, given the parameter estimates of the item on the test and the ability of the examinee. τ in this case is distinct from the classically-based τ (i.e., $\tau = x + \varepsilon$), because of the link to θ. The basis for IRT true-score equating is the assumption that the true scores $\tau_x(\theta_j)$ and $\tau_y(\theta_j)$ are equivalent, given θ_j. The underlying θ acts as a link between test X and test Y that allows equating to occur.

The form X equivalent of a form Y score is (from Kolen & Brennan, 2004, p. 176):

$$irt_x(\tau_y) = \tau_x(\tau_y^{-1}), 0 < \tau_y < K_y \ ,$$

where the τ_Y^{-1} is θ_j corresponding to τ_Y, and K_Y is the maximum total test score. Finding θ_j for each τ_Y^{-1} requires using the Newton-Raphson procedure. The general form of the Newton –Raphson procedure is:

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

where θ^{-} is a starting value, θ^{+} is a new value that becomes θ^{-} on the next iteration, $f(\theta)$ is a function of θ, and $f'(\theta)$ is the first derivative of $f(\theta)$. This algorithm is iterated until the difference between θ^{+} and θ^{-} meets a pre-specified stopping minimum value.

In IRT equating, the Newton-Raphson formula is specified as:

$$f(\theta_j) = \tau_y - \sum_{i:Y} p_{ij}(\theta_j, a_i, b_i) \text{ and,}$$

$$f'(\theta_j) = -\sum_{i:Y} p'_{ij}(\theta_j, a_i, b_i),$$

where Y is the set of all i items on form Y , and:

$$p'_{ij}(\theta_j, a_i, b_i) = 1.7 a_i p_{ij}(1 - p_{ij})$$

(adapted from Kolen & Brennan, 2004, p. 177).

Once the θ_j is found, it is substituted into the 2-PL model along with the form X item parameters to find $p_{ij}(\theta)$ for each item on form X, which are then summed to calculate τ_X. Because this procedure must be repeated for each τ_Y, with several iterations for each τ_Y, true-score equating is computationally intensive.

True-score equating is used often in research (e.g., Bolt, 1999; Camilli, Wang, & Fesq, 1995; De Champlain, 1996; Dorans & Kingston, 1985; Stocking & Eignor, 1986) possibly because the equating function is not score-distribution dependent. The main disadvantage of this method is that it is based on the true score, which cannot be directly found.

IRT scale transformation and equating methods are based on the same models and, thus, the same assumptions of the IRT parameter estimation models. The Stocking and Lord (1983) procedure and IRT true-score equating are often observed in the equating literature (e.g., Bolt, 1999; Camilli, Wang, & Fesq, 1995; De Champlain, 1996; Dorans & Kingston, 1985) making these methods appealing for further exploration.

Criteria for Assessing the Adequacy of Equating

Whenever equating is conducted, it is useful to be able to assess how effective it was in aligning test scores and meeting one or more of the requirements of equating: symmetry, invariance, and equity. While having an evaluative procedure seems a logical step, no guidelines exist for how to do so. Harris and Crouse (1993) conducted a review of the types of criteria that have been used to evaluate equating. They identifed four types that are relevant to the present research study: 1) simulated data, 2) indices, 3) circular equating, and 4) practical criteria.

Simulated data. Simulation studies are often used to assess the robustness of equating (e.g., Bolt, 1999; Jodoin & Davey, 2003; Skaggs & Lissitz, 1986). The criterion in this case is whatever is specified by the researcher when the data are generated. Unlike a study that uses real data, which only has estimated parameters, a simulation study has pre-specified true item (e.g., difficulty), test (e.g., correlation between dimensions), and population (e.g., mean ability) parameters, which can be used as criteria for assessing the accuracy of equating results. For example, using a simulation design, Bolt (1999) used expected scores on the target test as the criterion to which expected equated scores were compared for each examinee. Expected true scores in IRT equating can be created for each examinee by summing the probabilities of answering each item correctly that are computed when the known item and examinee parameters are used in the chosen IRT model.

Indices. Indices are a common type of criteria that are employed to attempt to quantify the magnitude of the equating error. Indices can be used to measure error both globally (i.e., one index value summarizes equating performance across the entire score scale) or locally (i.e., individual index values are calculated at many points along the score scale). Measures of differences (e.g., Root Mean Squared Error (RMSE) and Mean Absolute Difference (MAD)) are commonly used types of equating criteria (e.g., Bolt, 1999; Hwong, Im, Si, Seong, & Kim, 2005; Klein & Jarjoura, 1985; Ricker & Von Davier, 2006). Typically, these indices are used in studies where many methods or sets of conditions are being compared to one another. In other words, the interpretation of indices is generally norm-referenced.

Circular equating. Circular equating refers to equating scores on a test form to itself, either directly or indirectly. Equating a test form to itself can be used as a means for evaluating the magnitude of the equating error that is generated. For example if scores from a single group examinees on one form were equated to themselves, the resulting conversion would be an identity function (Harris & Crouse, 1993). Using this design as the baseline, it is possible to evaluate what proportion of observed equating error is computational error for the actual equating of Form Y to Form X (Wang, Hanson & Harris, 2000).

Practical criteria. Another way of assessing equating results is by using a practical definition of "good" equating and using it as a "yardstick" against which the goodness of equating results can be measured. One practical criterion is the definition of a "score difference that matters" developed by Dorans and Feigenbaum (1994). They argued that any score difference that was detectable was meaningful from an equity perspective. A score difference that "matters" is a difference in scores between an equated score and a criterion score that is greater that one half of a scale score reporting interval. For example, if a scale is on a 10-point reporting interval, then a difference of 5 points or more would be detectable once scores were rounded. Any score difference that was smaller than a score difference that matters would be inconsequential to the examinees and therefore not of concern for protecting equity.

A second type of practical criterion that can be used is to choose a specific type of equating data collection design and method as a "gold standard" to which other methods can be compared. For example, from a theoretical perspective, equating using an equivalent groups design is less prone to equating error than equating using a common-items nonequivalent groups design (Kolen & Brennan, 2004, p. 29). This type of criterion has been employed in evaluating the equating results of multiple types of equating methods under various equating conditions (von Davier et al., 2005; Ricker & von Davier, 2006).

Many other types of equating criteria exist. In addition to those discussed here, Harris and Crouse (1993) identified five additional types, but no definitive criterion class emerged as the best one to use. It was suggested that no one existing type of criterion is universally appropriate, but rather that the best criterion would depend on the purposes of the equating and the evaluation. For example, when comparing equating methods, Tong & Kolen (2005) found that the methods that performed best against one type of equity criterion would often perform poorest against another type of equity criterion. No criterion has to be used as a single evaluative measure. Instead, they can be used in conjunction with one another, to develop a more complete picture of the efficacy of equating.

Summary

Test equating is a statistical solution to the problems that arise from having multiple test forms in large-scale testing programs. There are desirable properties of test equating, in particular equity. These properties ensure that equating solves the issues surrounding scoring multiple test forms rather than creating more test scoring problems.

The data collection design is as critical to the equating process as the equating procedures themselves. The common-items nonequivalent groups design is used to overcome practical data collection issues, while still ensuring a link between test forms by means of the common items. However, the common-items nonequivalent groups design introduces its own set of issues about how the location of the multidimensional items on the test might affect equating outcomes

IRT equating is a popular method of equating large-scale assessments. It is a powerful type of equating must requires that several key assumptions be met. Despite the apparent simplicity of the unidimensionality assumption, it is not always clear when it has been met. Therefore, it is important to explore the conditions under which the violations of IRT equating assumptions might have a deleterious effect on outcomes that use IRT equating methods and to see when and where the violation of assumptions start to matter.

To evaluate the robustness of an equating method, a criterion or criteria are needed. Several types of criteria exist, and can be used together to help interpret the equating results. The type of criteria that should be used will depend on the type of question that is to be answered.

Review of Literature Specific to the Effects of Multidimensionality on IRT Equating

A small literature exists that examines specifically the robustness of IRT equating to violations of the unidimensionality assumption. The literature presented in this chapter represents what is currently known about this topic, as well as pointing to what is not known or what variables have not been examined.

One of the most important factors to consider when reviewing past literature in this area is the definition or conception of multidimensionality that was used in each study. Unidimensionality is a central tenet of IRT equating. The definition of unidimensionality and, by default, multidimensionality, and its potential consequences to IRT equating are central to understanding how each study was designed, which dependent variables were examined, and what results were obtained.

The following summary of the literature is specific to testing the robustness of unidimensional IRT equating using multidimensional data and includes the following for each study: 1) the authors' definition of multidimensionality; 2) data source, number of items used, and whether the data were real or simulated; 3) the correlation between dimensions; 4) a brief description of the method, including the equating design, the independent factors manipulated, and the scaling and equating techniques used; 5) key results; and 6) a summary of the importance or significance of the study results. The studies are presented in chronological order, as each researcher thus far has attempted to build on the research that has been conducted previously. A discussion of the emerging themes from this literature follows the review of the individual studies.

Dorans and Kingston (1985)

One of the earliest published studies that examined the effects of multidimensionality on unidimensional IRT equating was conducted by Dorans and Kingston (1985). This study was conducted prior to the work of Stout (1987, 1990) who defined essential unidimensionality operationally. They defined unidimensionality of a test as being able to model examinee performance using only one θ. Their approach to operationalizing multidimensionality was to use two substantively different subscales as part of the test forms. Dorans and Kingston (1985) reasoned that if multidimensionality affected the results of unidimensional IRT equating, then differences would be observed between calibrating the dimensions concurrently or separately.

Two highly correlated subscales of real items (discrete verbal and reading comprehension items) from four separate forms of the Graduate Readiness Examination (GRE, administration years not reported) were employed. The forms contained between 53 and 55 operational and 47 to 55 non-operational discrete verbal items and between 22 and 25 operational and 20 and 25 non-operational reading comprehension items. The correlation between subscales on each form was high (0.73-0.80). All other forms were equated to one form through either a random groups design or a common-items nonequivalent groups approach. Different sets of common items were used for each scale calibration, depending on the form pair being equated. Scale calibration for the anchor test (common-items nonequivalent groups design) was conducted using Stocking and Lord's (1983) method, using item parameter estimates calculated in three different ways: (a) one single calibration of each test form for all test items (i.e., both subscales together), (b) separate calibrations of each test form for each subscale, and (c) a single calibration of the reference form and separate calibrations for each subscale on the equated form. All test forms were equated using true-score equating.

Dorans and Kingston (1985) found that while equating results were somewhat different depending on the calibration procedure used, IRT equating might be robust to a violation of unidimensionality, at least in a case where the correlation between the dimensions was high. This result was both unexpected and significant based on previous theory that IRT models would fail if the test contained more than one dimension.

High correlations between the two subscales used in this study (0.73 to 0.80) might have resulted in more robust equating results than with more distinct (i.e., less highly correlated) dimensions because the test was essentially unidimensional.

Stocking and Eignor (1986)

Stocking and Eignor (1986) defined unidimensionality indirectly via their operationalization of multidimensionality in modeling examinee performance on different items. They used two sets of θ parameters to model examinee responses, one for each dimension. By inference, unidimensionality existed when one θ only was required to explain examinee performance. Like Dorans and Kingston (1985), they used substantively distinct dimensions in their research.

Pre-equating is an equating design that is used commonly in large-scale testing programs, including the SAT, which was the data source for this study. Examinees are given a set (usually a test section) of items in addition to the operational items on the test. The additional test section will be used as a real test section in a subsequent administration of the test. By administering the test in sections to a previous test group, the new test can be equated to old test forms prior to its actual administration, in a manner similar to the common-items design. It was observed that section pre-equating of the SAT was producing unexpected equating errors. Stocking and Eignor (1986) evaluated multidimensionality as a source of the equating error.

Multidimensionality, they argued, was a potential source of error because it violated the assumptions of the equating model. It was reasoned that multidimensionality might be arising on the pre-equated section because the students knew the pre-equated section did not count in their own test score and thus were less motivated during this section. In this case, motivation becomes a secondary construct (and also dimension) for the items in this section.

Groups of responses were simulated from real data for each of three testing conditions. In condition 1, 60 SAT items were used. In conditions 2 and 3, 30 non-overlapping subsets of the SAT items used in condition 1 were employed. All test forms also included the pre-equating section as a set of 24 common items. In conditions 1 and 2, one ability estimate was used for all item responses. In condition 3, one ability estimate was used to simulate responses for the first 30 items, then a second lower ability estimate was used to simulate responses to the last 24 pre-equating items (simulating reduced motivation among examinees on the pre-equated section). No correlation between motivation and SAT achievement was reported. The authors did not indicate which linear scale transformation technique was used, but the resulting transformed data was equated using true-score equating.

The results of the simulated conditions indicated a lack of model fit when a second dimension was introduced (motivation) when compared to conditions when only one dimension was present (Condition 3 versus Conditions 1 and 2). This result was in the same direction as results observed in real equating situations, and was much larger in magnitude than under simulated conditions.

The findings of Stocking and Eignor are significant because they separated the effects of multiple dimensions on a test from differences between groups in mean ability on a single attribute. Ostensibly, they partitioned the effects of impact from test bias on equating results, highlighting an important reason to pursue this line of research.

Camilli, Wang, and Fesq (1995)

Camilli, Wang, and Fesq (1995) were the first researchers to use the idea of essential unidimensionality as defined by Stout (1987) in studying IRT equating robustness to multidimensionality. However, they further defined the dimensionality of a test as a validation argument (Messick, 1989) related to test content. As such, they also worked from a substantive rather than a purely statistical approach to select items to represent each dimension in their test forms. However, they did use statistical techniques to confirm their judgments.

Camilli et al. (1995) used six sets of multidimensional data from the Law School Admission Test (LSAT) forms from 1989 and 1990. Each form included 94 and 98 items, respectively. The data were collected using a section pre-equating design. The multidimensionality of the test forms was confirmed using factor analysis. On each test form, the correlation between the two dominant dimensions ranged between 0.68 and 0.75. They used the same methods as Dorans and Kingston (1985), conducting separate and concurrent calibrations of the distinct item subsets. They compared the true scores of the separate and concurrent calibrations to see if differences occurred.

Similar to Dorans and Kingston (1985), Camilli et al. (1995) showed that score differences were small under multidimensional conditions, with the biggest differences occurring at the low and high ends of the performance scale. They concluded that for the purposes of equating, the LSAT might be essentially unidimensional. They also suggested that the robust results on this test might be due to good test construction, which ensures that tests are very similar in terms of both content and statistics from administration to administration. This result suggests that the degree of statistical matching between tests might be an important factor to explore.

DeChamplain (1996)

While never explicitly defined, De Champlain (1996) utilized essential unidimensionality in his research because both NOHARM and DIMTEST were used to assess dimensionality and because both procedures employ this concept. De Champlain (1996) examined the underlying multidimensional structure of real LSAT data to see if it exerted differential effects on the equating scores of different subpopulations of examinees. De Champlain argued that if the population of examinees was treated homogenously without accounting for differences in all salient dimensions, then some groups of examinees might be unfairly advantaged or disadvantaged by using unidimensional equating to get test scores.

A section pre-equating design was used to collect the data for two LSAT test forms (years not reported), containing 101 and 102 items each. Both NOHARM and DIMTEST tests rejected the null hypothesis of unidimensionality. The two-dimensional structure of each test was confirmed using NOHARM (Fraser & McDonald, 1988). The correlation between dimensions on each test was not reported. De Champlain randomly selected samples to create three equally sized populations of Caucasian, African American, and Hispanic examinees. Separate unidimensional and multidimensional models were fit to the data of each population and scale calibrations were conducted for each group using the Stocking and Lord procedure with scores equated using true-score equating. These scores were compared to the equated scores that were derived from calibrating and equating the population as a heterogeneous whole. In addition, the fit of the different dimensional models to item responses for each group was assessed using both NOHARM and DIMTEST.

The results of the study indicated that a two-dimensional test fit well for the Caucasian and African American samples, but not for the Hispanic sample. Despite differences in the adequacy of model fit for each group, equating resulted in small score differences for each group, with most of the differences occurring in the low tail of raw scores. De Champlain concluded that having one equating function does not penalize subgroups of examinees, even though models were not the same for each subgroup.

These results provide further evidence that the IRT equating, when employing the Stocking and Lord (1983) and IRT true-score equating procedures, is robust to violations of unidimensionality. However, this research is also based on real data. Consequently, the robustness of the results might also be due to the correlation between the dimensions. To this point in the literature review, no systematic examination of factors affecting unidimensional IRT equating under conditions of multidimensionality had been conducted.

Bolt (1999)

Bolt (1999) defined unidimensionality implicitly as essential unidimensionality by the use of NOHARM to confirm the dimensional structure of the LSAT data. Bolt (1999) conducted two studies. In the Study 1, Bolt compared the performance of IRT true equating to linear and equipercentile methods under a section pre-equating design. Forty items selected from each of the October 1992 and June 1993 LSAT forms were used to generate realistic data. Groups of response vectors for four groups were simulated.

Each section of the test was fit separately using the 2-PL model, then linked using Stocking and Lord's (1983) procedure for transformation, with the old form items as the anchor (or common) items. The forms were equated using true-score, equipercentile, and linear methods.

On measures of equity, IRT equating performed the best overall. However, as Bolt (1999) pointed out, in some areas of ability, particularly among examinees with either low ability on both dimensions or high ability on both dimensions, unidimensional IRT equating produced errors that indicated that while it performed the best of all of the methods under multidimensional conditions, it still might not be appropriate for some test score inferences.

In Study 2, using the same section pre-equating design and Stocking and Lord scaling and true-score equating, Bolt (1999) created two 50-item, two-dimensional tests, where the only difference between the items on the two forms was that the difficulty on the second dimension was designated to be easier on the Y form items, resulting in the *d*-parameter that was systematically higher on the Y form items than the X form items. The correlation between dimensions varied at 0.3, 0.5, 0.7 and 1.0 across the conditions.

IRT methods performed about as well as other methods under multidimensional conditions when the correlation between dimensions was 0.7 or higher, but poorer (though only slightly so than equipercentile) when the correlation between dimensions was less than 0.5. Interestingly, the relative contribution of multidimensionality to poor performance was minor relative to the equity that was present even when the correlation between dimensions was 1.0. However, Bolt suggested that differences in difficulty that are specific to one dimension (i.e., the tests are well matched on one dimension, but not on the other dimension) might be a condition where unidimensional equating would break down. What was not clear from the results was whether the equity was adequate under the various conditions. Bolt suggested that decisions about the adequacy of equity would depend on whether the practitioner requires equity at all combinations of levels of ability (local equity) or just good equity for the majority of examinees (global equity).

A strength of this study is the systematic approach to examining equating under conditions of multidimensionality. When compared to previous studies, all of which employed real data, it is much easier to rule out other possible explanations or sources of error that might also explain the results. These results also confirm the idea that the correlation between dimensions can have an impact on the degree to which multidimensionality will affect IRT equating.

Jodoin and Davey (2003)

In this study, Stout's (1987) dimension of essential unidimensionality was employed, and by default, those tests that did not meet this definition were multidimensional.

Jodoin and Davey (2003) used items from two separate non-identified standardized tests to simulate data for these studies. The dimensional structure was confirmed to be essentially unidimensional for one test and multidimensional for the second. In the simulations, they examined the robustness of unidimensional IRT equating techniques in the common-items nonequivalent groups design by introducing potential equating errors in two ways. In condition 1, each test form met conditions of essential unidimensionality, but each test form in the equated pair measured a different substantive dimension (two test form pairs, with 29 and 48 items each, respectively, were created). Condition 2 used two sets of test pairs (30 and 40 items each) that were well matched but each contained two dimensions. The correlation between dimensions for the test pairs was not reported.

Item parameters for each test form were estimated under nine different conditions. The first data set was unidimensional. Eight sets of multidimensional parameters (with between two and 25 dimensions specified) were also created. The test forms were scale transformed using the mean/sigma method (Marco, 1977). No equating was conducted. Equating robustness was evaluated by examining the resulting A and B scale transformation coefficients. If a scale transformation method is robust under the experimental conditions, then the expected mean value of A is 1.0, and mean B is 0.0. These are the expected values because it is generally standard practice to calibrate ability estimates on a N(0, 1) metric. The results of the simulated conditions revealed that the unidimensional IRT equating was robust to the tests measuring different dimensions, but not to conditions where the tests were well matched but contained more than on one dimension.

The main limitation of the Jodoin and Davey (2003) study is that equating robustness was a secondary consideration to other research questions and thus they failed to systematically investigate the effects of multidimensionality on IRT equating. As a result, while the parameters used to generate the data were realistic, the test specifications were not. The items that were chosen as the common items were selected for their dimensional structure, but were not assessed for how well these common items were representative of the unique items on each test form.

This study introduced the idea that the location of the multidimensional items may have an impact on equating results, but only examined the condition of multidimensionality occurring in the unique items. Location as an independent variable has not been systematically tested to date.

Emerging Themes from the Literature

The research examining unidimensional IRT equating under conditions of multidimensionality has varied little in context (including the tests used for investigation), equating designs, and equating techniques. Based on its confirmed two-dimensional structure (e.g., Douglas, Kim, & Roussos, 1999), the LSAT has been a popular choice for studying multidimensionality and its effects on equating (Bolt, 1999; Camilli et al., 1995; De Champlain, 1996). The section pre-equating design and the common-items nonequivalent groups designs are similar designs used in this research.

The popularity of the Stocking and Lord procedure and IRT true-score equating make them the most logical choices for further research, particularly for comparability to past work.

Based on this body of research, it would seem that violating the assumption of unidimensionality does not cause large or significant errors in equating, provided that the correlation between dimensions is high (i.e., 0.7 or higher). Bolt's (1999) systematic study on the strength of the correlation between the dimensions gives greater credence to the results of previous research where this variable was not controlled. Given that this variable is the only one that was the most commonly cited, it is logical to include this variable in future research, particularly in a simulation study where it and other variables could be varied systematically.

Another potential variable of interest relates to what happens when multidimensional data is modeled as unidimensional. Bolt (1999) and Camilli et al. (1995) refer to Wang's (1986) idea of a reference composite, whereby multiple dimensions are represented as a composite in one dimension. This idea suggests that when multiple dimensions are present, it is not just the most dominant dimension that is modeled in the unidimensional model, but a representation of all dimensions that are present. If all dimensions are present in the single dimension that is modeled, then it might be expected that a mismatch in difficulty on a dimension between test forms could produce greater equating error than a condition where multiple dimensions are present, but the tests match in difficulty on all dimensions. This idea has not been explored in the research to date.

