ABSTRACT

Title of Document: MIXED-FORMAT TEST EQUATING: EFFECTS OF TEST DIMENSIONALITY AND COMMON-ITEM SETS

Yi Cao, Doctor of Philosophy, 2008

Directed By: Professor Robert W. Lissitz
Department of Measurement, Statistics and Evaluation

The main purposes of this study were to systematically investigate the impact of representativeness and non-representativeness of common-item sets in terms of statistical, content, and format specifications in mixed-format tests using concurrent calibration with unidimensional IRT models, as well as to examine its robustness to various multidimensional test structures. In order to fulfill these purposes, a simulation study was conducted, in which five factors – test dimensionality structure, group ability distributions, statistical, content and format representativeness - were manipulated. The examinees' true and estimated expected total scores were computed and BIAS, RMSE and Classification Consistency indices over 100 replications were then compared. The major findings were summarized as follows:

First, considering all of the simulation conditions, the most notable and significant effects on the equating results appeared to be those due to the factor of group ability distributions. The equivalent groups condition always outperformed the nonequivalent groups condition on the various evaluation indices.

Second, regardless of the group ability differences, there were no statistically and practically significant interaction effects among the factors of the statistical, content and format representativeness.

Third, under the unidimensional test structure, the content and format representativeness factors showed little significant impact on the equating results. Meanwhile, the statistical representativeness factor affected the performance of the concurrent calibration significantly.

Fourth, regardless of the various levels of multidimensional test structure, the statistical representativeness factor showed more significant and systematic effects on the performance of the concurrent calibration than the content and format representativeness factors did. When the degree of multidimensionality due to multiple item formats increased, the format representativeness factor began to make significant differences especially under the nonequivalent groups condition. The content representativeness factor, however, showed minimum impact on the equating results regardless of the increase of the degree of multidimensionality due to different content areas.

Fifth, the concurrent calibration was not quite robust to the violation of the unidimensionality since the performance of the concurrent calibration with the unidimensional IRT models declined significantly with the increase of the degree of multidimensionality.

MIXED-FORMAT TEST EQUATING:
EFFECTS OF TEST DIMENSIONALITY AND COMMON-ITEM SETS

By

Yi Cao

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2008

Advisory Committee:

Professor Robert W. Lissitz, Chair/Advisor
Professor Paul J. Hanges
Professor Jeffrey R. Harring
Professor Hong Jiao
Professor Robert J. Mislevy

Acknowledgements

First, I would like to thank Dr. Robert W. Lissitz, my academic advisor, dissertation supervisor, and mentor. He was always there for me whenever I was in need. Without his cordial and constant guidance, assistance, and encouragement, I might not be able to complete this work in a timely manner.

I am also sincerely grateful to my other committee members, Dr. Robert J. Mislevy, Dr. Jeffrey R. Harring, Dr. Hong Jiao, and Dr. Paul J. Hanges, for their time and energy devoted to this study. Their careful review, insightful comments, and pertinent suggestions improved the quality of this study and made it more valuable to the field. I also would like to thank the department of measurement, statistics, and evaluation for providing me financial support throughout my five-year graduate studies.

At last, I would like to express my deepest appreciation to my parents, without their unconditional understanding, lifelong sacrifice and selfless love, I might not go this far. I also extended this gratitude to all my friends around the world and future colleagues in ACT, who encourage and support me along the whole journey.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Background

The multiple-choice (MC) item format continues to be the mainstay of standardized testing programs due to its broad content sampling, high reliability, and objective and efficient scoring. Meanwhile, in order to fulfill the federal calling for "multiple approaches with up-to-date measures of student achievement, including measures that assess higher-order thinking skills and understanding of challenging content" (U.S. Department of Education, as cited in Kirkpatrick, 2005, p. 3), constructed response (CR) items and mixed-format tests which consist of both MC and CR items have been earning increasing interest and popularity. The advocates believe that MC and CR items both have their own advantages and limitations, and the combination may allow the concatenation of their strengths while compensating for their weaknesses. Therefore, more and more large-scale testing programs and state assessment systems have embraced mixed-format tests. In fact, a survey from Lane (2005) declared that 63% of the state assessments contained both MC and CR items. In this study, a mixed-format test refers to a test that consists of dichotomously-scored MC items and polytomously-scored CR items. Among the examples of using mixed-format tests are the College Board's Advanced Placement (AP) examinations, the National Assessment of Educational Progress (NAEP), the Test of English as a Foreign Language (TOEFL), the Massachusetts Comprehensive Assessment System, the California's Learning Assessment System, the Indiana's Performance Assessment for School Success, and the Michigan's High School Proficiency Test.

Partly due to state and federal legislation, especially the authorization of the *No Child Left Behind* Act, the stakes and consequences associated with test scores have been more of a concern than ever across all levels of clients, including students, teachers, parents, and district principles. High-stakes tests normally require strict test security. One of the general practices to ensure test security is to administer multiple test forms on the same or different test dates. When multiple test forms are used, an equating process should be applied so that examinees' proficiencies obtained across forms and across occasions can be compared on the common scale, which further addresses the fairness concern.

According to Kolen & Brennan (2004), equating refers to a statistical process that is widely used to adjust scores on different forms so that scale scores can be used interchangeably. Various equating procedures are available. Classical equating methods (e.g., the mean, linear and equipercentile methods) are commonly used in many testing programs. Existing literature has offered both theoretical and practical guidance on these classical methods (Holland & Doran, 2006; Kolen & Brennan, 2004). Meanwhile, with the increasing advancement of item response theory (IRT) and the availability of sophisticated computer software, IRT equating methods have become more and more appealing. The equating methods that will be implemented in this study are IRT equating methods under common-item nonequivalent groups (CINEG) design. IRT equating under CINEG design is a multi-step process, which includes item calibration, scale transformation and/or raw-to-scale score conversion. These steps have been well investigated for dichotomously-scored item only tests and also in recent years have been fully extended to polytomously-scored item only tests (see Holland & Doran, 2006;

Kolen & Brennan, 2004; Muraki, Hombo, & Lee, 2000 for detailed discussions). However, the use of mixed-format tests would greatly complicate the IRT equating process and pose a number of new challenges to the IRT equating under CINEG design.

One challenge for mixed-format test equating using IRT methods with CINEG design is how to extend traditional IRT equating procedures that were originally developed for single-format tests to those appropriate for mixed-format tests. As mentioned above, IRT equating procedures have been well developed for single-format tests. Until recent years, researchers started to extend various IRT equating procedures from single-format tests to mixed-format tests and compare the relative performance of these extended equating procedures. Detailed literature review of this issue will be presented in Chapter Two. However, compared to a large body of comparison studies for single-format tests, the studies in this field for mixed-format tests are limited in number and in their coverage of the important issues.

Another challenge related to mixed-format test equating is brought by the use of raters to score items. For mixed-format tests where CR items are included and scored by raters, systematic changes in rater judgments from year to year in terms of rater severity and leniency may influence the accuracy of equating (Kim & Kolen, 2006; Tate, 1999). The task of equating MC-only tests is to disentangle the variation of form difficulty and group ability. When mixed-format tests are used and rater effect occurs, it adds another source of systematic variations that is intertwined with form difficulty and group ability. To adjust and evaluate the impact of rater effects, especially on the accuracy of equating results, several studies have been conducted and various statistical procedures have been proposed (Fitzpatrick, Ercikan, Yen, & Ferrara, 1998; Linacre, 1988; Tate, 1999, 2000,

and 2003). There will be a brief description of several studies about the issue of rater effects on mixed-format test equating in Chapter Two. However, it is not the primary subject of this study.

Multidimensionality is another serious issue associated with mixed-format test equating using IRT methods. Multidimensionality could be caused by a variety of reasons and it might exist in single-format tests as well as mixed-format tests. However, there is a unique source that could lead to multidimensionality in mixed-format tests, that is, the use of multiple item formats in a single assessment. Multidimensionality associated with item formats could occur when examinees process MC and CR items in different ways (Traub, 1993). In other words, if multidimensionality associated with item formats occurs, MC and CR items in a mixed-format test that assesses the same content will measure examinees' different proficiencies (Kim & Kolen, 2006). The multidimensionality due to item format and especially how it will influence the mixed-format test equating will be one of the primary concerns in this study. According to the previous research on dimensionality structure underlying mixed-format tests, the findings appeared to be mixed and largely varying in different contexts. Unidimensionality exists in some large-scale operational mixed-format tests, and multidimensionality exists in others. When multidimensionality occurs, unidimensional IRT models might no longer be appropriate for equating. In that case multidimensional IRT models for dichotomously-scored and polytomously-scored items might be applied. However, since the use of unidimensional IRT models is still dominant in most testing programs, it seems to be more valuable for this study to focus on the robustness of unidimensional IRT equating methods to various dimensionality structures underlying mixed-format tests.

When CINEG design is involved in mixed-format test equating, another challenge cannot be ignored, which is the composition of common-item sets. Although there is no universal conclusion that has been reached on the impact of the characteristics of the common-item sets on test equating (see research findings in Chapter Two for details), for MC-only tests, there is a widely accepted opinion that the characteristics of the common-item sets play an important role and that the common-item set should be a mini version of the whole test in terms of content and statistical specifications (Kolen & Brennan, 2004). The similar but more complicated scenario occurs when mixed-format tests need to be equated under CINEG design. In addition to considering content and statistical representativeness of the common-item sets, to develop a proportionally representative set of common items for mixed-format tests one has to take the item format effect into account. Kirkpatrick (2005) provides a very good example in which CINEG design is not even a desirable design to implement in mixed-format test equating in the first place because of the impossibility to develop a proportionally representative common-item set. So far, very little research has been conducted on the impact of the composition of common-item sets on mixed-format test equating. Among them, most of the researchers only focused on separate aspect of the common-item set such as whether CR items should be included or excluded, not the collective impact of the characteristics of the common-item sets in terms of content and statistical representativeness as well as mixture of item formats on the CINEG equating results. That will be the other primary focus of this study.

Research Purposes and Questions

As described earlier, quite a few challenges exist in mixed-format test equating using IRT methods with CINEG design. In this study, the composition of common-item

sets under various multidimensional test structures will be the primary concern. Although research on the effects of multidimensionality and the composition of common-item sets for mixed-format test equating exists and is informative, it still leaves room for improvement in our knowledge base. Previous research on multidimensionality (Kim & Kolen, 2006; Kirkpatrick, 2005; Sykes, Hou, Hanson, & Wang, 2002; Tate, 2000) only simulated data for the correlated-format-specific factorial model. However, the underlying test structure could be far more complicated in reality in a way that more than one factor could affect an examinee's correct response to each item and different items may require different combinations of factors for an examinee to respond correctly. Moreover, the relative influential power of various factors on examinees' item responses could vary. In this case, the multidimensional IRT models might better capture the test structure and the results might be generalizable to more realistic situations. However, no current research has investigated the impact of more general multidimensional test structures on mixed-format test equating. Meanwhile, previous research on the composition of common-item sets (Kim & Lee, 2006; Kirkpatrick, 2005; Sykes et al., 2002; Tate, 2000) only focused on whether to include or exclude CR items in the common-item sets. They all assumed content and statistical representativeness of the common-item sets. No research has explored the interactive effects of representativeness and non-representativeness of content, statistical and format specifications on mixed-format test equating.

There are two main purposes of this study. First, to systematically investigate the impact of representativeness and non-representativeness of common-item sets in terms of content, statistical and format specifications on mixed-format test equating using IRT

methods. Second, to investigate the robustness of unidimensional IRT equating methods under various simulated conditions, especially the multidimensionality due to item format. More specifically, this study will attempt to address the following three questions:

1) In an ideal situation where the unidimensionality assumption is satisfied, what are the effects of content, statistical and format representativeness of common-item sets on mixed-format test equating?
2) In hypothetical but possibly practical situations where multidimensionality exists, what are the effects of content, statistical and format representativeness of common-item sets on mixed-format test equating?
3) How robust is the unidimensional IRT equating method to the presence of different multidimensional situations?

In order to address the above research questions, a simulation study will be conducted. The three-parameter logistic (3PL) model and the graded response model (GRM) will be used to generate and calibrate MC and CR items, respectively. Concurrent calibration will be conducted for placing parameters onto a common scale. The factors that will be manipulated are dimensionality structure of mixed-format tests, content, statistical, and format representativeness, and group ability distributions. Details about the simulation study will be presented in Chapter Three.

Organization of the Study

This study will be presented in five chapters. In Chapter One, the background of this study as well as the research purposes and questions associated with this study have been addressed. In Chapter Two, a review of relevant literature will be presented. The review will concentrate on four aspects: the comparison of MC and CR items, the

dimensionality differences of mixed-format tests, the equating design components for mixed-format test equating, and comprehensive research on mixed-format test equating. In Chapter Three, the methodology for this study, especially the simulation design employed in this study will be addressed. In Chapter Four, the results of this study will be summarized and reported. In Chapter Five, the major findings will be discussed, the implications for practice will be addressed, the limitations of this study and some suggestions for future research will be provided.

Chapter 2: Literature Review

This chapter consists of four major sections. In the first section, multiple-choice (MC) items and constructed response (CR) items which comprise the mixed-format tests considered in this study are defined and their advantages and limitations are compared. The second section provides a discussion of relevant literature on the dimensionality structure of mixed-format tests. The third section presents and discusses all the equating design components employed in this study, which include data collection designs, mixture of item response models, and IRT equating methods. Finally, detailed reviews of previous comprehensive research on mixed-format test equating, which are suggestive of this study, are provided. Each section ends with a short description of relevant factors of this study.

Item Formats in Mixed-format Tests

For the purpose of this study, a mixed-format test is defined as a test consisting of two item formats in a single assessment. Two general categories of item formats that will be used to comprise mixed-format tests in this study are MC and CR items. MC items require the examinee to select an answer from a relatively small set of response options (e.g., four or five) and are often dichotomously scored. As a competing and complementary alternative for MC item format, CR items require the examinee to generate his/her own answer rather than selecting among alternative options. CR items could be either dichotomously or polytomously scored. However, this study will limit the CR items to those polytomously scored.

A wealth of research has been devoted to understanding the distinctions between MC and CR items from both psychometric and cognitive perspectives. The psychometric

research compares these two formats in terms of content coverage, reliability and validity evidence, scoring objectivity and efficiency, and so on. The cognitive research emphasizes the format distinctions in terms of cognitive skills they can elicit and cognitive ranges they might sample. Seven major differences that were typically addressed in the literature are summarized in this section.

First, MC items allow evaluation of a greater breadth of content coverage in a fixed testing time period under limited budget. In contrast, since CR items require examinees to spend a certain amount of time generating responses, usually fewer CR items can be administered during a particular period. The use of a limited number of items usually results in an inadequate sample of the content domain (Kolen, 1999-2000; Linn, 1995). Oosterhof (1996) states that when there are only ten or fewer items in a test, the content sampled by these items largely determines an examinee's score. An examinee who knows only a small fraction of the content could possibly do well on a test if the test items happen by chance to sample content with which the examinee is knowledgeable. On the other hand, if the test items sample what the examinee is not familiar with, the examinee would perform very poorly. Consequently, it could substantially reduce the reliability of the test and hence constrain its generalizability.

Second, the scoring of MC items is inexpensive, efficient and objective. In contrast, the scoring of CR items nearly always involves a group of judges and requires fairly detailed scoring rubrics, which substantially increase the amount of time and the cost. Experimentation with computer based scoring is a current interest of some testing firms, although the systems are not widely applied yet. However, even though the judges are well trained, the scoring rubrics are clearly specified, and the scoring processes are

strictly monitored, the scoring of CR items is still subjective and may vary across judges and occasions. All the efforts intended to help judges score reliably are only moderately successful (Bennett, 1993).

Third, CR items increase content validity by providing a more direct measure of content and skill objectives than MC items do. In addition, CR items might offer higher construct validity than MC items do. MC items, constrained by their nature, are less likely to elicit certain types of cognition like divergent thinking. If adequate assessment of a construct requires these sorts of cognitions, the construct might be under-represented (Messick, 1995). Furthermore, MC items are more susceptible to some test-taking strategies, also known as test-wiseness. For example, they are sensitive to a response elimination strategy (Burton, 2001), in which examinees could use secondary cues to rule out several implausible response options without knowing the correct answer and thus increase the probability of guessing right. If test scores partly reflect examinees' ability to use this strategy, then construct-irrelevant variance will be introduced into scores and construct validity will be compromised.

Fourth, CR items require examinees to generate the answers rather than to choose from a set of options, and therefore eliminate the random guessing effect. An examinee who has not acquired the necessary knowledge would be unlikely to generate the correct response simply by guessing. In contrast, MC items are too vulnerable to guessing, which has even been nicknamed by Oosterhof (1996) as the "multiple-guess" format. As a result, guessing often leads to an inflated score.

Fifth, both MC and CR items could evoke lower-level and/or higher-level cognitive skills. MC items are often written to assess the lower-level cognitive skills,

such as recognition or recall. However, researchers (Haladyna, 1994, and 1997; Hamilton, Nussbaum, & Snow, 1997; Wainer & Thissen, 1993) have demonstrated that MC items can also be written to evoke complex cognitions, such as understanding, evaluation, and problem solving. On the other hand, CR item format encompasses a wide variety of tasks ranging from filling in blanks, sentence completions, short answers, to a multi-page essay writing or a multi-step solution to a mathematical problem. Accordingly, the cognitive skills elicited by different varieties of CR items vary substantially. Therefore, as Martinez (1999) pointed out that characterizing CR and MC as a simple dichotomy disguises the diversity of forms and the cognitive demands within these two item formats.

Sixth, although MC items can elicit complex cognitions, the range of cognitions within the reach of CR items is broader. Martinez (1999) indicated that MC items, by design, largely limit the examinees' behavior so that it is extremely difficult for MC items to capture two qualities of response: complex performance and divergent production. For example, MC items might assess most aspects of English language proficiency such as listening, reading, and even writing, but they cannot directly measure examinees' ability of speaking. Another limitation of MC items is that cognition must eventually lead to convergence since a single answer must be chosen from a set of alternatives and thus divergent production is excluded. Such cognitive tasks involving elaborating creative ideas prompted by a stimulus, or generating original applications of a scientific principle, cannot be accomplished via MC items. In conclusion, MC items, to some extent, cannot reach the full spectrum of complex cognition represented in CR items.

Seventh, CR items may be more useful to facilitate students' learning and teachers' instruction than MC items. There is some evidence (Snow, 1993) showing that students who expect a CR test generally work harder and prepare more than those who anticipate a MC test. Furthermore, CR items can elucidate much richer responses than MC items. Therefore, CR items might be more helpful than MC items to record students' cognitive trace of their solution processes and provide teachers with more informative diagnostic messages about students' learning errors and misconceptions (Lukhele, Thissen, & Wainer, 1994; Martinez, 1999), which would in turn help teachers to realize their teaching problems and thereby improve their teaching efficiency.

Implications

Table 2.1 summarizes the distinctions between MC and CR item formats discussed above.

Table 2.1 A Comparison of MC and CR Item Formats

	MC items	CR items
Psychometric Perspective		
Content sampling in unit time	Broad (+)	Narrow
Reliability	Generally high (+)	Generally low
Scoring	Objective, efficient and inexpensive (+)	Subjective, time-consuming and expensive
Validity	Generally low	Generally high (+)
Robustness to guessing	Low	High (+)
Cognitive Perspective		
Cognitive skills	Lower- or higher-level	Lower- or higher-level
Cognitive range	Narrow	Broad (+)
Facility in learning and teaching	Generally low	Generally high (+)

Note: Features marked with a plus (+) are usually considered to be desirable.

After expatiating and weighing the desirable and undesirable features of MC and CR item formats, one implication can be reached. That is one item format would never be superior to another in all respects and for all purposes. Therefore, "sometimes the best policy decision will not be a matter of either/or, but of what mixture of item formats will

yield the best possible combined effect" (Martinez, 1999, p. 216). The combination may allow the concatenation of their strengths while compensating for their weaknesses. In the case of MC and CR item formats, the benefits of broader content sampling, high reliability, objective and efficient scoring offered by MC item formats can be complemented by those of high content and construct validity, robustness to guessing, and integrated reach of complex cognitions offered by CR item format. In conclusion, a mixture of MC and CR item formats may provide a more appropriate measure of learning and teaching.

Base upon these arguments, the use of mixed-format tests seems to be more and more popular in large-scale testing programs and state assessment systems. In fact, a survey from Lane (2005) declared that 63% of the state assessments currently include mixed-format tests. Therefore, this study will focus on mixed-format tests, rather than single format tests in order to provide meaningful guidance and fulfill the increasing needs of better constructing the mixed-format tests and employing common-item non-equivalent groups design on it.

Dimensionality Structures among Mixed-format Tests

Given the desirable features that the combination of MC and CR items could carry, a mixed-format test has been earning increasing interest and popularity in recent years. However, a mixed-format test may also introduce additional complexity into the test dimensionality. The dimensionality of a mixed-format test has been examined by several researchers using approaches that include correlation analysis, factor analysis, and the Poly-DIMTEST. The findings appear to be mixed, indicating that dimensionality

varies greatly in different contexts. In this section, previous research on the dimensionality of mixed-format tests is summarized.

Bennett and his colleagues (1990, and 1991) conducted two studies to examine the relationship of MC and CR items on the College Board's Advanced Placement (AP) Computer Science examination. In the 1990 study, they assessed the relationship of a constrained CR item with both MC and CR items. Confirmatory factor analysis was used to test the fit of four different factorial models (three-factor, two-factor, single-factor, and null) to two approximately random samples. They found that a single-factor model was sufficient for one sample, but a two-factor model hypothesizing separate MC/CR and constrained CR factors was required for the other sample. These factors were highly correlated. In the 1991 study, they again employed confirmatory factor analysis, and compared the data-model fit of the two-factor model hypothesizing separate but correlated MC and CR factors with the single-factor model in two random samples. They found that 1) In the two-factor model, the disattenuated correlation coefficient between two format-specific factors was significantly, but not substantially different from unity; and 2) the single-factor model provided a more parsimonious fit than the two-factor model did.

Several other studies using AP data in the areas of mathematics, computer science, chemistry, history, and so on (Lukhele et al., 1994; Thissen et al., 1994; Wainer & Thissen, 1993; Wainer, Wang, & Thissen, 1994) investigated the question "how much additional or new information about a construct was gained when CR items were added to MC items in a single test". The findings from these studies appear to be inconsistent.