An interesting question also arises from the work of Jodoin and Davey (2003). While Jodoin and Davey examined the common-items design, they only evaluated the effects of multidimensionality occurring among the unique items. Are there differences in equating outcomes when the multidimensional items occur among the common items, unique items, or in both sets of items? Given the importance of matching the characteristics of the common and unique items under unidimensional conditions, it might be expected that less equating error would be observed in the common-items nonequivalent groups design when multidimensionality was present in both sets of items (and thus the common and unique items match).

A major theme that emerges from this body of literature is the need for the use of simulation studies to explore the effects of different variables on IRT equating. With the exception of Dorans and Kingston (1985), all of the studies were conducted, at least in part, using a simulation design. However, most of these studies have focused on using realistic item parameter estimates from a particular test (e.g., Camilli, Wang, & Fesq, 1995; De Champlain, 1996; Stocking & Eignor, 1986), or the major focus of the study was some other variable or aspect of the study (Jodoin & Davey, 2003). As a result, most factors that have the potential to interact between multidimensionality and IRT equating have not been examined systematically.

The general focus of previous research has been assessing the appropriateness of IRT equating on particular sets of unidimensional data. In the present study, the focus was on systematically exploring when and under what conditions unidimensional IRT equating is still appropriate for multidimensional data. The approach used was similar to that of Bolt (1999). As such, the aim was to document where and under what conditions unidimensional IRT true score equating using a common-items nonequivalent groups design begins to break down.

Summary

Little research exists that explores systematically the effects of specific variables on the robustness of IRT equating using multidimensional data. Thus far, only the correlation between dimensions has been shown to modulate the effects of multidimensionality on equating error. Other variables, including the matching on each dimension between test forms, and the location of the multidimensional items in the common-items nonequivalent groups design have not yet been explored. These variables need to be examined methodically so that their effects can be better understood. In Chapter 3, the methods used to conduct a study to examine the effect of these variables on equating equity are described.

Chapter 3: Methods

Introduction

Chapter 3 is divided into three main sections: 1) procedures for the simulation study, 2) procedures for the equating of actual test data, and 3) a brief description of the specialized computer programs used to conduct both of the studies. The simulation design described in Section 1 was used so that group equivalence and correlation between dimensions could be tightly controlled. The real data analysis described in Section 2 was conducted to supplement the results of the simulation study. That is, if the results of the real data analysis corroborate the results obtained in the simulation study, then an argument can be made that while the simulated data is artificial, the simulated equating results are a reasonable representation of what could be observed in an actual testing scenario.

Section 1: Simulated Data

The present section describes the procedures required to conduct the simulation study. The section is organized into the following seven subsections: 1) research design, 2) independent variables, 3) description of the data source, as well as procedural information regarding, 4) test construction, 5) data generation, 6) data processing, and, 7) the dependent variables. Using this design, the independent variables, and the procedures described in these subsections, the following questions about IRT equating using the common-items nonequivalent groups design when two dimensions are present were addressed:

1) What is the baseline error that is associated with equating? That is, how much error would be present if scores on a test form were equated to scores on the same test form?

2) How does the correlation between dimensions affect both the magnitude of equating error and the proportion of examinees with error in their equated scores that is large enough to matter?

3) Is the magnitude of equating error (or the proportion of examinees with equating error that is large enough to matter in their equated scores) different if the groups are randomly equivalent versus nonequivalent?

4) Does the location of the items (unique, common, both unique and common locations) measuring the second dimension have an effect on the magnitude of equating error, or the proportion of examinees who are affected by equating error?

Research Design

As shown in Table 1, a 2 (form parallelism) x 4 (correlation between the dimensions) x 3 (group equivalence) x 3 (location of the items measuring the second dimension) factorial design was employed. This factorial design produced a total of 72 conditions.

Independent Variables

Form parallelism. This independent variable arose from a need to try to put the effects of the other independent variables into context. In operational situations, one test form is equated to another test form so that the scores are comparable across forms. However, it stands to reason that part of the error associated with equating under different conditions is related to the computational error associated with equating, and not just the conditions under which the equating occurs. In other words, even under perfect equating conditions, there will always be non-zero error. The question is, how large does equating error have to be to be considered meaningful?

Table 1

Experimental Design –Form Parallelism (2 Levels) x Correlation Between Dimensions (4 Levels) x Group Equivalence (3 levels) x Location of Dimension 2 Items (3 levels)

Form Parallelism	Correlation Between Dimensions	Group Equivalence[a]	Location of Dimension 2 Items[b]		
			BL	UI	CI
Parallel	Perfect (1.0)	EG (0.0)			
		NEGS (0.1)			
		NEGM (0.3)			
	High (0.7)	EG (0.0)			
		NEGS (0.1)			
		NEGM (0.3)			
	Low (0.3)	EG (0.0)			
		NEGS (0.1)			
		NEGM (0.3)			
	No (0.0)	EG (0.0)			
		NEGS (0.1)			
		NEGM (0.3)			
Nonparallel	Perfect (1.0)	EG (0.0)			
		NEGS (0.1)			
		NEGM (0.3)			
	High (0.7)	EG (0.0)			
		NEGS (0.1)			
		NEGM (0.3)			
	Low (0.3)	EG (0.0)			
		NEGS (0.1)			
		NEGM (0.3)			
	No (0.0)	EG (0.0)			
		NEGS (0.1)			
		NEGM (0.3)			

[a] EG = Equivalent Groups, NEGS = Nonequivalent Groups Small, NEGM = Nonequivalent Groups Moderate
[b] BL = both locations, UI = unique items only, CI = common items only

Two levels of this independent variable were employed. First, in an attempt to gain a sense of "baseline" equating error, equating was conducted to equate Group Q Form X scores to Group P Form X scores. Equating Form X to itself was identified as the

Parallel forms condition. The second level of the independent variable is the Nonparallel forms condition, where Group Q Form Y scores were equated to Group P Form X scores.

Correlation between dimensions. Bolt (1999) selected correlations of 0.3, 0.5, 0.7 and 1.0 to systematically explore the relationship of the correlation between dimensions and equating outcomes. He did not, however, include a condition where the dimensions were uncorrelated. Given that no large differences in performance were reported between the 0.3 and 0.5 conditions by Bolt (1999), correlations between the dimensions were set in the present study at 0.0, 0.3, 0.7, and 1.0 to represent No (N), Low (L), High (H) and Perfect (P) correlation conditions, respectively.

Group equivalence. Equating is intended to align test scores on different test forms for a given population. In cases where the common-items nonequivalent groups design is used, it is not assumed that the sample of examinees at a given administration are equivalent to each other, even though they are assumed to have been drawn from the same population. Further, there are limits to how effective equating methods are at dealing with large group differences. Kolen and Brennan (2004, p. 286) suggest that group ability differences of ±0.3 standard deviations or larger will introduce error into equating using any method, and also identify IRT equating methods as among the most susceptible to this source of error.

In the equivalent groups (EG) conditions, the groups taking each X (Group P) and Y form (Group Q) were specified to have bivariate normal distributions of ability, with mean ability equal to zero θ, standard deviation equal to one for each dimension, and the correlation between the dimensions for a given condition. For the nonequivalent groups (NEG) conditions, the group taking Form X (Group P) remained unchanged, while the group taking Form Y (Group Q) was specified to have mean ability 0.1θ, and a standard deviation of one for both dimensions in the NEG Small difference conditions; and mean ability 0.3θ, and a standard deviation of one for both dimensions in the NEG Moderate difference conditions. These three conditions, Equivalent Groups (EG), Nonequivalent Small (NEGS), and NEG Moderate (NEGM), were intended to represent points on a continuum between ideal conditions and the most extreme group differences that might reasonably be accepted as meeting the condition of group difference needed to conduct score equating.

Location of the items measuring the second dimension. In the present study, in the Both Locations (BL) condition, 16 Dimension two items were placed among the unique items and eight Dimension two items were placed in the common items (see Table 2). For the Unique Items (UI) conditions, 16 Dimension two items were placed among the unique items on the test form, while in the Common Items (CI) conditions, eight Dimension two items were placed among the common items. This variable is very important because if differences among the conditions exist, it points to test construction procedures as a means of controlling equating error.

Table 2

Test Specifications for Simulation Study

Location of Dimensionality	Unique Items		Common Items	
	Dimension One	Dimension Two	Dimension One	Dimension Two
Both Locations (BL)	36	16	8	8
Unique Items (UI)	36	16	16	0
Common Items (CI)	52	0	8	8

Description of the Data Source

Item parameter estimates from the 1992 administration of the Law School Admission Test (LSAT) were used as realistic item parameters to improve the generalizability of the simulation results (Harwell, Stone, Tsu, & Kirisci, 1996). The 2-PL compensatory multidimensional (M2PL) model (Reckase, 1985) was used to model the parameter estimates for test construction.The LSAT is a realistic data set that is used frequently in studying dimensionality (e.g., Bolt, 1999; Camilli et al., 1995; Douglas, Kim & Roussos, 1999; Walker, Gierl, Ackerman, Ricker, & Gosz, 2003). The test contains four separate subtests: logical reasoning (LR) 1 and 2, analytical reasoning (AR), and reading comprehension (RC), and contain 51, 24, and 27 items, respectively. In statistical/substantive assessments of dimensionality, the LSAT has been demonstrated to have two dominant dimensions, one composed of items from AR, and one from the combination of LR and RC (Stout et al., 1996). For the purposes of this research, only the statistical dimensionality of the test items was considered.

Test Construction

For the purposes of this study, parameter estimates for 103 LSAT items were taken from a previous research study (Walker et al., 2003). In the previous study, the parameters were estimated using NOHARM (Fraser & McDonald, 1988) in confirmatory mode with two dimensions specified. The items were fit to a 2-PL compensatory MIRT (M2PL) model (Reckase, 1985) using the computer program NOHARM (Fraser, 1988). Though the LSAT is typically modeled operationally as having a pseudo-guessing (or *c*-) parameter in addition to *a*- and *b*-parameters, a 2-parameter model was chosen for two main reasons. First, in IRT equating the *c*-parameter is set to be invariant and therefore would not change as a result of equating (Kolen & Brennan, 2004, p. 180). Second, models with more parameters require larger sample sizes for stable estimates, which would increase the computing demands of the analyses conducted in this study, making it less manageable.

Once the parameters were estimated, the items were categorized into Dimension one items, Dimension two items, and items that were not used for the purposes of this study. The categorization was based on an operational definition of an angular direction between 0 and 20 degrees, and 70 and 90 degrees for Dimensions one and two, respectively (Walker et al., 2003). This angular separation (that the mean angular direction of the items measuring Dimension one and items measuring Dimension two are distinct from each other) between the dimensions of this magnitude is considered to be approximate simple structure (Gierl, Leighton & Tan, 2006; Stout et al., 1996). The relationship between the angular separation between dimensions and the correlation expressed as a -1.0 to 1.0 value is that the cosine of the angular separation between dimensions is approximately equal to the correlation between the two dimensions (Leucht & Miller, 1992). In this case, if an angular separation of approximately 70 degrees (assuming an average angular direction of about 10 degrees for Dimension one and 80 degrees for Dimension two, the correlation between the dimensions is approximately 0.34.[3] Based on the parameter estimates from the LSAT items, an additional 103 realistic items were created for the present study with similar item parameters to the Dimension one items and Dimension two items. These extra items were necessary to create two test

[3] The correlation that is used for the purposes of the study is specified when the data are simulated. This correlation merely describes the original angular orientation of the items measuring each dimension as distinct from each other.

forms that met the test specifications in terms of dimensional content and statistical difficulty and discrimination.

For the present study, two test forms for each condition were created, Form X and Form Y, according to the following specifications. Each Form X and Y consisted of 68 items, chosen from the sets of items generated using the LSAT items. The 68 items were divided such that 52 items were unique (or non-common) to each form and 16 items were common to both test forms in the pair of forms that were equated (i.e., the common items for equating) A set of common items were specified by including a set of items with the identical mean a_1-, a_2-, and d-parameters on both forms. For the unique items, Form X and Y were constructed to have closely parallel mean a_1- and a_2- parameters, but with a 0.2θ mean difference in the d-parameter. This difference was intentionally built in so that there were form differences that necessitated score equating. A mean difference of 0.2θ was selected in order to be certain that the forms were not parallel. The common items were set to have nearly identical mean a_1- and a_2-parameter to the means of the unique items, with the mean d-parameter set to be 0.1θ from the mean d-parameters of the unique items of both forms. Table 2 previously presented the test form specifications for each condition (see p. 50), and Table 3 presents the means and standard deviations for the parameters for each pair of test forms. Appendixes A to C contain the specific item parameters for each test form.

Table 3

Mean a_1-, a_2-, and d- parameters Dimensions One and Two for Test Forms.

Location of Second Dimension/ Test Form	a_1	a_2	*D*
BLX[a]	0.410	0.373	-0.494
BLY	0.415	0.371	-0.342
UIX	0.466	0.272	-0.483
UIY	0.468	0.270	-0.330
CIX	0.520	0.170	-0.474
CIY	0.520	0.170	-0.321

[a] BL = both locations, UI = unique items only, CI = common items only

The 16 common items constituted 23.5% of the entire test length on all test forms, which was close to the suggested 20% guideline. That is, Kolen and Brennan (2004, p. 313) advocate having the common items represent at least 20% of the total items on the test forms to ensure that the number of common items is sufficient to adequately represent the unique items on the test form in content (relevancy and representation), as well as statistical properties. A review of experimental IRT equating literature suggest that equating results may be adequate with between 5 and 15 common items, without consideration of the total number of test items (Cook & Petersen, 1987). With 16 common items, both of these suggested criteria were met. The items were selected for each section according the specifications in Table 2.

During the simulation, the sample of examinees specified as Group P was used to simulate responses to Form X. Group Q was used to simulate responses to Form Y. In addition, simulated responses were generated for Group Q to the 52 unique items from Form X, for a total of 120 items simulated for Group Q (52 unique Form Y items + 16 common items + 52 unique Form X items = 120). For test scoring and equating purposes, a Form X score and a Form Y score (each scored out of 68) for all members of Group Q was

calculated (see Figure 1).

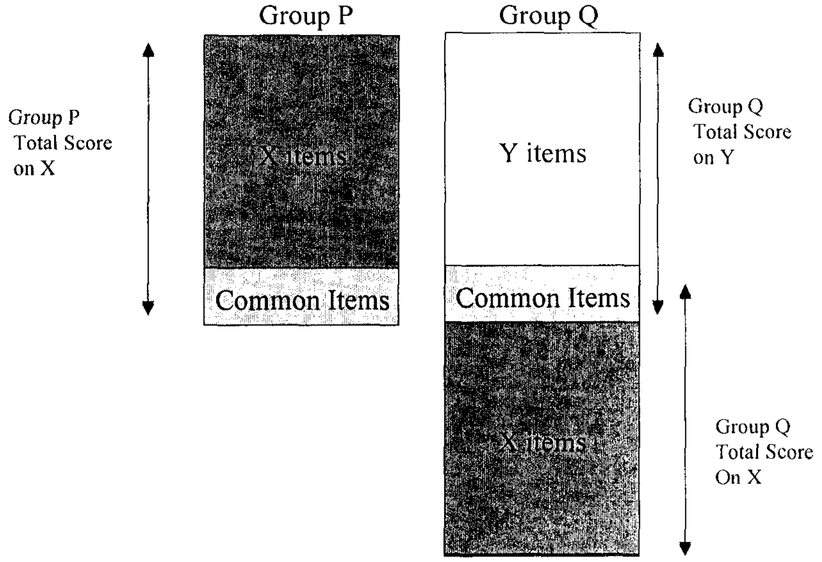

Figure 1. Diagram of Test Form for Group P and Group Q

The 16 common items acted as an internal anchor to Forms X and Y (i.e., the common items scores were included in the calculation of the total score for each form). Further, for Group Q, the raw scores on the 16 common items were used in the calculation of both Form X and Y total scores (i.e., responses to the common item set were generated once, but used in both total scores, so the anchor test scores for Group Q on Form X and Y were always the same).[4] Generating a score on both forms for Group Q was necessary to create variables that compare the Form Y score equated to the scale of Form X to the actual score (or target score) on Form X. In relation to real world testing

[4] The use of the same common item responses in calculating total score for both forms was due to limitations of the data generation software, which could only simulate 120 items at a time. Please see chapter 6 for a discussion of the potential limitation of this procedure.

conditions, this procedure was equivalent to having one group of examinees take both Form X and Form Y. Thus, for the parallel forms conditions, Group Q Form X scores were equated to Group P Form X scores, while in the nonparallel forms conditions, Group Q Form Y scores were equated to Group P Form X scores.

Data Generation

In a simulation study, replications of each experimental condition are conducted to provide more stable and reliable parameter estimates (Harwell et al., 1996). For the purposes of this study, 100 replications of each condition were generated.

Data for 100 sets of 2,000 simulated responses were generated for each test form for each condition. The main consideration in specifying sample size is to ensure that there are sufficient numbers to ensure stable IRT parameter estimates. Hanson and Béguin (2002) found significantly higher squared bias, variance, and mean squared error in samples of 1,000 versus samples of 3,000 using a 3-PL unidimensional IRT model. Bolt and Lall (2003) reported similar trends (between sample sizes of 1,000 and 3,000) in root mean squared error in the M2PL model, but also found that increasing the number of test items from 25 to 50 items improved parameter estimate precision. A practical consideration in selecting sample size is that simulation processing times increase rapidly with increasing sample size. Given that the number of items on the tests is large (68, 120), a sample size of 2,000 was adequate for reasonable measurement precision. As a result, 2,000 simulated responses were used for all conditions. *M2gen2* (Ackerman, 2004) was used to generate examinee response data matrices for each replication.

Data Processing

Once the data were simulated, the Group Q data were split into two 68-item data sets, one for Form Y and one for Form X (with the 16 common item responses included in both data sets). The Group Q Form X data were set aside for calculating dependent variables, while the Group Q Form Y data were used for equating to the Group P Form X data.

Unidimensional item parameters were estimated for each data set using *BILOG* (Mislevy & Bock, 1990). For this step, the unidimensional IRT 2-PL model was specified. At this point, the common items from the forms were calibrated relative to the other items within the respective test forms, and were thus on separate scales. The Stocking and Lord (1983) procedure was used to calibrate the forms. The program *ST* (v. 1.0, Hanson & Zeng, 1995b) was used to calculate scale transformation coefficients. These scale transformations were applied to the parameters of all items of the test form to be equated. The transformed parameters were used to run the program *PIE* (v.1.0, Hanson & Zeng, 1995a) in order to conduct IRT true-score equating.

Dependent Variables

The dependent variables measure the performance of IRT true score equating with reference to the central aspect of IRT equating, equity. Bolt (1999) argued that equity was the most important aspect of IRT equating because the principle guiding IRT-based equating methods is Lord's (1980) property of equity. In his book, Lord makes the statement, that in order to have true equity, it should be "a matter of indifference to applicants at every given ability level θ whether they are to take test X or test Y" (p. 195).

The measures proposed here are slightly different from those proposed by Bolt (1999) and Thomasson (1993), who have previously measured first- and higher order-equity. Rather than using expected values of *X* given $\boldsymbol{\theta}$ (the vector containing θ_1 and θ_2) and equated *Y* to *X* (*x(Y)*) given $\boldsymbol{\theta}$ values, it is possible to use the actual values for both of these tests for each simulated examinee who took form Y, which contained both the X and Y test items. This adjustment is advantageous because it allows the comparison between equated and actual test scores, rather than between equated and estimated test scores. For the purposes of defining the dependent variables, the term "target score" refers to the actual score to which the equated score was compared for each examinee. It should be noted that the use of a target score, which is an observed score (as opposed to a true score) on the actual target test, will contain a certain amount of error associated with measurement that will make it unlikely for the equating errors to ever be exactly zero.

All dependent variable calculations and intermediate steps were conducted using *Visual BASIC* in *Microsoft Excel.*

As a second means of gauging the importance or meaningfulness of the equating error associated with equating Form Y scores to Form X scores, the dependent variables for this set of conditions were also calculated using the raw, unequated Y scores. Calculating the dependent variables in this way indicated the magnitude of the error associated with the Y scores if they were used in place of the X scores, but not equated.

Mean absolute difference. The mean absolute difference was calculated by using absolute value of the difference between their target score and their equated score for each simulated examinee in calculating the condition mean:

$$MAD_{xy} = \frac{\sum_{j=1}^{n} \left| x_j - {}_{irt_X}(y_j) \right|}{n},$$

where x_j is the Form X score for examinee j , ${}_{irt_X}(y_j)$ is the equated Form X score for examinee j, and n is the sample size. Similarly, when Group Q Form X scores were equated to Group P Form X scores, the formula is given as:

$$MAD_{xx(equated)} = \frac{\sum_{j=1}^{n} \left| x_j - {}_{irt_X}(x_j) \right|}{n},$$

where ${}_{irt_X}(x_j)$ is the equated Form X score for examinee j. This dependent variable assesses the magnitude of the differences between equated and target scores when positive and negative differences do not cancel each other out in the calculation of the mean. When the raw Group Q Form Y scores were compared to Group P Form X scores, a mean absolute difference was also found between the target score and the unequated Form Y scores using the formula:

$$MAD_{xy(unequated)} = \frac{\sum_{j=1}^{n} \left| x_j - y_j \right|}{n},$$

where y_j is the Form Y score for examinee j. This measure represents the magnitude of the error associated with interpreting the Y scores as interchangeable with the Form X scores *before* equating was conducted.

As a final measure of equity using MAD, the difference between the unequated and equated MADs was calculated. This measure is referred to as the MAD gain value, because it represents the amount of error that is added or subtracted from the MAD when scores are equated. In other words, if MAD gain is positive, equating was beneficial to improving equity of the scores, whereas if MAD gain is negative, more error is present in the score after equating, and therefore, equating was detrimental to score equity.

Percent of examinees with a score difference that matters. As a final, more practical measure of equity, the score difference that matters (SDTM) (Dorans & Feigenbaum, 1994) was employed. Dorans and Feigenbaum (1994) argue that from a conceptual standpoint, only a difference that is greater than one-half of a reporting scale unit matters to equating outcomes. Equated scores differences that are smaller than this criterion would be lost when final scores were rounded to the nearest full reporting scale unit. For the purposes of this research, the percentage of examinees with an equated score on Form Y that differs by more than 0.5 from their actual test score on test form (or whose equated Form X score differs from their unequated Form X score) were calculated for each condition.

Just like MAD, percent SDTM was calculated for the equated score and unequated scores. The gain percent SDTM was calculated by finding the difference between the unequated SDTM and equated SDTM. A positive gain percent SDTM represents proportionally fewer people affected by a score difference that matters, while a negative gain score represents proportionally more people affected by a score difference that matters.

Section 2: Real Data Study

In this section, the data and procedures used for the real data analysis are described. The section is organized into the following four subsections: 1) description of the data sources, 2) procedures, 3) data processing, and 4) dependent variables.