Wainer & Thissen (1993) examined the AP data in seven different areas and suggested that when CR items were combined with MC items in a single test, little new information was obtained about any of the areas. In other words, these mixed-format tests were unidimensional up to the limits of the ability to examine them. This conclusion has been supported by Lukhele and his colleagues (1994). They used the three parameter logistic (3PL) model for MC items and graded response model (GRM) for CR items to examine the amount of information obtained from different item formats on the AP Chemistry and AP US History tests. They reached the same conclusion that for both tests, adding CR items provided little information beyond what the MC items yielded.

Thissen et al. (1994) analyzed the AP Computer Science data that Bennett, Rock, & Wang (1991) used, but proposed a different factorial model, in which a general factor for all items plus orthogonal factors specific to CR items were assumed. They then repeated this general-plus-specific-factors model onto the AP Chemistry test. They compared their model to Bennett et al.'s model and concluded that the general-plus-specific-factor model provided a more parsimonious way to depict the dimensionality of mixed-format tests than the format-specific-factor model did. Moreover, they found that MC and CR items both heavily loaded on the general factor, which indicated that MC and CR items measured the same construct for most part of the tests. However, they also found small but significant loadings on the CR-specific factors, which indicated that the CR items actually measured something unique from the MC items. Wainer et al. (1994) also found the same dimensionality structure in a separate study.

Manhart (1996) applied confirmatory factor analysis to compare whether the MC and CR science tests measured the same construct. The MC tests they used were the Tests

of Achievement and Proficiency (TAP) in science and the CR tests were the TAP Performance Assessments for science. Each test was divided into several parcels of items. A single-factor model and a two-factor model assuming two correlated format-specific factors were compared in terms of data-model fit using chi-square values and standardized residuals. The results showed that the two-factor model generally fit the data better than the single-factor model. In conjunction with other evidence based on the content and cognitive skills analysis of the two tests, he concluded that MC and CR tests measured different constructs.

Perkhounkova & Dunbar (1999) employed the Poly-DIMTEST procedure in a confirmatory way to explore the dimensionality structure of three kinds of achievement tests in two subject areas. The three kinds of achievement tests were MC tests (specifically, Form M of the Iowa Tests of Basic Skills, ITBS), CR tests (specifically, the CR supplement to ITBS), and a test combining both item formats. The two subject areas were Language Arts and Mathematics. They also compared the results across two grades, grade 7 and 8. The results differed for two subject areas. For Language Arts tests, the analysis showed that the MC tests and CR tests possibly assessed the same dimension, that is, language achievement. For Mathematics tests, the analysis indicated that both the MC test and the mixed-format test were not essentially unidimensional. Moreover, the CR test appeared to measure dimensions differing from the MC test.

Sykes et al. (2002) used the Poly-DIMTEST procedure and principal factor analysis to investigate the effects of multidimensionality due to item format by using data from a mixed-format state math field test. The results demonstrated multidimensionality

related to item format. The first factor was found to be a common dimension, while only MC items loaded heavily on the second factor.

Implications

There has been a steady increase in the use of mixed-format tests in various assessment settings. When dealing with mixed-format tests, test developers frequently face the need to obtain a meaningful composite score for each examinee. Aggregating scores from a mixture of different item formats naturally raises the question about the dimensionality of the tests mainly because the traditional IRT applications, including test equating and linking, assume unidimensionality. However, a mixture of item formats in a single test may increase the chance of violating the unidimensionality assumption because it introduces an additional source of multidimensionality. The multidimensionality due to mixed item formats, especially about its effect on equating will be one of the primary concerns in this study. Previous research on dimensionality differences among mixed-format tests is summarized in Table 2.2.

Table 2.2 Dimensionality Differences among Mixed-format Tests

Unidimensionality	Multidimensionality due to Item Formats
Bennett, Rock, Braun, Frye, Spohrer, & Soloway (1990)	
Bennett, Rock, Wang (1991)	Thissen, Wainer, & Wang (1994)
Wainer & Thissen (1993)	Wainer, Wang, & Thissen (1994)
Lukhele, Thissen, & Wainer (1994)	Manhart (1996)
Perkhounkova & Dunbar (1999)	
	Sykes, Hou, Hanson, & Wang (2002)

Evidence from previous research is equivocal. Many reported that essential unidimensionality exists in some large-scale operational mixed-format tests, that is, MC and CR items measure nearly the same construct. However, others found the existence of multidimensionality due to different item formats in mixed-format tests. In the former situation, a single-factor model is expected to be sufficient to capture the only dimension

underlying the test. In the latter case, a more general multi-factor model is often anticipated.

Since the test dimensionality structure is one of the crucial factors which will influence the composition of common-item sets and the appropriate selection and employment of IRT equating methods, these two factorial models will be simulated (a single-factor model will be used as a baseline for comparison) and their effects on equating will be investigated in this study.

Data Collection Designs, Item Response Models, and Equating Procedures for Mixed-format Tests

In this section, the data collection design including the common-item nonequivalent groups design, various item response models, and IRT equating procedures that will be employed in this study are presented and defined.

Common-item Nonequivalent Groups Design

Three data collection designs are widely used in test equating and scaling. They are single group (SG) design, random groups (RG) design and common-item nonequivalent groups (CINEG) design (Kolen & Brennan, 2004). In the SG design, the same examinees take both test forms X and Y, usually in counterbalanced order. In the RG design, examinees are randomly assigned to take either form X or form Y. In the CINEG design, there are two usually nonequivalent groups of examinees. One group takes form X and the other takes form Y. Form X and form Y have a set of items in common. When the score on the common-item set are counted in the examinee's total scores, it is referred to as an internal set. When the score on the common-item set does not contribute to the examinee's total scores, it is referred to as an external set. The

CINEG design improves the flexibility upon RG design by allowing nonequivalent groups to take form X and Y. Meanwhile, it improves upon SG design by not requiring examinees to take both form X and Y. In this study, attention will be restricted to the CINEG design with an internal common-item set.

For traditional MC-only tests, the crucial step in CINEG design is to develop a proportionally representative set of common items which should resemble a mini-version of the overall test in terms of content and statistical specifications (Kolen & Brennan, 2004). This is because the common-item set is the only means to disentangle group differences from form differences in CINEG design. Researchers have examined the effects of the characteristics of the common-item set, such as the length, content and statistical representativeness of the common-item sets on the accuracy of equating and concluded that the composition of common-item sets affected equating results.

Petersen, Marco, and Stewart (1982) examined the effects of content and statistical representativeness on traditional linear equating. They found that differences in difficulty between the total test and the common-item set led to greater equating error than did moderate disparity in content representativeness of the common-item set. Klein and Jarjoura (1985) compared shorter but content representative common-item sets with longer but content non-representative common-item sets. They found that the longer, non-representative common-item sets produced less accurate equating results than the shorter, representative ones did and thus concluded that content representativeness of the common-item set was vital to equating accuracy. Cook and Peterson (1987) found that inadequate content representativeness of the common-item set created serious problems especially when the two examinee groups differed considerably in levels and dispersions

of ability. Harris' study (1991) and Yang's study (2000) later supported these findings. However, Gao, Hanson and Harris (1999) reached somewhat different conclusions. They examined the effect of content and statistical representativeness on the CINEG design and found that content itself did not greatly impact equating results, but the interaction between content and statistical specifications did have effects. Specifically, on one hand, if the common-item set was not statistically representative, a content non-representative common-item set may produce more equating error than a content representative set. On the other hand, if the common-item set was not content representative, a statistical non-representative common-item set may produce more equating error than a statistical representative set. Hanick and Huang (2002) later confirmed Gao et al.'s results that content non-representativeness of the common-item set has minimal effect on equating accuracy. Furthermore, they found that if the equating plan was "well designed", statistical non-representativeness did not greatly influence the equating results either, although larger numbers of discarded items increased equating error. So far, no universal conclusions on the effects of the characteristics of the common-item sets on test equating have been identified, but there is a widely accepted opinion that the characteristics of the common-item sets play an important role in traditional MC-only test equating, such that the common-item set should be a mini version of the whole test, especially when examinee groups differ considerably in their ability level.

The composition of common-item sets is also crucial for successfully equating multiple mixed-format tests under the CINEG design. In addition to considering content and statistical representativeness of the common-item sets, mixed-format test equating has to take item format effect into account, which in turn dramatically increases the

complexity and difficulty of developing appropriate common-item sets for mixed-format tests. In practice, many practitioners have suggested using only MC items as a common-item set (Baghi, Bent, DeLain, & Hennings, 1995; Livingston, 1994). The use of MC items only as a common-item set could be defensible only when MC and CR items measure the same construct(s). If MC and CR items measures somewhat different constructs, the use of MC items only as a common-item set might be "dangerous" and will lead to serious linking bias. Several researchers therefore advocate that common-item sets should include the appropriate proportion of MC and CR items in order to make the equating reasonably robust to the violation of assumptions (Kim & Kolen, 2006; Kirkpatrick, 2005; Sykes et al., 2002; Tate, 2000). Details about these studies will be provided in the next section.

Item Response Models

For mixed-format tests, since both MC and CR items are included in a single test, it seems intuitive and reasonable that the mixture of dichotomous and polytomous IRT models can be applied to estimate parameters. For MC items, dichotomous IRT models, such as one, two, and three-parameter logistic models (i.e., 1PL, 2PL and 3PL) can be applied. For CR items, polytomous IRT models, such as graded-response model (GRM), generalized partial credit model (GPCM), and nominal response model (NRM) can be applied. The concepts, assumptions, and features of these models have been well-documented in many book-length references (Baker & Kim, 2004; Embretson & Reise, 2000; Hambelton & Swaminathan, 1985; van der Linden & Hambleton, 1996).

As Baker and Kim (2004) pointed out, any combination of dichotomous and polytomous IRT models can be chosen for the analysis of item responses in the mixed-

format test. In this study, 3PL/GRM model combination will be used to estimate IRT parameters. The same model combination has also been applied in Lukhele et al. (1994), Bastari (2000), and Rosa, Swygert, Nelson, & Thissen (2001).

The reasons to choose the 3PL/GRM model combination for parameter estimation are as follows. For MC items, previous studies have showed great flexibility of the 3PL model over the 1PL and 2PL models since the 3PL model enables us to take guessing into account. Meanwhile, many well-known assessment programs have applied the 3PL model to MC items, such as the National Assessment of Educational Progress (NAEP), the Comprehensive Tests of Basic Skills (CTBS), the Armed Services Vocational Aptitude Battery (ASVAB), the Law School Admissions Test (LSAT), the Graduate Management Admissions Test (GMAT), the Scholastic Assessment Test (SAT), the Graduate Record Exam (GRE), and the Tests of English as a Foreign Language (TOEFL). Model-data fit for these examples has been proven to be excellent (Hambleton, as cited in Bastari, 2000). For CR items, GRM and GPCM are two polytomous IRT models widely-used and fully-examined. Several studies have been conducted to evaluate the recovery of model parameters and to compare the applied results of GRM and GPCM. For example, Dodd (1984) has compared and contrasted the Rasch versions of GRM and PCM, and found that although the two models were conceptually and mathematically different, the applied results were very similar. Maydeu-Olivares, Drasgow, & Mead (1994) compared the GRM with the GPCM and found that the model-data fit was equally good. Some other studies reached the same conclusion (e.g., Cao, Yin, & Gao, 2007; Tang & Eignor, 1997). Since neither the GRM nor GPCM consistently exhibits superiority over the other, the choice of GRM is basically built upon the previous research conducted by the author

in which the SAS code for generating polytomous item response data using GRM has been developed and applied.

IRT Equating Procedures

Only IRT equating procedures will be considered in this study. The IRT equating procedures typically incorporate three steps. The first step is item calibration, in which appropriate IRT models are used to estimate parameters. The second step is scale transformation in order to place the estimated parameters from one form onto the baseline form scale. The third step is raw-to-scale score conversion if number-correct scores need to be reported. In this study, only the first two steps will be investigated.

In the item calibration step, as mentioned earlier, the 3PL model will be used to estimate MC item responses and the GRM will be used to estimate CR item responses. Theoretically, the item parameters for a mixed-format test can be estimated separately by format on a one-format-at-a-computer-run basis (a.k.a., format-wise calibration) or jointly across formats on an all-formats-at-a-computer-run basis (a.k.a., simultaneous calibration). When a common score scale needs to be created and a total test score needs to be reported, simultaneous calibration has been recognized as more justifiable than format-wise calibration because it provides a statistically optimal way to solve the weighting selection problem of the format-wise calibration so that different item formats could be placed on the same scale and the performance on different item formats could be compared directly (Ercikan et al., 1998; Sykes & Yen, 2000).

In the scale transformation step, two alternative procedures – separate calibration and concurrent calibration - can be applied for placing IRT parameter estimates on a common scale. In separate calibration, the parameters for the two mixed-format test

forms are estimated separately in two computer runs. It is then followed by scale transformation methods, which can in turn transform the parameter estimates of one form to the scale of the other form through a common-item set. The commonly-used scale transformation methods for dichotomous IRT models are two moment methods: mean/mean (Loyd & Hoover, 1980), mean/sigma (Marco, 1977), and two characteristic curve methods: Haebara (Haebara, 1980) and Stocking-Lord (Stocking & Lord, 1983). They have also been extended to different polytomous IRT models (Baker, 1992, and 1993; Cohen & Kim, 1998; Kim & Cohen, 1995). Kim & Lee (2006) also extended these four scale transformation methods to mixed-format tests using any mixture of five dichotomous and polytomous IRT models: 3PL model, GRM, GPCM, NRM, and MC model. In concurrent calibration, the parameters on both forms are estimated simultaneously in one run which guarantees that all parameter estimates are on the same scale. This is done by combining data from both examinee groups and treating items not taken by a particular group as not reached or missing.

Considerable attention has been given to the relative performance of concurrent calibration and separate calibration with different scale transformation methods. Concurrent calibration generally outperformed separate calibration under various conditions because it is believed that concurrent calibration makes complete use of the available information and may remove some equating errors yielded by potentially inaccurate scale transformation procedures that are used by separate calibration. Kim & Cohen (1998) used simulation procedures in which dichotomously scored data were generated to compare separate calibration with the Stocking-Lord method and concurrent calibration using different software. For small numbers of common items, they found that

concurrent calibration yielded more accurate results than did the separate calibration with Stocking-Lord method. When the number of common items was sufficiently large, separate and concurrent calibration yielded similar results. Hanson & Beguin (2002) simulated dichotomous response data to compare the relative performance of separate calibration (mean/mean, mean/sigma, Haebara, and Stocking & Lord methods) versus concurrent calibration. They found that concurrent calibration generally resulted in lower error than separate calibration, although not universally so. In separate calibration, the two characteristic curve methods yielded more accurate results than the mean/mean and mean/sigma methods. Kim & Cohen (2002) simulated polytomous item response data and compared separate calibration with the Stocking-Lord method to concurrent calibration under GRM. They found that concurrent calibration yielded consistently though only slightly more accurate results than separate calibration did.

In the above studies, the same unidimensional IRT model which was used to generate data was also used to estimate parameters. In other words, in these studies, data were simulated to fit the IRT model. Beguin, Hanson, & Glas (2000) and Beguin & Hanson (2001) purposefully simulated multidimensional data which did not fit the unidimensional dichotomous IRT model. They compared separate calibration with Stocking-Lord method to concurrent calibration and found that multidimensionality affected the relative performance of separate and concurrent calibration. Estimates from the correctly specified multidimensional model generally resulted in less error than those from the unidimensional model. In general, unidimensional concurrent calibration resulted in slightly less or equivalent total error than separate calibration did. Kim & Kolen (2006) also simulated multidimensional data that reflects the format effects in

mixed-format tests. They then compared the relative performance of concurrent calibration to separate calibration with four scale transformation methods under the unidimensional IRT model. They found that concurrent calibration generally outperformed separate calibration in terms of linking accuracy and robustness to multidimensionality. Therefore, only concurrent calibration is considered in this study.

Implications

IRT equating for mixed-format tests is a very complicated process and includes multiple steps, such as designing a data collection, calibrating items using appropriate IRT models for each item format, and conducting scale transformation. In each step, the design conclusion which will be employed in this study is reached mainly based on previous research results and summarized in Table 2.3.

Table 2.3 Current Research Design Alternatives and Conclusions (in *Italic*)

Data Collection Design	Dichotomous IRT models for MC items	Polytomous IRT models for CR items	Calibration Methods
SG design with counterbalancing	1PL model	*GRM*	Separate calibration with scale transformation methods
RG design	2PL model	GPCM	*Concurrent calibration*
CINEG design	*3PL model*	NRM	

The impact of common-item set configuration has long been recognized and emphasized, but recently its importance has been advanced to a very public level by being addressed in the presidential speech at NCME annual meeting (Fizpatrick, 2008). In CINEG design, the composition of common-item sets is a key step and worth more systematic investigation, especially in mixed-format test equating. The common-item sets should be proportionally representative of the whole test with regard to the content and statistical specifications as well as the mixture of item formats. The impact of these

characteristics of the common-item sets on the mixed-format test equating will be another focus of this study.

Meanwhile, this study will only examine the unidimensional IRT equating method and how robust it will be to the various multidimensional test structures. The focus on the robustness study is driven by the practical need since the use of unidimensional IRT models is still dominant in most testing programs. Therefore, it seems to be more valuable for this study to focus on the robustness of the unidimensional IRT equating method to various multidimensional structures.

Comprehensive Research on Mixed-format Test Equating

Little research has been conducted on equating and linking for mixed-format tests. Existing research on mixed-format test equating has mainly focused on extending various linking procedures from single format tests to mixed-format tests and comparing the relative performance of these extended linking procedures, and on evaluating the impact of rater severity across occasions on the accuracy of equating results. Very few other studies investigated the effects of multidimensionality and the composition of common-item sets on equating.

Li, Lissitz & Yang (1999) investigated the performance of the extended Stocking-Lord method in mixed-format test equating. Furthermore, they examined the impact of the proportion of different item formats in a common-item set on the accuracy of equating. A simulation study was conducted, in which several other factors such as sample size, equating situations, and group ability differences were also manipulated. They estimated item parameters by using the 3PL model for dichotomously scored items and GPCM for polytomously scored items. They then used the average differences

between the true parameters and the corresponding estimates (BIAS) and the root mean square errors (RMSE) as the evaluation criterion. The results showed that the BIAS index of the recovery linking coefficients across all simulation conditions was close to zero and the RMSE index was relatively small, which indicated that the extended Stocking-Lord method was able to produce accurate linking for mixed-format tests.

Bastari (2000) investigated the effects of six factors on the linking accuracy for mixed-format tests under CINEG design. He conducted a simulation study, in which test length, proportion of MC and CR items in the whole test, the length of a common-item set, sample size, group ability distributions, and scale transformation methods were considered. The 3PL/GRM model combination was used to generate and estimate item responses. The mean square differences (MSD), bias, variance, and the RMSE based on the differences between estimated and true item characteristic curves were used as evaluation criteria. The results showed that overall, the recovery of CR item parameters was relatively worse than that of MC item parameters, hence resulting in larger linking errors. Among all the factors examined, sample size and scale transformation methods were most important. Larger sample size and concurrent calibration consistently produced less RMSE, less MSD, less bias, and less variance for both item formats. Other factors tended to influence the linking accuracy of CR items more than that of MC items. But in general, a longer test, larger proportion of MC items in the test, more common items, larger sample size, equivalent groups, and concurrent calibration led to more effective linking.

For tests where CR items are included and scored by raters, which is the typical case for mixed-format tests, systematic changes in rater judgments from year to year in

terms of rater severity and leniency may influence the accuracy of equating and linking (Tate, 1999). Tate and his colleagues (2000, 2003, and 2005) conducted a series of studies to investigate rater effect on the accuracy of equating and proposed an equating design that incorporates rater effect into the mixed-format test equating. In the proposed design, a common-item set was still used to disentangle form difference from group ability difference; meanwhile, the same group of raters from second year scored examinee response samples for CR items in a common-item set from both first and second year in order to disentangle rater effect from group ability difference.

Tate (2000) simulated a total of 11 conditions, in which various factors such as sample size, group ability differences, proportion of MC items and CR items in the total test (30/10, 60/0, 0/20), the length of common-item sets (1/5 and 1/10 of the total test), the types of common-item sets (MC+CR, and MC only), and the multidimensionality, were manipulated. The 2PL/GRM model combination was used to generate and calibrate unidimensional response data, as the two-factor model with separate but moderately correlated factors associated with each item format was used to generate multidimensional data. To evaluate the robustness of the proposed linking procedure and compare it to traditional procedures (in which rater severity was assumed constant), the linking coefficients were estimated and compared to the true values to compute the estimated bias and estimated standard error. The results showed that the linking error yielded by the proposed linking procedure was acceptable. As expected, the longer common-item set and/or larger sample size produced more accurate estimates of the linking coefficients. Another result that is more relevant to the current study was that the use of MC items only in a common-item set yielded satisfactory linking accuracy when

the test was unidimensional, but resulted in serious linking error when the test was multidimensional. In the latter case, the use of proper proportions of MC and CR items in a common-item set was more robust to the violation of unidimensionality assumption.

Tate (2000)'s conclusions about the composition of common-item sets were later partly supported by Sykes et al. (2002). In their study, real data from a math field test of large-scale state assessment were used as the data pool, in which there were two dimensions one with both item formats heavily loaded on and the other with only MC items heavily loaded on. From it, they constructed four operational forms, each containing the same set of items but with different combinations of common-item sets. Two common-item sets were balanced with items approximately evenly loaded on both dimensions (referred to as B1 and B2). The other two sets were unbalanced with items heavily loaded on one of each of the dimensions (referred to as F1 and F2). All the common-item sets were content and difficulty representative. The 3PL/GPC model combination was used to estimate IRT parameters simultaneously. Separate calibration with Stocking-Lord method was used to place each of the four operational forms separately onto the field test scale. The equating result using the common-item set B1 was arbitrarily set as the criterion. A weighted mean square difference (WMSD) was then computed to evaluate the other three equatings. The results showed that the equating using two unbalanced common-item sets resulted in more equating errors than those using two balanced common-item sets. Between the two equatings with the unbalanced common-item sets, surprisingly, the one with common items heavily loaded on the common factor yielded considerably larger discrepancy.