Description of the Data Sources

Both of the real datasets were taken from tests that are part of the Praxis Series^{TM} testing program. These tests are used as partial requirements for teacher licensure in several U.S. states. Both tests contained 120 multiple-choice items. All multiple-choice items had four response options. Normally, scores for these tests would be reported on a scale of 100-200, but for the purposes of this analysis, only raw scores were used.

Test A was a test of health and physical education content knowledge. This test contains several content categories, but can be substantively considered to contain two major dimensions: 1) health (50 items), and 2) physical education (70 items). A sample size of 3,877, accumulated from administrations held between November 2003 and March 2005, was used for this analysis. One item that was identified as problematic during operational administration was not included in the analysis.

Test B was a test of English literature and composition content knowledge. This test contains three content categories that can substantively be considered three major dimensions: 1) reading and understanding text (66 items), 2) language and linguistics (21 items), and 3) composition and rhetoric (34 items). The two largest categories, reading and understanding text, and composition and rhetoric, were chosen for the purposes of this analysis. Only the items related to these two dimensions (100 items) were included. A sample of 4,226, accumulated from administrations of one test form held between April and August 2005 were used for the analysis.

Procedures

This analysis was designed so that it would be possible to use the same dependent variables that were used in the simulation study. Therefore, instead of equating two different forms of the same test, each test was split into two forms and equated, so that equated scores could be compared to target scores (on the form to which the scores were equated) within examinees. All 120 items from Test A and 100 items for Test B were separately fit to the M2PL model using *NOHARM* (Fraser, 1988). Then, the items for each form were split to create two separate forms (Form X and Form Y). The two new forms from each original form were created in three different ways: 1) Dimension two items among the unique items only, 2) Dimension two items among the common items only, and 3) Dimension two items among both the unique and common items. Thus, three test pairs were created from each original test form. The test specifications are outlined in Table 4. The test specifications for the number of items representing each dimension were limited by the number of items on the original test form, as well as the number of items in each substantive dimension. The number of items in each dimension was also constrained by an attempt to keep the relative proportion of items representing each substantive dimension similar to the proportions used in the simulation study (see Table 2 in Section 1 of this chapter). In order to fulfill the test specifications for each set of conditions, not all items from the original test forms were included in the new forms. The item parameters for each test pair for Test A can be found in Appendixes D-F, and for Test B in Appendixes G-I.

Table 4

Test Specifications for Real Data Test Forms.

		Unique Items		Common Items		Total Items
		Dimension		Dimension		
	Location of Second Dimension	1	2	1	2	
Test A	Unique items	21	10	10	0	41
	Common items	31	0	5	5	41
	Both locations	21	10	5	5	41
Test B	Unique items	20	10	10	0	40
	Common items	30	0	5	5	40
	Both locations	20	10	5	5	40

The selection process for the items for each test form was conducted using the following approach. Once the M2PL parameters were obtained, the items were separated into Dimension one and Dimension two items according to the substantive dimension to which the item was originally designated. Each dimension for each test was sorted according to the *d*-parameter estimates, from highest to lowest difficulty. Then the items were assigned to each form of each test by placing the first item (the item with the highest difficulty) in each sorted dimension into Form X, then the second and third items (the second and third highest difficulties) into Form Y, the fourth and fifth items into Form X, swapping back and forth until the test specifications for that dimension were met. Then, using a more arbitrary approach, items from the remaining items in the sorted category were traded in and other items traded out, to ensure that the easier items were represented on the forms. Table 5 presents the resulting mean item parameter estimates for each form of Test A, while Table 6 presents the mean items parameter estimates for each form of Test B.

Table 5

Mean a_1-, a_2-, and d- parameter Estimates and Correlation between Dimensions One and Two for Test A Test Forms.

Location of Second Dimension/ Test Form	a_1	a_2	d	$r_{1,2}$[a]
BLX[b]	0.223	0.209	0.530	0.399
BLY	0.255	0.257	0.725	0.395
UIX	0.240	0.201	0.562	0.377
UIY	0.275	0.25	0.757	0.371
CIX	0.215	0.167	0.495	0.213
CIY	0.292	0.248	0.819	0.212

[a] $r_{1,2}$ denotes correlation between substantive dimensions on each test
[b] BL = both locations, UI = unique items only, CI = common items only

As Table 5 illustrates, the Test A form pairs used for equating had similar mean a_1- and a_2-parameter estimates, with differences ranging 0.03θ to 0.05θ between forms for the BL and UI pairs, and a 0.08θ difference between forms for the CI pair, but differed on the mean *d*-parameter estimates by about 0.20θ for BL and UI pairs and about 0.30θ for the CI pair. Therefore, the test pairs for Test A were similar in form parallelism to the Form X and Y pairs that were created for the simulation study. The correlation between the substantive dimensions on each form of Test A was low (see Table 5). In Table 6, the Test B form pairs used for equating are much more closely parallel, with differences of between 0.01θ and 0.05θ for the a_1- and a_2-parameter estimates. The *d*-parameter differences between forms were in the range of 0.01θ for the BL and UI form pairs, but were larger for the CI form pairs with a difference of 0.10θ between forms. The Test B form pairs fall somewhere between the parallel forms and nonparallel forms conditions. The correlation between substantive dimensions for each form of Test B was about halfway between low and high.

Table 6

Mean a_1-, a_2-, and d- parameter Estimates and Correlation between Dimensions One and Two for Test B Test Forms.

Location of Second Dimension/ Test Form	a_1	a_2	D	$r_{1,2}$[a]
BLX[b]	0.326	0.342	0.681	0.498
BLY	0.360	0.310	0.676	0.548
UIX	0.326	0.301	0.733	0.466
UIY	0.361	0.291	0.747	0.522
CIX	0.344	0.255	0.628	0.403
CIY	0.394	0.265	0.742	0.442

[a] $r_{1,2}$ denotes correlation between substantive dimensions on each test

[b] BL = both locations, UI = unique items only, CI = common items only

Sample

Once the tests forms were created, the sample of examinees was split into two groups, Group P and Group Q. For Test A, Group P had 1,939 examinees, while Group Q had 1,938 examinees. For Test B, both groups had 2,113 examinees. For the purposes of this study, Group P was “administered” Form X and Group Q was “administered” Form Y, but because data were available for both test forms, Group Q Form X data were retained as target scores for those examinees. The groups for each test were set by simply splitting the data files into half. This arbitrary splitting of the data allowed for, but did not guarantee, some level of group nonequivalence for the purposes of comparison to the simulated data conditions.

Ability parameters for each group were found for each test form by combining groups P and Q, then calibrating the response data using the program 2DEAP version 1.0 (Luecht, 1992). Table 7 displays the means and standard deviations of the ability parameter estimates on both Dimension one and Dimension two for each group for Test A. Group P and Group Q were very similar in ability on both dimensions, with the largest difference between groups of 0.040 θ on Dimension one on the UIX form. These groups were considered to be very close to randomly equivalent. Therefore, the results of the Test A analysis were most directly comparable to the equivalent groups (EG) conditions of the simulation study. It should be noted that while the means for each sample for each θ are near zero, as would be expected, the standard deviations are less than the expected value of one. Luecht and Miller (1992) note a similar pattern of smaller standard deviations than expected in some of the initial documentation for this program.[5] They attribute this result to bias in the variance-covariance matrix when expected a priori (EAP) estimation procedures are used in the program 2DEAP, and might be particularly noticeable in these test forms with very few items measuring each dimension. The bias can be considered a limitation of the method. The raw scores on Dimension One and Two for Populations P and Q on Test A are presented in Appendix J. Comparison of these raw scores between populations provides additional evidence that the assertion that the groups are comparable is a reasonable one.

[5] Several test runs of this program with known simulated data produced similar results to those observed with the real data. Two dimensional data with sample means and standard deviations of 0 and 1, respectively produced ability estimates for each θ with means near 0 but standard deviations near 0.7 and 0.8. Even with a correlation between dimensions specified as 1.0, standard deviations did not go over 0.8.

Table 7

Mean and standard deviations of θ_1, θ_2 for Groups P and Q for Test A forms.

Location of Second Dimension	Group *P*		Group *Q*	
	θ_1[a]	θ_2	Θ_1	θ_2
BLX[b]	0.012 (0.691)	-0.001 (0.673)	-0.013 (0.669)	-0.002 (0.678)
UIX	0.020 (0.706)	0.001 (0.667)	-0.020 (0.687)	-0.002 (0.560)
CIX	0.014 (0.677)	0.013 (0.560	-0.015 (0.660)	-0.016 (0.556)

[a] Standard deviations are presented in brackets ()
[b] BL = both locations, UI = unique items only, CI = common items only

Table 8 presents the mean and standard deviations of the ability estimates for Dimensions one and two for both groups on Test B. On all forms, there was a group difference of approximately 0.1 θ favouring Group Q on both dimensions. Therefore, the results of the Test B analysis were most directly comparable to the nonequivalent small (NEGS) conditions of the simulation study. Similar to that observed in Table 7, the standard deviations are not near one as would be expected. The raw scores on Dimension One and Two for Populations P and Q on Test B are presented in Appendix K. Comparison of these raw scores between populations provides additional evidence that the assertion that the groups are nonequivalent (with small differences) is a reasonable one.

Table 8

Mean and standard deviations of θ_1, θ_2*for Groups P and Q for Test B forms.*

Location of Second Dimension	Group *P*		Group *Q*	
	θ_1[a]	θ_2	θ_1	θ_2
BLX[b]	-0.053	-0.044	0.066	0.056
	(0.770)	(0.811)	(0.769)	(0.792)
UIX	-0.051	-0.040	0.062	0.056
	(0.766)	(0.787)	(0.743)	(0.765)
CIX	-0.057	-0.044	0.070	0.055
	(0.776)	(0.631)	(0.785)	(0.635)

[a] Standard deviations are presented in brackets ()
[b] BL = both locations, UI = unique items only, CI = common items only

Equating Procedures

For each test, Group Q Form Y scores were equated to Group P Form X scores using the Stocking and Lord scale transformation (through the common items) and equated using IRT true-score equating. The same software that was used for the simulation study was used for these analyses.

Dependent Variables

To calculate the dependent variables, Group Q equated and unequated Form Y scores were compared to Group Q X scores for each individual. Mean absolute difference (MAD) and percent score difference that matters (percent SDTM) were both calculated, as well as the gain (difference between unequated and equated values) for each dependent variable for each test.

Section 3: Computer Programs

NOHARM

NOHARM can fit unidimensional or multidimensional IRT models to response data, and can be specified to work in either confirmatory or exploratory mode. In confirmatory mode, the user provides the vector **c** containing the fixed guessing estimates (which in the case of the M2PL model are all set to 0), as well as the initial pattern matrices F and P. F is an *n* x *m* factor matrix which specifies which *m* items belong to which *n* dimension(s), while P is an *n* x *n* correlation matrix which specifies the correlation matrix among the n dimensions. The numbers in each matrix are set to 0 if the corresponding estimated value is to be fixed, 1 if the corresponding value is to be estimated. Once these inputs are provided, NOHARM estimates the parameters in F and P using the procedure outlined in McDonald (1982). The procedure uses a least squares minimization algorithm. It also produces a residual matrix and root mean square residual which can be used to assess the goodness of fit of the model to the data.

M2GEN2

M2GEN2 was used to generate the data for the simulation study. Once the desired means, variances, and correlation of the bivariate normal distributions are specified, the program randomly selects thetas from the bivariate normal distribution according to those specifications. The seed for all random selections is based on the internal clock of the computer CPU that is used to run the program. The program uses several International Mathematics and Statistical Library (IMSL) subroutines (e.g., IMSL, 1994). After two thetas are selected for an individual examinee the program goes through each of the items and calculates a response vector of zeroes and ones. The first step is for the program to calculate probabilities of a correct response for each item for each examinee, based on the specified item parameters and the randomly selected examinee parameters. The second step is that the program again uses an IMSL subroutine to generate a deviate from a uniform distribution bound by zero and one. The program compares the probability of a correct response to the deviate. If the probability of a correct response is greater than or equal to the probability of the deviate, the examinee is assigned a 1 (denoting a correct response) for the item. If the probability is less than the probability of the deviate, the examinee is assigned a 0 (denoting an incorrect response.) For each item, a new uniform deviate is selected for the purposes of comparison. The program cycles through each item for an examinee and then repeats the process beginning with generating a second set of ability θs until the desired number of response vectors (corresponding to number of examinees) has been generated (T. A. Ackerman, personal communication, May 24, 2006).

BILOG

BILOG is a program that can estimate IRT item and parameter estimates for unidimensional data. Operationally, BILOG approximates the integration across θ that is required for MMLE (described in Chapter 2) by applying a prior approximate distribution of examinee abilities (a set of quadrature points and corresponding density weights) to the data (Hambleton et al., 1991; Mislevy & Bock, 1990). Next, parameter estimates of *a* and *b* are found such that the likelihood of the examinee responses to a given item is maximized. BILOG uses MMLE to jointly estimate the *a* and *b* parameters. The process that is used to estimate the parameters is called the expectation-maximization (EM) algorithm. The EM algorithm contains two phases. In the first expectation phase, expected values of the parameter estimates are calculated. These estimates are adjusted iteratively in the second maximization phase using the Newton-Gauss procedure that is used to find the root of nonlinear functions (Bock & Aitkin, 1981). The Newton-Gauss procedure is very similar to the Newton-Raphson procedure (Lee & Jennrich, 1979) that is used in IRT true-score equating (described in Chapter 2). Examinee ability estimates are created based on the item parameter estimates.

ST

The program ST calculates scale transformations to align two sets of parameter estimates on common items so that the transformation can in turn be applied to one set of parameter estimates to bring them onto a common scale with the other set of parameter estimates (Hanson & Zeng, 1995b). ST can conduct the mean/mean (Loyd & Hoover, 1980) and mean/sigma (Marco, 1977) methods, as well as the test characteristic curve methods the Haebara (1980) and the Stocking and Lord transformation (1983). The formulas used to calculate the Stocking and Lord transformation function are identical to those presented in Chapter 2. The loss minimization routine DFPMIN (Press, Teukolsky, Vetterling & Flannery, 1994, p. 428) is used to find the transformation function.

PIE

PIE is a computer program that can be used for IRT true or observed score equating (Hanson & Zeng, 1995a). The program is based on the 3-PL model (Lord, 1980), but can be used for the 2-PL model if all *c*-parameters are set to 0. PIE uses true score equating formulas identical to those identified in Chapter 2. The program uses the routine RTSAFE (Press et al., 1994, p. 366), which combines the Newton-Raphson procedure with a bisection routine to ensure that the Newton-Raphson stays within reasonable bounds.

2DEAP

To calculate two-dimensional MIRT parameter estimates for the data used in the real data study, the program 2DEAP version 0.1 (Luecht, 1992) was used. Similar to

BILOG, 2DEAP approximates the item parameter estimates using MMLE. The program first applies a bivariate normal prior distribution to get the posterior likelihood distribution (2DEAP assumes orthogonal dimensions). The likelihood function used is an extension of the same likelihood used for the unidimensional case, but is based on a MIRT probability function (Reckase, 1985, 1997).

Chapter Summary

In Chapters 1 and 2, an introduction to the research questions, as well as background information and a rationale for the study was provided. In this chapter, the definitions, procedures and programs used to conduct the simulation and real data studies were presented. In Chapter 4, the results of the simulation study will be described and discussed, followed by Chapter 5 in which the results of the real data study and their relation to the results in the simulation study are provided.

Chapter 4: Results

Introduction

The results of the simulations are presented in this chapter in three sections. Section 1 contains results from all conditions using parallel forms. Section 2 contains the results from all conditions using nonparallel forms, while Section 3 contains all results examining the gain of equating versus not equating. A chapter summary follows the presentation of all of the simulated results.

The rationale for this organization was two-fold. First, it was logical to split the results using parallel forms from nonparallel forms because it was only possible to calculate the gain values of equating versus not equating for the nonparallel forms. For the nonparallel forms, the criterion for comparing the raw unequated Y score for each examinee was their respective raw unequated X score. In contrast, for the parallel forms, the criterion and the score being compared are the same unequated X score and therefore the unequated difference for parallel forms would always be zero.

Second, upon preliminary examination of the results it became apparent that, of the independent variables considered in this study, form nonparallelism made the largest contribution to equating error. When forms were not parallel[6], the error associated with the conditions was much larger than the error associated with conditions where the forms were parallel. Given these findings, separating the results for the two levels of this independent variable made describing the results more clear and concise.

[6] It is important to note that the identical (and therefore perfectly parallel) forms in this study are used for the purposes of establishing an ideal baseline for study purposes and should not be viewed as a realistic condition in practice. In reality it is virtually impossible to generate perfectly parallel forms, and identical forms do not need to be equated.

With this organization the results also make a logical progression. In the parallel forms results, the effects of the correlation between dimensions, the location of the items measuring the second dimension, and group equivalence under ideal conditions were evident. In the nonparallel forms results, the effect of varying the forms systematically in mean difficulty from each other is introduced and, as explained above, the effect of this additional variable dominated the effects attributable to the correlation between dimensions, the location of the items measuring the second dimension, and group equivalence. Finally, the gain results describe whether equating under each set of conditions removes error from or adds error to the scores when compared to the magnitude of error that is present in the scores prior to equating.

Criteria for Assessing the Meaningfulness of Equating Error and Gain Values

The purpose of this study was to examine the robustness of IRT equating using a common-items nonequivalent groups design under a set of conditions that varied the multidimensional character of the test forms, the location of the items measuring the second dimension, the equivalence of the groups, and the degree of parallelism between the forms. As mentioned in the previous chapter, the robustness was assessed using the mean absolute deviation (MAD) between the pairs of unequated and equated scores and the percent of examinees who have a difference greater than a score difference that matters, which in this study was half a score point (0.5). Given that equating error would likely never be exactly zero, it was necessary to develop criteria that could be used to decide if equating under each set of conditions was robust. For each dependent variable, a set of criteria was developed to classify the magnitude of errors into small, medium, and large error. These categories can be thought of as analogous to the A-, B- and C-level effects used to interpret differential item functioning (DIF) (Roussos & Stout, 1996). Small error parallels A-level DIF, which is considered negligible. In this study, conditions with small error are considered robust. Medium error parallels B-level DIF, which is meaningful and should be addressed if it is possible to do so, but might be acceptable in some situations. Large error parallels C-level DIF, which warrants immediate attention and suggests that if found in the equated scores of this study, the IRT equating under those conditions was not a satisfactory method for creating scores that are interchangeable for interpretation.

Mean Absolute Difference

The Mean Absolute Difference (MAD) represents the average magnitude of the error associated with an examinee's score for a given sample, regardless of the direction of the error. For test equating to be robust at the sample level, the equating error needs to be small enough that it would not be detectable once the equated score was rounded for reporting. This idea comes from Dorans and Feigenbaum's (1994) concept of a *score difference that matters*. Dorans and Feigenbaum (1994) suggested that any difference between equated and true scores that is "observable" once scores are rounded to the nearest reporting scale unit is meaningful and matters because it is visible. For example, if test scores are on a scale where scores are reported in units of 10, then any score difference of 5 or greater would matter because it is larger than one-half of a reporting scale unit, and would therefore result in a different reported score, once the scores were rounded. In this study, where the scores are reported on a 0 - 68 scale with a score interval of 1, any MAD that was less than 0.5 was considered not meaningful. Therefore, a MAD of less than 0.50 was set as the criterion for small error. The boundary between medium and large error was more difficult to establish. A medium level of error might be anything that was 0.5 score points or greater but less than two percent of the total possible score. Two percent error was an arbitrary decision but seemed like a reasonable error size that would not be strictly considered robust according to the *score difference that matters* criterion, but might still be considered an error of acceptable size. For this study, two percent of a total possible score of 68 is 1.36 score points. Any error that was two percent or greater would be large enough to be problematic when trying to compare examinees. Therefore, any MAD that was 1.36 or greater was considered large error.

Upon reviewing the data, discernible patterns within size categories (small, medium, large) also became evident. In order to interpret these patterns consistently, another criterion needed to be developed for determining when the MAD "increased" or "decreased" across conditions. A rule of 0.05 score points, or one-tenth of the 0.5 score point difference that matters criterion, was adopted for this purpose. For example, in Table 9, in the 0.3 correlation BL conditions, there was a 0.05 score point difference between EG (0.29) and NEGS (0.34), but only a 0.04 difference between NEGS and NEGM (0.38). In these two cases, the MAD increased between EG and NEGS, but there was no systematic difference between NEGS and NEGM.

Percent of Examinees with Score Difference that Matters

The proportion (expressed as a percent) of examinees with a score difference that matters (percent SDTM) examines equating error differently than MAD, though it is based on the same set of raw data and draws upon the same concept of a difference that matters. MAD reflects the *average* magnitude of all of the examinee differences to express the degree of error in the sample overall. In contrast, the percent SDTM is a *count* of all examinee differences that have detectable error, based on the same "difference that matters" criterion described for MAD. Because of what each dependent variable focuses on, their metrics are also different. For MAD, the metric is score points. For percent SDTM, the metric is percent of examinees. Percent SDTM is a more examinee-centered dependent measure because it makes a decision about the robustness of equating for each individual first, which is then summarized in order to be able to make inferences about equating robustness at the sample level. By contrast, MAD also calculates the difference for each examinee, but when they are averaged, information is lost about how many people are affected.

Because the percent SDTM differs from the MAD in its underlying focus and metric, the percent SDTM requires a different operational definition of robustness at the sample level. Ideally, the percent SDTM should be zero, meaning that no examinee has a score difference that mattered. However, this definition was not useful for interpretative purposes in this study because none of the conditions was likely to meet this ideal criterion; therefore no condition would have been classified as robust. Instead, the criterion for robustness was developed based on the results for conditions where the test equating performance was expected to be the best using a common-items design (Bolt, 1999; Kolen & Brennan, 2004, p. 294), and where the unidimensionality assumptions of IRT equating were not violated (Kolen & Brennan, 2004, pp. 156-7). The results of the perfect (1.0) correlation, equivalent groups (EG), parallel forms conditions were examined, and based upon these results, small error was defined as any value less than 25.0 percent. For a boundary between medium and large error, any value of percent SDTM that was 25.0 percent or greater, but less than 50.0 percent was defined as medium and any value that was 50.0 percent or greater was defined as large. The rationale for the 50 percent cut-off was that an equating function that produced equated scores where more than half of the examinees had a detectable level of error could not be considered robust under any circumstances because the majority of examinees had meaningful error in their scores.

An additional criterion was developed to frame changes in percent SDTM between conditions. Just like the rule used for MAD, one-tenth of the "small" categorization was adopted as this value. A change of 2.5 percent or greater must have occurred between conditions in order for the values to be systematically different.

Gain Values

Gain values, for both MAD and percent SDTM, are the differences between the unequated and equated errors for each respective condition. Gain values represent the magnitude of error that was either added or subtracted from scores by equating. Establishing criteria for assessing gain values was approached differently than the approach used to establish criteria for assessing error because there are two factors to consider, the magnitude and direction of gain. First, the magnitude of gain is categorized into small, medium and large gain based on the original criterion corresponding to that dependent variable. In other words, for MAD the small criteria was less than 0.5, medium was 0.5 or greater but less than 1.36, and large was 1.36 or greater. For the percent SDTM, small gains were change of less than 25 percent, medium gains were 25 percent or greater but less than 50 percent and large gains were 50 percent or greater.

Subsequent to determining the magnitude, the direction of the gain must be considered: positive gains indicate that there was less error present in the scores after equating and therefore equating was beneficial, while negative gains indicate that there was more error present in the scores after equating and therefore equating was detrimental. Changes to gains across conditions were also assessed relative to the same criteria that were established for the corresponding dependent variable. The sign of the gain was also taken into consideration when examining differences in gain across conditions.

Section 1: Parallel Forms

The results for the parallel forms conditions are presented in Table 9. For this table and all the tables that follow in this chapter, the classifications for small, medium and large error are identified by the font style. Values classified as small are presented in normal font, medium values are in bold normal font, and large values are in bold italics. Further, the structure of each table is the same. The correlations between dimensions are listed in column 1 followed by the group equivalence in column 2. The next three columns correspond to the location of the dimension 2 items: included in both the common and unique item sets (column 3), only in the unique item sets (column 4), and only in the set of common items (column 5). Given the purpose of this study, the discussion is organized in terms of the three locations of the dimension 2 items.

Mean Absolute Difference

As indicated above, the third column of data in Table 9 corresponds to the location of the items measuring the second dimension in both the common and unique items sets (BL). When the correlation between dimensions was 1.0 or 0.7, the values for MAD were small at all levels of group equivalence. When the correlation was 0.3, the MAD increased from EG to NEGS and NEGM, but again the MAD values for all three conditions were small. When the correlation between dimensions was 0.0, the MAD values for EG and NEGS were small, while the MAD value for NEGM was medium.

In the fourth column of Table 9, the items measuring the second dimension were among the unique items only (UI). When the correlation was 1.0, the MAD values were all small. When the correlation between dimensions was 0.7, all three MAD values were all small, but there was an increase in MAD from EG and NEGS to NEGM. When the correlation between dimensions was 0.3, the MAD increased from EG to NEGS to NEGM, but the MADs for EG and NEGS were small, while the MAD for NEGM was medium. When the correlation between dimensions was 0.0, the pattern of increasing MAD as the groups became more nonequivalent was similar to, but more marked than the pattern when the correlation between dimensions was 0.3. In this case, the EG MAD was small, while the NEGS and NEGM MADs were both medium.

Table 9

Mean Absolute Difference (MAD) for Parallel Forms, by Correlation Between Dimensions, Group Equivalence, and Location of the Second Dimension Items

Correlation Between Dimensions	Group Equivalence	Location of Second Dimension Items		
		BL[a]	UI	CI
Perfect (1.0)	EG (0.0)	0.25	0.32	0.29
	NEGS (0.1)	0.27	0.32	0.24
	NEGM (0.3)	0.28	0.34	0.28
High (0.7)	EG (0.0)	0.28	0.34	0.30
	NEGS (0.1)	0.31	0.31	0.34
	NEGM (0.3)	0.30	0.40	0.32
Low (0.3)	EG (0.0)	0.29	0.38	0.47
	NEGS (0.1)	0.34	0.43	0.48
	NEGM (0.3)	0.38	**0.50**	**0.68**
No (0.0)	EG (0.0)	0.32	0.42	0.49
	NEGS (0.1)	0.32	**0.53**	**0.61**
	NEGM (0.3)	**0.56**	**0.82**	**1.33**

Note: Values in normal text represent small errors (<0.50), values in **bold** represent medium errors (0.5 ≤ x < 1.36), values in ***bold italics*** represent large errors (≥1.36).

[a] BL = both locations, UI = unique items only, CI = common items only

In the fifth column, the items measuring the second dimension were present in only the common items (CI). When the correlation between dimensions was 1.0, MADs for EG, NEGS, and NEGM were all small, but MAD decreased from EG to NEGS. When the correlation between dimensions was 0.7, the MAD for EG, NEGS, and NEGM were also all small and did not differ systematically. When the correlation between dimensions was 0.3, the MADs for EG and NEGS were small and lower than the MAD for NEGM, which was medium. When the correlation between dimensions was 0.0, the MADs increased systematically as group nonequivalence increased. The MAD for EG was small, while MADs for NEGS and NEGM were both medium.

Percent of Examinees with a Score Difference that Matters

The results for percent SDTM are presented in Table 10. For the BL location (column 3), all levels of group equivalence had small percent SDTMs when the correlation between dimensions was 1.0. When the correlation between dimensions was 0.7, the percent SDTM increased from EG and NEGS to NEGM, but all percent SDTM values were still small. When the correlation between dimensions was 0.3, there was a more distinct trend of increasing percent SDTM for EG and NEGS to NEGM. For this correlation, the percent SDTM increased from EG to NEGS but were both small, while the percent SDTM for NEGM was medium. Lastly, when the correlation between conditions was 0.0, percent SDTM for EG and NEGS were small and did not differ, while the NEGM percent SDTM was large.

For the UI location (Table 10, Column 4), when the correlation between dimensions was 1.0, the percent SDTMs for the EG and NEGS conditions were small, while the NEGM percent SDTM value was medium. When the correlation between dimensions was 0.7, the percent SDTMs were small for EG and NEGS and medium for NEGM. When the correlation between dimensions was 0.3, the percent SDTM for all three levels of group nonequivalence was medium, with an increasing pattern from EG to NEGS to NEGM. When the correlation between dimensions was 0.0, the pattern of increasing percent SDTM was more marked as the groups became more nonequivalent, with increases in percent SDTM from EG to NEGS to NEGM. Further, while the percent SDTMs for the EG and NEGS conditions were medium, the percent SDTM for the NEGM was large.

In the CI location (Table 10, column 5), when the correlation between dimensions was 1.0, percent SDTM for all group equivalence conditions was small, but there was a decrease in percent SDTM from EG to NEGS and then an increase from NEGS to NEGM. When the correlation between dimensions was 0.7,the percent SDTMs for all three group equivalence conditions were small. However, the percent SDTM increased from EG to NEGS, and then decreased from NEGS to NEGM. When the correlation between dimensions was 0.3, the percent SDTMs were medium for EG and NEGS and did not differ, while the percent SDTM was large for NEGM. When the correlation between dimensions was 0.0, there was a noticeable increase in the percent SDTM from EG to NEGS to NEGM. The percent STDM for EG was medium, while percent STDMs were large for both NEGS and NEGM.

Table 10

Percent of Examinees with Score Difference That Matters (SDTM) for Parallel Forms, by Correlation between Dimensions, Location of the Second Dimension, and Group Equivalence

Correlation Between Dimensions	Group Equivalence	Location of Dimension 2 Items		
		BL[a]	UI	CI
Perfect (1.0)	EG (0.0)	13.4	21.9	16.7
	NEGS (0.1)	14.3	21.2	11.5
	NEGM (0.3)	15.0	**25.4**	14.9
High (0.7)	EG (0.0)	16.2	23.3	18.8
	NEGS (0.1)	15.5	21.9	23.6
	NEGM (0.3)	18.8	**31.7**	19.0
Low (0.3)	EG (0.0)	16.7	**25.2**	**38.0**
	NEGS (0.1)	24.7	**34.8**	**37.9**
	NEGM (0.3)	**28.7**	**40.1**	***57.4***
No (0.0)	EG (0.0)	19.2	**31.1**	**36.7**
	NEGS (0.1)	20.4	**43.4**	***54.8***
	NEGM (0.3)	***51.6***	***70.4***	***84.1***

Note: Values in normal text represent small errors (<25%), values in **bold** represent medium errors (25% ≤ x < 50%), values in ***bold italics*** represent large errors (≥50%).

[a] BL = both locations, UI = unique items only, CI = common items only

Summary

A pattern of results begins to emerge from the MAD and percent SDTM results. In general, the errors were smaller when the correlations were higher, and increased as the correlations decreased. While the differences between the MADs and percent STDMs among the different levels of group equivalence were not clear for the higher levels of correlation, they became much sharper for the two lower correlations, with a general tendency of increased error with increased group nonequivalence. The errors were also generally smaller if the items measuring the second dimension were present in both the common and unique items sets, particularly at the lower correlations.

Section 2: Nonparallel Forms

Mean Absolute Difference

The MAD results for the nonparallel forms are reported in Table 11. All MADs were large for all locations and, within locations, for all correlations between dimensions and levels of group equivalence. The MAD values ranged from 3.25 in the 1.0 correlation, EG, BL conditions to 3.89 in the 0.0 correlation, NEGM, CI condition.

Differences among the MAD values emerged in the nonparallel forms results. For the BL location (Table 11, column 3), the MADs increased from EG and NEGS to NEGM when the correlation between dimensions was 1.0. When the correlation between dimensions was 0.7, the MAD for EG was smaller than the MAD for NEGS and NEGM. When the correlation between dimensions was 0.3, the MAD did not differ across conditions. When the correlation between dimensions was 0.0, the MAD did not change between EG and NEGS, but increased from EG to NEGM.

In the UI location (Table 11, column 4), when the correlation between dimensions was 1.0, the MADs increased from EG and NEGS to NEGM. The MADs did not differ when the correlation between dimensions was 0.7 and 0.3. Finally, when the correlation between dimensions was 0.0, the MADs increased from EG to NEGS, then decreased between NEGS and NEGM.

In the CI location (Table 11, column 5), when the correlation between dimensions was 1.0, the MADs increased from EG and NEGS to NEGM. When the correlation between dimensions was 0.7, the MAD increased from EG to NEGM. Lastly, when the correlation between dimensions was either 0.3 or 0.0, the MADs increased from EG to NEGS to NEGM.

Table 11

Mean Absolute Difference (MAD) for Nonparallel Forms, by Correlation between Dimensions, Location of the second dimension, and Group Equivalence

Correlation Between Dimensions	Group Equivalence	Location of Dimension 2 Items		
		BL[a]	UI	CI
Perfect	EG (0.0)	***3.25***	***3.29***	***3.32***
(1.0)	NEGS (0.1)	***3.28***	***3.29***	***3.33***
	NEGM (0.3)	***3.35***	***3.35***	***3.40***
High	EG (0.0)	***3.26***	***3.33***	***3.33***
(0.7)	NEGS (0.1)	***3.31***	***3.33***	***3.37***
	NEGM (0.3)	***3.34***	***3.37***	***3.41***
Low	EG (0.0)	***3.31***	***3.32***	***3.35***
(0.3)	NEGS (0.1)	***3.34***	***3.34***	***3.45***
	NEGM (0.3)	***3.35***	***3.37***	***3.60***
No	EG (0.0)	***3.35***	***3.35***	***3.41***
(0.0)	NEGS (0.1)	***3.37***	***3.67***	***3.55***
	NEGM (0.3)	***3.41***	***3.39***	***3.89***

Note: Values in normal text represent small errors (<0.50), values in **bold** represent medium errors (0.5 ≤ x < 1.36), values in ***bold italics*** represent large errors (≥1.36).

[a] BL = both locations, UI = unique items only, CI = common items only

Percent of Examinees with a Score Difference that Matters

Table 12 presents the results for the percent SDTM for the non-parallel forms. The percent SDTMs were large for all locations and within locations, for all conditions. The values ranged from 89.8 percent in the 1.0 correlation, EG, BL condition to 91.9 percent in the 0.0 correlation, NEGM, CI condition. The differences in the percent SDTM between conditions were too small to claim any systematic differences.

Table 12

Percent of examinees with a score difference that matters (SDTM) for nonparallel forms, by correlation between dimensions, location of the second dimension, and group equivalence

Correlation Between Dimensions	Group Equivalence	Location of Second Dimension Items		
		BL[a]	UI	CI
Perfect (1.0)	EG (0.0)	***89.8***	***90.0***	***90.2***
	NEGS (0.1)	***90.1***	***89.9***	***90.4***
	NEGM (0.3)	***90.4***	***90.2***	***90.4***
High (0.7)	EG (0.0)	***89.9***	***90.2***	***90.1***
	NEGS (0.1)	***90.2***	***90.3***	***90.4***
	NEGM (0.3)	***90.3***	***90.4***	***90.6***
Low (0.3)	EG (0.0)	***90.1***	***90.2***	***90.5***
	NEGS (0.1)	***90.2***	***90.3***	***90.7***
	NEGM (0.3)	***90.5***	***90.6***	***91.0***
No (0.0)	EG (0.0)	***90.4***	***90.3***	***90.4***
	NEGS (0.1)	***90.6***	***90.5***	***91.2***
	NEGM (0.3)	***90.5***	***90.7***	***91.9***

Note: Values in normal text represent small errors (<25%), values in **bold** represent medium errors (25% ≤ x < 50%), values in ***bold italics*** represent large errors (≥50%).

[a] BL = both locations, UI = unique items only, CI = common items only

Summary

In the nonparallel forms conditions, the effect of the form nonparallelism was large relative to the effects of the other independent variables. All locations and group equivalence conditions within location had large errors. The patterns of differences in the MADs within location and condition were similar, though much more dampened, than patterns observed in the parallel forms case. In general, the errors were smaller when correlations were higher, and increased as the correlation between dimensions decreased. The errors were generally smallest within the conditions for the BL location, and were largest within the conditions for the CI location. The errors were also generally smallest when the two groups were equivalent, and largest when the groups were moderately nonequivalent (NEGM). No patterns were observed in percent SDTM.

Section 3: Gain from Equating Nonparallel Forms

In this section, the gain realized by equating scores is reported. For each dependent variable, the unequated results are presented first in Table 13. The unequated results were calculated by finding the differences between the raw *Y* and raw *X* scores for each examinee in Group Q (the group whose scores were equated). If the raw *X* score for each examinee represents the target score for that examinee, then the unequated difference represents the error that is inherent in the raw *Y* score as a representation of the target score prior to equating.

The gains realized from equating are presented next in Table 14. The gain value is equal to the difference between the unequated results and equated results for each condition. These values describe the degree to which equating added or removed error from the *Y* score as a representation of the target *X* score. Positive gain values occur when the unequated error was larger than the equated error, and indicate that there was a benefit to equating because it reduced the magnitude of the difference (MAD) or the percentage of examinees with a difference that matters (percent SDTM). Negative gain values indicate that equating increased the magnitude of the difference (MAD) or increased with percentage of examinees with detectable score differences (percent SDTM).

Mean Absolute Difference

Unequated results. The unequated MAD results are presented in Table 13. The unequated MADs for all locations and conditions within location without equating were large. The unequated MAD values ranged from 4.01 in the 1.0 correlation, UI, EG condition to 4.33 in the 0.0 correlation, NEGM, CI condition.

While the values are all classified as large, there are still patterns of increasing MAD present in results. For the BL location (Table 13, column 3), when the correlation between dimensions was either 1.0 or 0.7, the unequated MADs increased from EG to NEGS to NEGM. When the correlation between dimensions was 0.3, the unequated MAD increased from EG and NEGS to NEGM. When the correlation between dimensions was 0.0, the unequated MADs increased from EG to NEGS to NEGM.

In the case of the UI location (Table 13, column 4), when the correlation between dimensions was 1.0, the unequated MADs increased from EG and NEGS to NEGM. When the correlation between dimensions was 0.7, the unequated MADs increased from EG to NEGS to NEGM. When the correlation between dimensions was 0.3, the unequated MADs increased from EG and NEGS to NEGM. When the correlation between dimensions was 0.0, the unequated MADs increased from EG to NEGS to NEGM.

Lastly, for the CI conditions (Table 13, column 5), when the correlation between dimensions was 1.0 or 0.7, the unequated MADs increased from EG and NEGS to NEGM. When the correlation between dimensions was 0.3, unequated MAD increased from EG to NEGS and NEGM. Finally, when the correlation between dimensions was 0.0, the unequated MADs increased from EG to NEGS to NEGM.

Table 13

Unequated Mean Absolute Difference (MAD) for Nonparallel Forms, by Correlation between Dimensions, Location of the Second Dimension, and Group Equivalence

Correlation Between Dimensions	Group Equivalence	Location of Second Dimension Items		
		BL[a]	UI	CI
Perfect (1.0)	EG (0.0)	***4.02***	***4.01***	***4.14***
	NEGS (0.1)	***4.07***	***4.05***	***4.17***
	NEGM (0.3)	***4.13***	***4.13***	***4.26***
High (0.7)	EG (0.0)	***4.04***	***4.03***	***4.17***
	NEGS (0.1)	***4.10***	***4.09***	***4.21***
	NEGM (0.3)	***4.16***	***4.16***	***4.28***
Low (0.3)	EG (0.0)	***4.10***	***4.07***	***4.18***
	NEGS (0.1)	***4.13***	***4.10***	***4.24***
	NEGM (0.3)	***4.19***	***4.22***	***4.28***
No (0.0)	EG (0.0)	***4.11***	***4.09***	***4.22***
	NEGS (0.1)	***4.16***	***4.16***	***4.28***
	NEGM (0.3)	***4.24***	***4.23***	***4.33***

Note: Values in normal text represent small errors (<0.50), values in **bold** represent medium errors ($0.5 \leq x < 1.36$), values in ***bold italics*** represent large errors (≥1.36).

[a] BL = both locations, UI = unique items only, CI = common items only

Gain values. Gain values are presented in Table 14. All of the gain values were positive, indicating that there was a benefit to equating under all conditions. The smallest gain was 0.44 in the CI, 0.0 correlation NEGM condition, and the largest gain was 0.87 in the CI, 0.7 NEGM condition. With the exception of the CI, 0.0 correlation NEGM and UI, 0.0 correlation NEGS conditions, which had small gains (0.44 and 0.49, respectively), all of the gains were larger than a score difference that matters, but smaller than two percent of the score scale (1.36 points) and were therefore medium gains.

The change in gain pattern is somewhat different than what was observed for the MAD and unequated MAD values. In the BL conditions (Table 14, column 3), when the correlation between dimensions was either 1.0 or 0.7, no changes in MAD gain were observed between changing group equivalence conditions. When the correlation between dimensions was either 0.3 or 0.0, MAD gain increased from EG and NEGS to NEGM.

Table 14

Gain Mean Absolute Difference (MAD) for Nonparallel Forms, by Correlation between Dimensions, Location of the Second Dimension, and Group Equivalence

Correlation Between Dimensions	Group Equivalence	Location of Second Dimension Items		
		BL[a]	UI	CI
Perfect (1.0)	EG (0.0)	*0.77*	*0.72*	*0.82*
	NEGS (0.1)	*0.79*	*0.76*	*0.84*
	NEGM (0.3)	*0.78*	*0.78*	*0.86*
High (0.7)	EG (0.0)	*0.78*	*0.70*	*0.84*
	NEGS (0.1)	*0.79*	*0.76*	*0.84*
	NEGM (0.3)	*0.82*	*0.79*	*0.87*
Low (0.3)	EG (0.0)	*0.79*	*0.75*	*0.83*
	NEGS (0.1)	*0.79*	*0.76*	*0.79*
	NEGM (0.3)	*0.84*	*0.85*	*0.68*
No (0.0)	EG (0.0)	*0.76*	*0.74*	*0.81*
	NEGS (0.1)	*0.79*	0.49	*0.73*
	NEGM (0.3)	*0.83*	*0.84*	0.44

Note: Positive gain values indicate a benefit to equating, negative gain values indicate detriment attributable to equating. Values in normal text represent small gains (<0.50), values in **bold** represent medium gains ($0.5 \leq x < 1.36$), values in ***bold italics*** represent large gains (≥ 1.36).
[a] BL = both locations, UI = unique items only, CI = common items only

In the UI conditions (Table 14, column 4), when the correlation between dimensions was 1.0, MAD gain decreased from EG and NEGS to NEGM. When the correlation between dimensions was 0.7, MAD gain decreased from EG to NEGS and NEGM. When the correlation between dimensions was 0.3, MAD gain decreased from EG and NEGS to NEGM. When the correlation between dimensions was 0.0, MAD gain decreased from EG to NEGS, then increased from NEGS to NEGM.

In the CI conditions (Table 14, column 5), when the correlation between dimensions was either 1.0 or 0.7, MAD gain did not change with changing group equivalence. When the correlation between dimensions was 0.3, MAD gain decreased from EG and NEGS to NEGM. Finally, when the correlation between dimensions was 0.0, MAD gain decreased from EG to NEGS to NEGM.

Percent of Examinees with a Score Difference that Matters

Unequated results. The unequated percent SDTM results are presented in Table 15. In all conditions, the unequated percent SDTM was classified as large. The range of unequated percent SDTM was from 91.6 percent in the 1.0 correlation, EG, UI condition to 93.0 in the 0.0 correlation, NEGM, CI condition. The differences in unequated percent SDTM between conditions were too small to claim any systematic differences in the results.

Table 15

Unequated Percent of Examinees with a Score Difference that Matters (SDTM) for Nonparallel Forms, by Correlation between Dimensions, Location of the Second Dimension, and Group Equivalence

Correlation Between Dimensions	Group Equivalence	Location of Second Dimension Items		
		BL[a]	UI	CI
Perfect (1.0)	EG (0.0)	***91.8***	***91.6***	***92.3***
	NEGS (0.1)	***92.0***	***92.0***	***92.4***
	NEGM (0.3)	***92.1***	***92.1***	***92.5***
High (0.7)	EG (0.0)	***91.9***	***91.8***	***92.4***
	NEGS (0.1)	***92.1***	***92.0***	***92.5***
	NEGM (0.3)	***92.4***	***92.5***	***92.8***
Low (0.3)	EG (0.0)	***92.2***	***92.2***	***92.5***
	NEGS (0.1)	***92.3***	***92.3***	***92.6***
	NEGM (0.3)	***92.5***	***92.9***	***92.6***
No (0.0)	EG (0.0)	***92.3***	***92.4***	***92.6***
	NEGS (0.1)	***92.5***	***92.5***	***92.8***
	NEGM (0.3)	***92.5***	***92.7***	***93.0***

Note: Values in normal text represent small errors (<25%), values in **bold** represent medium errors (25% ≤ x < 50%), values in ***bold italics*** represent large errors (≥50%).

[a] BL = both locations, UI = unique items only, CI = common items only

Gain values. The gain values for the percent SDTM are presented in Table 16. All gain values were positive, indicating that equating reduced the percent of examinees with a score difference that matters in all conditions. The values ranged from 1.2 percent in the 0.0 correlation NEGM CI condition, to 2.3 percent in the 0.3 correlation, NEGM, UI condition. All gain values were classified as small, and no systematic patterns were observed across conditions.