The effects of the composition of common-item sets on mixed-format test equating have also been investigated by Kirkpatrick (2005). Both empirical and simulation studies were conducted. In the empirical study, he used the real data from a large-scale state testing program that utilized mixed-format tests (MC, grid-in and restricted response items) in both reading and math subject areas across various grade levels. The unidimensionality assumption was met for the data. Two common-item sets with similar score points, content and statistical representativeness were selected. The only major difference between them was whether or not a restricted-response item was included or excluded. The 3PL/2PL/GPC model combination was used to estimate item and ability parameters. Separate calibration with Stocking-Lord method was applied to place the new form onto the old form scale. And the true score, observed score and estimated ability equating were used to obtain final equivalent scores. Three criteria were used to demonstrate the differences between equating results using two different common-item sets under each equating technique. They are 1) plots of equivalent scores; 2) the classification consistency using fictional cutoff scores; and 3) the distribution of linking coefficients computed by a jackknife-like sampling of the common items. The results presented a mixed picture and differed across subject areas and grade levels. The general finding showed that equating using the common-item set with a restricted response item included or excluded resulted in different equivalent scores. Based on the findings from the empirical study, he further conducted a simulation study to investigate the effects of the multidimensionality related to item formats on equating results. The dimensionality due to item format was quantified as the correlation between two format-specific factors. The correlation was set as 1.0, 0.8, and 0.5 to represent

unidimensionality, low multidimensionality, and severe multidimensionality. The group ability difference was also manipulated. The same equating scenario applied in the empirical study was also followed in the simulation study. One of the results that was most relevant to the current study showed that with different correlations between item formats, whether including a restricted response item in a common-item set or not had different patterns of differences on equating results. If there was a strong correlation between item formats, then the effects of item format alone were minimal. The weaker the correlation between item formats (i.e., tend to violate the unidimensionality assumption), the stronger the effects of item formats in a common-item set on equating results.

Kim & Lee (2006) extended four widely used scale transformation methods: mean/mean, mean/sigma, Haebara, and Stocking-Lord methods to handle mixed-format tests using any mixture of the following five unidimensional IRT models: the 3PL model, the GRM, the GPCM, the NRM and the MC model. They then conducted a simulation study to compare their performance. In their simulation study, random group design was assumed; in which only one mixed-format test was administered to two different groups. The 3PL/GPC model combination was used to generate item response data. Further, this model combination with simultaneous calibration across formats was employed to estimate parameters in the mixed-format test, which guaranteed that the unidimensionality assumption held and the model fit the data well. Four factors were manipulated in their simulation study: group ability differences, sample size, proportion of dichotomously scored (DS) items to polytomously scored (PS) items in the mixed-format test (10/10, 20/5, 30/2), and the composition of common-item sets (DS+PS, DS

only, and PS only). To evaluate the relative performance of the four linking methods across simulation conditions, the category characteristic curve criterion was used. The results showed that with few exceptions, the characteristics curve methods generally produced more accurate linking results than the moment methods. Keeping other factors constant, linking using equivalent groups resulted in less linking error than the nonequivalent groups linking. Linking using both DS and PS as a common-item set usually yielded more accurate results than linking using only DS or PS as a common-item set. Furthermore, the comparison between linking using DS only and linking using PS only showed that linking using the "dominant" item format in the mixed-format test led to lower linking error.

Kim & Kolen (2006) further investigated the above extended linking methods proposed by Lee & Kim (2006) to see whether they were robust to multidimensionality due to item formats in comparison to the concurrent calibration under CINEG design. The item parameters of the science assessment from the 1996 NAEP were used as source for their simulation study. Multidimensional data were then generated in such a way that two correlated format-specific factors were assumed. Other factors investigated were three levels of correlation between two factors (1, 0.8 and 0.5), two types of mixed-format tests (wide-range and narrow-range), and three levels of nonequivalence between two groups (group mean = 0, 0.5 and 1). The common-item set was constructed to include both MC and CR items proportionally, also to be sufficiently long, content and statistical representative. The 3PL/GPC model combination was then used to simulate the multidimensional data. Both concurrent calibration and separate calibration with four linking methods were used to place the new form onto the old form scale. Finally, the

equating results were evaluated using the observed score distribution (OSD) criteria, which was based on the difference between the estimated and true OSDs. The results showed that the concurrent calibration generally outperformed the separate calibration with various linking methods in terms of linking accuracy and robustness to multidimensionality due to item formats. Among linking methods, the characteristic curve methods resulted in more linking accuracy than the moment methods, regardless of the degrees of multidimensionality. However, the differences in linking results using the concurrent calibration and separate calibration with the characteristic curve methods were small.

Implications

As mentioned in the beginning of this section, only a few studies have been conducted in the area of mixed-format test equating. Existing literature in this area is mainly concentrated on the following aspects: 1) technical review and extension of current linking methods for single item format to the mixture of MC and CR items; 2) comparison of relative performance of different linking methods; 3) the new procedure proposed to handle the impact of rater severity on the linking accuracy; 4) investigation of the effects of several factors (e.g., test length, sample size, group ability distributions, the number of common items, and the proportion of MC and CR items in the total test, etc.) on the accuracy of mixed-format test equating; 5) examination of multidimensionality due to item format on equating results; and 6) evaluation of the influence of the composition of common-item sets on equating (Kim & Lee, 2006; Kirkpatrick, 2005; Sykes et al., 2002; Tate, 2000).

The composition of common-item sets under the multidimensional test situation due to multiple item formats will be of primary concern in this study. Although research on the effects of multidimensionality and the composition of common-item sets for mixed-format test equating exists and is informative, it still leaves room for improvement in our knowledge base. Previous research on multidimensionality (Kim & Kolen, 2006; Kirkpatrick, 2005; Sykes et al., 2002; Tate, 2000) only simulated data based on the two-factor model with correlated format-specific factors. However, as pointed out by Kim & Kolen (2006), the underlying test structure could be far more complicated in reality in a way that more than one factors could affect an examinee's correct response to each item and different items may require different combinations of factors for an examinee to respond correctly. Moreover, the relative influential powers of various factors on examinees' item responses could vary to a different extent. In this case, the multidimensional IRT models might better capture the test structure and the results might be generalized to more realistic situations. However, no current research investigates the impact of this more general multidimensional test structure on mixed-format test equating.

Meanwhile, previous research on the composition of common-item sets (Kim & Lee, 2006; Kirkpatrick, 2005; Sykes et al., 2002; Tate, 2000) only focused on whether to include or exclude CR items in the common-item sets. They all assumed content and statistical representativeness of the common-item sets, and furthermore, if both item formats were included in the common-item sets, both item formats were constructed proportionally representative of the total test. No research has explored the interactive effects of representativeness and non-representativeness of content, statistical and format specifications on mixed-format test equating.

In all, the current literature does not provide adequate direction for test developers in designing mixed-format tests, especially in terms of constructing representative common-item sets under different multidimensional test structures for equating purposes. This study is intended to contribute useful information to this area of inquiry.

Chapter 3: Methods

The main purpose of this study is to investigate the impact of representativeness and non-representativeness of common-item sets in terms of content, statistical and format specifications on mixed-format test equating using concurrent calibration with unidimensional IRT models and how robust is the procedure under various conditions of multidimensional test structure. In order to fulfill this purpose, a simulation study is conducted, which not only allows for assessing the effects of factors of interest in the ideal situation but also provides true population values that can be used as baseline for evaluation. In this chapter, the methodological framework for the simulation study is described. It is divided into five sections. The first section specifies the configuration of the mixed-format test forms simulated in this study. It is followed by detailed descriptions of the factors of investigation. Next, a step-by-step procedure of generating response data and conducting quality control is provided. Then the equating scenario is described. Finally, the criteria for evaluating the results are presented.

Test Configuration

The mixed-format test simulated in this study is specified to reflect one reasonable configuration for a large-scale high-stakes assessment. It considers two test forms for equating.

Each test form consists of items from two content areas: Content Area 1 and Content Area 2 (see examples of the Medical College Admission Test (MCAT), Childs & Oppler, 2000). These two content areas are represented by two distinct but correlated content factors. Each content area contains one half of the items in the total test, which

assumes that each content area occupies the same weight in the total test. In each content area, there are both MC and CR items.

Test length is an important issue in test configuration to ensure accurate item parameter estimation and successful equating (Fitzpatrick & Yen, 2001; Hambleton, & Cook, 1983). Bastari (2000) found that the longer the test, the more accurate the equating result. Since a mixed-format test is considered in this study, test length is defined in terms of both the number of items and the number of score points per item. Therefore, in order to ensure that the test length will not be a potential factor influencing equating results, each test form in this study is comprised of 54 items including 48 dichotomous MC items (0/1) and 6 five-category CR items (0/1/2/3/4), which results in the number-correct score of 72 in the total test. Meanwhile, the ratio of MC and CR items in the total test in terms of item numbers is 8:1, and the ratio in terms of score points is 2:1. These ratio settings of MC and CR items are similar to some state assessments, such as the Florida statewide assessment programs and the Arkansas 2002 field test. This arrangement of MC and CR items results in 27 items (24 MC items and 3 CR items) per content area.

Furthermore, the number of common items to use in CINEG design will also influence the equating results (Bastari, 2000; Hanick & Huang, 2002). Kolen and Brennan (2004) suggested a rule of thumb regarding the number of common items, that is, "a common-item set should be at least 20% of the length of a total test containing 40 or more items" (p. 271). To ensure that the number of common items is sufficient for accurate equating, 18 common items are used in this study, which represents one third of the total test. This proportion of common items was also used by Kim & Kolen (2006).

Factors of Investigation

Test Dimensionality Structure (5 levels)

Unidimensionality

The test scenario under this condition is that the test is truly unidimensional and measures examinees' general ability within a certain domain. It further indicates that although the test consists of items from two content areas, these two content areas are perfectly correlated (ρ=1.0). Meanwhile, various item formats measure the same ability. Figure 3.1 demonstrates this unidimensional test structure.

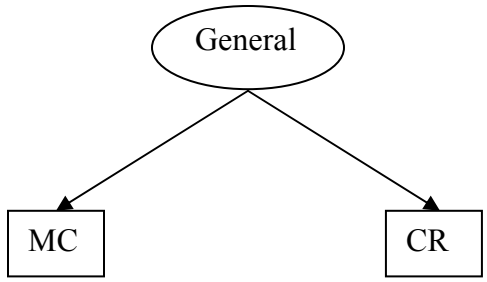


Figure 3.1 Unidimensional Test Structure

Multidimensionality

A more realistic and general test scenario is described using the multidimensional test structure. Under this condition, the test consists of items from two content areas within a certain domain and these two content areas are distinct but highly correlated. Two content factors ($\theta_{Content1}$ and $\theta_{Content2}$) are used to represent these two content areas and the levels of correlation between them (ρ) are set as 0.9 and 0.75, which are reasonably high values as might be found in many real assessments (Bolt, 1999; Childs & Oppler, 2000; Wu & Adams, 2006). Meanwhile, a common hypothesis to use a mixed-format test is that even though MC and CR items are used to assess the same content area, they might measure examinees' different abilities (Traub, 1993). Therefore, in addition to two content factors, two orthogonal format factors with one specific for MC items (θ_{MC})

and the other specific for CR items (θ_{CR}) are introduced. These two format factors are set to be orthogonal because they only represent their unique contributions to examinees' response above and beyond those attributed by the content factors. Furthermore, the relationship between content and format factors is also set to be orthogonal. Figure 3.2 demonstrates this multidimensional test structure.

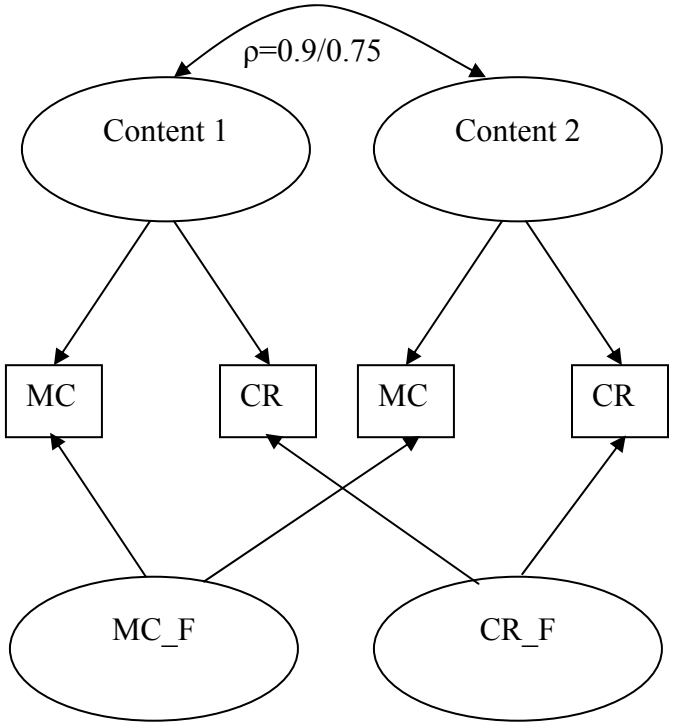


Figure 3.2 Multidimensional Test Structure

Figure 3.2 also indicates that an examinee's correct response to each item is determined by both content and format factors. However, different items may require different combinations of content and format factors for an examinee to respond correctly. Moreover, the relative importance of content and format factors on examinees' item responses is also manipulated through controlling for the item discrimination parameters associated with them. More specifically, the magnitude of the item discrimination parameters is manipulated using the angle (α) between the multidimensional item discrimination parameter (*MDISC*) and the $\theta_{content}$-axis (see Figure 3.3). The smaller the

angle, the higher the item discrimination parameter for the content factor, and the more important is the content factor in responding to the item correctly. The angle of 45 degrees indicates that the content and format factors are equally important. In this study, the content factors are always set to have stronger influence on item responses than the format factors do, which is a realistic assumption in practice since the main purpose of test construction is to measure examinees' ability to master content knowledge. Therefore, two angle degrees are set: 10° and 35°, in which 10° is in the angle range (0° -15°) representing items that highly load on the content factors and 35° is in the angle range (25° -40°) representing items that are sensitive to the composite of the content and format factors (Min, 2003; Yon, 2006).

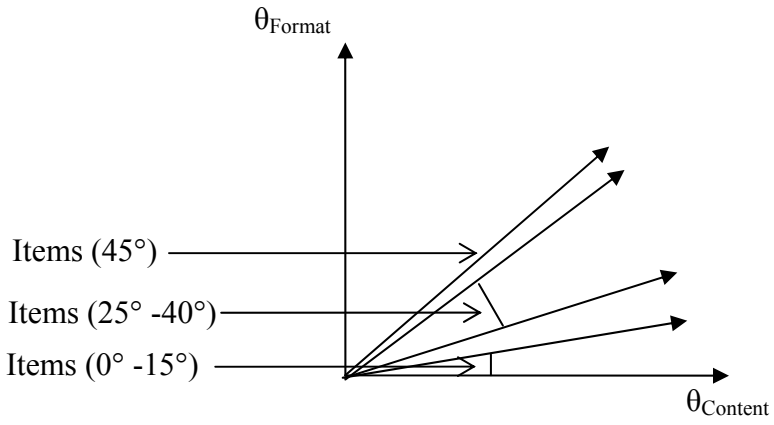


Figure 3.3 Demonstration of the relative importance of content and format factors

In all, two levels of correlation between content factors and two angle degrees yield four conditions of multidimensional test structure.

Format Representativeness (2 levels)

Format representativeness means that the ratio of MC items to CR items in the common-item set corresponds to the similar ratio of MC items to CR items in the total test in terms of the item number and score points. Two conditions are considered:

1. Format representativeness: The ratio of MC items to CR items in the total test is 8:1 in terms of item numbers and 2:1 in terms of score points. The similar ratios are reflected in the common-item set, which results in 16 MC items and 2 CR items.
2. Format non-representativeness: This study only considers one situation of format non-representativeness, in which only MC items are included in the common-item set. This is a frequently-used but questionable strategy to construct a common-item set especially when mixed-format test equating is applied and the unidimensionality assumption is violated. In this case, 18 common items are exclusively MC items.

Content Representativeness (3 levels)

This study adapts the definition of content representativeness proposed by Klein & Jarjoura (1985). In the common-item set, "the proportions of items from each content area correspond to the proportion of items from those content areas in the total tests" (Klein & Jarjoura, 1985, p. 198). Three conditions are considered:

1. Content representativeness: In this study, two content areas are designed to have equivalent weights in the total test. Therefore, in the common-item set, two content areas occupy 1/2 and 1/2 as in the total test, which means 9 items per content area in the common-item set.
2. Content non-representativeness (partially under-representative): In the common-item set, one content area is partially under-represented, which is a more typical case in real assessments. In this study, Content Area 1 occupies nearly 2/3 and Content Area 2 occupies nearly 1/3 in the common-item set. Therefore, 12 items represent Content Area 1 and 6 items represent Content Area 2.

3. Content non-representativeness (completely under-representative): In the common-item set, one content area is completely under-represented, which is an extreme case. In this study, one content area (Content Area 2) is missing in the common-item set. Therefore, 18 items represent Content Area 1 and 0 item represents Content Area 2.

Format and content representativeness combinations result in the following six conditions:

	Format Representativeness (MC:CR=8:1)	Format Non-Representativeness (MC items only)
Content Representativeness (1/2, 1/2)	8:1, 8:1	9, 9
Partially Under-representativeness (nearly 2/3, 1/3)	11:1, 5:1	12, 6
Completely Under-representativeness (1, 0)	16:2, 0	18, 0

Statistical Representativeness (2 levels)

Statistical representativeness refers to the average item difficulty in the common-item set being similar to that in the total test (Gao et al., 1999; Petersen, Marco, & Stewart, 1982). For MC items, the computation of average item difficulty is straightforward. While for CR items, no research has yet addressed this issue. Therefore, in this study, the between category threshold parameters for each category in each item (the GRM and corresponding parameters will be discussed in the data generation section) are computed as the index of item difficulty for CR items. Two conditions are considered:

1. Statistical representativeness: The average item difficulty in the common-item set is similar to that in the total test.
2. Statistical non-representativeness: The average item difficulty in the common-item set is 0.3 mean difficulty different from that in the total test (See Appendix B and C for detailed parameter settings).

Group Ability Distributions (2 levels)

Two groups of examinees are considered in this study: Group 1 and Group 2. The equating is conducted so that ability parameters of examinees in Group 2 are on the Group 1 scale. Therefore, the ability distribution of Group 1 is always set as a standard normal distribution with mean of 0 and standard deviation of 1. The ability distribution of Group 2 could differ from that of Group 1 in terms of the mean and/or standard deviation. In this study, only the mean differences are considered and the standard deviations are held constant, which is one of the typical treatments in many simulation studies (Kim & Kolen, 2006; Kim & Lee, 2006; Kirkpatrick, 2005). Two conditions are considered:

1. Equivalent groups: It is assumed that the ability distributions of Group 1 and Group 2 are equivalent, that is, the two groups have the same normal distributions with mean of 0 and standard deviation of 1. Although CINEG design does not require the ability distributions of two groups to be equivalent, this study still considers the equivalent groups as a baseline level, against which the level of nonequivalence is compared.
2. Nonequivalent groups: It is assumed that the ability distributions of Group 1 and Group 2 are not equivalent. Specifically, examinees in Group 2 are more competent than those in Group 1. Therefore, the ability distribution of Group 2 has a higher population mean of 0.5 but unchanged standard deviation of 1. A difference of 0.5 in the mean proficiency between the two groups is chosen because it has been shown to be big enough to reflect the effect of group difference on equating (Li & Lissitz, 2000).

Implications

Five factors of investigation can be further categorized into three groupings.

First is the test dimensionality structure factor, which has a total of five levels with a unidimensional test structure and four multidimensional test structures. It should be noted that for the practitioners in the field of test equating, once the test is developed, it is not a factor that could be manipulated.

The second grouping includes the format, content, and statistical representativeness. When the CINEG design is applied, these are the three most important characteristics in the composition of common-item sets. How to construct a most representative and efficient common-item set is one of the top concerns for the equating practitioners and should be handled with great caution.

The last factor of interest is the group ability distributions. Two levels are included, either the equivalent groups or the nonequivalent groups. It should be noted that the CINEG design itself does not require the use of equivalent groups. As a matter of fact, in many operational settings especially in licensure testing, the groups of examinees taking different forms are usually not considered to be equivalent. Therefore, the composition of common-item sets under nonequivalent groups condition should be worthy of more attention. In addition, this factor also cannot be controlled by the equating practitioners.

Data Generation

The item responses of examinees in Group 1 taking test form 1 and those in Group 2 taking test form 2 are generated separately. For each group under each design condition, 3,000 examinees' responses are simulated using the appropriate IRT models. This sample size is chosen as it has been shown to be generous enough to yield accurate equating results (Hanson & Beguin, 2002; Kim & Lee, 2004; Kirkpatrick, 2005).

Three steps are followed to accomplish the data generation process and all the steps are executed using SAS. Detailed procedure for each step is described as follows:

Step 1: Ability Parameter Generation

To generate examinee ability parameters for Group 1 and Group 2 separately, two factors need to be considered: test dimensionality structure and group ability distributions.

Unidimensional test structure

Under the unidimensional condition, one factor (say, θ_G) affects examinees' responses to both MC and CR items. The examinees' ability parameters (θ_G) in Group 1 are randomly drawn from a standard normal distribution (N(0, 1)) and those in Group 2 are from pre-specified normal distributions (N(0, 1) for equivalent groups, and N(0.5, 1) for nonequivalent groups) which can be achieved from a random normal deviation generator (RANNOR) in SAS.

In all, a total of two sets of population ability parameters are generated based on the distributions shown in Appendix A.

Multidimensional test structure

Under the multidimensional condition, the examinees' ability parameters (say, $\theta_{Content1}$, $\theta_{Content2}$, θ_{MC} and θ_{CR}) in Group 1 and Group 2 are randomly drawn from a multivariate normal distribution with pre-specified mean and variance-covariance matrix. The relationship between content and format factors and that between two format factors are orthogonal. The correlation between two content factors is set at 0.9 and 0.75. The mean difference between equivalent Group 1 and Group 2 is 0 and that between nonequivalent groups is 0.5. To obtain the correlated ability parameters, VNORMAL CALL command in SAS is used.

In all, a total of four sets of population ability parameters are generated based on the distributions shown in Appendix A.

Step 2: Item Parameter Generation

Two test forms are generated for equating. Test form 1 consists of a unique item set and a set of common items. Test form 2 consists of a different set of unique items specific to form 2 and the same common-item set as that in form 1. These three item sets (a unique item set for form 1, a unique item set for form 2, and a common-item set) are generated separately.

Test dimensionality structure, format representativeness, content representativeness and statistical representativeness all influence the process of generating item parameters.

Unidimensional test structure

Under the unidimensional condition, the unidimensional 3PL model for MC items can be expressed as:

$$P_i(\theta) = c_i + (1 - c_i)\frac{\exp[Da_i(\theta - b_i)]}{1 + \exp[Da_i(\theta - b_i)]} \tag{3.1}$$

where D is a scaling constant (1.7), a_i is the item discrimination parameter, b_i is the item difficulty parameter, and c_i is the guessing parameter. The MC items in the unique item set for form 1 are generated as follows. The item discrimination parameters (a_i) in Content Area 1 and 2 are sampled from a log-normal distribution with mean of 0 and standard deviation of 0.5 of the logarithm, which is the default distribution setting of item discrimination parameters in BILOG-MG. In addition, the range for the item discrimination parameters is set from 0.5 to 2.5, which was used by Spence (1996). The item difficulty parameters (b_i) are sampled from a normal distribution with mean of 0 and

standard deviation of 1. In addition, the range for the item difficulty parameters is set from -2.0 to 2.0, which was used in many studies (Finch, 2006; Spence, 1996). The guessing parameters (c_i) are sampled from a beta distribution with $\alpha=8$ and $\beta=32$, which was used by Kim & Lee (2006).

Meanwhile, the unidimensional GRM for CR items can be expressed as:

$$P_{ij}^{*}(\theta)=\begin{cases}1 & j=1\\ \dfrac{\exp[Da_i(\theta-b_{ij})]}{1+\exp[Da_i(\theta-b_{ij})]} & 2\le j\le J\\ 0 & j>J\end{cases} \qquad (3.2)$$

where category $j=1, 2,..., J$, D is a scaling constant (1.7), a_i is the item slope parameter, each item has its own a_i. b_{ij} is the between category threshold parameter of category j in item i whose value represents the point on the θ continuum where individuals have a 50% chance of responding in or above category j. The CR items in the unique item set for form 1 are generated as follows. The item slope parameters (a_i) are generated from the same log-normal distribution as for the 3PL model (*LN (0, 0.5)* with the range from 0.5 to 2.5). Since all the CR items are assumed to have five categories, four between category threshold parameter (b_{ij}) are sequentially sampled from *N(-1.5, 0.2), N(-0.5, 0.2), N(0.5, 0.2),* and *N(1.5, 0.2)*, which was used by Kim & Lee (2006). It should be noted that the between category threshold parameters in GRM are always in the ascending order ($b_{i1}< b_{i2}< b_{i3}< b_{i4}$). If the order is reversed, the item parameters will be flagged and re-sampled from the normal distributions.

When the statistical representativeness of the common-item set is satisfied, the same item parameter distributions that are used to generate the unique item set for form 1 are also used to generate item parameters for the common-item set. However, when the statistical representativeness is not met, the distributions of item difficulty parameters (b_i)

in the 3PL model and between category threshold parameters (b_{ij}) in GRM for the common-item set are shifted in increments of 0.3 over the corresponding parameter distributions in the unique item set for form 1.

The item discrimination parameters and guessing parameters in the unique item set for form 2 are generated from the same item parameter distributions as in the unique item set for form 1. However, since the purpose of equating is to statistically adjust difficulty differences across forms, the distributions of item difficulty parameters (b_i) in 3PL model and between category threshold parameters (b_{ij}) in GRM for the unique item set for form 2 are shifted in increments of 0.5 over the corresponding parameter distributions in the unique item set for form 1, which indicates that form 2 is more difficult than form 1.

In all, under the unidimensional condition, a total of 12 (2 Format Representativeness × 3 Content Representativeness × 2 Statistical Representativeness) sets of population item parameters are generated. Detailed item parameter distributions and the number of MC and CR items across each item set and each content area under each of 12 simulation conditions are illustrated in Appendix B.

Multidimensional test structure

Under the multidimensional condition, an examinee's response to each item is determined by two factors (one content factor and one format factor), and different items require different combinations of content and format factors for an examinee to respond correctly. The multidimensional 3PL (M-3PL) model (Reckase, 1985) for MC items can be expressed as:

$$P_i(\theta) = c_i + (1 - c_i)\frac{\exp[D(\sum_{h=1}^{H} a_{ih}\theta_h + d_i)]}{1 + \exp[D(\sum_{h=1}^{H} a_{ih}\theta_h + d_i)]} \quad (3.3)$$

where D is a scaling constant (1.7), θ_h is the latent ability on dimension h, a_{ih} is the parameter related to the discriminating power of item i on dimension h, d_i is the parameter related to the difficulty of item i, and c_i is the guessing parameter. $MDISC_i$ and $MDIFF_i$ are two parameters derived from M-3PL model. They represent the overall item discrimination and item difficulty for item i and thus can be interpreted in the same way as those in the unidimensional 3PL model.

$$MDISC_i = \sqrt{\sum_{h=1}^{H} a_{ih}^2} \quad (3.4)$$

$$MDIFF_i = \frac{-d_i}{MDISC_i} \quad (3.5)$$

In this study, $MDISC_i$ and $MDIFF_i$ in the unique set for form 1 are generated first. $MDISC_i$s in Content Area 1 and 2 are sampled from a log-normal distribution with mean of 0 and standard deviation of 0.5 of the logarithm[1] as that under the unidimensional condition. In addition, the range of $MDISC_i$s is set from 0.5 to 2.5, which is chosen according to the results of empirical studies reported by Ackerman (1988), Doody-Bogan & Yen (1983), Spence (1996), and Roussos, Stout, & Marden (1998). $MDIFF_i$s are sampled from a standard normal distribution with the range from -2.0 to 2.0 which is determined based on the previous studies so that it is reasonable for published tests

[1] There is no universal agreement on the distribution of $MDISC_i$. However, most of the researchers believed that $MDISC_i$ follows the lognormal distribution (Finch, 2006; Min, 2003; Spence, 1996; Tate, 2003).

(Finch, 2006; Spence, 1996). The values of a_{i1}, a_{i2}, and d_i are then determined using the following equations:

$$a_{i1} = MDISC_i \times \cos(\alpha_i) \tag{3.6}$$

$$a_{i2} = MDISC_i \times \sin(\alpha_i) \tag{3.7}$$

$$d_i = -MDISC_i \times MDIFF_i \tag{3.8}$$

where α_i is the angle between $MDISC_i$ and $\theta_{content}$-axis and takes the values of 10° and 35°. The guessing parameters (c_i) have the same meaning as in the unidimensional 3PL model and are sampled from a beta distribution with $\alpha=8$ and $\beta=32$.

Meanwhile, the multidimensional graded response model (M-GRM) for CR items can be expressed as:

$$P_{ij}^{*}(\theta) = \begin{cases} 1 & j = 1 \\ \dfrac{\exp[D(\sum_{h=1}^{H} a_{ih}\theta_h + d_{ij})]}{1 + \exp[D(\sum_{h=1}^{H} a_{ih}\theta_h + d_{ij})]} & 2 \le j \le J \\ 0 & j > J \end{cases} \tag{3.9}$$

where all the parameters have similar meanings to the M-3PL model except for d_{ij}, which is the parameter related to the between category threshold of category j in item i. The same item parameter generation process is followed for CR items in the unique item set for form 1 as that for MC items using M-3PL model. $MDISC_i$ and $MDIFF_{ij}$ ($= \dfrac{-d_{ij}}{MDISC_i}$) are generated first. $MDISC_i$ follows the same log-normal distribution as for the M-3PL model (*LN (0, 0.5)* with the range from 0.5 to 2.5). Since all the CR items are assumed to have five categories, $MDIFF_{ij}$ are sequentially sampled from *N(-1.5, 0.2), N(-0.5, 0.2), N(0.5, 0.2),* and *N(1.5, 0.2)*. Then Equations (3.6) and (3.7) along with the same angel

degrees (10° and 35°) are used to generate a_{i1} and a_{i2}, and the equation of $d_{ij} = -MDISC_i \times MDIFF_{ij}$ is used to generate d_{ij}.

When the statistical representativeness of the common-item set is satisfied, the same item parameter distributions used to generate the unique item set for form 1 are used to generate item parameters for the common-item set. However, when the statistical representativeness is not met, the means of the distributions of $MDIFF_i$ in M-3PL model and $MDIFF_{ij}$ in M-GRM for the common-item set are increased by 0.3 over the corresponding parameter distributions in the unique item set for form 1.

The $MDISC_i$s and guessing parameters in the unique item set for form 2 are generated from the same item parameter distributions as in the unique item set for form 1. However, the form 2 is assumed to be more difficult than the form 1 by increasing the means of the distributions of $MDIFF_i$ in M-3PL model and $MDIFF_{ij}$ in M-GRM for the unique item set for form 2 by 0.5 compared to those in the unique item set for form 1.

In all, under the multidimensional condition, in addition to the format, content and statistical representativeness, two angle degrees between $MDISC_i$ and the content factor also need to be taken into account, which yield a total of 24 (2×3×2×2) sets of population item parameters. Detailed item parameter distributions and the number of MC and CR items across each item set and each content area are illustrated in Appendix C.

Step 3: Response Data Generation

Given the item parameters for each test form and the ability parameters for each group, applicable IRT models are used to generate appropriate correct response probabilities. Then we compare these values of correct response probability to the values of the uniform random number in the range (0, 1) to assign the item responses. Different

IRT models are used to generate item responses under the various conditions of test dimensionality structure.

Unidimensional test structure

Under the unidimensional condition, two sets of population ability parameters and 12 sets of population item parameters have been generated in Step 1 and Step 2. Therefore, a total of 24 sets of examinees' response data are simulated.

The unidimensional 3PL model expressed in Equation (3.1) is used to compute the probability that an examinee with ability θ correctly responds to MC item i ($P_i(\theta)$). Then compare the value of the correct probability ($P_i(\theta)$) to a value of the uniform random number (U_i) to generate a dichotomous item response of an examinee with ability of θ to MC item i (R_i) based on the following rule:

$$R_i = \begin{cases} 0, & P_i(\theta) \le U_i \\ 1, & P_i(\theta) > U_i \end{cases}$$

The GRM requires a two-step procedure. The first step is to estimate operating characteristic curves ($P^*_{ij}(\theta)$) using Equation (3.2), which represents the conditional probability of an examinee's response falling in or above a given item category. Once the $P^*_{ij}(\theta)$ are estimated, the second step is to compute the actual category response curves using the following equation,

$$P_{ij}(\theta) = P^*_{ij}(\theta) - P^*_{i,j+1}(\theta) \qquad (3.10)$$

They represent the conditional probabilities of an examinee responding to a particular category. Then compare the value of the operating probability ($P^*_{ij}(\theta)$) to a value of the uniform random number (U_i) to generate a polytomous item response of an examinee with ability of θ to CR item i (R_i) based on the following rule:

$$R_i = \begin{cases} 0, & P_{i2}^*(\theta) \le U_i < 1 \\ 1, & P_{i3}^*(\theta) \le U_i < P_{i2}^*(\theta) \\ 2, & P_{i4}^*(\theta) \le U_i < P_{i3}^*(\theta) \\ 3, & P_{i5}^*(\theta) \le U_i < P_{i4}^*(\theta) \\ 4, & 0 \le U_i < P_{i5}^*(\theta) \end{cases}$$

Multidimensional test structure

Under the multidimensional condition, four sets of population ability parameters and 24 sets of population item parameters have been generated in Step 1 and Step 2. Therefore, a total of 96 sets of examinees' response data are generated. Compensatory multidimensional IRT models (Way, Ansley, & Forsyth, 1988), in which an examinee's high ability on one factor can potentially compensate for deficient or lower ability on the other factor, are used to generate response data.

The M-3PL model expressed in Equation (3.3) is used to compute the probability that an examinee with multiple abilities ($\theta_{content}$ and θ_{format}) correctly responds to MC item i ($P_i(\theta)$). Then compare the value of the $P_i(\theta)$ to a value of the uniform random number (U_i) to generate a dichotomous item response to MC item i (R_i) based on the same rule as demonstrated under the unidimensional condition.

The M-GRM expressed in Equation (3.9) is used to compute the operating characteristic curves ($P^*_{ij}(\theta)$), and Equation (3.10) is used to compute the actual category response curves ($P_{ij}(\theta)$). Then compare the value of the $P^*_{ij}(\theta)$ to a value of the uniform random number (U_i) to generate a polytomous item response to CR item i (R_i) based on the same rule as demonstrated under the multidimensional condition.

Procedure for Quality Control

Following each step of the data generation process, a strict quality control procedure was conducted. Additional SAS codes were written to test whether the

generated ability and item parameters follow the pre-specified population distributions. Furthermore, EXCEL spreadsheets were created to check whether examinees' correct response probabilities were properly computed and their item responses were correctly assigned.

Equating Scenario

Following the three-step procedure as demonstrated in the Data Generation section, item response data for group 1 examinees taking test form 1 and group 2 examinees taking test form 2 are generated. Concurrent calibration is then conducted to put the two sets of parameters estimates onto the common scale through the common-item set. To conduct concurrent calibration, the first step is to combine the response data from both groups of examinees and treating items not taken by a particular group as not reached or missing. Meanwhile, specify group membership, 1 or 2, in the first column of the combined data. Second, pre-select a scale before running the calibration (here, group 1 examinees' ability scale), using the common items as the anchor. Then, the unidimensional 3PL model is used to calibrate the MC items and the GRM is used to calibrate the CR items. The computer program, MULTILOG (Thissen, 1991), is used to estimate the parameters of MC and CR items on both forms simultaneously in one run which guarantees that all parameter estimates are on the same pre-selected scale. Marginal Maximum Likelihood (MML) method is used to estimate item parameters and Maximum A Posteriori (MAP) method is used to estimate ability parameters. They are default estimation methods in MULTILOG.

Cautions should be taken on several issues when running MULTILOG. First, MULTILOG provides researchers with tools to handle multi-group data as well as mixed-

format tests. However, MULTILOG is not well suited to conduct non-equivalent groups equating via concurrent calibration, because item and ability parameter estimates from the two forms cannot be simultaneously placed on the scale for the old group without specifying the standard deviation of the population distribution for the new group. The population distribution of the new group is rarely known in practical situations. However, in a simulation study like the present one, the standard deviation of the population distribution for the new group is pre-specified. Then the default arrangement of MULTILOG is used in which the mean of the old group is fixed as 0 and the standard deviations of both groups are fixed at 1. The mean of the new group needs to be estimated. Second, when MULTILOG was used to conduct concurrent calibration, group membership was specified in the combined response data. This is an important step especially when MAP method is chosen to estimate examinees' ability parameters from nonequivalent groups. Then the MAP estimates for each group will shrink to its own distribution mean instead of to the combined group distribution mean. Third, all models estimated with MULTILOG except for 3PL model are truly "logistic", which means that there is no D=1.7 scaling factor. However, the 3PL model is estimated in normal metric. Therefore, more attention is paid when generating and estimating item parameters using the 3PL model and GRM.

Examples of MULTILOG syntaxes for item and ability calibrations are provided in Appendix D.

Replications

Test dimensionality structure, format representativeness, content representativeness, statistical representativeness, and group ability distributions are fully

crossed, leading to a total of 120 (5 × 2 × 3 × 2 × 2) simulation conditions. Data generation and equating process under each condition is replicated 100 times.

Evaluation Criteria and Data Analysis

The population parameters of the test form 1 (taken by Group 1) and the test form 2 (taken by Group 2) are generated based on the same scale. Then in order to fix the indeterminacy of the scale and to keep the estimated parameters on the same scale as their population parameters, the mean and standard deviation of estimated ability parameters in test form 1 (taken by Group 1) are set to their population parameters during the calibration. Since the purpose of equating in this study is to put the test form 2 estimated parameters onto the test form 1 scale, it is expected that in each simulation condition, through concurrent calibration, the estimated parameters for the test form 2 (taken by Group 2) should be on the same scale as parameters for the test form 1 (taken by Group 1), which is also the scale for the population parameters.

The examinees' expected total scores on all items in form 2 computed using the population parameters ($E(X_k \mid \theta_k)$) are compared with those computed using the estimated parameters ($E(X_k \mid \widehat{\theta}_k)$). The $E(X_k \mid \theta_k)$ can be expressed as:

$$E(X_k \mid \theta_k) = \sum_{i=1}^{n_{MC}} P_i(\theta_k) + \sum_{i=1}^{n_{CR}} \sum_{j=1}^{J} u_{ij} P_{ij}(\theta_k) \qquad (3.11)$$

where θ_k is the ability or ability vector for an examinee k, u_{ij} represents the category scores, *0, 1, 2, 3, 4*. Under the unidimensional condition, equation (3.1) is used to compute the $P_i(\theta_k)$, and equations (3.2) and (3.10) are used to compute the $P_{ij}(\theta_k)$. Under the multidimensional condition, equation (3.3) is used to compute the $P_i(\theta_k)$, and equations (3.9) and (3.10) are used to compute the $P_{ij}(\theta_k)$.

Correspondingly, the $E(X_k \mid \widehat{\theta}_k)$ can be expressed as:

$$E(X_k \mid \widehat{\theta}_k) = \sum_{i=1}^{n_{MC}} P_i(\widehat{\theta}_k) + \sum_{i=1}^{n_{CR}} \sum_{j=1}^{J} u_{ij} P_{ij}(\widehat{\theta}_k) \tag{3.12}$$

Since only the unidimensional IRT models are used to calibrate response data, equation (3.1) is always used to compute the $P_i(\theta_k)$, and equations (3.2) and (3.10) are used to compute the $P_{ij}(\theta_k)$.

Several summary criteria are then used to evaluate the accuracy of the equating results: (1) the BIAS; (2) the RMSE (Root Mean Squared Error); and (3) the classification consistency.

The BIAS shows the differences between the average of true expected total scores and the average of estimated expected total scores. It can be expressed as

$$BIAS_r = \frac{\sum_{k=1}^{3000} (E(X_k \mid \widehat{\theta}_k) - E(X_k \mid \theta_k))}{3000} \tag{3.13}$$

Where 3,000 is the number of examinees in each group. The average of the BIASs is taken over the 100 replications. The average BIAS indicates the accuracy of the equating results as well as the direction.

The RMSE shows the extent to which the estimated expected total scores match the true expected total scores. It can be expressed as

$$RMSE_r = \sqrt{\frac{\sum_{k=1}^{3000} (E(X_k \mid \widehat{\theta}_k) - E(X_k \mid \theta_k))^2}{3000}} \tag{3.14}$$

The average of the RMSEs is taken over the 100 replications. The smaller the average RMSE, the better the equating result is.

The classification consistency index is defined as the percentage of time that the same decision is reached based on the true scores and the estimated scores. The total score range for test form 1 and 2 is from 0 to 72. Two cut scores (24 and 48) are used to classify examinees' performance into three proficiency categories: Basic ([0, 24)), Intermediate ([24, 48)), and Proficient ([48, 72]). The classification consistency index then can be expressed as $P=P_{00}+P_{11}+P_{22}$ (Equation (3.15) as shown in the 3×3 contingency table below.

		Estimated scores		
		Basic	Intermediate	Proficient
True scores	Basic	P_{00}	P_{01}	P_{02}
	Intermediate	P_{10}	P_{11}	P_{12}
	Proficient	P_{20}	P_{21}	P_{22}

Then, the average of the classification consistency is taken over the 100 replications. The higher the average proportion, the better the equating result is.

Plots for the overall BIAS, RMSE and the classification consistency are separately drawn to better demonstrate the differences under various simulation conditions of interest. Moreover, a series of statistical tests including two sample t tests for independent groups and/or analysis of variance (ANOVA) are applied to further indicate whether the differences of factors of interest are statistically significant or not. Finally, multiple comparison procedure is used if necessary.

Chapter 4: Results

In this chapter, the results from the simulation study described in Chapter Three are summarized. The presentation of the results is divided into three sections in response to the three research questions proposed in Chapter One. Each section is completed by a summary of the findings. For succinctness of presentation, supporting tables of ANOVA results not discussed elaborately are provided in Appendix E and F.

Research Question 1

To answer research question 1: "In an ideal situation where the unidimensionality assumption is satisfied, what are the effects of content, statistical and format representativeness of common-item sets on mixed-format test equating using concurrent calibration with unidimensional IRT models?" The unidimensional test structure was simulated and then the data generation process was strictly followed as demonstrated in Chapter Three. The various evaluation criteria (BIAS, RMSE, and Classification Consistency) were computed and compared.

Table 4.1 presents the average BIAS, RMSE, and classification consistency proportion over 100 replications under the unidimensional test structure. There are a total of 24 simulation conditions in this category which are presented in a $2 \times 2 \times 3 \times 2$ contingency table. The four factors of investigations are listed using abbreviations, which are noted right below Table 4.1. In details, they are Group Ability Distributions (EQ represents equivalent groups, and NEQ represents nonequivalent groups.), Format Representativeness (FR represents format representativeness, and FNR represents format non-representativeness.), Content Representativeness (CR represents content representativeness, CPU represents content partially under-representativeness, and CCU

represents content completely under-representativeness.), and Statistical Representativeness (SR represents statistical representativeness, and SNR represents statistical non-representativeness.).

Table 4.1 Evaluation Criteria under Unidimensional Structure

			EQ		NEQ	
			SR	SNR	SR	SNR
CR	FR	BIAS	-0.95	-1.87	-3.65	-4.39
		RMSE	6.36	6.62	7.03	7.52
		CONSISTENCY	0.74	0.73	0.73	0.71
	FNR	BIAS	-0.96	-1.84	-3.81	-4.46
		RMSE	6.31	6.50	7.12	7.49
		CONSISTENCY	0.74	0.74	0.72	0.72
CPU	FR	BIAS	-0.93	-1.83	-3.62	-4.49
		RMSE	6.35	6.59	6.96	7.54
		CONSISTENCY	0.73	0.73	0.73	0.71
	FNR	BIAS	-1.16	-1.86	-3.84	-4.42
		RMSE	6.32	6.52	7.14	7.46
		CONSISTENCY	0.74	0.74	0.73	0.72
CCU	FR	BIAS	-1.01	-1.82	-3.66	-4.50
		RMSE	6.32	6.54	7.01	7.54
		CONSISTENCY	0.73	0.74	0.73	0.71
	FNR	BIAS	-1.15	-1.82	-3.87	-4.40
		RMSE	6.32	6.51	7.16	7.46
		CONSISTENCY	0.74	0.74	0.72	0.72

Note: Group Ability Distributions: EQ – Equivalent groups, NEQ – Nonequivalent groups.
Format Representativeness: FR – Format representativeness, FNR – Format non-representativeness.
Content Representativeness: CR – Content representativeness, CPU – Content partially under-representativeness, CCU – Content completely under-representativeness.
Statistical Representativeness: SR – Statistical representativeness, SNR – Statistical non-representativeness.