Table 16

Gain Percent of Examinees with a Score Difference that Matters (SDTM) for Nonparallel Forms, by Correlation between Dimensions, Location of the Second Dimension, and Group Equivalence

Correlation Between Dimensions	Group Equivalence	Location of Second Dimension Items[a]		
		BL[a]	UI	CI
Perfect (1.0)	EG (0.0)	2.0	1.6	2.1
	NEGS (0.1)	1.9	2.1	2.0
	NEGM (0.3)	1.8	1.9	2.1
High (0.7)	EG (0.0)	2.0	1.7	2.2
	NEGS (0.1)	1.9	1.7	2.1
	NEGM (0.3)	2.1	2.1	2.2
Low (0.3)	EG (0.0)	2.1	2.0	2.0
	NEGS (0.1)	2.1	2.0	1.9
	NEGM (0.3)	2.0	2.3	1.6
No (0.0)	EG (0.0)	1.9	2.1	2.2
	NEGS (0.1)	1.9	2.0	1.6
	NEGM (0.3)	2.0	2.1	1.2

Note: Positive gain values indicate a benefit to equating, negative gain values indicate detriment attributable to equating. Values in normal text represent small errors (<25%), values in **bold** represent medium errors (25% ≤ x < 50%), values in ***bold italics*** represent large errors (≥50%).

[a] BL = both locations, UI = unique items only, CI = common items only

Summary

Unequated MAD and unequated percent SDTM displayed patterns similar to the patterns observed for the equated nonparallel forms. The gain results indicate that there was only a small benefit to equating, regardless of location of the common items and the conditions within location.

Chapter Summary

The results of this simulation study suggest that of all the factors manipulated in this study, form parallelism produced the largest effect. When forms were parallel, equating error tended to be small, but equating was sensitive to the effects of the location of the second dimension items, correlation between dimensions, and group equivalence. When forms were nonparallel, the effects of these other dependent variables were still evident, but the magnitude of their effect was largely overshadowed by the effect of form nonparallelism. In the nonparallel forms case, there was evidence of a benefit to equating across all conditions, but the benefit was modest relative to the size of the error that was present. The benefit of equating to reducing error was diminished somewhat by lower correlations and larger group differences when the second dimension was present among the common items only (CI location).

These results suggest several interesting points for discussion. But do the results reflect what would be observed with real data? In Chapter 5, the results of real data analyses will be presented. These results will provide evidence that the results of this simulation are realistic.

Chapter 5: Real Data Study

Introduction

To augment the simulation results presented in chapter 4, analyses on two real data sets were conducted. These analyses illustrate the effect that the location of the items measuring the second dimension exerts on the robustness of IRT equating when using a common-items nonequivalent groups design. This chapter is organized into the following sections: 1) the criteria for assessing the meaningfulness of the equating errors and gains, 2) real data results, 3) discussion of the relation of the real data results to the simulation study and, 4) a summary of the real data analyses.

Criteria for Assessing the Meaningfulness of Equating Error and Gain Values

The criteria used to assess the meaningfulness of equating error and gain values were based on the same rationales that were used in the simulation study. The criterion for small MAD was less than 0.5 score points, because it represents a score difference that is undetectable. The criterion for medium MAD error was less than 2 percent of the score scale (0.82 for test A and 0.80 for test B, respectively). The criterion for large MAD for Test A was 0.82 score points or greater and for Test B, 0.80 score points or greater. The same criteria that were used for percent SDTM in the simulation study were used for these analyses. These criteria were less than 25 percent for small, 25 percent to less than 50 percent for medium, and 50 percent or greater for large, respectively.

Criteria were also used to quantify shifts in MAD and percent SDTM that were noticeable, but not large enough to cause changes in the small, medium, and large error categorizations. The criterion for MAD was 0.05 score points, which represents one-tenth of the score difference that matters (and the cut-off between small and medium errors).

Similarly, the criterion for the percent SDTM was 2.5 percent, also one-tenth of the criterion for small percent SDTM error. Both of these criteria are identical to what was used in the simulation study.

The gain attributable to equating was also examined in the real data. Just as in the simulation study, a positive gain reflects the amount of error that was removed from the scores as a result of equating. Conversely, a negative gain reflects error that was introduced into the scores as a result of equating. The magnitudes of the gains follow the same criteria for assessing error as their respective dependent variables. Further, the criteria for assessing differences in gain across conditions correspond to the "increasing" and "decreasting" criteria set for the dependent variable. Additionally, because it is possible to have a negative gain, shifts in gain are assessed relative to their sign as well as their magnitude. For example, in Table 19, the gain MAD for Test A for the BL location is 0.16 and -0.18 for the UI condition, which is a difference in absolute magnitude of only 0.02, but a relative difference of -0.34. Therefore, a large negative difference is observed between the BL and UI locations.

Results

Test A

Mean absolute difference. The MAD results for Test A are presented in Table 17. For this table and all tables that follow in this chapter, the classifications for small, medium, and large error are identified using the font style used in the previous chapter. Small values are presented in normal font, medium values are in bold normal font, and large values are in bold italics. Further, the structure of each table is the same. Each row represents a different derivation of the dependent variables; the unequated values are presented in row 2, the equated values are presented in row 3, and finally, the gain values are presented in row 4.

For the unequated results (Table 17, row 1), the MADs were large for all locations. However, the MAD for the CI location was larger than the MADs for both the BL and UI locations by over 0.5 score points. In the equated results (Table 17, row 2), the MADs were still large for all locations, but the difference between the MAD for the CI location and the MADs for the other locations of the common items was much greater (1.31 and 1.35 score points than the BL and UI locations, respectively). When the gain of equating was examined (Table 17, row 3), the gain scores were small and positive for the BL and UI locations, indicating that equating reduced error in the scores. In the CI location, the gain was also small, but was negative, indicating that more error was introduced by equating than was present prior to equating.

Table 17

Unequated, Equated and Gain MAD for Test A.

	Location of Second Dimension Items		
	BL[a]	UI	CI
Unequated MAD	***2.96***	***2.94***	***3.52***
Equated MAD	***2.68***	***2.64***	***3.99***
Gain MAD	0.28	0.30	-0.47

Note: Values in normal text represent small errors or gains (<0.50), values in **bold** represent medium errors or gains (0.5 ≤ x < 0.82), values in ***bold italics*** represent large errors or gains (≥0.82). Positive gain values indicate a benefit to equating; negative gain values indicate detriment attributable to equating.

[a] BL = both locations, UI = unique items only, CI = common items only

Percent of examinees with a score difference that matters. Table 18 presents the percent SDTM results for Test A. For the unequated results (Table 18, row 1), the unequated percent SDTMs for all locations were large. However, the unequated percent SDTM was larger for the CI location than for the UI or BL locations. The percent SDTMs for the equated results (Table 18, row 2) were also all large. However the percent

STDM for the CI location was greater than the percent STDMs for both the BL and UI locations. The percent SDTM gains (Table 18, row 3) were small and positive for both the BL and UI locations, indicating a benefit of equating (reduction in error) in these conditions. The gain was and small and negative for the CI location, indicating that equating introduced a small amount of error in this location.

Table 18

Unequated, Equated and Gain Percent SDTM for Test A.

	Location of Second Dimension Items		
	BL[a]	UI	CI
Unequated Percent SDTM (%)	***88.9***	***89.2***	***91.6***
Equated Percent SDTM (%)	***88.2***	***88.0***	***93.0***
Gain Percent SDTM (%)	0.7	1.2	-1.4

Note: Values in normal text represent small errors or gains (<25%), values in **bold** represent medium errors or gains (25% ≤ x < 50%), values in ***bold italics*** represent large errors or gains (≥50%). Positive gain values indicate a benefit to equating; negative gain values indicate detriment attributable to equating.

[a] BL = both locations, UI = unique items only, CI = common items only

Test B

*Mean absolute difference.*The MAD results for Test B are presented in Table 19. When the scores were unequated (Table 19, row 1), the MADs were large across the three locations. As with Test A, the unequated MAD for the CI location was greater than the unequated MADs for the BL and UI locations. For the equated MAD scores (Table 19, row 2), the equated MADs for all locations were large. When equated, the MADs increased from the BL location to the UI location to the CI location, with the MAD for the CI location being much larger (2.25 and 1.91 score points, respectively). The MAD gain (Table 19, row 3) was small and positive for the BL location, but small and negative for the UI location and large and negative for the CI location.

Table 19

Unequated, Equated and Gain MAD for Test B.

	Location of Dimension 2 Items		
	BL[a]	UI	CI
Unequated MAD	***2.63***	***2.63***	***2.97***
Equated MAD	***2.47***	***2.81***	***4.72***
Gain MAD	0.16	-0.18	***-1.75***

Note: Values in normal text represent small errors or gains (<0.50), values in **bold** represent medium errors or gains (0.5 ≤ x < 0.80), values in ***bold italics*** represent large errors or gains (≥0.80). Positive gain values indicate a benefit to equating; negative gain values indicate detriment attributable to equating.

[a] BL = both locations, UI = unique items only, CI = common items only

Percent of examinees with a score difference that matters. The percent SDTM results for Test B are presented in Table 20. When the scores were unequated (Table 20, row 1), the percent SDTMs were large across the three locations. The unequated percent SDTM for the CI location exceeded the unequated percent SDTMs for both the BL and UI locations. The values of the percent STDMs for the equated scores (Table 20, row 2) followed a similar trend, with the largest value, to a greater degree, for the CI location. The percent SDTM gain (Table 20, row 3) for the BL location was small and positive, indicating that equating was beneficial (i.e., less error was present after equating), while it was small and negative for the UI and CI locations, indicating that equating was detrimental. The negative impact of equating was greater for the CI location than it was for the UI location.

Table 20

Unequated, Equated and Gain percent SDTM for Test B.

	Location of Dimension 2 Items		
	BL[a]	UI	CI
Unequated Percent SDTM (%)	***86.9***	***86.9***	***90.6***
Equated Percent SDTM (%)	***86.7***	***87.9***	***95.3***
Gain Percent SDTM (%)	0.2	-1.0	-4.7

Note: Values in normal text represent small errors or gains (<25%), values in **bold** represent medium errors or gains ($25\% \leq x < 50\%$), values in ***bold italics*** represent large errors or gains ($\geq$50%). Positive gain values indicate a benefit to equating; negative gain values indicate detriment attributable to equating.

[a] BL = both locations, UI = unique items only, CI = common items only

Relationship of Real Data Results to Simulated Data Results

The Test A and Test B mean item, test, and examinee parameters align most closely with the nonparallel forms, 0.3 correlation equivalent and nonequivalent groups (for Test A and Test B, respectively) conditions from the simulation study. Therefore, these simulated conditions are the most logical conditions to use as a point of comparison between the simulated and real data results. Several points arise from this comparison.

First, the nonparallel forms simulated results produced large MAD and percent SDTM errors. A similar magnitude of errors was also observed for both Test A and Test B. This similarity suggests that the large errors observed in the simulation were comparable to what is observed in real data.

Second, when examining the effect of location, there are some differences between the simulated and real data results. In the simulated conditions (Table 14 and Table 16 for MAD and percent SDTM, respectively), the gains across locations were approximately equal. In the real data conditions, the BL and UI locations produced similar results, which were less error-prone than the CI location.[7] BL and UI location

[7] In saying there is an advantage to the BL and UI locations, it is important to qualify the assertion by also observing that all three locations still produced large errors in both MAD and percent SDTM, in both unequated and equated results. For Test B, the UI location had small negative gains.

gain results for Test A (Table 17 and Table 18 for MAD and percent SDTM, respectively), are similar to each other and are small but still positive, while the CI location gain results are negative. For Test B, the gain results are slightly different (Table 19 and Table 20 for MAD and percent SDTM, respectively). For MAD gain, the BL location showed a small reduction in error through equating, while the UI location produced a small increase in error and the CI location produced a large increase in error. A similar pattern was observed in the percent SDTM gain, the only difference being that the CI location produced only a small increase in error.

Interestingly, while the real data results are not similar in pattern to the simulated data results to which their parameters most closely match, the results do look very similar to the simulated 0.3 correlation, nonequivalent moderate (NEGM) conditions, as well as to the 0.0 correlation, nonequivalent small (NEGS) and NEGM conditions. This similarity suggests that the simulated data results are realistic, though the results might overestimate the robustness of the equating.

Third, when compared to the simulated data, the real data gains were also much smaller, again suggesting that although the patterns in the simulated data were similar to those observed in the real data, they demonstrated a greater level of robustness for low correlations than the real data results did. Test B had a higher correlation between dimensions than did Test A. However, the gains for Test A were still better (i.e., larger positive and smaller negative gains) than the gains for Test B.

Fourth, the Test A groups were nearly equivalent while the Test B groups had a small group nonequivalence (see Table 8, p. 68), which might explain why the gains for Test A were better than these gains for Test B. This result aligns with the simulated data results that revealed an advantage of improved group equivalence, particularly at the lower correlations.

Summary

The MAD for all conditions was large, in both the unequated and equated conditions. Once equated, however, the MAD was decreased for the BL location for both Test A and Test B, and for the UI location for Test B. MAD was increased for the CI location for both Test A and Test B, and for the UI location on Test B. The same pattern of results was observed for percent SDTM. The characteristics of these test forms were most similar to the nonparallel forms, 0.3 correlation, EG and NEGS conditions of the simulated data. The results of the real data analysis of Test A and B produced similar patterns of results to those conditions, suggesting that the simulated data results were reasonable, although they might have overestimated robustness of IRT equating using a CI-NEG design, particularly when the correlations were low. These results also suggest that the link between the correlation between dimensions and equating robustness is not as straight forward as what has been demonstrated previously (e.g., Bolt, 1999), but rather is mediated by form parallelism, group equivalence, and the location of the items containing the second dimension.

Chapter 6: Discussion, Conclusions, and Recommendations

Introduction

The preceding two chapters provided the results of analyses for both simulated and real data. This chapter contains a summary of the study, a discussion of the results, and the conclusions and recommendations that arise from the findings. First, a brief summary of the design and procedures for both studies will be outlined, followed by a listing of the key results. Then, the limitations of the study design and the results of the analyses in light of these limitations will be discussed. A general discussion of the results will follow, which will consider the likely causes of equating error in relation to the observed patterns in the results. It will also include where this research fits in relation to what was previously known about the influence of multidimensionality on the robustness of IRT equating, what questions this research answered, and what questions either arise or remain. Finally, conclusions, implications for practice, and directions for future research are presented.

Summary of Purpose and Procedure

Simulation Study

The purpose of the simulation study was to systematically study the effects of multidimensionality on the robustness of IRT equating when using a common items nonequivalent groups design. The research design for this study included four fully-crossed independent variables: form parallelism (parallel forms-0.0θ, nonparallel forms-0.2θ), the correlation between the dimensions present on the test forms (None-0.0, Low-0.3, High-0.7, Perfect-1.0), group equivalence (Equivalent-0.0θ, Nonequivalent Small-0.1 θ, Nonequivalent Moderate-0.3 θ), and the location of the items measuring the second dimension (common items, unique items, both common and unique items locations).The LSAT was used to determine the values of the item parameters used to generate the data. Test Forms X and Y were constructed so they were well matched statistically, with each form including a set of common items that were representative of the unique items on each form. Data were generated for two samples: Group P (for Form X) and Group Q (for Forms X and Y). Each condition was replicated 100 times with 2,000 examinees per test form. Data were processed using previously established statistical programs that are used by current testing agencies, including *BILOG* (Mislevy & Bock, 1990) to fit the 2-PL model to the data, and *ST* (Hanson & Zeng, 1995b) and *PIE* (Hanson & Zeng, 1995a) to conduct the Stocking and Lord scale transformation and IRT true score equating, respectively. The dependent variables were designed to capture error that contributes to loss of equity in true score equating, both in terms of the magnitude of the error (Mean Absolute Difference), as well as the proportion of examinees that are affected by a score difference that is meaningful (Percent of Examinees with a Score Difference that Matters), by comparing equated scores to actual target test scores. Gain values for each dependent variable were also calculated to determine if equating made the error inherent in scores better or worse by comparing equated scores to unequated scores for each examinee.

Real Data Study

The real data study used the results from two teacher licensure tests (Test A and Test B) to see if the results were similar to those observed in the simulated data. Data from one group of examinees were split and treated as two test forms, so that the equated scores on one form could be compared to the actual scores on the other form. The Test A form pairs had close to the same form differences as in the Nonparallel forms conditions, and the correlations between the substantive dimensions were low (0.2-0.4). Test A examinee samples were similar in ability (with the largest observed group difference 0.04 in the UI condition), like the Equivalent groups conditions. The Test B form pairs had differences between the forms that were somewhere in between the Parallel and Nonparallel forms conditions (0.00θ -0.10θ) and the correlations between substantive dimensions fell between low and high (0.4-0.5), according to the definitions used in the simulated conditions. The Test B examinee samples had a small difference in mean ability between the groups (about 0.1θ), similar to the Nonequivalent Small conditions in the simulation study. The same statistical methods, procedures, programs, and dependent variables that were used in the simulation study were used in the real data study.

Summary of Findings

The simulation and real data analyses yielded several results. The key results were, as follows.

- In the simulation study, when the forms were perfectly parallel, equating tended to be robust under most conditions, but when the forms were not parallel (as set in this study), equating tended not to be robust even under unidimensional conditions, with errors exceeding the large error criteria for both MAD ($\geq$1.36) and percent SDTM ($\geq$50.0%).
- In both the simulated and real data studies, where the items measuring the second dimension were located in the test forms mattered. When the second dimension items were among both the unique and common items, equating tended to be more robust than when the items were in the other two locations considered. More

specifically, when the items were among only the unique items, equating was somewhat compromised, but equating was least robust when the items measuring the second dimension were represented only in the common items.

- In the simulation study, the previously established relationship between the correlation between the dimensions on a test form and the robustness of IRT equating was confirmed; as the correlation increases, so does the robustness of the equating.
- In both the simulated and real data studies, IRT equating using a CI-NEG design tended to be more robust when the groups were equivalent.
- In both the real data and simulation studies, the benefit of equating was somewhat limited when the forms were not parallel. In many nonparallel conditions, the error associated with the scores was larger when equating was performed than before the scores were equated.

Limitations of the Studies

Form Parallelism

Only two levels of this variable were used: a parallel forms and nonparallel forms. The parallel forms were selected because it was the ideal case. The nonparallel forms difference was set to 0.2θ because the difference was large enough to ensure that the effects of the variable would be observable. However, unlike the correlation between dimensions, which included several levels of the variable, there are limits to the conclusions that can be drawn about what effect form difficulty differences have on the robustness of IRT equating. A more systematic approach to varying the degree of form parallelism is necessary to draw conclusions about what level of form parallelism is required to ensure robustness under differing conditions.

Confound between Number and Location of Second Dimension Items

In this study, the location of the items measuring second dimension results must be interpreted with care to acknowledge that the location of the second dimension items is confounded by the number (and by extension, proportion) of Dimension two items. Referring to Table 2 (chapter 3), in the BL location, eight Dimension one common items represented 36 unique Dimension one items, while eight Dimension two common items represented 16 unique Dimension two items. In the UI location, 16 Dimension one common items represented 36 Dimension one items, while the 16 unique Dimension two items had no representation in the common items. In the CI location, eight Dimension one common items represented 52 unique Dimension one items, while eight Dimension two items made up the remainder of the common items. In this set of test specifications, the CI location had the fewest number of items representing Dimension two, while in the UI location the Dimension two items were not represented in the common items at all. The BL location had neither of these issues. It is not clear if location of the second dimensions itself made a difference, or if the number of items representing each dimension was more important. Additional investigations of the effect of location of second dimension items where the number and proportion of common items are varied would give greater confidence in the inferences about the effects of location of the second dimension on equating robustness.

Real Data Limitations

An important limitation of the real data analyses is that the test forms were not equated in their “real” state. Rather, each real test form was split into two test forms, so that data were available for both test forms for all examinees. An ideal situation would have been a single-groups data collection design, whereby data for two entire forms of each test would be available for a single group. Another option would have been to use only test repeaters who had data available for multiple forms. However, the number of examinees would have been severely limited, and the ability level of the examinees would have been unrepresentative of the population of first time examinees. This is because only examinees with low test scores (who failed the test) would be repeat test-takers. Despite this limiting condition, the results of the real data analyses suggest that the outcomes in the simulation study were realistic.

A second important point to consider is the possibility of model misfit in the data. While NOHARM was run using a two dimensional M2PL model in confirmatory mode, no further models with additional dimensions were run to compare their goodness-of-fit to those of the two dimensional model. Model misfit might contribute to the larger equating errors that were observed in the real data conditions than were observed in the simulated conditions. However, the fact that similar patterns of results were observed despite the possibility of this misfit further suggests that the simulation results were realistic.

Discussion

Independent Variables and their Interactions

Form parallelism. This independent variable addressed the first research question outlined in chapter 1: What is the baseline error that is associated with equating? That is, how much error would be present if scores on a test form were equated to scores on the same test form?

This research question was addressed by the parallel forms analyses. In these analyses, scores from a single test form (Form X) were equated, which, as perfectly parallel forms, should not require equating. The magnitude of the difference between the equated X scores and the actual X scores provided the basis for assessing how much error was inherent in equating under ideal circumstances. It is important to note that while these results act as a baseline for the purposes of research, such an ideal as equating perfectly parallel (i.e., identical) forms would never occur in reality.

When the forms were parallel, the MAD was either small or medium, regardless of the levels of the other variables. While medium-sized errors are not robust according to the definitions used in this study, they might be considered an acceptable level of error, depending on the purpose and use of the test scores. However, the percent SDTM proved to be more sensitive to error than MAD, and large errors that would not be considered acceptable or are too large to ignore were observed in some conditions. What makes these results even more interesting is that when the forms are parallel, equating is not necessary; any error that was introduced by equating can be considered negative gain because more error was introduced by equating than existed in the unequated score.

The pattern of results was similar for both MAD and percent SDTM in the parallel forms conditions, with two exceptions: when the correlation between dimensions was either perfect (1.0) or high (0.7), equating produced small errors.[8] These results are in keeping with previous research (Bolt, 1999; Camilli et al., 1995; Dorans & Kingston, 1985) that suggests IRT equating will be robust to multiple dimensions, as long as the correlation between the dimensions is 0.7 or higher (i.e., the dimensions are closely related to one another). The results were more complicated as the correlation between dimensions decreased from high to low (0.3) and to no correlation (0.0). When the second dimension items were in the BL location and the groups were either equivalent (EG) or had small group differences (NEGS), equating was robust (i.e., errors were small), but were larger when the groups had moderate differences (NEGM). In the UI and CI conditions, the MAD was still robust at the 0.0 and 0.3 correlation levels, as long as the groups were equivalent (EG). But as the correlations grew smaller and the groups more nonequivalent, errors became increasingly larger for both the MAD and percent SDTM. The errors due to the CI location tended to be larger than in the UI location.