As shown in Table 4.1, all the values of BIAS are negative which indicates that the examinees' true expected total scores are always underestimated. The differences in BIAS, RMSE and classification consistency proportion between two levels of the group ability distributions are most noticeable. The average values of BIAS, RMSE and classification consistency proportion under the equivalent groups condition are about -

1.43, 6.44 and 0.74, respectively. In contrast, -4.09, 7.29, and 0.72 are the average values under the nonequivalent groups condition. Therefore, the equivalent groups condition, on average, yields closer to zero BIAS, smaller RMSE and higher classification consistency proportion than the nonequivalent groups condition does. The two sample t test for independent groups further shows that the differences between these two levels of the group ability distributions are statistically significant. Moreover, two levels of the statistical representativeness also show large differences in BIAS and RMSE. It seems that there are no obvious differences in all three evaluation criteria across various levels of the content representativeness and format representativeness.

Several three-way ANOVAs were conducted to further investigate not only the main effects but also the interaction effects among the statistical, content and format representativeness factors under the unidimensional test structure. Overall, there are no statistically significant two-way or three-way interaction effects on all three evaluation criteria. Therefore, later discussion will be limited to the main effects only. In order to better explore the patterns underlying the values of evaluation criteria, the main effects which are found statistically significant (p-value ≤ 0.05) and practically meaningful (effect size $\omega^2 \geq 0.0099$ [2]) in the three-way ANOVA results (refer to Appendix E for more details) are plotted.

Statistical Representativeness

The three-way ANOVA results summarized in Table 4.2 show that compared to the statistical non-representativeness condition, the statistical representativeness

[2] According to Cohen (1988)'s rule of thumb, the cutoff point for small effect size is set as η^2 or $\omega^2 = .0099$; for medium effect size is set as η^2 or $\omega^2 = .0588$; and for large effect size is set as η^2 or $\omega^2 = .1379$. ω^2 is computed using the following equation:

$$\text{Effect Size } \omega^2 = (SS_{effect} - (df_{effect})(MS_{error})) / (MS_{error} + SS_{total})$$

condition generally yields closer to zero BIAS, smaller RMSE, and similar or slightly higher classification consistency proportion. Moreover, these differences between the statistical representativeness and non-representativeness are statistically significant with low to medium effect sizes (range from .033 to .099). There is one exception when equivalent groups are studied and the classification consistency proportion is used as the evaluation criterion.

Table 4.2 Three-way ANOVAs: Main Effects of Statistical Representativeness under the Unidimensional Test Structure

Simulation Condition	Evaluation Criterion	Mean (SR/SNR)	Standard Error (SR/SNR)	F	Effect Size ω^2
UNI-EQ	BIAS	-1.026/-1.842	.05/.05	132.843**	.099
	RMSE	6.329/6.546	.019/.019	63.218**	.049
	CONSISTENCY	.735/.737	.001/.001	2.362	-
UNI-NEQ	BIAS	-3.741/-4.444	.054/.054	85.090**	.066
	RMSE	7.071/7.500	.036/.036	69.322**	.054
	CONSISTENCY	.728/.716	.001/.001	42.063**	.033

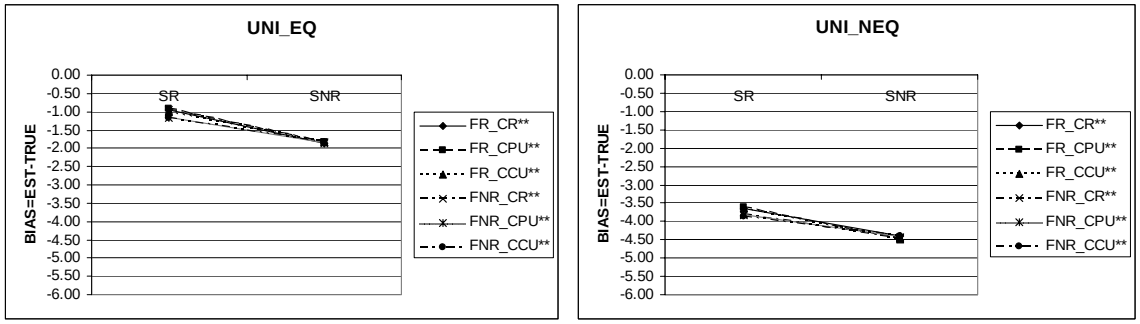
Note: UNI – Unidimensional test structure; EQ – Equivalent groups; NEQ – Nonequivalent groups

SR – Statistical representativeness; SNR – Statistical non-representativeness

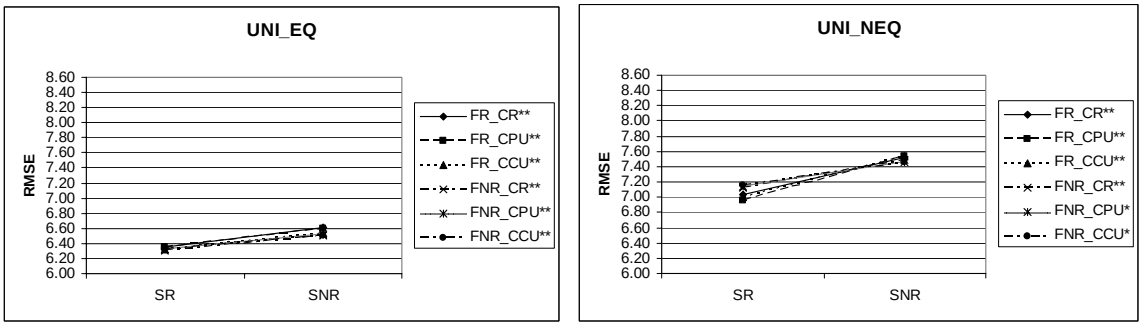
** represents statistical significance at $p \leq .01$

Figure 4.1 further echoes these findings from a slightly different perspective. It graphically compares the statistical representativeness versus non-representativeness condition across all the combinations of the content and format representativeness factors. It divides the comparisons into three parts. Part (a) compares the BIAS differences. Part (b) compares the RMSE differences, and Part (c) compares the differences in the consistency classification proportions. In each part, two graphs which represent two levels of the group ability distributions are presented side by side to demonstrate the substantial differences between the equivalent groups condition and the nonequivalent groups condition. The X-axis of each graph specifies two levels of the statistical representativeness and the Y-axes specify the values of evaluation criteria. Each line in

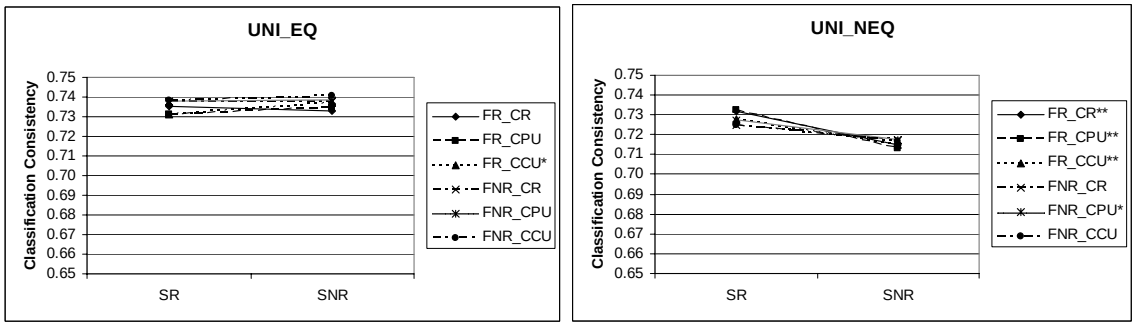
the graph represents one possible combination of the content and format representativeness and is shown in the legend along with the two independent samples t test result by using the asterisks to indicate whether the differences between the statistical representativeness and non-representativeness conditions are statistically significant or not in each case.



(a) BIAS



(b) RMSE



(c) Classification Consistency

Note: * represents statistical significance at $p \leq 0.05$;
** represents statistical significance at $p \leq 0.01$.

Figure 4.1 Comparisons of Statistical Representativeness VS. Statistical Non-representativeness under the Unidimensional Test Structure

Figure 4.1 (a) shows that mixed-format test equating using concurrent calibration with unidimensional IRT models (later, shortened as Concurrent Calibration) under the statistical representativeness condition consistently yields significantly closer to zero BIAS values across all six combinations of the content and format representativeness factors than it does under the statistical non-representativeness condition. This finding exists regardless of the levels of the group ability distributions. Figure 4.1 (b) shows that the RMSE values under the statistical representativeness condition are always significantly smaller than those under the statistical non-representativeness condition. Figure 4.1 (c) is based on the classification consistency proportions. Two levels of the statistical representativeness do not differ a lot when equivalent groups are used. On the other hand, when nonequivalent groups are used, concurrent calibration under the statistical representativeness condition mostly yields significantly higher classification consistency proportions than it does under the statistical non-representativeness condition.

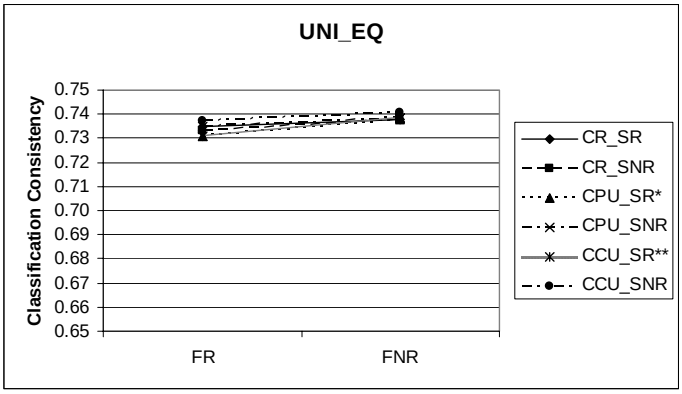
Content Representativeness

The three-way ANOVA results (refer to Appendix E for details) on the differences among the content representativeness, the content partially under-representativeness, and the content completely under-representativeness indicate that the concurrent calibration yields no statistically significant differences in BIAS, RMSE, and classification consistency proportion among all three levels of the content representativeness factor.

Format Representativeness

The three-way ANOVA results (refer to Appendix E for details) on the main effects of the format representativeness factor show that with one exception, there are no

significant differences between the format representativeness and non-representativeness conditions based on the values of BIAS, RMSE, and classification consistency proportion. The exception happens when equivalent groups are applied and the classification consistency proportion is used as the evaluation criterion. In this case, the average classification consistency proportions for the format representativeness and non-representativeness conditions are .734 and .738, respectively. The main effect of the format representativeness is statistically significant with a small effect size of .014. Figure 4.2 shows virtually horizontal lines indicating slight not big differences.



Note: * represents statistical significance at $p \leq 0.05$;
** represents statistical significance at $p \leq 0.01$.

Figure 4.2 Comparisons of Format Representativeness VS. Format Non-representativeness under the Unidimensional Test Structure

Summary

To answer research question 1, truly unidimensional test structure was simulated. As described in Chapter Three, under this condition, the test scenario indicates that although the test consists of items from two content areas, these two content areas are perfectly correlated. Meanwhile, various item formats including MC and CR items measure essentially the same ability. Therefore, the expectation for simulation results in this scenario is that the content and format representativeness factors will have no to

minimum impact on the equating results while the statistical representativeness factor might affect the equating results to unspecified degree. The following findings support the hypothesis.

First, the examinees' true expected total scores were underestimated through the concurrent calibration in this study.

Second, under the unidimensional test structure, the factor of group ability distributions imposed a most significant impact on the equating results. The equivalent groups condition yielded closer to zero BIASs, smaller RMSEs, and higher classification consistency proportions compared to the nonequivalent groups condition.

Third, under the unidimensional test structure, the statistical representativeness condition produced significantly closer to zero BIASs and smaller RMSEs compared to the statistical non-representativeness condition. When nonequivalent groups were used, it also yielded significantly higher classification consistency proportions.

Fourth, with few exceptions, various levels of the content representativeness and the format representativeness factors displayed no significant differences under the unidimensional test structure.

Fifth, there are no statistically and practically significant interaction effects among the statistical, content and format representativeness factors.

Research Question 2

To answer research question 2: "In hypothetical but possibly practical situations where multidimensionality exists, what are the effects of content, statistical and format representativeness of common-item sets on mixed-format test equating using concurrent calibration with unidimensional IRT models?" A total of four multidimensional test

structures were simulated. Two sources – different content areas and multiple item formats – which might cause multidimensionality were manipulated. The data generation process described in Chapter Three was strictly followed, and the various evaluation criteria (BIAS, RMSE, and Classification Consistency) were then computed and compared.

Tables and figures are first categorized into four portions based on the four multidimensional test structures: 1) the multidimensional structure in which the correlation between two content factors is 0.9 and the angle between *MDISC* and the content factor is 10° (later, summarized as Multidimensionality (ρ=0.9, α=10°)); 2) the multidimensional structure in which the correlation between two content factors is 0.9 and the angle between *MDISC* and the content factor is 35° (later, summarized as Multidimensionality (ρ=0.9, α=35°)); 3) the multidimensional structure in which the correlation between two content factors is 0.75 and the angle between *MDISC* and the content factor is 10° (later, summarized as Multidimensionality (ρ=0.75, α=10°)); and 4) the multidimensional structure in which the correlation between two content factors is 0.75 and the angle between *MDISC* and the content factor is 35° (later, summarized as Multidimensionality (ρ=0.75, α=35°)). In each of the four multidimensional test structures, a total of 24 simulation conditions are included. The layouts and formats of tables and figures in each portion are similar to those used in response to research question 1. It should be noted that in this section, three evaluation criteria (BIAS, RMSE, and classification consistency proportion) will only be compared under each multidimensional test structure independently, not across various multidimensional test

structures. To answer research question 3, these three evaluation criteria across various levels of the test dimensionality structure factor will be compared.

Multidimensionality (ρ=0.9, α=10°)

Table 4.3 presents the average values of BIAS, RMSE, and classification consistency over 100 replications under the condition of Multidimensionality (ρ=0.9, α=10°).

Table 4.3 Evaluation Criteria under Multidimensionality (ρ=0.9, α=10°)

			EQ		NEQ	
			SR	SNR	SR	SNR
CR	FR	BIAS	-1.10	-1.78	-4.01	-4.96
		RMSE	6.45	6.64	7.21	7.84
		CONSISTENCY	0.73	0.72	0.72	0.70
	FNR	BIAS	-1.03	-1.86	-4.15	-4.99
		RMSE	6.35	6.66	7.30	7.89
		CONSISTENCY	0.73	0.73	0.71	0.70
CPU	FR	BIAS	-0.95	-1.76	-4.06	-4.99
		RMSE	6.50	6.67	7.18	7.86
		CONSISTENCY	0.72	0.73	0.72	0.70
	FNR	BIAS	-1.10	-1.89	-4.18	-5.08
		RMSE	6.46	6.63	7.33	7.95
		CONSISTENCY	0.72	0.73	0.72	0.69
CCU	FR	BIAS	-1.03	-1.81	-4.14	-4.91
		RMSE	6.47	6.62	7.25	7.78
		CONSISTENCY	0.72	0.73	0.72	0.70
	FNR	BIAS	-1.11	-1.88	-4.21	-5.07
		RMSE	6.43	6.64	7.35	7.91
		CONSISTENCY	0.72	0.73	0.71	0.70

Note: Group Ability Distributions: EQ – Equivalent groups, NEQ – Nonequivalent groups.
Format Representativeness: FR – Format representativeness, FNR – Format non-representativeness.
Content Representativeness: CR – Content representativeness, CPU – Content partially under-representativeness, CCU – Content completely under-representativeness.
Statistical Representativeness: SR – Statistical representativeness, SNR – Statistical non-representativeness.

Table 4.3 indicates that the factor of the group ability distribution imposes the greatest influence on the equating results, followed by the statistical representativeness

factor. It seems that there is little impact of the content and format representativeness factors on the equating results. More specifically, the concurrent calibration under the equivalent groups condition obviously outperforms that under the nonequivalent groups condition on all of the three evaluation criteria. It yields overall negative but closer to zero BIAS, smaller RMSE, and higher classification consistency proportion than that under the nonequivalent groups condition. The concurrent calibration under the statistical representativeness condition also produces overall closer to zero BIAS and smaller RMSE. However, the statistical representativeness condition yields similar classification consistency proportion as the statistical non-representativeness condition does when equivalent groups are used and yields slightly higher classification consistency proportion than the statistical non-representativeness condition does when nonequivalent groups are applied.

Next, three-way ANOVAs were conducted to investigate the main and interaction effects of the statistical, content and format representativeness factors. The results are shown in Appendix E. In general, there are no statistically significant (p-value ≤ 0.05) and practically meaningful (effect size $\omega^2 \geq 0.0099$) main effects of the content and format representativeness factors, as well as the two-way or three-way interaction effects based on the values of BIAS, RMSE and classification consistency proportion. Therefore, later discussions and figures will only involve the main effects of the statistical representativeness factor.

Statistical Representativeness

Table 4.4 presents the three-way ANOVA results on the main effects of the statistical representativeness factor. The results are similar to those under the

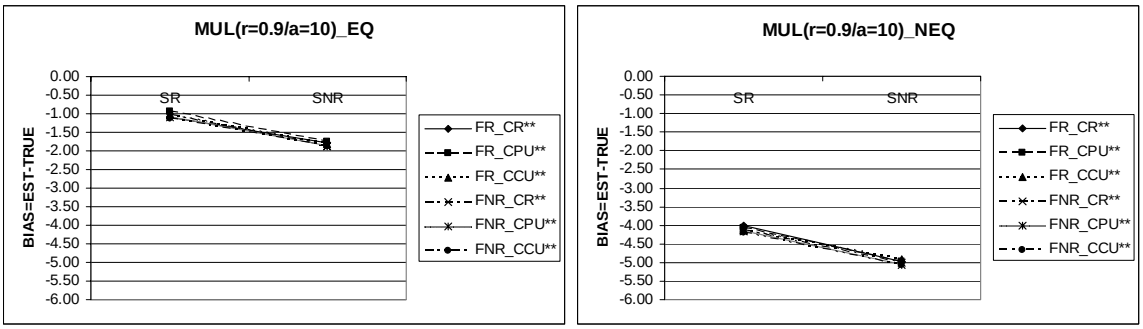
unidimensional test structure found in Table 4.2. The statistical representativeness condition always yields significantly negative but closer to zero BIAS and smaller RMSE than the statistical non-representativeness condition does. The differences in classification consistency proportion under two levels of the statistical representativeness factors are less obvious. When equivalent groups are used, the statistical representativeness condition yields .5% lower classification consistency proportion. In contrast, when nonequivalent groups are utilized, the statistical representativeness condition yields 2.1% higher classification consistency proportion. These two differences are both statistically significant, but under the equivalent groups condition one has small effect size (.020) while the other, under the nonequivalent groups condition, having medium effect size (.084).

Table 4.4 Three-way ANOVAs: Main Effects of Statistical Representativeness under Multidimensionality (ρ=0.9, α=10°)

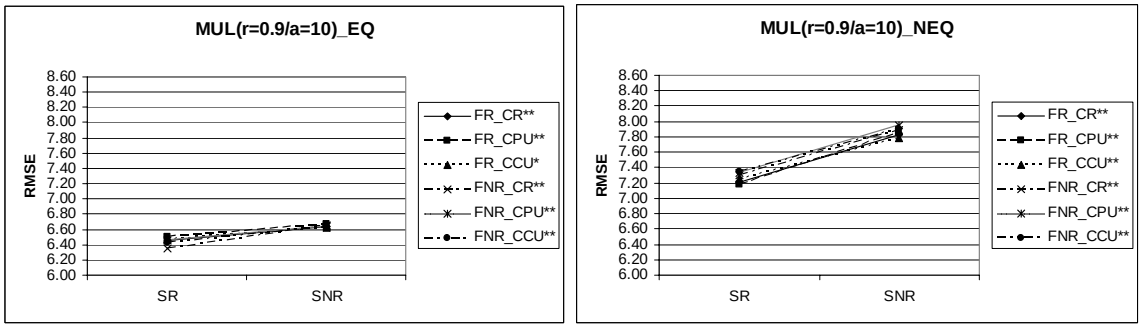
Simulation Condition	Evaluation Criterion	Mean (SR/SNR)	Standard Error (SR/SNR)	F	Effect Size ω^2
MUL0.9/10-EQ	BIAS	-1.052/-1.831	.049/.049	127.598**	.096
	RMSE	6.444/6.645	.018/.018	62.139**	.049
	CONSISTENCY	.724/.729	.001/.001	23.462**	.020
MUL0.9/10-NEQ	BIAS	-4.125/-5.000	.051/.051	146.902**	.109
	RMSE	7.270/7.871	.036/.036	138.536**	.103
	CONSISTENCY	.718/.697	.001/.001	110.506**	.084

Note: MUL0.9/10 – Multidimensional test structure with ρ=0.9 and α=10°;
EQ – Equivalent groups; NEQ – Nonequivalent groups
SR – Statistical representativeness; SNR – Statistical non-representativeness
** represents statistical significance at $p \leq .01$

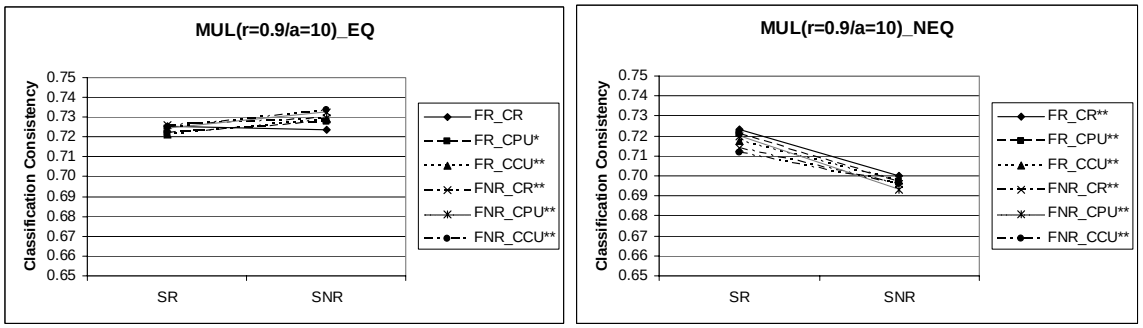
Figure 4.3 further confirms these findings. Figure 4.3 uses the same layout as those used in Figure 4.1 to compare the statistical representativeness versus non-representativeness condition across all the combinations of the content and format factors.



(a) BIAS



(b) RMSE



(c) Classification Consistency

Note: * represents statistical significance at $p \leq 0.05$;
** represents statistical significance at $p \leq 0.01$.

Figure 4.3 Comparisons of Statistical Representativeness VS. Statistical Non-representativeness under Multidimensionality (ρ=0.9, α=10°)

Figure 4.3 (a) shows the BIAS differences between the statistical representativeness and non-representativeness conditions when the two groups involved are either equivalent or nonequivalent. The statistical representativeness condition produces significantly closer to zero BIAS compared to the statistical non-representativeness condition regardless of the group ability distribution factor. Figure 4.3

(b) indicates that the RMSE values under the statistical representativeness condition are significantly smaller than those under the statistical non-representativeness condition no matter equivalent or nonequivalent groups are applied. However, the RMSE differences between the statistical representativeness and non-representativeness conditions are larger under the nonequivalent groups condition than those under the equivalent groups condition which can be proven from the steeper slopes under the nonequivalent groups condition. Figure 4.3 (c) depicts the differences of classification consistency proportion. Two levels of the statistical representativeness factor show significant differences across almost all the combinations of the content and format representativeness factors except for the FR-CR combination. Interestingly, the statistical non-representativeness condition yields higher classification consistency proportions under the equivalent groups condition. For the nonequivalent groups condition, the statistical representativeness condition yields significantly higher values.

Multidimensionality (ρ=0.9, α=35°)

Table 4.5 presents the values of BIAS, RMSE, and classification consistency under the condition of Multidimensionality (ρ=0.9, α=35°). As shown in Table 4.5, in the comparison of the equivalent versus nonequivalent groups condition, the equivalent groups condition keeps outperforming on all three evaluation criteria. In the comparison of the statistical representativeness versus non-representativeness condition, the statistical representativeness condition yields closer to zero BIAS, smaller RMSE regardless of the group ability distribution factor, and higher classification consistency proportion when nonequivalent groups are applied. Various levels of the content representativeness factor still show few differences on all three evaluation criteria. However, the format

representativeness factor starts to impose its influence on the equating results, especially when nonequivalent groups are used. Under the nonequivalent groups condition, the format representativeness condition produces closer to zero BIAS, smaller RMSE, and slightly higher classification consistency proportion compared to the format non-representativeness condition.

Table 4.5 Evaluation Criteria under Multidimensionality (ρ=0.9, α=35°)

			EQ		NEQ	
			SR	SNR	SR	SNR
CR	FR	BIAS	-0.98	-1.42	-4.18	-4.93
		RMSE	6.56	6.60	7.35	7.85
		CONSISTENCY	0.72	0.73	0.72	0.70
	FNR	BIAS	-1.01	-1.80	-4.71	-5.58
		RMSE	6.54	6.70	7.65	8.28
		CONSISTENCY	0.72	0.73	0.71	0.68
CPU	FR	BIAS	-0.91	-1.56	-4.18	-5.02
		RMSE	6.54	6.64	7.38	7.93
		CONSISTENCY	0.72	0.72	0.71	0.69
	FNR	BIAS	-1.03	-1.72	-4.76	-5.63
		RMSE	6.52	6.68	7.71	8.34
		CONSISTENCY	0.72	0.73	0.71	0.68
CCU	FR	BIAS	-0.91	-1.64	-4.31	-4.99
		RMSE	6.53	6.74	7.46	7.87
		CONSISTENCY	0.72	0.72	0.71	0.69
	FNR	BIAS	-0.97	-1.79	-4.85	-5.56
		RMSE	6.46	6.71	7.75	8.25
		CONSISTENCY	0.72	0.72	0.70	0.68

Note: Group Ability Distributions: EQ – Equivalent groups, NEQ – Nonequivalent groups.
Format Representativeness: FR – Format representativeness, FNR – Format non-representativeness.
Content Representativeness: CR – Content representativeness, CPU – Content partially under-representativeness, CCU – Content completely under-representativeness.
Statistical Representativeness: SR – Statistical representativeness, SNR – Statistical non-representativeness.

Three-way ANOVAs were then conducted (refer to Appendix E for more details). The results show that in addition to the main effects of the statistical representativeness factor, some main effects of the format representativeness factor are also statistically

significant (p-value ≤ 0.05) and practically meaningful (effect size $\omega^2 \geq 0.0099$), which will be the main focus of the discussion in this portion.

Statistical Representativeness

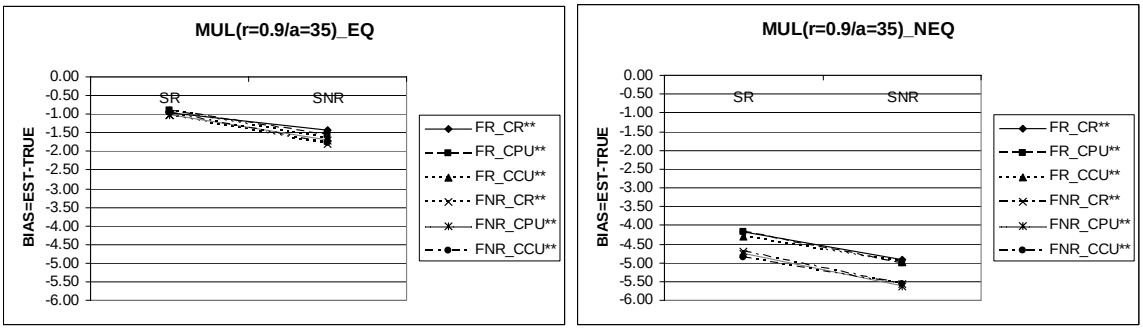
Table 4.6 Three-way ANOVAs: Main Effects of Statistical Representativeness under Multidimensionality (ρ=0.9, α= 35°)

Simulation Condition	Evaluation Criterion	Mean (SR/SNR)	Standard Error (SR/SNR)	F	Effect Size ω^2
MUL0.9/35-EQ	BIAS	-.968/-1.655	.045/.045	116.122**	.088
	RMSE	6.525/6.678	.016/.016	47.732**	.037
	CONSISTENCY	.722/.724	.001/.001	3.960*	(.003)
MUL0.9/35-NEQ	BIAS	-4.497/-5.284	.050/.050	122.197**	.088
	RMSE	7.550/8.087	.036/.036	108.725**	.080
	CONSISTENCY	.709/.687	.002/.002	78.038**	.060

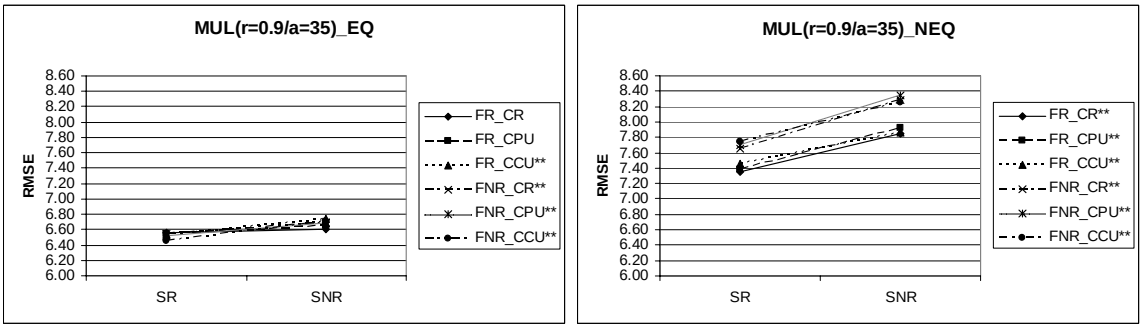
Note: MUL0.9/35 – Multidimensional test structure with ρ=0.9 and α=35°;
EQ – Equivalent groups; NEQ – Nonequivalent groups
SR – Statistical representativeness; SNR – Statistical non-representativeness
* represents statistical significance at p ≤ .05; **represents statistical significance at p ≤ .01

Table 4.6 summarizes the main effects of the statistical representativeness factor under Multidimensionality (ρ=0.9, α=35°). The three-way ANOVA results show that with one exception, the differences between the statistical representativeness and non-representativeness conditions are significant with low to medium effect sizes range from .037 to .088. The results from the "Mean (SR/SNR)" column indicate that compared to the statistical non-representativeness condition, the statistical representativeness condition yields negative but closer to zero BIAS, smaller RMSE, and higher classification consistency proportion.

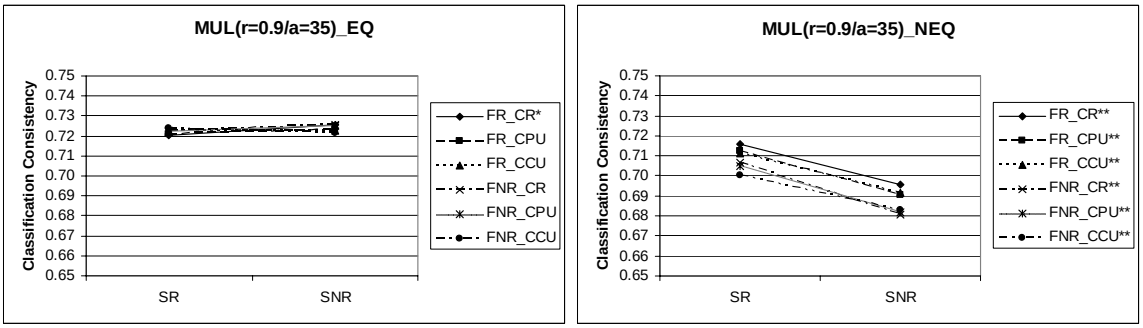
Figure 4.4 is plotted to better demonstrate these differences across all the combinations of the content and format representativeness factors. It further supports the above findings.



(a) BIAS



(b) RMSE



(c) Classification Consistency

Note: * represents statistical significance at $p \leq 0.05$;
** represents statistical significance at $p \leq 0.01$.

Figure 4.4 Comparisons of Statistical Representativeness VS. Statistical Non-representativeness under Multidimensionality (ρ=0.9, α=35°)

As shown in Figure 4.4, the statistical representativeness condition consistently shows superiority over the statistical non-representativeness condition in terms of BIAS (as shown in Figure 4.4 (a)) and RMSE (as shown in Figure 4.4 (b)). As to its performance on the classification consistency proportion, it outperforms the statistical non-representativeness condition only when nonequivalent groups are applied (as shown in Figure 4.4 (c)). It should be noted that when nonequivalent groups are used (right plots

in Part (a) – (c)), six lines representing six combinations of the content and format representativeness factors are further split into two parts, three overlapping lines under the format representativeness condition and the other three under format non-representativeness condition. These plots are different from those shown in Figure 4.1 and 4.3 and display a new trend in the comparison of the format representativeness versus non-representativeness condition.

Format Representativeness

Table 4.5 and Figure 4.4 above both show that the format representativeness and non-representativeness conditions affect the equating results differently especially under the nonequivalent groups condition. Table 4.7 further demonstrates this trend by providing three-way ANOVA results.

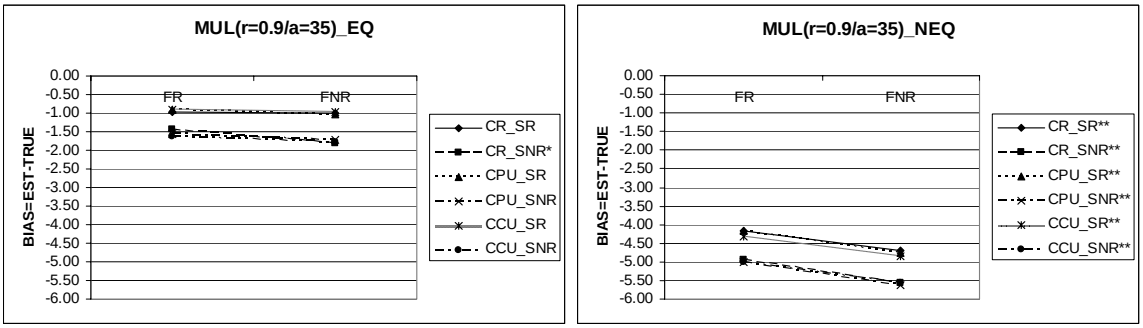
Table 4.7 Three-way ANOVAs: Main Effects of Format Representativeness under Multidimensionality (ρ=0.9, α= 35°)

Simulation Condition	Evaluation Criterion	Mean (FR/FNR)	Standard Error (FR/FNR)	F	Effect Size ω^2
MUL0.9/35-EQ	BIAS	-1.236/-1.387	.045/.045	5.602*	(.004)
	RMSE	6.602/6.602	.016/.016	.000	-
	CONSISTENCY	.723/.723	.001/.001	.812	-
MUL0.9/35-NEQ	BIAS	-4.601/-5.180	.050/.050	66.057**	.047
	RMSE	7.640/7.997	.036/.036	47.749**	.035
	CONSISTENCY	.703/.693	.002/.002	16.800**	.012

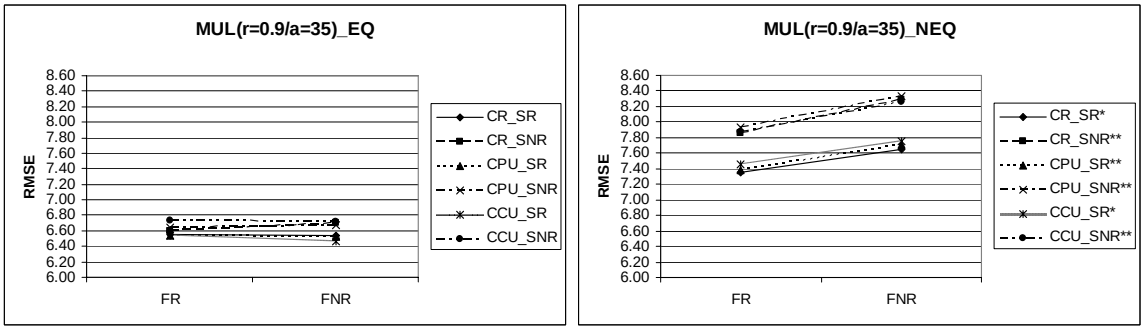
Note: MUL0.9/35 – Multidimensional test structure with ρ=0.9 and α=35°;
EQ – Equivalent groups; NEQ – Nonequivalent groups
FR – Format representativeness; FNR – Format non-representativeness
* represents statistical significance at $p \leq .05$; **represents statistical significance at $p \leq .01$

Table 4.7 shows that under the nonequivalent groups condition, the format representativeness condition produces negative but closer to zero BIAS, smaller RMSE, and slightly higher classification consistency proportion compared to the format non-representativeness condition. Under the nonequivalent groups condition, the average values of BIAS, RMSE and classification consistency proportion for the format

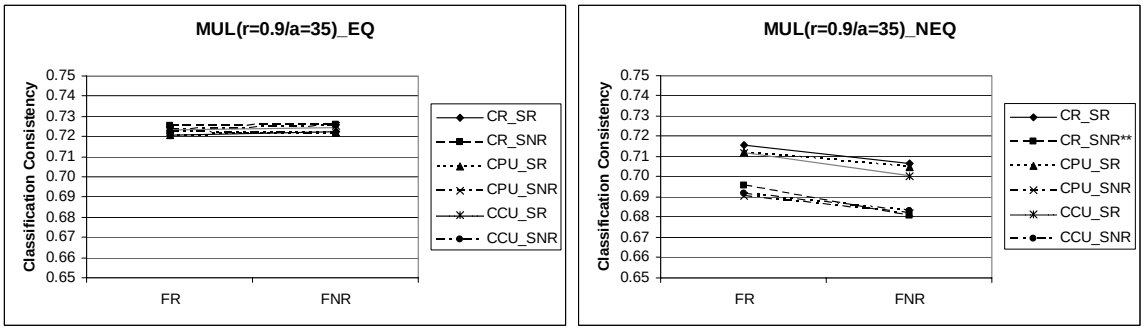
representativeness condition are -4.601, 7.640, and 0.703, respectively. While for the format non-representativeness condition, they are -5.180, 7.997, and 0.693, respectively. These differences are statistically significant at .01 level but with small effect sizes (from .012 to .047) compared to the statistical representativeness factor.



(a) BIAS



(b) RMSE



(c) Classification Consistency

Note: * represents statistical significance at $p \leq 0.05$;
** represents statistical significance at $p \leq 0.01$.

Figure 4.5 Comparisons of Format Representativeness VS. Format Non-representativeness under Multidimensionality (ρ=0.9, α=35°)

Figure 4.5 compares the format representativeness versus the format non-representativeness condition using the same layout as that used in the Figure 4.4. The X-

axes specify the two levels of the format representativeness factor and the lines in the figure represent various combinations of the statistical and content representativeness factors. It should be noted that Figure 4.5 is displayed to facilitate comparisons between the two levels of the format representativeness factor, but the information conveyed in this figure is partially redundant to that in Figure 4.4.

As shown in Figure 4.5, under the equivalent groups condition (left plots), there are seemingly horizontal lines which indicate no significant differences of BIAS, RMSE, and classification consistency proportion between the format representativeness and non-representativeness conditions with one exception. When nonequivalent groups are used (right plots), the format representativeness condition exhibits significantly closer to zero BIAS, significantly smaller RMSE, and observably higher classification consistency proportion across six combinations of the statistical and content representativeness factors.

Multidimensionality (ρ=0.75, α=10°)

Table 4.8 summarizes the average BIAS, RMSE, and classification consistency proportion over 100 replications under Multidimensionality (ρ=0.75, α=10°). As shown in Table 4.8, these evaluation criteria repeat very similar patterns to those found in Table 4.3 under the condition of Multidimensionality (ρ=0.9, α=10°). First, all the values of BIAS are negative which indicates that the examinees' true expected total scores are always underestimated. Second, the concurrent calibration under the equivalent groups condition consistently performs better than under the nonequivalent groups condition by producing lower absolute value of BIAS, smaller RMSE, and higher classification consistency proportion. Third, among three important characteristics of common-item

sets, only the statistical representativeness factor shows observable impact on equating results in terms of the BIAS and the RMSE; while there are no big differences among various levels of the content and format representativeness factors.

Table 4.8 Evaluation Criteria under Multidimensionality (ρ=0.75, α=10°)

			EQ		NEQ	
			SR	SNR	SR	SNR
CR	FR	BIAS	-0.93	-1.54	-4.32	-5.06
		RMSE	6.54	6.66	7.48	7.94
		CONSISTENCY	0.72	0.72	0.70	0.68
	FNR	BIAS	-0.92	-1.75	-4.45	-5.13
		RMSE	6.47	6.70	7.56	8.06
		CONSISTENCY	0.72	0.72	0.69	0.68
CPU	FR	BIAS	-0.89	-1.76	-4.50	-5.18
		RMSE	6.51	6.70	7.60	8.05
		CONSISTENCY	0.72	0.72	0.70	0.68
	FNR	BIAS	-1.06	-1.72	-4.50	-5.24
		RMSE	6.52	6.70	7.64	8.15
		CONSISTENCY	0.72	0.72	0.69	0.68
CCU	FR	BIAS	-0.96	-1.79	-4.72	-5.34
		RMSE	6.53	6.71	7.74	8.21
		CONSISTENCY	0.71	0.72	0.69	0.68
	FNR	BIAS	-1.06	-1.78	-4.60	-5.35
		RMSE	6.51	6.70	7.68	8.19
		CONSISTENCY	0.72	0.72	0.69	0.68

Note: Group Ability Distributions: EQ – Equivalent groups, NEQ – Nonequivalent groups.
Format Representativeness: FR – Format representativeness, FNR – Format non-representativeness.
Content Representativeness: CR – Content representativeness, CPU – Content partially under-representativeness, CCU – Content completely under-representativeness.
Statistical Representativeness: SR – Statistical representativeness, SNR – Statistical non-representativeness.

Three-way ANOVAs are then conducted to further examine the main effects and interaction effects of the statistical, content and format representativeness factors. The results (see Appendix E for details) show that there are no statistically significant interaction effects. The main effects of the statistical representativeness factor are mostly statistically significant (p-value ≤ 0.05) and practically meaningful (effect size $\omega^2 \geq$

0.0099). There are also several significant main effects of the content representativeness factor but with very low effect sizes (below .0099).

Statistical Representativeness

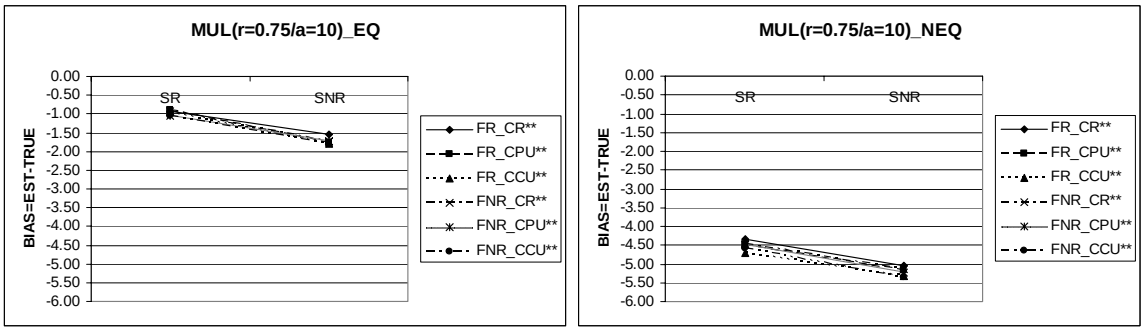
Table 4.9 Three-way ANOVAs: Main Effects of Statistical Representativeness under Multidimensionality (ρ=0.75, α= 10°)

Simulation Condition	Evaluation Criterion	Mean (SR/SNR)	Standard Error (SR/SNR)	F	Effect Size ω^2
MUL0.75/10-EQ	BIAS	-.969/-1.724	.046/.046	136.875**	.102
	RMSE	6.515/6.695	.016/.016	64.543**	.051
	CONSISTENCY	.717/.720	.001/.001	10.038**	(.007)
MUL0.75/10-NEQ	BIAS	-4.515/-5.217	.050/.050	97.427**	.074
	RMSE	7.615/8.100	.037/.037	87.702**	.067
	CONSISTENCY	.694/.681	.001/.001	46.151**	.030

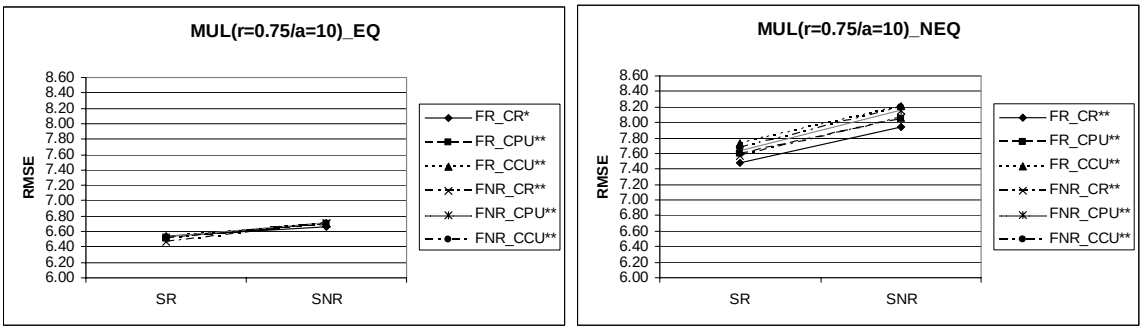
Note: MUL0.75/10 – Multidimensional test structure with ρ=0.75 and α=10°;
EQ – Equivalent groups; NEQ – Nonequivalent groups
SR – Statistical representativeness; SNR – Statistical non-representativeness
** represents statistical significance at $p \leq .01$

Table 4.9 presents the main effects of the statistical representativeness factor under Multidimensionality (ρ=0.75, α= 10°). Overall, the main effects of the statistical representativeness factor are statistically significant with small to medium effect sizes from .030 to .102. Compared to the statistical non-representativeness condition, the statistical representativeness condition yields negative but closer to zero BIAS, smaller RMSE, and higher classification consistency proportion. There is one exception when equivalent groups are applied and classification consistency proportion is used as the evaluation criterion.