The results of the parallel forms analyses suggest that under some conditions, IRT equating is robust to the presence of multiple dimensions. The results of the analyses suggest that even under ideal circumstances where forms are perfectly parallel, the presence of distinct dimensions can only be tolerated with good test design (BL location) and even then only if the groups taking each form do not differ substantially from each other. That is, the errors associated with these conditions were small enough that equity would be preserved in the reported scores according to the definitions used in this study.

[8] The exceptions to this statement were the percent SDTM for the 1.0 and 0.7 correlation, NEGM, UI conditions, where medium-sized errors were observed.

As such, these results suggest that under some conditions of multidimensionality, group equivalence, and location of items measuring the second dimension, perfectly parallel forms that should not require equating fail to meet Lord's strong equity criteria (Lord, 1980, p. 195).

In the simulation study, the form parallelism variable made, by far, the largest contribution to the variability of both the MAD and percent SDTM error. When the forms were strictly parallel (i.e., the baseline condition), the equating errors were several times smaller than when forms were nonparallel. However, these analyses also indicated that even under ideal circumstances where forms are perfectly parallel, the presence of distinct dimensions can only be tolerated with good test design (BL location) and even then only if the groups taking each form do not differ too much from each other. The results of the analyses of the real data for Test B also suggested that near parallel forms is not a guarantee of robust equating, even when the second dimension is represented in the common and unique items (BL location).

The magnitude of the equating errors were so much larger in the nonparallel conditions that it was much more difficult to see the patterns produced by the other variables that were observed in the parallel forms analyses. While the patterns that were observed in the nonparallel forms were similar to those observed in the parallel forms results, they were less obvious because the changes in error were small relative to the amount of error contributed by form nonparallelism.

The nonparallel forms results indicate that, when forms are not perfectly parallel (which is a likely scenario, given the form differences in Test A and Test B), IRT equating will not be robust even when the forms are unidimensional (i.e., correlation =

1.0). In the real data analyses, the Test B forms were more closely parallel than the Test A forms. Despite this, Test A had smaller equating errors overall than Test B. However, the results were not robust for both Test A and Test B. Taken together, these results cannot be used to make specific recommendations about how closely parallel tests have to be in order to obtain robust equating results. However, the results of the analyses of both the simulation and real data sets did reveal that when forms differed in difficulty by 0.2θ, equating was not robust, regardless of the degree of group differences, the location of the items measuring the second dimension, or whether the test was multidimensional or unidimensional. These results suggest that form differences equal to or greater than 0.2θ would not be recommended for IRT equating because equating actually introduces more error than was present before equating.

These results demonstrate that other factors in addition to the multidimensional character of the test forms mediate equating robustness. Besides form differences, group differences and the test design (in terms of location of items measuring the second dimension) also played a role in the degree of equating errors that were observed. Their respective roles are discussed in greater detail next.

Correlation between dimensions. The second research question addressed was: How does the correlation between dimensions affect both the magnitude of equating error and the proportion of examinees with error in their equated scores that is large enough to matter? This research question is at the heart of the research study because the degree of the two-dimensional character of the test forms was operationalized by the correlation between the dimensions. When the correlation between dimensions is perfect, the test is unidimensional because the dimensions cannot be distinguished from one another. As the correlation between dimensions decreases, the dimensions become more distinct from one another, and therefore the lower the correlation, the more clearly the violation of the unidimensionality assumption.

As discussed previously in the section on form parallelism, the results provide further evidence of decreasing robustness with increasing multidimensionality (i.e., lower correlations), particularly in the parallel forms conditions. The same patterns were also observed to a lesser extent in the nonparallel forms conditions. These results are consistent with previous research: Bolt (1999), Camilli et al. (1995), Dorans and Kingston (1985), and DeChamplain (1996) found that IRT equating was robust to multidimensionality when the correlation between dimensions was 0.7 or greater. However, the results of this study also suggest that the relationship between multidimensionality and robustness of IRT equating is mediated by the other factors explored in the study.

First, while form parallelism played a major role in the overall size of error, its interconnection with the correlation between dimensions appears to be, at most, minor. The change in the magnitude of errors as correlation between dimensions decreased was reasonably similar between the parallel and nonparallel form conditions of the simulation study. In fact, the relationship was more evident in the parallel forms conditions than in the nonparallel forms conditions, but this observation might be due to the large effect of form differences in the nonparallel forms conditions, which seemed to minimize the effect of all other variables.

Second, group differences worked together with the changes in correlation between dimensions such that the effects of a lower correlation on the magnitude of equating error appeared to be amplified by increasing group nonequivalence. When groups were equivalent, equating was more likely to be robust, even at lower correlations. As the groups became increasingly nonequivalent, proportionally larger equating errors at each correlation between dimensions were evident.

Third, the negative effect of lower correlations on equating robustness was dampened if the test forms were designed with second dimension items placed among both the common and unique items (BL). Conversely, the effect of decreasing correlation between dimensions was much more evident when the second dimension items were among the common items only (CI). These results provide evidence of the need for the common items (anchor test) to be representative of all dimensions of the test form in order to minimize equating error.

Group equivalence. Inclusion of this independent variable addressed the third research question: Is the magnitude of equating error or the proportion of examinees with equating error that is large enough to matter in their equated scores different if the groups are randomly equivalent versus nonequivalent? The common-items nonequivalent groups (CI-NEG) design is intended to disentangle test form differences from group differences. For this reason common items need to be a surrogate of the rest of the test form so performance on the common items can be compared between the groups to make inferences about the groups' relative abilities. However, while the CI-NEG is a data collection design that can be employed to account for group differences when group equivalence cannot be assumed, previous research indicates that there are limits as to how different the groups can be before the robustness of equating is threatened (see Cook & Peterson, 1987, for a review). Kolen and Brennan (2004, p. 286) suggested that group differences of 0.3θ or higher can cause significant problems to equating, and that IRT equating is particularly sensitive to group differences. Perhaps it is because large group differences give greater opportunity for error in determining form differences. IRT might be more susceptible to group differences because this factor plays a key role in how the θ scales are aligned in the transformation step.

The results of this research provide further evidence on the limits of IRT equating to handle group differences, at least beyond a difference of 0.1θ, even when the forms are perfectly parallel. Group equivalence interacted with both the correlation between dimensions and the location of the items measuring the second dimension. Among both the parallel and nonparallel forms results, errors generally increased[9] in size at any given correlation between dimensions as the group nonequivalence got larger.

The effect of the location of the items measuring the second dimensions also seemed to be amplified by group differences, with the CI location having the largest errors, followed by the UI location, and then the BL location. The differences between the locations tended to be larger as the groups became increasingly more nonequivalent. Given the limitations of IRT equating at handling group differences, it is not surprising that having a more balanced and representative set of common items is advantageous. Klein and Jarjoura (1985) found a similar advantage of a well-designed common items set when groups differed from one another when equating using linear and equipercentile methods.

Location of items measuring second dimension. This variable addresses the fourth research question: Does the location of the items (unique, common, both unique and

[9] There was no trend in percent SDTM observed in the nonparallel forms results, but the general trend was observed in the parallel percent SDTM, and parallel and nonparallel MAD results.

common locations) containing the second dimension have an effect on the magnitude of equating error, or the proportion of examinees that are affected by equating error? In this study, the location of items measuring the second dimension was really a proxy for test design. Does test design influence the robustness of IRT equating when multiple dimensions are present? The results for both simulated and real data suggest that the design of the common item set does make a difference to the amount of equating error observed. In general, the smallest errors were associated with the BL conditions, where the second dimension items were present among both the common and unique items. The results of this study are similar to those observed by Camilli et al. (1995).

However, this variable was evaluated in greater detail in this study and, therefore, provides further information about the implications of test design. In the parallel forms conditions, the UI location tended to produce similar results to the BL location until the correlation between dimensions was low, and the differences became larger as the groups became increasingly more nonequivalent. The CI location tended to have more error at all conditions, but appeared to be more sensitive to the correlation between dimensions and group nonequivalence. In the nonparallel forms conditions, the differences between BL and UI were much less obvious, perhaps because the overall magnitude of the error was so much larger. However, the CI conditions still had larger errors, particularly when the correlation between dimensions was low.

The relationship between the correlation between dimensions and the location of the second dimension items was even more evident in the results for the real data sets. The equated error tended to be smaller for the BL location than for the UI location which, in turn, had smaller errors than the CI location. In fact, the gain for the CI locations for both tests was negative, indicating that more error was introduced than eliminated during equating. In contrast, the gain for the UI location was positive for Test A and negative for Test B, while the gain for the BL location was positive for both tests. The difference in results between the simulated and real data may be due, in part to the correlation between dimensions on Test A and Test B, as these two forms tended to be low, which decreased the robustness of equating.

The results to both the simulated and real data analyses indicated that the location of the second dimension items became more important as the correlation between dimensions decreased, and to a lesser extent, as the groups became increasingly nonequivalent. When the correlation between dimensions was higher (test was more unidimensional), the differences between the location conditions were not as great as when the correlation between dimensions were lower.

The advantage of the BL location, particularly when the correlation between dimensions was low, appears to be due to the need for representation of the unique items by the common items. When both dimensions were represented in the common item set, IRT equating was robust to the violation of the unidimensionality assumption. The most likely reason would be that less error is generated in the scale transformation step, where the examinee performance on the common items is used to find a transformation that is applied to all of the items.

It is interesting to note that the UI location tended to perform better than the CI location even though in both circumstances the common items set was not representative of the test as a whole. In the UI location, the common items fail to represent the second dimension. In the CI location, the common items represent a second dimension that is *not* present among the unique items. One reason why the CI performance was poorer is because the common items are the bridge between the forms. The common items allow the separation of score variability due to form difficulty from score variability due to differences in group ability or abilities. The more representative the common items are of the remaining test items, the more accurately the differences between the groups and the forms can be estimated. When the number of common items was held constant, the presence of items measuring Dimension two among the common items reduced the number of items representing the Dimension one in the CI location. In contrast, in the UI location, Dimension one was well represented in the common item set, while Dimension two, constituting a relatively smaller portion of the unique items, was not represented at all. The results of these analyses suggest the latter scenario is less susceptible to error, but must be interpreted with care given the confound between location and number of Dimension two items.

General Discussion

The results of the two studies conducted in this dissertation indicate that it is difficult, if not impossible, to describe the effect of one variable on IRT equating error in the presence of multiple dimensions without discussing the interaction of that variable with all of the other variables considered in the study design. Form parallelism, the correlation between dimensions, group equivalence, and the location of items measuring the second dimension interact to create a complicated set of outcomes. That the pattern of results observed are complex seems reasonable, given that the main purpose of the study was to more fully map the effects of multidimensionality on IRT equating results under different conditions. These new findings were intended to contribute to and build on what was already known based on previous research in this area.

Previous Research

A brief summary of the results of previous studies provides us with the following picture: past research points to the robustness of IRT equating to multidimensionality under some, but not all, conditions, including when the correlation between dimensions on the test is high (Camilli et al., 1995; DeChamplain, 1996; Dorans & Kingston, 1985; Stocking & Eignor, 1986). The location of the items measuring the second dimension was identified as a variable of interest, although it had not been studied systematically (Jodoin & Davey, 2003). What other variables might play a role in robustness under conditions of multidimensionality were not clear because the studies (with the exception of Jodoin & Davey, 2003) were conducted with real data. A simulation study conducted by Bolt (1999) confirmed that equating results had less error associated with them when the correlation between dimensions was high versus when it was low. His results also suggested that the degree of multidimensionality that test forms contain makes a relatively small contribution to the total amount of equating error that is observed. What the other factors might contribute to the error was not studied in any systematic fashion. Bolt (1999) stated that it was not clear under which conditions that equity was sufficient, and that sufficiency of equating would depend on the type and purpose of the test.

Present Study

The results of the present studies suggest that all of the variables that were studied play a role in determining the magnitude of equating error. This study replicated the results observed by Bolt (1999), but then expanded our scope of understanding about why equating might not be robust, even when the correlation between dimensions was high. It also demonstrated conditions under which IRT equating might be robust even when the correlation between dimensions was low. In the following paragraphs, the discussion will also turn to trying to explain why the levels of these variables produce the results that were observed.

When IRT true score equating using the CI-NEG design, there are three main steps in the process where error might be introduced. The first is in the parameter estimation step, where the raw data are converted to numeric representations of the item characteristics. The second step is the scale transformation, where the metrics of the two sets of parameters (one set from each form) are aligned onto a common scale via comparison of performance on the common items. The final step is the actual equating, where the scores from one form are adjusted to be equivalent to the scores on the other form. Because these steps are linked, it is also possible that error can be cumulative. That is, small errors introduced in one step can be added to small errors introduced in another step, so that the final equating results have larger errors.

In the present study, the conditions with parallel forms were intended to act as control conditions, to gain a sense of how much equating error observed in the studies was random error. What was not expected was that under some conditions, equating error would be large enough to be classified as not robust. Kolen and Brennan (2004, p. 11), stated, when referring to the equity property of equating, that "...if identical forms could be constructed, there would be no need for equating.". Perfectly parallel (identical) forms do not require equating because they are equitable. Rather, any difference in performance between separate administrations of parallel forms is attributable solely to group differences.

When equating using parallel forms in an IRT design, it would be expected that the equating errors would be small, because no adjustments to scores should be required once the scales are transformed onto a common metric. Any error that is observed under these conditions would have originated in either the parameter estimation or the scale transformation step. The error associated with parallel forms conditions in the simulation seem to have come from several different variables that acted cumulatively. The results reproduced the outcomes in Bolt (1999), where equating results were robust when the dimensions on the test were highly correlated. But equating results were also robust when the correlation was low so long as the groups were equivalent and the anchor set was representative of both dimensions on the test. Each of the other three variables seemed to exert its own effect.

As the dimensions on a test become more distinct from each other, parameter estimation is more prone to error because the model that is fit to the items will not force the creation of a reference composite (Wang, 1986). Ackerman (1987, 1989) demonstrated that greater dimensional distinctness increases error associated with the parameter estimates. Presumably this is because the parameter estimation procedures attempt to capture the overall picture of all of the items in the composite, when the items measuring Dimension two clearly do not fit. This misfit would cause errors in the estimation of all items, and the error would be greater the more distinct the dimensions were from each other.

The effects of increasingly distinct dimensions was modulated by group equivalence. When groups were equivalent, equating errors were smaller and equating was more likely to be robust. As the groups became more nonequivalent, errors tended to be larger. But why does group nonequivalence matter in an equating design that is intended to account for such differences? The answer seems to be that just like equating is intended to make minor adjustments to scores, the common-items nonequivalent groups design is only intended to account for minor differences between groups within a specific population. If group differences are large enough, then the groups are no longer considered to be selected from the same population. Kolen and Brennan (2004, p. 286) suggest that differences of 0.3θ or greater between groups are large enough to introduce error into equating . Group equivalence would play a significant role in the scale transformation step, because larger differences between groups would make estimating the transformation coefficient A and B to transform scale Q to scale P more prone to error because larger adjustments would be required.

Jodoin and Davey (2003) found IRT equating to not be robust to multidimensionality when the second dimension was present only in the unique items. The present studies supports the Jodoin and Davey (2003) results, but further refined the relationship between location of the second dimension items and the presence of multiple dimensions on a test. The effects of multidimensionality were modulated by the location of the items that contained the second dimension. This result is significant because it reinforces the importance of good test design to robust equating. When tests were designed to account for both dimensions among both the unique and common items (i.e., the BL conditions), equating results tended to be robust at lower correlations than when tests included a second dimension, but only in either the unique (UI conditions) or the common items (CI conditions). Further, equating results tended to have more error, and were not robust at higher correlations than other location conditions when the second dimension was only present in the common items. The possible mechanism is that the common item set is directly involved in the estimation of the *A* and *B* scale transformation coefficients. If the common item set contains a second dimension that the unique items do not, then there are items on the anchor test that are used in assessing the relative performance of the groups that are not representative of the remaining items on the test. With fewer items to use for estimating the scale transformation coefficients, there is more likely to be error in the estimates. If the dimensions on the test are closely related, then examinee performance on the common items measuring the second dimension are likely to have some utility for predicting performance on the remaining test items. But as the dimensions become more distinct, performance on the Dimension two items are not related to performance on the dimension one items, so error is introduced into the scale transformation when they are used in the estimation step. When only the unique items contains a second dimension that the common items do not, the error in the scale transformation would arise from the common items not fully representing the remaining items on the test, meaning that estimation of performance on the Dimension two items is not adjusted appropriately. Presumably, this error is less because it affects fewer items. Having the Dimension two items in both location protects equity (creates less error) because the contribution of the Dimension two items to the scale transformation is appropriate. It is appropriate because Dimension two items exist among the unique items on the test, and all of the unique items have representation in the calculation of the scale transformation coefficients.

The form differences in the nonparallel forms conditions made the greatest contribution to equating error, based on the difference in the magnitude of error when forms were nonparallel versus parallel. The likely reason for the increase in error is two-fold. First, when forms are not parallel, they do not have an identical relationship between the common items and the remaining unique items. The difference in the relationship can contribute to error during the scale transformation. Second, when forms are not parallel, there is a greater difference between scores that must be overcome in equating.

Benefits of Equating

One of the unique aspects of this research was the use of calculating gain values for each condition. Beyond simply quantifying the magnitude of error associated with each condition, the gain values indicated how much benefit (or cost) was attributable to equating. That is, did equating provide any benefit to reducing the discrepancy between the equated test score and the target score and, therefore, improve equity?

The results of the simulation study indicated that regardless of the conditions, there was always some benefit to equating when forms were not parallel; that is, on average, scores were always closer to the target scores with equating. However, the magnitude of the benefit was small relative to the magnitude of the error, especially in terms of reducing the percent SDTM. The degree of the gain tended to get only slightly larger as the groups became more nonequivalent, with no systematic patterns attributable to changes in the correlation between dimensions or the location of the items measuring the second dimension. In contrast, the real data results demonstrated that in some cases, equating can make errors larger. When the common items were representative of the unique items on the test (BL), equating decreased error, while having a second dimension present in the common items only (CI) tended to cause scores to be further from the target scores after equating.

A further examination of parallel forms conditions also suggests that when forms are very closely parallel, additional error could be introduced by equating that would pose a greater threat to equity than not equating, depending on the correlation between dimensions, the location of the items measuring the second dimension, and the equivalence of the groups who are administered each form. If the tests are nearly parallel but these characteristics are at undesirable levels, then equating could actually make equity worse, because the errors introduced by a faulty equating conversion are larger than the original score differences between closely parallel forms. Combined, these results suggest that caution should be used when equating test forms. It should not be assumed that equating is always beneficial, especially if the dimensions on the test are distinct from each other, the groups taking each form might differ in ability, and if the forms are built such that the dimensions are not proportionately represented among both the unique and common items.

Definitions/Criteria for Robustness

As reflected by the nonparallel forms simulation and real data results, no conditions were robust according to the definitions of robustness established for the present studies. This lack of robustness extended to the 1.0 correlation conditions of the simulation, where the test was strictly unidimensional and therefore the unidimensionality assumption of IRT equating was not violated. When examining these results, a question arises: If not even the unidimensional conditions met the definitions of robustness, were the criteria set out for the analyses too stringent?

The criteria that were established for this study were based on two main considerations. First, was there a theoretical or at least rational guideline that existed in the equating literature? Dorans and Feigenbaum's (1994) score difference that matters fit that description and also seemed a reasonable expectation for equating error to be robust. Second, it also seemed likely that a more lenient definition, one that would not be considered robust, strictly speaking, but that might still be acceptable in a testing standard is that while testing programs might program was necessary. The rationale for this second be constantly striving for an ideal condition, often factors that are not controllable prevent meeting ideal conditions. Consequently, psychometricians and other testing staff are left to decide what level of error might be acceptable, given the purpose and stakes of the test. In the absence of a specific guideline, an error that was less than two percent of the total possible score was selected because less than two percent seemed like an error size that was small enough to be at least tolerable in most testing situations.

Selecting criteria for the percent score difference that matters (SDTM) was a more challenging task. No criteria for this variable existed because no published reports could be found that had employed percent SDTM. It was also a more difficult set of criteria to develop because it meant that decision had to be made as to what percentage of examinees it is acceptable to give inequitable scores. To solve this issue, the errors for the parallel forms, 1.0 correlation, equivalent groups (EG) conditions were examined to select the definition of small error that implies robustness, which resulted in the selection of 25% or less criterion. The second criterion, between 25% and 50% for medium-sized error, was selected because of the rationale that no percentage higher than 50% could possibly be considered acceptable.

An alternative, less conservative approach to selecting and developing criteria for interpreting both dependent variables would have been to use the nonparallel forms, 1.0 correlation, equivalent groups conditions as starting point. The rationale for using the results of these conditions as the basis for making this decision is that in these conditions, the assumptions of IRT equating still hold. However, upon examining the results, it was decided that even in the cases where the test forms were unidimensional and the groups were equivalent, a MAD over 3.25 score points, which translates into 4.8% difference on the 68-point scale, and particularly percent SDTM in the range of 90% and over did not seem reasonable definitions of "robust" or even "tolerable."

Conclusions

The results of this study indicate that IRT equating using a common-items nonequivalent groups design under conditions of multidimensionality can be robust, but under very restricted conditions. The forms must be parallel or nearly parallel, the groups must be very nearly randomly equivalent, and the common items must represent both dimensions that are present in the test forms to be equated. As soon as any of these variables start to deviate from these near-perfect conditions, robustness is threatened. The results of the real data analyses suggest that as conditions deteriorate, equating can actually begin to contribute more error than would be present if scores were not equated.

The results of this dissertation support the findings of Bolt (1999) who reported that the effects of multidimensionality on equating error are minor relative to other variables. Certainly in the present study, the magnitude of the differences between forms played the largest role in the equating error. When forms deviated from perfect parallelism, the magnitude of error increased markedly. Further, the group differences also played an important role. The more nonequivalent the groups, the larger the errors associated with equating. Finally, the design of the test forms, in relation to where the items measuring each dimension are located, had an influence on the degree of error observed. When the test forms were designed to have both dimensions represented in both the common and unique items, the amount of error observed was much smaller than when the items were not placed in both item sets. The increase in error associated with a decreasing correlation between dimensions was attenuated by ensuring that the common item set was a surrogate of the remaining items on the form.

Overall, equating did a poor job of removing the error that was present between the observed score and the target score. In the simulation study, errors were always reduced by equating. However, the results of the analyses of the real data sets were not so positive, as equating actually introduced more error in some of the real data conditions. Clearly, IRT equating under conditions of multidimensionality is not a "silver bullet" that can align test scores, but rather is constrained by many variables, including form differences, group differences, the correlation between dimensions, and the location of the items measuring each dimension.