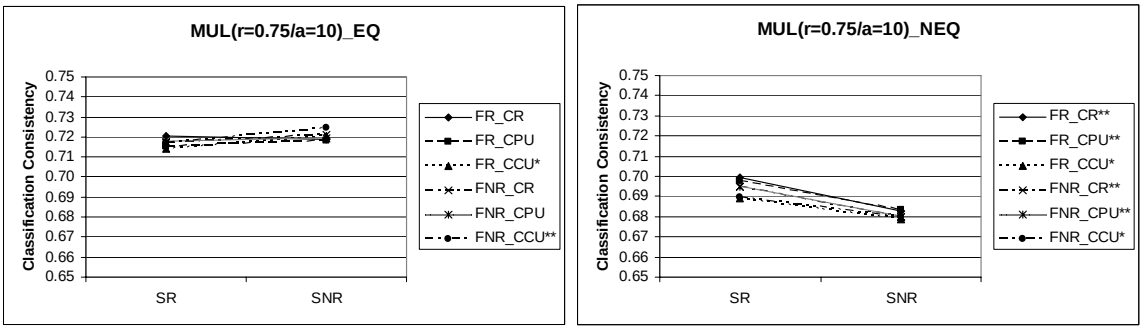
Figure 4.6 confirms the above findings by graphically demonstrating these differences between the statistical representativeness and non-representativeness conditions across all combinations of the content and format representativeness factors.



(a) BIAS



(b) RMSE



(c) Classification Consistency

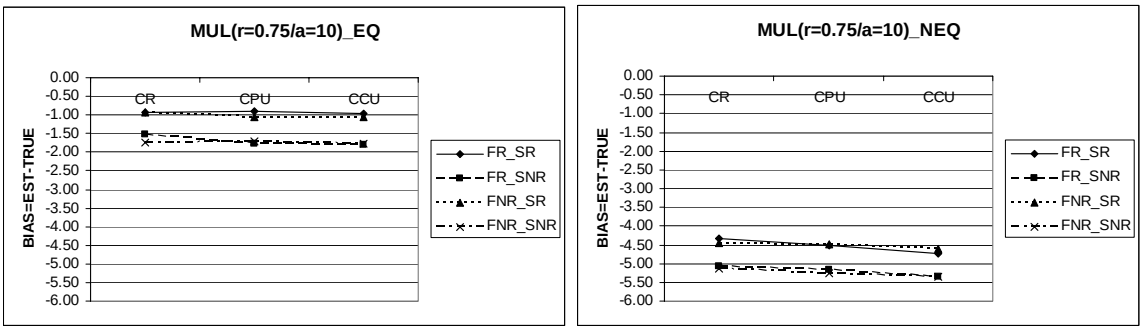
Note: * represents statistical significance at $p \leq 0.05$;
** represents statistical significance at $p \leq 0.01$.

Figure 4.6 Comparisons of Statistical Representativeness VS. Statistical Non-representativeness under Multidimensionality (ρ=0.75, α=10°)

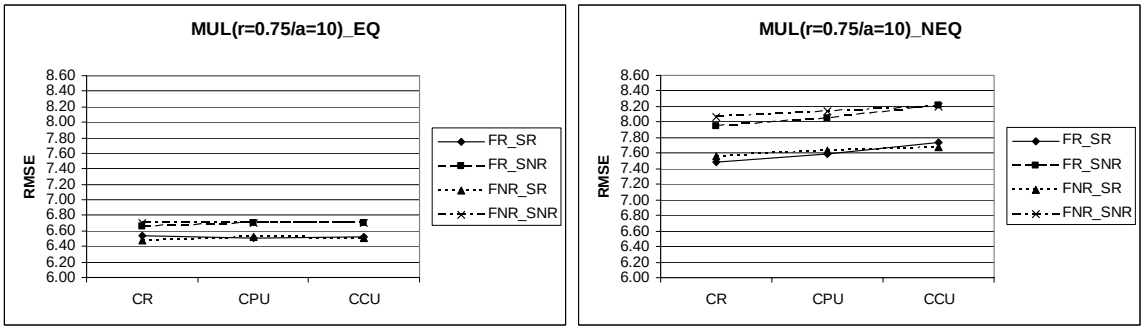
Content Representativeness

It is worth mentioning (shown in Appendix E) that under the nonequivalent groups condition, the main effects of the content representativeness factor on BIAS and RMSE are statistically significant but with very small effect sizes (below .0099). The values of average BIAS under three levels of the content representativeness are -4.742, -4.855, and -5.000, respectively, while the values of RMSE are 7.761, 7.857, and 7.955,

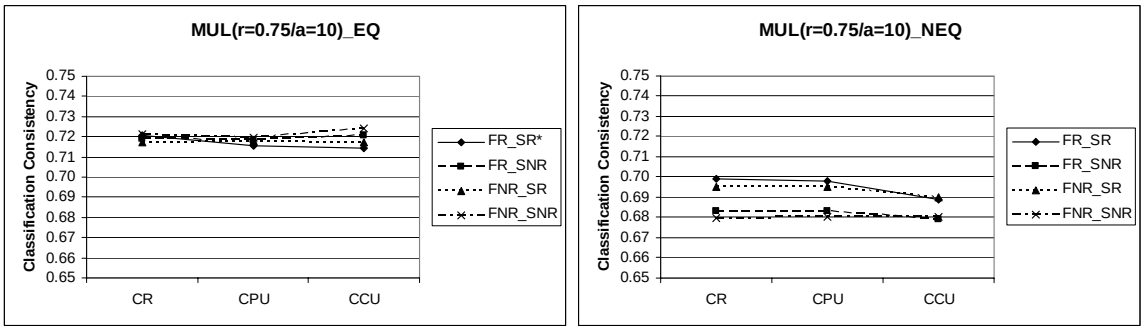
respectively. The content representativeness condition yields the least BIAS and RMSE and the content completely under-representativeness condition produces the most BIAS and RMSE.



(a) BIAS



(b) RMSE



(c) Classification Consistency
Note: * represents statistical significance at $p \leq 0.05$.

Figure 4.7 Comparisons of Content Representativeness VS. Content Partially Under-representativeness VS. Content Completely Under-representativeness under Multidimensionality (ρ=0.75, α=10°)

As Figure 4.7 (right plots) demonstrated, although the differences among three levels of the content representativeness factor are not significant across various combinations of the statistical and format representativeness factors, they do show an

ambiguous trend when nonequivalent groups are utilized. There are no longer horizontal lines in the plots. Instead, lines with slight slopes indicate that the concurrent calibration under the content representativeness condition performs best, followed by the content partially under-representativeness condition, and the content completely under-representativeness condition performs worst.

Multidimensionality (ρ=0.75, α=35°)

Table 4.10 displays the values of BIAS, RMSE, and classification consistency proportion under Multidimensionality (ρ=0.75, α=35°).

Table 4.10 Evaluation Criteria under Multidimensionality (ρ=0.75, α=35°)

			EQ		NEQ	
			SR	SNR	SR	SNR
CR	FR	BIAS	-0.85	-1.61	-4.28	-4.98
		RMSE	6.56	6.66	7.42	7.91
		CONSISTENCY	0.72	0.72	0.71	0.68
	FNR	BIAS	-0.90	-1.67	-4.79	-5.43
		RMSE	6.51	6.73	7.71	8.20
		CONSISTENCY	0.72	0.72	0.70	0.67
CPU	FR	BIAS	-0.95	-1.63	-4.36	-5.07
		RMSE	6.56	6.74	7.49	7.96
		CONSISTENCY	0.72	0.72	0.70	0.68
	FNR	BIAS	-1.06	-1.70	-4.85	-5.52
		RMSE	6.58	6.73	7.78	8.27
		CONSISTENCY	0.72	0.72	0.70	0.67
CCU	FR	BIAS	-0.95	-1.69	-4.47	-5.20
		RMSE	6.61	6.74	7.57	8.07
		CONSISTENCY	0.72	0.72	0.69	0.67
	FNR	BIAS	-1.03	-1.79	-4.87	-5.61
		RMSE	6.57	6.78	7.78	8.32
		CONSISTENCY	0.72	0.72	0.70	0.67

Note: Group Ability Distributions: EQ – Equivalent groups, NEQ – Nonequivalent groups.

Format Representativeness: FR – Format representativeness, FNR – Format non-representativeness.

Content Representativeness: CR – Content representativeness, CPU – Content partially under-representativeness, CCU – Content completely under-representativeness.

Statistical Representativeness: SR – Statistical representativeness, SNR – Statistical non-representativeness.

Table 4.10 indicates the following findings about the comparisons of the group ability distributions, statistical, content, and format representativeness factors. First, compared to the nonequivalent groups condition, the equivalent groups condition performs better on all three evaluation criteria. Second, compared to the statistical non-representativeness condition, the statistical representativeness condition yields negative but smaller BIAS, smaller RMSE regardless of the group ability distributions factor. It also produces higher classification consistency proportion when nonequivalent groups are applied. Third, no noticeable differences are found among three levels of the content representativeness factor. Fourth, when nonequivalent groups are involved, it seems that the format representativeness condition outperforms the format non-representativeness condition on the BIAS and RMSE.

Three-way ANOVAs were then conducted to investigate not only the independent effects of the statistical, content and format representativeness factors but also the combined interactions among these three most important characteristics of common-item sets. Detailed tables are presented in Appendix E. No two-way or three-way interactions are found statistically significant (p-value ≤ 0.05) and practically meaningful (effect size $\omega^2 \geq 0.0099$). Based on the same selection rule, only the main effects of the statistical representativeness factor and the format representativeness factor will be discussed next in details.

Statistical Representativeness

Table 4.11 summarizes some important results about the main effects of the statistical representativeness factor from the three-way ANOVA tables in Appendix E. The results show that the concurrent calibration under the statistical representativeness

condition, in general, outperforms the statistical non-representativeness condition by yielding closer to zero BIAS, smaller RMSE, and higher classification consistency proportion when nonequivalent groups are used. These findings are proved to be statistically significant with low to medium effect sizes (.052 to .102).

Table 4.11 Three-way ANOVAs: Main Effects of Statistical Representativeness under Multidimensionality (ρ=0.75, α= 35°)

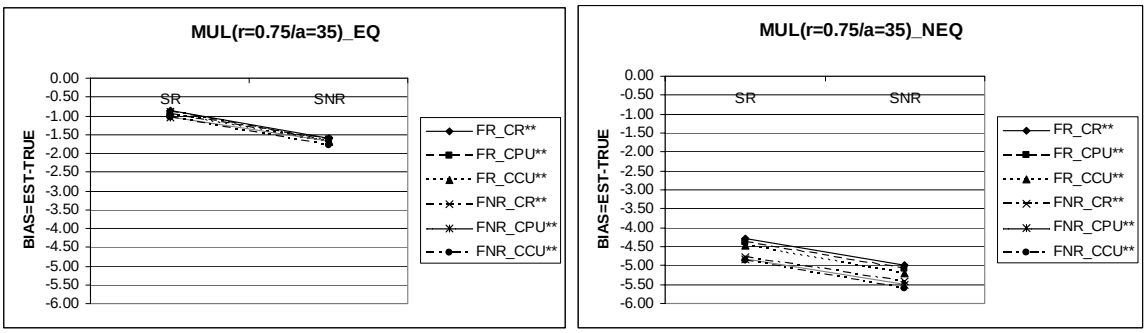
Simulation Condition	Evaluation Criterion	Mean (SR/SNR)	Standard Error (SR/SNR)	F	Effect Size ω^2
MUL0.75/10-EQ	BIAS	-.957/-1.681	.044/.044	137.201**	.102
	RMSE	6.564/6.730	.014/.014	67.336**	.052
	CONSISTENCY	.720/.722	.001/.001	7.041**	(.007)
MUL0.75/10-NEQ	BIAS	-4.604/-5.301	.048/.048	103.427**	.076
	RMSE	7.625/8.122	.036/.036	96.335**	.072
	CONSISTENCY	.698/.676	.002/.002	78.730**	.061

Note: MUL0.75/35 – Multidimensional test structure with ρ=0.75 and α=35°;
EQ – Equivalent groups; NEQ – Nonequivalent groups
SR – Statistical representativeness; SNR – Statistical non-representativeness
** represents statistical significance at $p \leq .01$

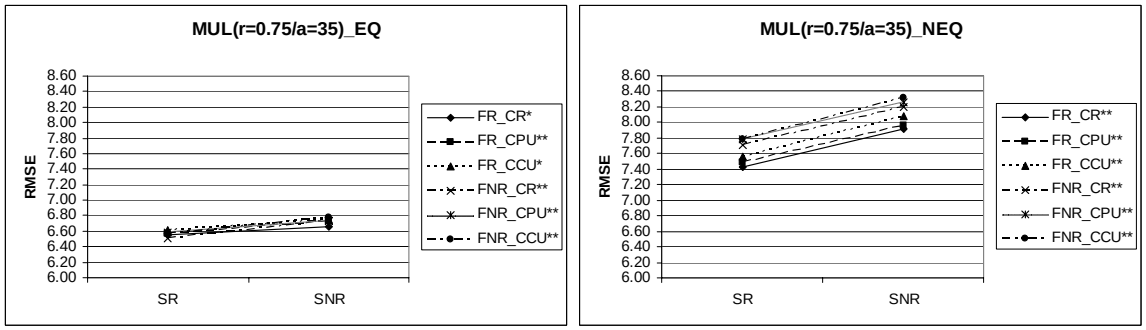
Figure 4.8 graphically plots the differences of BIAS, RMSE, and classification consistency proportion between two levels of the statistical representativeness factor across all six combinations of the content and format representativeness factors. The patterns appear in the figure are consistent with the findings in Table 4.11.

Figure 4.8 (a) shows that the concurrent calibration under the statistical representativeness condition consistently yields significantly closer to zero BIAS values than it does under the statistical non-representativeness condition regardless of the levels of the group ability distributions. Figure 4.8 (b) shows that the RMSE values under the statistical representativeness condition are significantly smaller than those under the statistical non-representativeness condition. Figure 4.8 (c) is based on the classification consistency proportions. Two levels of the statistical representativeness do not differ a lot when equivalent groups are used. On the other hand, when nonequivalent groups are used,

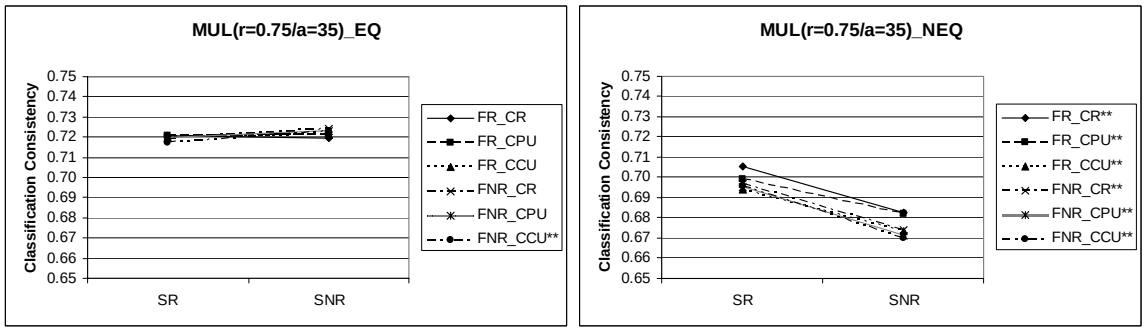
the concurrent calibration under the statistical representativeness condition yields significantly higher classification consistency proportions than it does under the statistical non-representativeness condition.



(a) BIAS



(b) RMSE



(c) Classification Consistency

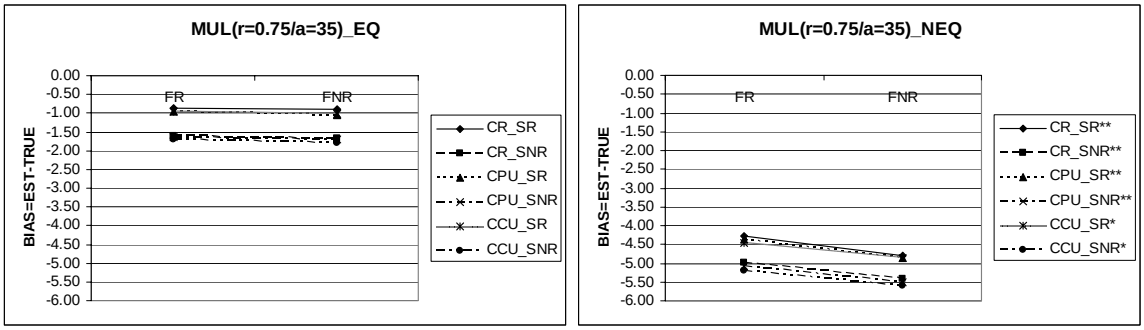
Note: * represents statistical significance at $p \leq 0.05$;
** represents statistical significance at $p \leq 0.01$.

Figure 4.8 Comparisons of Statistical Representativeness VS. Statistical Non-representativeness under Multidimensionality (ρ=0.75, α=35°)

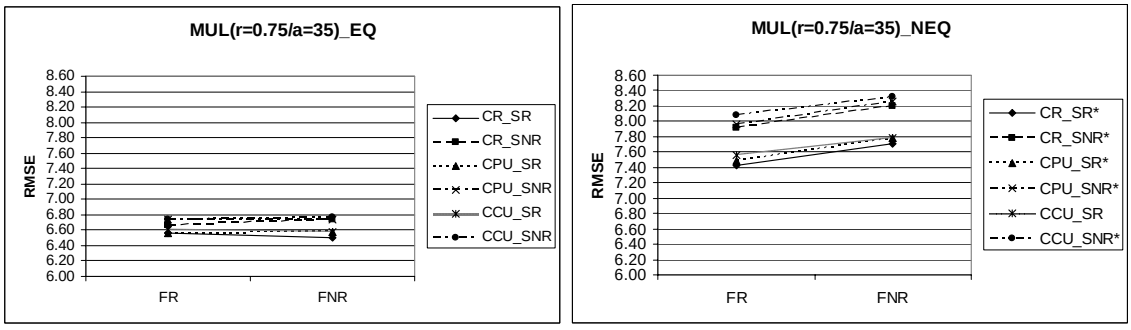
Format Representativeness

When nonequivalent groups are used and the differences are compared based on the values of BIAS and RMSE, the main effects of the format representativeness factor

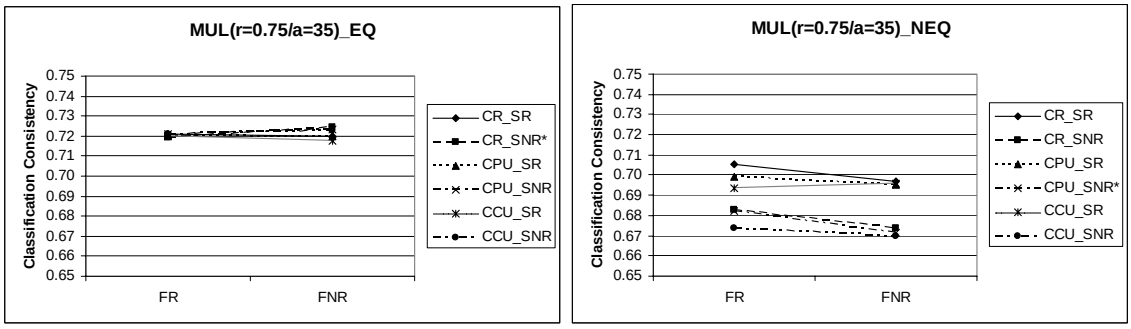
are also found to be statistically significant and practically meaningful under the condition of Multidimensionality (ρ=0.75, α=35°). Their F values are 43.142 (sig. value = .000) and 29.020 (sig. value = .000), respectively, while their effect sizes are .031 and .021, respectively.



(a) BIAS



(b) RMSE



(c) Classification Consistency

Note: * represents statistical significance at $p \leq 0.05$;
** represents statistical significance at $p \leq 0.01$.

Figure 4.9 Comparisons of Format Representativeness VS. Format Non-representativeness under Multidimensionality (ρ=0.75, α=35°)

The main effects of the format representativeness factor are graphically demonstrated in Figure 4.9. On the left part of the figure, all the lines representing

various combinations of the statistical and content representativeness factors are virtually horizontal, which indicates that under the equivalent groups condition, there are no significant differences of BIAS, RMSE, and classification consistency proportion between the format representativeness and non-representativeness conditions. However, the right part of the figure tells a different story. All the lines have slight slopes, which indicates that under the nonequivalent groups condition, the format representativeness condition exhibits significantly closer to zero BIAS, significantly smaller RMSE, and observable but most non-significantly higher classification consistency proportion compared to the format non-representativeness condition.