Implications for Practice

Of the four variables examined in this study, the correlation between dimensions is determined by the test content (laid out in the test specifications), group differences by the examinees, and only form parallelism and location of the items measuring the second dimension by the test developer, once the test specifications are in place. Hopefully, while only two of these variables can be controlled, it is possible to be aware of the values of the other two variables so that good decisions can be made about whether or not IRT equating with multiple test dimensions is feasible.

If test specifications require that a test contain multiple distinct dimensions that are not highly correlated, then IRT equating might be suitable, provided that other variables, including form parallelism, group equivalence, and location of the items measuring each dimension, fall within reasonable parameters. However, if there is an expectation that groups will differ between form administrations (i.e., groups are not expected to be equivalent), then IRT equating is not a good solution for a test that contains multiple dimensions. For example, if it is expected that a population will improve performance over time due to improved instruction or some other factor, then the group differences that result might prove problematic for IRT equating, particularly if the correlation between dimensions on the test is also low (less than 0.7).

It is important to assess the dimensional structure of test forms and to identify which items are measuring each dimension. As these results indicate, the location of the items measuring the second dimension matters. Knowing which items measure the second dimension can help prevent unnecessary equating error by enabling the test developer to assemble the form, including the common item set, while being mindful of the dimensional structure. Testing dimensional structure can be especially helpful if the multiple dimensions do not fall exactly along the same lines as the test specifications.

While it might seem unlikely that a test developer would inadvertently include items measuring a second dimension into only the common item set, the results from the location measuring the second dimension items conditions could have implications for vertical equating applications. Briefly, vertical equating is used across different levels of abilities, as an attempt to capture growth and development of examinees over time. For example, to be able to monitor progress of students as they move through the education system, it is important to be able to compare results of standardized tests at one grade level to the results at the next grade level to see if improvements in performance are occuring. One of the issues in this type of equating is that it would be very easy for additional unintentional dimensions to be added because the populations at each grade level have different characteristics. For example, if a set of math word problems that were designed for fifth graders were used as a set of common items to a test for fourth graders, the items might measure one dimension on the fifth grade test (math ability) , but two dimensions on the fourth grade test (math and reading abilities). This difference would occur because the fifth graders, as a group, have sufficient reading ability, but the same item separates both weak and strong math and reading abilities in the fourth graders. This is an issue that would require empirical study to confirm the similarity in data structure across grade levels.

The ability to design parallel forms will be limited by the availability and quality of pilot test data. Data collected from well-designed pilot studies can be used to construct tests that are sensitive to statistical characteristics of the items if the data are used to conduct dimensional assessments. Pilot testing can be an expensive addition to a testing program, but would be highly recommended for those wishing to use IRT equating on complex assessments.

The results of this study, while informative, provide little in the way of specific guidelines when deciding whether or not IRT equating would be an appropriate method for a given testing situation. Instead, the results provide some points for consideration when making a decision and suggest that when tests are well-designed, forms closely parallel, and groups are randomly or at least close to equivalent, multidimensionality is generally well tolerated. The results also suggest that if these factors are not well controlled, multidimensionality is a relatively small problem compared to the errors contributed by the other factors. Additionally, there are many other potential factors, such as the test length, anchor test length (number of common items), as well as other as yet unidentified factors that might also complicate the issue of robustness. Much research is required before specific guidelines for IRT equating test forms with multiple dimensions will be possible.

Implications for Further Research

Form Parallelism

The results of the current study suggest that when forms are parallel, multiple test dimensions are reasonably well tolerated. One simple way in which test forms can be made more parallel is to add proportionally more common items. That is, to the extent possible, a new form should contain as many of the same items as the reference form as possible. In the current studies, only about one-quarter of the items were common between the new and reference forms, but perhaps better equating results would be possible with a larger proportion of common items. A more systematic investigation of this variable using a simulation design, varying form differences between 0.0 and 0.2θ would be helpful to determine how closely parallel forms must be in order for IRT equating to be robust. A guide to the number of common items relative to the total number of test items required to be robust, depending on the degree of test multidimensionality, would be helpful in determining if IRT equating would be feasible for a specific testing purpose. For example, if equating was robust to multidimensionality under a set of conditions, but only if the forms contained 90% common items, IRT equating might not be desirable because it requires the reuse of too many items to be useful for security reasons, whereas a 50% requirement might be more reasonable.

Group Equivalence

In the simulation and real data for Test B, the groups were nonequivalent on both Dimensions one and two. Different results may occur if the group differences were located on only one dimension. A systematic examination of this variable would be useful, especially considering that if dimensions are distinct it is unlikely that the groups would differ from one another in the same manner on each dimension.

Anchor Test Structure

A potential confound in this study that should be addressed is the number of items measuring the second dimension on each form confounded with the location of those items on the form. A systematic investigation of the format of the common items set could be designed (with variables like anchor test length, number of items in each dimension, and proportional representativeness of each dimension on the common items set) in a simulation design. This study could help better understand why the location of the Dimension two items mattered to equity under conditions of multidimensionality.

Dimensional Structure

The dimensional structure employed in this research was limited to two dimensions and test forms consisted only of items that distinctly measured either Dimension one or Dimension two, but not both. This simplistic dimensional structure is not realistic. When a statistical dimensionality assessment is conducted, many items will be identified that measure multiple dimensions simultaneously. For example, 26 of of 90 original LSAT items used as a basis for generating the simulated data in the present studies had complex 2-dimensional structure on the basis of their angular direction (between 20 and 70 degrees). The effect of these items on the robustness of IRT equating has yet to be explored.

Bibliography

Ackerman, T. (1987). *The use of unidimensional item parameter estimates of multidimensional items in adaptive testing* (Research Rep. No. 87-13). Iowa City, IA: American College Testing Program.

Ackerman, T. (1989). Unidimensional IRT calibration of compensatory and noncompensatory multidimensional items. *Applied Psychological Measurement*, *13*, 113-127.

Ackerman, T. A. (1991). The use of unidimensional parameter estimates of multidimensional items in adaptive testing. *Applied Psychological Measurement, 15,* 13-24.

Ackerman, T. A. (2004). M2GEN2. [computer software]. University of North Carolina-Greensboro.

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

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

Bock, R.D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46,* 443-459.

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

Bolt, D. M., & Lall, V. F. (2003). Estimation of Compensatory and Noncompensatory Multidimensional Item Response Models Using Markov Chain Monte Carlo. *Applied Psychological Measurement, 27,* 395-514.

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

Camilli, G., Wang, M-m., & Fesq, J. (1995). The effects of dimensionality on equating the Law School Admission Test. *Journal of Educational Measurement, 32*, 79-96.

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.

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

Divgi, D. R. (1981). Model-free evaluation of equating and scaling. *Applied Psychological Measurement*, *5,* 203-208.

Dorans, N. J. (1990). Equating methods and sampling designs. *Applied Measurement in Education, 3,* 3-17.

Dorans, N. J., & Feigenbaum, M. D. (1994). Equating issues engendered by changes to the SAT and PSAT/NMSQT. (ETS Research Memorandum 94-10). Princeton, NJ: Educational Testing Service.

Dorans, N. J., & Kingston, N. M. (1985). The effects of violations of unidimensionality on the estimation of item and ability parameters and on item response theory equating of the GRE verbal scale. *Journal of Educational Measurement, 22*, 249-262.

Douglas, J., Kim, H., & Roussos, L.(1999). LSAT dimensionality analysis for the December 1991, June 1992 and October 1992 administration. *LSAC Statistical Report 95-05*. Newtown, PA: Law School Admission Council.

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Finch, H. (2002). Comparison of the performance of NOHARM and conditional covariance methods of dimensionality assessment: Type I, power and item dimension clustering. Unpublished doctoral dissertation, University of South Carolina.

Fraser, C. (1988). NOHARM. [computer software]. Armidale, NSW: The University of New England.

Fraser, C., & McDonald, R. P. (1988). NOHARM: Least squares item factor analysis. *Multivariate Behavioral Research, 23*, 267-269.

Gierl, M. J., Leighton, J. P., & Tan, X. (2006, April). *Evaluating DETECT classification accuracy and consistency when data display complex structure.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

Gulliksen, H. (1950).*Theory of mental tests.* Hillsdale, NJ,: Lawrence Erlbaum.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research, 22,* 144-49.

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

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

Hanson, B., & Zeng, L. (1995a). PIE: A Computer Program for IRT Equating (Version 1.0). [computer software]. Iowa City, IA: ACT Inc.

Hanson, B., & Zeng, L. (1995b). ST: A Computer Program for IRT Scale Transformation (Version 1.0). [computer software]. Iowa City, IA: ACT Inc.

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

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

Hattie, J. A. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement. 9*, 139-164.

Hwong, S., Im, H., Si, K., Seong, T., & Kim, K. (April, 2005). *Comparison of scale linking results in a mixed-format test under the different conditions of common item format and test dimensionality.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Montreal, QC.

IMSL (1994). IMSL MATH/LIBRARY User's Manual (Version 3.0). [computer software]. Houston, TX: IMSL.

Jodoin, M .J. & Davey, T. (2003, April). *A multidimensional simulation approach to investigate the robustness of IRT common item equating.* Paper presented at the annual meeting of the American Educational Research Association and the National Council on Measurement in Education, Chicago, IL.

Klein, L. W., & Jarjoura, D. (1985). The importance of content representation for common-item equating with nonrandom groups. *Journal of Educational Measurement, 2,*197-206.

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

Lee, S-y., & Jennrich, R. I. (1979). A study of algorithms for covariance structure analysis with specific comparisons using factor analysis. *Psychometrika, 44*, 99-113.

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

Livingston, S. L., Dorans, N. J., & Wright, N. K. (1990). What combination of sampling and equating methods works best? *Applied Measurement in Education, 3,* 73-95.

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

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17,* 179-93.

Luecht, R. (1992). 2DEAP Version 0.1. [computer software]. University of North Carolina-Greensboro.

Luecht, R., & Miller, T. R. (1992). Unidimensional calibrations and interpretations of composite traits for multidimensional tests. *Applied Psychological Measurement, 16,* 279-93.

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

McDonald, R. P. (1982). Linear versus nonlinear models in item response theory. *Applied Psychological Measurement, 6,* 379-96.

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

Meara, K., Robin, F., & Sireci, S. G. (2000). Using multidimensional scaling to assess the dimensionality of dichotomous item data. *Multivariate Behavioral Research, 35,* 229-259.

Messick, S. (1989). Validity. In R. L. Linn (Ed). *Educational measurement (3rd ed.).* (pp. 13-103). New York: Macmillan.

Mislevy, R., & Bock, R. D. (1990). BILOG [Computer software]. St. Paul, MN: Assessment Systems Corporation.

Nandakumar, R. (1991). Traditional dimensionality versus essential dimensionality. *Journal of Educational Measurement, 28,* 99-117.

Nandakumar, R. (1993). Assessing essential unidimensionality of real data. *Applied Psychological Measurement, 17,* 29-38.

Nandakumar, R., & Stout, W. F. (1993). Refinements of Stout's procedure for assessing latent trait unidimensionality. *Journal of Educational Statistics, 18,* 41-68.

Oshima, T. C, Davey, T., & Lee, K.(2000). Multidimensional linking: Four practical approaches. *Journal of Educational Measurement, 37*(4), 357-373.

Press, W. H., Teuklosky, S. A., Vetterling, W. T., & Flannery, B. P. (1994). *Numerical Recipes in C, 2nd ed.* New York: Cambridge University Press.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4,* 207-230.

Reckase, M. D. (1985).The difficulty of test items that measure more than one ability. *Applied Psychological Measurement 9,* 401-412.

Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement, 21,* 25-36.

Reckase, M. D.(2006, September 14). Ambiguities in the interpretations of dimensions in item response data: coordinates, dimensions, factors and hypothetical constructs. Paper presented at Educational Testing Service, Princeton, New Jersey.

Ricker, K. L., and Von Davier, A. A. (2006, April). *The impact of anchor test length on equating results in a non-equivalent groups design.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

Roussos, L. A., & Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel Type I error performance. *Journal of Educational Measurement. 33,* 215-230.

Skaggs, G., & Lissitz, R. W. (1986). An exploration of the robustness of four test equating models. *Applied Psychological Measurement, 10,* 303-317.

Stocking, M. L., & Eignor, D. R. (1986). The impact of difference ability distributions on IRT pre-equating. (Research Report 86-49). Princeton, NJ: Educational Testing Service.

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

Stout, W. F. (1987) A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, *52*, 589-617.

Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika, 55,* 293-325.

Stout, W., Habing, B., Douglas, J., & Kim, H. R. (1996). Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement, 20,* 331-354.

Thomasson, G. L. (1993). The asymptotic equating methodology and other testing evaluation procedures. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.

Tong, Y., & Kolen, M. J. (2005). Assessing equating results on different equating criteria. *Applied Psychological Measurement, 29*, 418-432.

Walker, C., Gierl, M., Ackerman, T., Ricker, K. L., & Gosz, J. K. (2003). *The effect of model misspecification on exploratory and confirmatory models in multidimensional item response theory.* In C. Walker and M. Gierl (Chairs), Applications and Practical Considerations of Multidimensional Item Response Theory. Symposium conducted at the annual meeting of the National Council of Measurement in Education, Chicago, IL.

von Davier, A. A., Holland, P. W., Livingston S. A., Casabianca, J., Grant, M. C., & Martin, K. 2005). *An evaluation of the kernel equating method: A special study with pseudo-tests constructed from real test data.* Manuscript in progress.

Wang, M. (1986). Fitting a unidimensional model to multidimensional item response data: The effects of latent space misspecification on the application of IRT. Unpublished doctoral dissertation, University of Iowa.

Wang, T., Hanson, B. A., & Harris, D. J. (2000). The effectiveness of circular equating as a criterion for evaluating equating. *Applied Psychological Measurement, 24,* 195-210.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187-213.

Appendix A

Item Parameters for simulation study Forms X and Y for test constructed with second dimension present among common and unique items (BL)

Common Items			*Unique "X" Items*			*Unique "Y" Items*		
a_1	a_2	d	a_1	a_2	d	a_1	a_2	d
0.840	0.073	-0.341	0.701	0.000	-0.818	0.401	-0.010	-0.618
0.572	0.026	-0.062	0.513	0.001	-0.087	0.613	-0.001	0.013
0.487	0.099	-0.646	0.431	0.004	-1.456	0.431	-0.002	-1.156
0.488	0.097	-0.024	0.611	0.014	-0.560	0.491	0.044	-0.369
0.740	0.075	0.300	0.725	0.017	-0.732	0.725	0.048	-0.632
0.454	0.045	-1.279	0.628	0.017	-0.621	0.728	-0.025	-0.321
0.554	0.083	-0.099	1.041	0.031	-0.712	0.921	0.031	-0.612
0.466	0.049	-0.949	0.525	0.016	-0.153	0.425	0.075	0.042
0.166	*1.316*	*0.353*	0.733	0.025	-0.696	0.780	0.075	-0.496
0.226	*0.866*	*-0.360*	0.605	-0.024	-0.300	0.505	0.064	-0.400
0.110	*0.899*	*-0.251*	0.543	0.025	-0.909	0.560	0.023	-0.629
0.125	*0.873*	*-0.656*	0.508	0.038	-0.149	0.368	0.067	-0.049
0.072	*1.175*	*-0.764*	0.681	-0.052	0.046	0.741	0.050	0.246
0.060	*0.536*	*-0.760*	0.321	0.025	-1.135	0.521	0.070	-0.935
0.121	*0.822*	*-0.714*	0.510	0.041	-0.463	0.610	0.091	-0.263
0.084	*0.949*	*-0.635*	0.371	0.034	0.425	0.471	0.046	0.625
			0.518	0.050	-0.394	0.718	0.060	-0.194
			0.608	0.059	-0.314	0.408	0.069	-0.114
			0.327	0.033	-0.946	0.427	0.093	-0.676
			0.609	0.063	-0.839	0.559	0.073	-0.639
			0.724	0.075	-0.998	0.824	-0.085	-0.798
			0.479	0.050	0.115	0.379	0.080	0.335
			0.505	0.053	-0.524	0.555	0.046	-0.324
			0.470	0.060	-1.349	0.390	0.050	-1.149
			0.624	0.086	-0.916	0.524	0.076	-0.716
			0.543	0.077	-0.109	0.443	0.087	0.091
			0.817	0.214	-0.511	0.613	0.075	-0.202
			0.534	0.143	-0.890	0.465	0.075	-1.093
			0.440	0.120	0.478	0.703	0.084	0.410
			0.640	0.175	-0.391	0.412	-0.074	0.165
			0.522	0.143	-0.220	0.668	0.131	-0.236
			0.337	0.093	-0.796	0.744	0.145	0.905
			0.566	0.167	-1.423	0.681	0.187	1.144
			0.808	0.249	-0.134	0.680	0.186	-1.112
			0.492	0.171	-0.468	0.658	0.151	-0.262
			0.628	0.197	1.400	0.560	0.186	-0.319
			0.115	*0.500*	*-0.775*	*0.127*	*0.300*	*-0.585*
			0.205	*0.891*	*-0.430*	*0.220*	*1.189*	*-0.221*
			0.161	*0.833*	*-0.639*	*0.163*	*0.688*	*-0.440*
			0.137	*0.741*	*-0.812*	*0.143*	*0.811*	*-0.611*
			0.202	*0.913*	*-0.029*	*0.235*	*1.111*	*0.165*
			0.119	*0.730*	*-0.705*	*0.089*	*0.753*	*-0.524*
			0.113	*0.735*	*-0.017*	*0.113*	*0.712*	*0.186*
			0.165	*1.182*	*-0.768*	*0.155*	*1.067*	*-0.554*
			0.182	*1.362*	*0.416*	*0.188*	*1.366*	*0.611*
			0.110	*1.174*	*-0.995*	*0.116*	*1.097*	*-0.800*
			0.128	*1.213*	*-0.470*	*0.056*	*1.258*	*-0.270*
			0.120	*1.202*	*-0.603*	*0.111*	*1.111*	*-0.403*
			0.032	*0.458*	*-1.145*	*0.073*	*0.471*	*-0.945*
			0.056	*0.944*	*-0.666*	*0.078*	*0.956*	*-0.466*
			0.051	*1.252*	*-0.828*	*0.038*	*1.155*	*-0.628*
			0.023	*0.738*	*-0.713*	*0.049*	*0.838*	*-0.513*

Note. Dimension two items are noted in *italics*.

Appendix B

Item Parameters for simulation study Forms X and Y for test constructed with second dimension present among unique items only (UI)

Common Items			*Unique "X" Items*			*Unique "Y" Items*		
a_1	a_2	d	a_1	a_2	d	a_1	a_2	d
0.459	0.005	-0.798	0.701	0.000	-0.818	0.401	-0.010	-0.618
0.847	0.123	-0.311	0.513	0.001	-0.087	0.613	-0.001	0.013
0.634	0.065	-0.790	0.431	0.004	-1.456	0.431	-0.002	-1.156
0.340	0.013	0.578	0.611	0.014	-0.560	0.491	0.044	-0.369
0.740	0.073	-0.291	0.725	0.017	-0.732	0.725	0.048	-0.632
0.472	0.056	0.088	0.628	0.017	-0.621	0.728	-0.025	-0.321
0.437	0.099	-0.596	1.041	0.031	-0.712	0.921	0.031	-0.612
0.488	0.097	0.026	0.525	0.016	-0.153	0.425	0.075	0.042
0.740	0.075	0.350	0.733	0.025	-0.696	0.780	0.075	-0.496
0.454	0.045	-1.229	0.605	-0.024	-0.300	0.505	0.064	-0.400
0.554	0.083	-0.049	0.543	0.025	-0.909	0.560	0.023	-0.629
0.466	0.059	-0.899	0.508	0.038	-0.149	0.368	0.067	-0.049
0.808	0.035	-0.023	0.681	-0.052	0.046	0.741	0.050	0.246
0.492	0.111	-0.268	0.321	0.025	-1.135	0.521	0.070	-0.935
0.628	0.167	-1.396	0.510	0.041	-0.463	0.610	0.091	-0.263
0.601	0.012	-0.518	0.371	0.034	0.425	0.471	0.046	0.625
			0.518	0.050	-0.394	0.718	0.060	-0.194
			0.608	0.059	-0.314	0.408	0.069	-0.114
			0.327	0.033	-0.946	0.427	0.093	-0.676
			0.609	0.063	-0.839	0.559	0.073	-0.639
			0.724	0.075	-0.998	0.824	-0.085	-0.798
			0.479	0.050	0.115	0.379	0.080	0.335
			0.505	0.053	-0.524	0.555	0.046	-0.324
			0.470	0.060	-1.349	0.390	0.050	-1.149
			0.624	0.086	-0.916	0.524	0.076	-0.716
			0.543	0.077	-0.109	0.443	0.087	0.091
			0.817	0.214	-0.511	0.613	0.075	-0.202
			0.534	0.143	-0.890	0.465	0.075	-1.093
			0.440	0.120	0.478	0.703	0.084	0.410
			0.640	0.175	-0.391	0.412	-0.074	0.165
			0.522	0.143	-0.220	0.668	0.131	-0.236
			0.337	0.093	-0.796	0.744	0.145	0.905
			0.566	0.167	-1.423	0.681	0.187	1.144
			0.808	0.249	-0.134	0.680	0.186	-1.112
			0.492	0.171	-0.468	0.658	0.151	-0.262
			0.628	0.197	1.400	0.560	0.186	-0.319
			0.115	*0.500*	*-0.775*	*0.127*	*0.300*	-0.618
			0.205	*0.891*	*-0.430*	*0.220*	*1.189*	0.013
			0.161	*0.833*	*-0.639*	*0.163*	*0.688*	-1.156
			0.137	*0.741*	*-0.812*	*0.143*	*0.811*	-0.369
			0.202	*0.913*	*-0.029*	*0.235*	*1.111*	-0.632
			0.119	*0.730*	*-0.705*	*0.089*	*0.753*	-0.321
			0.113	*0.735*	*-0.017*	*0.113*	*0.712*	-0.612
			0.165	*1.182*	*-0.768*	*0.155*	*1.067*	0.042
			0.182	*1.362*	*0.416*	*0.188*	*1.366*	-0.496
			0.110	*1.174*	*-0.995*	*0.116*	*1.097*	-0.400
			0.128	*1.213*	*-0.470*	*0.056*	*1.258*	-0.629
			0.120	*1.202*	*-0.603*	*0.111*	*1.111*	-0.049
			0.032	*0.458*	*-1.145*	*0.073*	*0.471*	0.246
			0.056	*0.944*	*-0.666*	*0.078*	*0.956*	-0.935
			0.051	*1.252*	*-0.828*	*0.038*	*1.155*	-0.263
			0.023	*0.738*	*-0.713*	*0.049*	*0.838*	0.625

Note. Dimension two items are noted in *italics*.

Appendix C

Item Parameters for simulation study Forms X and Y for test constructed with second dimension present among common items only (CI)

Common Items			*Unique "X" Items*			*Unique "Y" Items*		
a_1	a_2	d	a_1	a_2	d	a_1	a_2	d
0.840	0.073	-0.341	0.701	0.000	-0.818	0.401	-0.010	-0.618
0.572	0.026	-0.062	0.513	0.001	-0.087	0.613	-0.001	0.013
0.487	0.099	-0.646	0.431	0.004	-1.456	0.431	-0.002	-1.156
0.488	0.097	-0.024	0.611	0.014	-0.560	0.491	0.044	-0.369
0.740	0.075	0.300	0.725	0.017	-0.732	0.725	0.048	-0.632
0.454	0.045	-1.279	0.628	0.017	-0.621	0.728	-0.025	-0.321
0.554	0.083	-0.099	1.041	0.031	-0.712	0.921	0.031	-0.612
0.466	0.049	-0.949	0.525	0.016	-0.153	0.425	0.075	0.042
0.166	*1.316*	*0.353*	0.733	0.025	-0.696	0.780	0.075	-0.496
0.226	*0.866*	*-0.360*	0.605	-0.024	-0.300	0.505	0.064	-0.400
0.110	*0.899*	*-0.251*	0.543	0.025	-0.909	0.560	0.023	-0.629
0.125	*0.873*	*-0.656*	0.508	0.038	-0.149	0.368	0.067	-0.049
0.072	*1.175*	*-0.764*	0.681	-0.052	0.046	0.741	0.050	0.246
0.060	*0.536*	*-0.760*	0.321	0.025	-1.135	0.521	0.070	-0.935
0.121	*0.822*	*-0.714*	0.510	0.041	-0.463	0.610	0.091	-0.263
0.084	*0.949*	*-0.635*	0.371	0.034	0.425	0.471	0.046	0.625
			0.518	0.050	-0.394	0.718	0.060	-0.194
			0.608	0.059	-0.314	0.408	0.069	-0.114
			0.327	0.033	-0.946	0.427	0.093	-0.676
			0.609	0.063	-0.839	0.559	0.073	-0.639
			0.724	0.075	-0.998	0.824	-0.085	-0.798
			0.479	0.050	0.115	0.379	0.080	0.335
			0.505	0.053	-0.524	0.555	0.046	-0.324
			0.470	0.060	-1.349	0.390	0.050	-1.149
			0.624	0.086	-0.916	0.524	0.076	-0.716
			0.543	0.077	-0.109	0.443	0.087	0.091
			0.817	0.214	-0.511	0.613	0.075	-0.202
			0.534	0.143	-0.890	0.465	0.075	-1.093
			0.440	0.120	0.478	0.703	0.084	0.410
			0.640	0.175	-0.391	0.412	-0.074	0.165
			0.522	0.143	-0.220	0.668	0.131	-0.236
			0.337	0.093	-0.796	0.744	0.145	0.905
			0.566	0.167	-1.423	0.681	0.187	1.144
			0.808	0.249	-0.134	0.780	0.186	-1.112
			0.492	0.171	-0.468	0.658	0.151	-0.262
			0.628	0.197	1.400	0.560	0.186	-0.319
			0.451	0.081	-0.154	0.601	0.050	-0.151
			0.189	0.026	-1.233	0.388	0.066	-1.033
			0.847	0.072	-0.711	0.817	0.060	-0.311
			0.543	0.081	-0.109	0.534	0.080	-0.690
			0.461	0.082	-0.478	0.440	0.061	-0.336
			0.650	0.057	-0.591	0.640	0.060	-0.191
			0.672	0.061	-0.620	0.522	0.101	-0.020
			0.268	0.033	-0.896	0.337	0.093	-0.596
			0.377	0.071	-0.134	0.257	0.071	0.096
			0.610	0.076	0.160	0.590	0.034	0.280
			0.574	0.063	-1.529	0.514	0.021	-1.529
			0.784	0.083	-0.549	0.845	0.130	-0.049
			0.566	0.147	-1.523	0.566	0.060	-1.223
			0.878	0.089	-0.134	0.827	0.089	0.066
			0.552	0.011	-0.368	0.492	0.141	-0.168
			0.728	0.024	1.100	0.628	0.098	1.250

Note. Dimension two items are noted in *italics*.

Appendix D

Item Parameters for Test A Forms X and Y for test constructed with second dimension present among common and unique items (BL)

Common Items			*Unique "X" Items*			*Unique "Y" Items*		
a_1	a_2	d	a_1	a_2	d	a_1	a_2	d
0.189	*0.297*	*-0.339*	*0.247*	*0.144*	*1.081*	*-0.186*	*0.348*	*-0.019*
0.105	*0.263*	*1.095*	*0.477*	*0.453*	*1.526*	*0.189*	*0.125*	*0.875*
-0.008	*0.181*	*0.922*	*0.154*	*0.015*	*0.381*	*0.358*	*0.192*	*1.320*
0.281	*0.264*	*1.486*	*0.113*	*0.343*	*0.264*	*0.347*	*0.246*	*0.760*
0.038	*0.118*	*-0.714*	*0.372*	*0.291*	*1.388*	*-0.050*	*0.376*	*-0.323*
0.050	0.200	0.075	*0.318*	*0.233*	*1.171*	*0.122*	*0.132*	*0.861*
0.119	0.037	-0.438	*-0.036*	*0.691*	*-0.218*	*0.307*	*0.279*	*0.992*
0.308	0.305	0.166	*0.081*	*-0.013*	*-0.086*	*0.229*	*0.128*	*1.223*
0.486	0.443	1.248	*0.214*	*0.174*	*-0.435*	*0.309*	*0.308*	*0.819*
0.724	0.627	1.988	*0.080*	*0.532*	*0.722*	*0.079*	*0.340*	*-0.753*
			0.139	0.122	1.675	0.342	0.223	0.443
			0.235	0.189	0.051	0.259	0.301	0.872
			0.338	0.116	0.089	0.282	0.276	2.065
			0.089	-0.041	-0.282	0.474	0.370	0.243
			0.247	0.210	0.447	0.296	0.098	0.419
			0.162	-0.004	0.112	0.453	0.248	1.355
			0.259	0.158	0.363	0.212	0.145	1.329
			0.092	0.068	0.898	0.105	0.120	0.021
			0.338	0.162	1.176	0.106	0.140	1.404
			0.099	0.147	0.915	0.187	0.150	1.788
			0.196	-0.003	0.129	0.216	0.140	0.127
			0.422	0.201	1.247	0.384	0.225	1.271
			0.278	0.118	0.620	0.250	0.214	1.177
			0.244	0.115	0.872	0.435	0.389	0.369
			0.056	0.191	-0.058	0.486	0.451	1.997
			0.233	0.198	0.302	0.486	0.471	2.434
			0.230	0.008	0.423	0.229	0.283	-0.187
			0.434	0.355	0.738	0.137	0.312	0.474
			0.328	0.238	0.637	0.446	0.304	-0.226
			0.373	0.198	0.699	0.468	0.392	0.359
			0.051	0.212	-0.615	0.216	0.069	0.762

Note. Dimension two items are noted in *italics*. For the real data analyses, dimension one and two items were determined by content.

Appendix E

Item Parameters for Test A Forms X and Y for test constructed with second dimension present among unique items only (UI)

Common Items			*Unique "X" Items*			*Unique "Y" Items*		
a_1	a_2	d	a_1	a_2	d	a_1	a_2	d
0.119	0.037	-0.440	*0.247*	*0.144*	*1.081*	*-0.190*	*0.348*	*-0.020*
0.267	0.230	0.554	*0.477*	*0.453*	*1.526*	*0.189*	*0.125*	*0.875*
0.222	0.132	1.280	*0.154*	*0.015*	*0.381*	*0.358*	*0.192*	*1.320*
0.119	0.037	-0.440	*0.113*	*0.343*	*0.264*	*0.347*	*0.246*	*0.760*
0.275	0.290	0.101	*0.372*	*0.291*	*1.388*	*-0.050*	*0.376*	*-0.320*
0.223	0.000	0.436	*0.318*	*0.233*	*1.171*	*0.122*	*0.132*	*0.861*
0.489	0.460	2.267	*-0.036*	*0.691*	*-0.218*	*0.307*	*0.279*	*0.992*
0.308	0.305	0.166	*0.081*	*-0.013*	*-0.086*	*0.229*	*0.128*	*1.223*
0.137	0.312	0.474	*0.214*	*0.174*	*-0.435*	*0.309*	*0.308*	*0.819*
0.215	0.043	-0.040	*0.080*	*0.532*	*0.722*	*0.079*	*0.340*	*-0.750*
0.724	0.627	1.988	0.139	0.122	1.675	0.342	0.223	0.443
			0.235	0.189	0.051	0.050	0.200	0.075
			0.338	0.116	0.089	0.282	0.276	2.065
			0.089	-0.041	-0.282	0.474	0.370	0.243
			0.247	0.210	0.447	0.296	0.098	0.419
			0.162	-0.004	0.112	0.453	0.248	1.355
			0.259	0.158	0.363	0.212	0.145	1.329
			0.092	0.068	0.898	0.105	0.120	0.021
			0.338	0.162	1.176	0.106	0.140	1.404
			0.099	0.147	0.915	0.187	0.150	1.788
			0.196	-0.003	0.129	0.216	0.140	0.127
			0.422	0.201	1.247	0.384	0.225	1.271
			0.278	0.118	0.620	0.250	0.214	1.177
			0.244	0.115	0.872	0.435	0.389	0.369
			0.056	0.191	-0.058	0.486	0.451	1.997
			0.233	0.198	0.302	0.486	0.443	1.248
			0.230	0.008	0.423	0.486	0.471	2.434
			0.434	0.355	0.738	0.229	0.283	-0.187
			0.328	0.238	0.637	0.446	0.304	-0.226
			0.373	0.198	0.699	0.468	0.392	0.359
			0.051	0.212	-0.615	0.216	0.069	0.762

Note. Dimension two items are noted in *italics*. For the real data analyses, dimension one and two items were determined by content.

Appendix F

Item Parameters for Test A Forms X and Y for test constructed with second dimension present among common items only (CI)

Common Items			*Unique "X" Items*			*Unique "Y" Items*		
a_1	a_2	d	a_1	a_2	d	a_1	a_2	d
0.189	*0.297*	*-0.340*	0.139	0.122	1.675	0.342	0.223	0.443
0.105	*0.263*	*1.095*	0.027	-0.002	-0.371	0.282	0.276	2.065
-0.010	*0.181*	*0.922*	0.267	0.230	0.554	0.174	0.155	0.755
0.281	*0.264*	*1.486*	0.296	0.236	1.724	0.474	0.370	0.243
0.038	*0.118*	*-0.710*	0.235	0.189	0.051	0.222	0.132	1.280
0.050	0.200	0.075	0.338	0.116	0.089	0.275	0.110	1.123
0.119	0.037	-0.438	0.275	0.290	0.101	0.261	0.061	-0.262
0.308	0.305	0.166	0.089	-0.041	-0.282	0.259	0.301	0.872
0.486	0.443	1.248	0.247	0.210	0.447	0.223	0.000	0.436
0.724	0.627	1.988	0.162	-0.004	0.112	0.296	0.098	0.419
			0.259	0.158	0.363	0.453	0.248	1.355
			0.123	0.096	0.362	0.212	0.145	1.329
			0.092	0.068	0.898	0.105	0.120	0.021
			0.338	0.162	1.176	0.106	0.140	1.404
			0.099	0.147	0.915	0.187	0.150	1.788
			0.196	-0.003	0.129	0.216	0.140	0.127
			-0.029	0.030	1.671	0.457	0.446	2.568
			0.422	0.201	1.247	0.384	0.225	1.271
			0.125	0.184	-0.879	0.250	0.214	1.177
			0.278	0.118	0.620	0.435	0.389	0.369
			0.244	0.115	0.872	0.486	0.451	1.997
			0.258	0.061	0.586	0.432	0.320	1.212
			0.056	0.191	-0.058	0.489	0.460	2.267
			0.233	0.198	0.302	0.486	0.471	2.434
			0.128	-0.049	-0.153	0.258	0.172	0.639
			0.230	0.008	0.423	0.229	0.283	-0.187
			0.434	0.355	0.738	0.418	0.295	0.379
			0.328	0.238	0.637	0.137	0.312	0.474
			0.373	0.198	0.699	0.215	0.043	-0.038
			0.051	0.212	-0.615	0.446	0.304	-0.226
			0.216	0.069	0.762	0.468	0.392	0.359

Note. Dimension two items are noted in *italics*. For the real data analyses, dimension one and two items were determined by content.

Appendix G

Item Parameters for Test B Forms X and Y for test constructed with second dimension present among common and unique items (BL)

Common Items			*Unique "X" Items*			*Unique "Y" Items*		
a_1	a_2	d	a_1	a_2	d	a_1	a_2	d
0.709	*0.408*	*-0.403*	0.068	-0.035	-0.052	0.433	0.000	0.873
0.367	0.360	1.947	0.369	0.185	0.012	0.346	0.203	0.705
0.679	0.384	-0.042	0.255	0.160	0.209	0.398	0.390	1.724
0.323	0.226	-0.280	0.141	-0.020	-0.417	0.377	0.194	0.567
0.534	0.439	1.327	0.401	0.196	1.008	0.496	0.322	0.903
0.532	0.377	1.066	0.401	0.231	0.421	0.158	0.081	0.335
0.270	*0.027*	*0.431*	*0.426*	*0.407*	*1.779*	0.252	0.132	0.296
0.390	*0.569*	*1.869*	0.565	0.370	0.716	0.430	0.210	0.808
0.187	*0.633*	*0.764*	0.469	0.361	1.080	0.545	0.237	0.603
0.081	*0.366*	*0.420*	0.098	0.027	-0.427	0.338	0.291	1.034
			0.282	0.194	0.890	*0.247*	*0.167*	*1.234*
			0.346	0.295	0.986	0.494	0.147	0.496
			0.303	0.179	0.507	0.411	0.247	-0.317
			0.334	0.102	0.083	0.269	0.197	0.438
			0.250	0.215	1.100	*0.409*	*0.144*	*-0.233*
			0.464	0.308	0.959	0.471	0.380	0.749
			0.169	0.219	0.480	0.483	0.247	0.450
			0.178	0.164	-0.437	0.358	0.067	0.600
			0.362	0.287	1.145	0.556	0.244	0.166
			0.221	0.223	1.099	0.328	0.284	0.871
			0.346	0.371	0.829	0.409	0.186	0.034
			0.673	*0.733*	*1.342*	0.188	0.234	0.927
			0.296	0.413	1.386	0.165	0.239	1.518
			0.088	*0.375*	*0.731*	*0.272*	*0.571*	*1.103*
			0.174	*0.572*	*0.652*	*0.361*	*0.523*	*0.420*
			0.430	*0.768*	*1.465*	*0.152*	*0.362*	*0.290*
			0.288	*0.916*	*1.194*	*0.331*	*0.515*	*1.554*
			0.273	*0.873*	*1.190*	*0.584*	*0.741*	*0.280*
			0.045	*0.548*	*0.640*	*0.159*	*0.693*	*1.114*
			0.254	*0.250*	*-0.410*	*-0.082*	*0.373*	*0.399*

Note. Dimension two items are noted in *italics*. For the real data analyses, dimension one and two items were determined by content.

Appendix H

Item Parameters for Test B Forms X and Y for test constructed with second dimension present among unique items only (UI)

Common Items			*Unique "X" Items*			*Unique "Y" Items*		
a_1	a_2	d	a_1	a_2	d	a_1	a_2	d
0.437	0.236	0.835	0.068	-0.035	-0.052	0.433	0.000	0.873
0.241	0.251	1.938	0.369	0.185	0.012	0.346	0.203	0.705
0.566	0.474	1.804	0.255	0.160	0.209	0.398	0.390	1.724
0.268	0.136	0.431	0.141	-0.020	-0.417	0.377	0.194	0.567
0.496	0.404	1.056	0.401	0.196	1.008	0.496	0.322	0.903
0.351	0.247	0.664	0.401	0.231	0.421	0.158	0.081	0.335
0.323	0.207	1.235	*0.426*	*0.407*	*1.779*	0.252	0.132	0.296
0.341	0.265	-0.432	0.565	0.370	0.716	0.430	0.210	0.808
0.534	0.439	1.327	0.469	0.361	1.080	0.545	0.237	0.603
0.532	0.377	1.066	0.098	0.027	-0.427	0.338	0.291	1.034
			0.282	0.194	0.890	*0.247*	*0.167*	*1.234*
			0.346	0.295	0.986	0.494	0.147	0.496
			0.303	0.179	0.507	0.411	0.247	-0.317
			0.334	0.102	0.083	0.269	0.197	0.438
			0.250	0.215	1.100	*0.409*	*0.144*	*-0.233*
			0.464	0.308	0.959	0.471	0.380	0.749
			0.169	0.219	0.480	0.483	0.247	0.450
			0.178	0.164	-0.437	0.358	0.067	0.600
			0.362	0.287	1.145	0.556	0.244	0.166
			0.221	0.223	1.099	0.328	0.284	0.871
			0.270	0.027	0.431	0.409	0.186	0.034
			0.346	*0.371*	*0.829*	0.188	0.234	0.927
			0.673	0.733	1.342	*0.165*	*0.239*	*1.518*
			0.296	*0.413*	*1.386*	*0.272*	*0.571*	*1.103*
			0.088	*0.375*	*0.731*	*0.361*	*0.523*	*0.420*
			0.174	*0.572*	*0.652*	*0.152*	*0.362*	*0.290*
			0.430	*0.768*	*1.465*	*0.331*	*0.515*	*1.554*
			0.273	*0.873*	*1.190*	*0.584*	*0.741*	*0.280*
			0.045	*0.548*	*0.640*	*0.159*	*0.693*	*1.114*
			0.254	*0.250*	*-0.410*	*-0.082*	*0.373*	*0.399*

Note. Dimension two items are noted in *italics*. For the real data analyses, dimension one and two items were determined by content.

Appendix I

Item Parameters for Test B Forms X and Y for test constructed with second dimension present among common items only (CI)

Common Items			*Unique "X" Items*			*Unique "Y" Items*		
a_1	a_2	d	a_1	a_2	d	a_1	a_2	d
0.709	*0.408*	*-0.403*	0.068	-0.035	-0.052	0.433	0.000	0.873
0.367	0.360	1.947	0.437	0.236	0.835	0.411	0.396	1.900
0.679	0.384	-0.042	0.369	0.185	0.012	0.346	0.203	0.705
0.323	0.226	-0.280	0.255	0.160	0.209	0.251	0.173	1.050
0.534	0.439	1.327	0.141	-0.020	-0.417	0.432	0.136	1.047
0.532	0.377	1.066	0.401	0.196	1.008	0.398	0.390	1.724
0.270	*0.027*	*0.431*	0.241	0.251	1.938	0.377	0.194	0.567
0.390	*0.569*	*1.869*	0.401	0.231	0.421	0.496	0.322	0.903
0.187	*0.633*	*0.764*	0.566	0.474	1.804	0.158	0.081	0.335
0.081	*0.366*	*0.420*	0.565	0.370	0.716	0.252	0.132	0.296
			0.469	0.361	1.080	0.546	0.159	0.248
			0.268	0.136	0.431	0.430	0.210	0.808
			0.361	0.148	-0.155	0.426	0.285	1.241
			0.098	0.027	-0.427	0.545	0.237	0.603
			0.496	0.404	1.056	0.408	0.395	1.892
			0.351	0.247	0.664	0.409	0.356	1.459
			0.323	0.207	1.235	0.338	0.291	1.034
			0.282	0.194	0.890	0.494	0.147	0.496
			0.346	0.295	0.986	0.411	0.247	-0.317
			0.303	0.179	0.507	0.269	0.197	0.438
			0.334	0.102	0.083	0.383	0.076	-0.411
			0.341	0.265	-0.432	0.161	0.155	1.388
			-0.051	-0.350	-0.062	0.471	0.380	0.749
			0.250	0.215	1.100	0.483	0.247	0.450
			0.464	0.308	0.959	0.358	0.067	0.600
			0.169	0.219	0.480	0.556	0.244	0.166
			0.178	0.164	-0.437	0.328	0.284	0.871
			0.362	0.287	1.145	0.409	0.186	0.034
			0.221	0.223	1.099	0.188	0.234	0.927
			0.673	0.733	1.342	0.533	0.385	0.490

Note. Dimension two items are noted in *italics*. For the real data analyses, dimension one and two items were determined by content.

Appendix J

Mean raw scores (and standard deviations) for Dimensions One and Two for Groups P and Q on Forms X forms of BL, UI and CI on Test A.

Location	Group *P*		Group *Q*	
of Second	Dimension	Dimension	Dimension	Dimension
Dimension	One[a]	Two	One	Two
	17.14	9.92	17.06	9.93
BL[b,c]	(3.10)	(1.97)	(3.14)	(2.01)
	20.73	6.70	20.56	6.74
UI	(3.54)	(1.60)	(3.59)	(1.59)
	23.44	3.22	23.27	3.19
CI	(3.82)	(0.89)	(3.87)	(0.91)

[a] Standard deviations are presented in brackets ()

[b] BL = both locations, UI = unique items only, CI = common items only

[c] BL, UI, and CI tests had different numbers of Dimension One and Dimension Two items. BL tests had 26 Dimension One and 15 Dimension Two items, UI tests had 31 Dimension One and 10 Dimension Two items, and CI tests had 36 Dimension One and 5 Dimension Two items.

Appendix K

Mean raw scores (and standard deviations) for Dimensions One and Two for Groups P and Q on X Forms of BL, UI and CI on Test B.

Location	Group *P*		Group *Q*	
of Second	Dimension	Dimension	Dimension	Dimension
Dimension	One[a]	Two	One	Two
	16.47	10.83	17.13	11.11
BL[b,c]	(3.57)	(2.60)	(3.53)	(2.55)
	20.68	7.51	21.41	7.70
UI	(4.05)	(1.92)	(3.95)	(1.85)
	23.87	3.32	24.67	3.42
CI	(4.65)	(1.13)	(4.54)	(1.13)

[a] Standard deviations are presented in brackets ()

[b] BL = both locations, UI = unique items only, CI = common items only

[c] BL, UI, and CI tests had different numbers of Dimension One and Dimension Two items. BL tests had 25 Dimension One and 15 Dimension Two items, UI tests had 30 Dimension One and 10 Dimension Two items, and CI tests had 35 Dimension One and 5 Dimension Two items.