Summary

To answer research question 2, four different multidimensional test structures were simulated. As described in Chapter Three, these multidimensional test structures attempt to mimic the reasonable test configurations in real assessments. Two main sources of multidimensionality were manipulated: different content areas and multiple item formats. Two content areas are considered to be distinct but highly correlated by setting the correlation coefficients between two content factors at .90 and .75 which are realistic high values as found in previous literature. In addition to the content factors, two format factors with one representing MC items and the other representing CR items are set to be orthogonal by assuming that they only make unique contributions to examinees' responses above and beyond those attributed to the content factors. Furthermore, the relative importance of format factors is controlled by the angle (α=10° or 35°) between *MDISC* and content factor in the multidimensional space. Therefore, the initial expected findings on the statistical, content and format representativeness factors in this scenario

are that with the decrease of the correlation coefficient between two content factors from .90 to .75, the degree of multidimensionality due to multiple content areas increases, and thus the content representativeness factor will increase its influence on equating results. Meanwhile, with the increase of the angle α from 10° to 35°, the degree of multidimensionality due to various item formats increases, and thus the format representativeness factor will impose more impact on equating results. The statistical representativeness factor will keep playing an important role like it did under the unidimensional test structure. The following findings partially confirm these hypotheses.

First, the examinees' true expected total scores were always underestimated through the concurrent calibration in this study.

Second, the factor of group ability distributions always imposed most significant impact on the equating results. The equivalent groups condition yielded closer to zero BIASs, smaller RMSEs, and slightly higher classification consistency proportions compared to the nonequivalent groups condition.

Third, under all four multidimensional test structures, the statistical representativeness condition produced significantly closer to zero BIASs and smaller RMSEs compared to the statistical non-representativeness condition. When nonequivalent groups were used, it also yielded higher classification consistency proportions.

Fourth, under all four multidimensional test structures, three levels of the content representativeness factor did not show statistically significant and practically meaningful differences in the equating results. Only under Multidimensionality (ρ=0.75, α= 10°) and when nonequivalent groups were used, did it seem that the content representativeness

condition yielded the least BIAS and RMSE and the content completely under-representativeness condition produced the most BIAS and RMSE.

Fifth, with the increase of the relative importance of the format factors compared to the content factors on examinees' item responses (from Multidimensionality (ρ=0.90/0.75, α=10°) to Multidimensionality (ρ=0.90/0.75, α=35°)), the format representativeness factor made significant differences on all three evaluation criteria especially under the nonequivalent groups condition.

Sixth, there are no significant and practically meaningful interaction effects among the statistical, content and format representativeness factors.

Research Question 3

To answer research question 3: "How robust is the unidimensional IRT equating method to the presence of various degree of multidimensionality?" Three evaluation criteria – the BIAS, RMSE, and Classification Consistency - are compared across various levels of the test dimensionality structure.

Table 4.12 One-Way ANOVA: Test Dimensionality Structure

Simulation Condition	Evaluation Criterion		SS	DF	F	Effect Size ω^2
EQ	BIAS	Effect	18.988	4	3.291*	(0.002)
		Error	8647.506	5995		
	RMSE	Effect	31.559	4	44.901**	0.056
		Error	1053.409	5995		
	CONSISTENCY	Effect	.221	4	170.812**	0.028
		Error	1.943	5995		
NEQ	BIAS	Effect	614.309	4	89.702**	0.055
		Error	10263.982	5995		
	RMSE	Effect	307.276	4	89.113**	0.101
		Error	5167.945	5995		
	CONSISTENCY	Effect	1.042	4	172.646**	0.103
		Error	9.043	5995		

Note: EQ – Equivalent groups; NEQ – Nonequivalent groups

* represents statistical significance at $p \leq .05$; **represents statistical significance at $p \leq .01$

A series of one-way ANOVAs were conducted and the results along with the effect sizes are provided in Table 4.12. As shown in Table 4.12, the one-way ANOVA results indicate that the differences among various levels of the test dimensionality structure are statistically significant with low to medium effect sizes except for the comparisons of BIAS under the equivalent groups condition in which the effect size is nearly neglectable.

After the omnibus tests, was the examination of the differences between which levels of the test dimensionality structure factor are significant and in what order. Hypothetically, it is expected that performance of the concurrent calibration would decline as the degree of multidimensionality increases (i.e., the correlation between two content factors goes from .90 to .75, and the angle between *MDISC* and content factor in the multidimensional space goes from 10° to 35°). Therefore, although the performance of the concurrent calibration under each level of test dimensionality structure factor is a concern, the focus in this section is on examining the change in the three evaluation criteria with the increase in the degree of multidimensionality.

Two measures are taken. First, the average values of the BIAS, RMSE, and classification consistency proportion, as well as the index η^2 which was proposed by Kim & Kolen (2006) and is adapted in this study are reported in Table 4.13. The index η^2 is to reflect the practical significance of differences between the unidimensionality and various levels of the multidimensionality. It can be expressed as:

$$\eta^2 = \left| \frac{\text{Evaluation Criteria in Multidimensionality - Evaluation Criteria in Unidimensionality}}{\text{Evaluation Criteria in Multidimensionality}} \right|$$

In this study, η^2 are taken in the absolute value and might be roughly interpreted as the proportion of the total equating error explained by the multidimensionality. In general,

the smaller the value of η^2, the more robust to the degree of the multidimensionality. Second, the Turkey HSD method is used to conduct pairwise multiple comparisons. The detailed results are presented in Appendix F. But the homogenous subsets are summarized in Table 4.13.

Table 4.13 Evaluation Criteria under the Levels of Test Dimensionality Structure

Simulation Condition	Evaluation Criterion	UNI	MUL0.90/10 (η^2)	MUL0.90/35 (η^2)	MUL0.75/10 (η^2)	MUL0.75/35 (η^2)
EQ	BIAS	-1.434	-1.442 (-)	-1.311 (-)	-1.346 (-)	-1.319 (-)
	Homogeneous Subsets	1	1	1	1	1
	RMSE	6.438	6.544 (.016)	6.602 (.025)	6.605 (.025)	6.647 (.032)
	Homogeneous Subsets	1	2	3	3	3
	CONSISTENCY	.736	0.727 (.013)	0.723 (.018)	0.719 (.024)	0.721 (.021)
	Homogeneous Subsets	1	2	3	5	4
NEQ	BIAS	-4.093	-4.562 (.103)	-4.891 (.163)	-4.866 (.159)	-4.953 (.174)
	Homogeneous Subsets	1	2	3	3	3
	RMSE	7.285	7.570 (.038)	7.819 (.068)	7.858 (.073)	7.874 (.075)
	Homogeneous Subsets	1	2	3	3	3
	CONSISTENCY	.7219	0.707 (.021)	0.698 (.034)	0.688 (.050)	0.687 (.051)
	Homogeneous Subsets	1	2	3	4	4

Note: EQ – Equivalent groups; NEQ – Nonequivalent groups

UNI – Unidimensionality; MUL0.90/10 – Multidimensional test structure with ρ=0.90 and α=10°; MUL0.90/35 – Multidimensional test structure with ρ=0.90 and α=35°; MUL0.75/10 – Multidimensional test structure with ρ=0.75 and α=10°; MUL0.75/35 – Multidimensional test structure with ρ=0.75 and α=35°

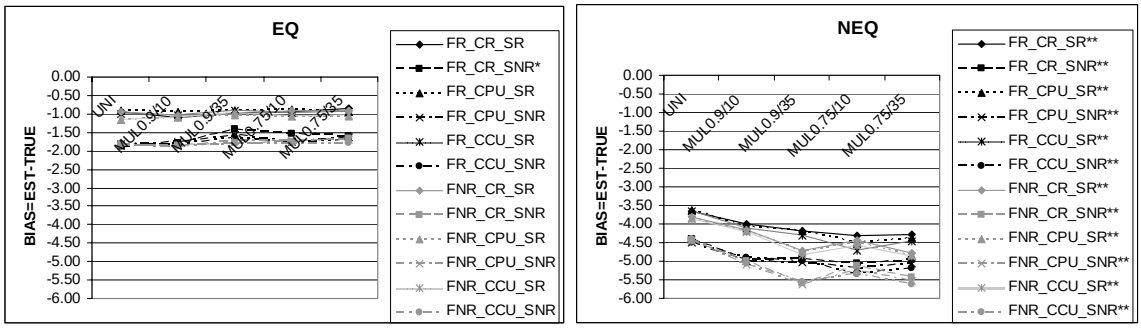
- represents not applicable because the test result on the evaluation criterion is not statistically significant (p-value $\leq$ 0.05) and practically meaningful (effect size $\omega^2 \geq$ 0.0099).

The following findings are found in Table 4.13. First, with one exception, the concurrent calibration under the unidimensionality condition yields the closest to zero BIAS, the smallest RMSE, and the highest classification consistency proportion as

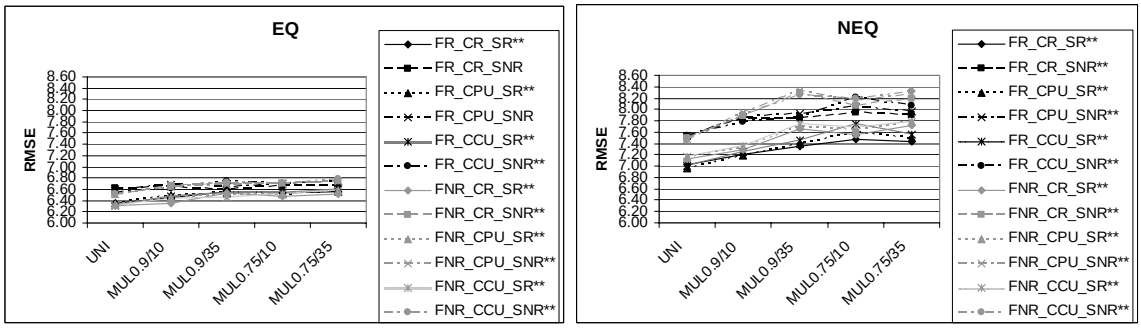
expected. Second, when the degree of multidimensionality caused by various content areas increases (i.e., the correlation between two content factors decreases from .90 to .75), the value of η^2 increases which indicates that the robustness of the concurrent calibration declines. Third, when the degree of the multidimensionality due to multiple item formats increases (i.e., the angle between *MDISC* and content factor in multidimensional space increases from 10° to 35°), the value of η^2 increases. The larger the value of η^2, the less robust the concurrent calibration. There is an exception when the classification consistency proportion is used as the evaluation criterion in comparison of the Multidimensionality (ρ=0.75, α=10°) and the Multidimensionality (ρ=0.75, α=35°).

From the homogenous subsets shown in Table 4.13, consistent conclusions can be reached. Under the equivalent groups condition, five levels of the test dimensionality structures yield no significant differences in BIAS. But under the nonequivalent groups condition, there are three homogenous subsets in BIAS. The concurrent calibration under Multidimensionality (ρ=0.90, α=10°) is more robust to the violation of the unidimensionality assumption than that under the Multidimensionality (ρ=0.90, α=35°), the Multidimensionality (ρ=0.75, α=10°), and the Multidimensionality (ρ=0.75, α=35°). There are no significant differences among the last three levels of the Multidimensionality. The same homogenous groupings are found when using RMSE as the evaluation criterion regardless of the group ability distributions factor. When the differences of classification consistency proportion are examined, the last three levels of the Multidimensionality are further divided. The classification consistency proportion under Multidimensionality (ρ=0.90, α=35°) is significantly higher than those under the Multidimensionality (ρ=0.75, α=10°) and the Multidimensionality (ρ=0.75, α=35°). The

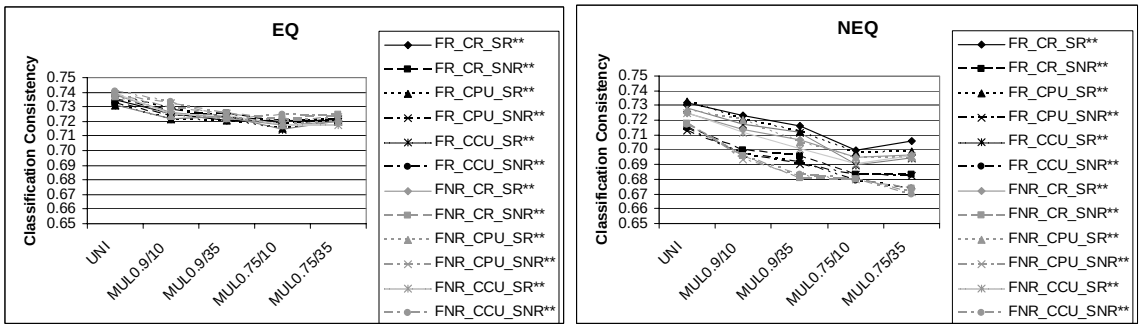
difference between the Multidimensionality (ρ=0.75, α=10°) and the Multidimensionality (ρ=0.75, α=35°) is significant when equivalent groups are used but is not significant when nonequivalent groups are used. However, it should be noted that all these differences in classification consistency proportion among four levels of the Multidimensionality are numerically small.



(a) BIAS



(b) RMSE



(c) Classification Consistency

Note: * represents statistical significance at $p \leq 0.05$;
** represents statistical significance at $p \leq 0.01$.

Figure 4.10 Comparisons of Test Dimensionality Structures

Figure 4.10 graphically compares the values of BIAS, RMSE, and classification consistency proportion among the five levels of the test dimensionality structure factor across all combinations of the statistical, content and format representativeness factors.

When equivalent groups are used, the patterns of the test dimensionality structure factor for all three evaluation criteria tend to be consistent across all 12 combinations of the statistical, content and format representativeness factors by having nearly overlapping or very close lines. The values of BIAS in five levels of the test dimensionality structure factor show no significant differences. As to the RMSE, the Unidimensionality condition always yields the smallest values. With the increase of the degree of multidimensionality, the values of RMSE increase. As to the classification consistency proportion, the Unidimensionality condition always yields the highest proportions. As the degree of multidimensionality increases, the classification consistency proportion generally declines. One exception happens, instead of decreasing classification consistency proportion from the Multidimensionality (ρ=0.75, α=10°) to the Multidimensionality (ρ=0.75, α=35°), it increases.

Under the nonequivalent groups condition, there are more dramatic patterns, which vary largely from one combination of the statistical, content and format representativeness factors to another. However, after thorough investigation, general patterns can be found. The Unidimensionality condition yields the closest to zero BIAS, the smallest RMSE, and the highest classification consistency proportion. Moreover, as the degree of multidimensionality increases, the performance of the concurrent calibration on all three evaluation criteria declines. In other words, the concurrent calibration becomes less and less robust. The greatest inconsistency occurs in the

comparison between the Multidimensionality (ρ=0.90, α=35°) and the Multidimensionality (ρ=0.75, α=10°), where the degree of multidimensionality cannot be determined.

Summary

In response to the research question about the robustness of the concurrent calibration, various levels of the multidimensional test structure must be discussed first. There are a total of four levels of the multidimensional test structure, which are determined by two sources – content areas and item formats. These four levels of the multidimensional test structure are as follows: 1) the multidimensional structure in which the correlation between two content factors is 0.9 and the angle between *MDISC* and content factor in the multidimensional space is 10°; 2) the multidimensional structure in which the correlation between two content factors is 0.9 and the angle between *MDISC* and content factor is 35°; 3) the multidimensional structure in which the correlation between two content factors is 0.75 and the angle between *MDISC* and content factor is 10°; and 4) the multidimensional structure in which the correlation between two content factors is 0.75 and the angle between *MDISC* and content factor is 35°. These four levels of the multidimensional test structure can be sorted in the ascending order of the degree of multidimensionality as the Multidimensionality (ρ=0.90, α=10°) shows the lowest degree of multidimensionality, followed by the Multidimensionality (ρ=0.90, α=35°) and the Multidimensionality (ρ=0.75, α=10°), and the Multidimensionality (ρ=0.75, α=35°) shows the highest degree of multidimensionality. The degree of multidimensionality between the Multidimensionality (ρ=0.90, α=35°) and the Multidimensionality (ρ=0.75, α=10°) is undetermined.

More attention is paid in this study to examining change in the values of BIAS, RMSE, and classification consistency proportion as a function of the increase in the degree of multidimensionality since it is expected that the performance of the concurrent calibration would decline as the degree of the multidimensionality increases. In other words, the higher the degree of multidimensionality, the less robust is the concurrent calibration. However, there is no universal answer in response to the question "how robust is robust enough?" Therefore, all the findings in this section are based on their relative comparisons to the unidimensional condition. The findings are as follow.

First, in the ideal unidimensional condition in which the unidimensionality assumption was met and model fits data well, the concurrent calibration usually yielded the closest to zero BIAS, the smallest RMSE, and the highest classification consistency proportion.

Second, when the degree of multidimensionality caused by various content areas increases (i.e., the correlation between two content factors decreases from .90 to .75), the robustness of the concurrent calibration generally declines.

Third, with few exceptions, when the degree of the multidimensionality caused by multiple item formats increases (i.e., the angle between *MDISC* and content factor in multidimensional space increases from 10° to 35°), the robustness of the concurrent calibration generally declines.

Chapter 5: Summary and Discussion

This chapter provides a summary and discussion of this study. It begins with a brief restatement of the research questions and a summary of the methodology used in this study. This is followed by discussion of the major findings of this study. Then the implications for practice, the limitations of this study, and some suggestions for future research are provided.

Restatement of Research Questions

As mentioned in Chapter One, the central focus of this study was to systematically investigate the impact of representativeness and non-representativeness of common-item sets in terms of statistical, content, and format specifications in mixed-format tests using concurrent calibration with unidimensional IRT models, as well as to examine its robustness to various multidimensional test structures. More specifically, this study attempted to provide information in response to the following research questions:

1) In an ideal situation where the unidimensionality assumption is satisfied, what are the effects of statistical, content and format representativeness of common-item sets on mixed-format test equating?
2) In hypothetical but reasonable practical situations where multidimensionality exists, what are the effects of statistical, content and format representativeness of common-item sets on mixed-format test equating?
3) How robust is the unidimensional IRT equating method to the presence of different multidimensional test structures?

These questions were answered by a simulation study which is briefly summarized next.

Summary of Methodology

In this simulation study, the common-item nonequivalent groups (CINEG) design was used for data collection. Two forms of a 54-item mixed-format test were used. Each test form was comprised of 48 dichotomous MC items and 6 five-category CR items. They were then split evenly into two content areas. Furthermore, these two test forms shared a set of 18 common items. Two groups each with 3,000 examinees were involved. The ability scale of group 1 taking test form 1 was set as a reference scale, and then the concurrent calibration was conducted to transform the scale of group 2 taking test form 2 to the reference scale. The unidimensional three-parameter logistic (3PL) model was used to calibrate the MC items and the unidimensional graded-response model (GRM) was employed to calibrate the CR items. The computer program MULTILOG was used for the test calibration process in each case.

Five factors were manipulated and they can be categorized into three groupings. First was the test dimensionality structure factor. Five levels were considered, which included the unidimensional test structure and four levels of multidimensional test structure due to the combination of various content areas and multiple item formats. The second grouping included the three important characteristics in the composition of common-item sets. There were two levels of the statistical representativeness factor, three levels of the content representativeness factor, and two levels of the format representativeness factor. The last factor of interest was whether the ability distributions of two examinee groups were equivalent or not. These five factors were fully crossed which resulted in a total of 120 simulation conditions. Each condition was repeated 100 times.

To evaluate the accuracy of the equating results under various simulation conditions, the group 2 examinees' true and estimated expected total scores were computed. Based upon these scores, the summary indices – BIAS, RMSE and Classification Consistency over 100 replications were computed and used for final comparisons. Some statistical tests were also conducted to identify the significant performance differences of the concurrent calibration among various simulation conditions.

Discussion of Major Findings

The major findings of this study are discussed as follows.

First, considering all of the simulation conditions, the most notable and significant effects on the equating results among the five factors of investigation appeared to be those due to the factor of group ability distributions. The equivalent groups condition always outperformed the nonequivalent groups condition on the various evaluation indices. This finding is supported by many of the previous simulation studies (Beguin, Hanson, & Glas, 2000; Kim & Kolen, 2006; Kim & Lee, 2006; Kirkpatrick, 2005; Li, Lissitz, & Yang, 1999; Tate, 2000). One plausible explanation is that the concurrent calibration with nonequivalent groups in CINEG design can only utilize common items to equate two forms. In contrast, when equivalent groups are utilized in CINEG design, the concurrent calibration makes better use of both common items and equivalent groups to place the parameters from a new group onto the reference group scale. However, it should be noted that the CINEG design itself does not require the use of equivalent groups. As a matter of fact, in many operational settings, such as when only one form can be administrated per test date due to test security or other concerns, the groups of

examinees taking different forms are usually not considered to be equivalent. Therefore, the composition of common-item sets under nonequivalent groups condition should be worthy of more attention.

Second, regardless of the group ability differences, there were no statistically and practically significant interaction effects among the factors of the statistical, content and format representativeness. It indicated that instead of impacting interactively on the performance of the concurrent calibration, these three important characteristics of common-item sets tended to affect the equating results independently.

Third, as expected, under the unidimensional test structure, the content and format representativeness factors showed little significant impact on the equating results except for a few conditions. Meanwhile, the statistical representativeness factor affected the performance of the concurrent calibration significantly. This result is determined by the nature of the unidimensional test structure. Under this condition, different content areas and multiple item formats truly measure the same ability. Therefore, whether content and item format are representative to the total test no longer becomes an issue in the composition of common-item sets. However, the statistical specification is least influenced by the test dimensionality structure and thus it becomes the most important characteristic when constructing common-item sets from unidimensional tests.

Fourth, regardless of the various levels of multidimensional test structure, the statistical representativeness factor showed more significant and systematic effects on the performance of the concurrent calibration than the other two important characteristics of common-item sets – the content and format representativeness factors. When the degree of multidimensionality due to multiple item formats increased (from Multidimensionality

(ρ=0.90/0.75, α=10°) to Multidimensionality (ρ=0.90/0.75, α=35°)), the format representativeness factor began to make significant differences on all three evaluation criteria especially under the nonequivalent groups condition. The content representativeness factor, however, showed minimum impact on the equating results regardless of the increase of the degree of multidimensionality due to different content areas (from Multidimensionality (ρ=0.90, α=10°/35°) to Multidimensionality (ρ=0.75, α=10°/35°)). There is no surprise that the statistical representativeness factor consistently imposed its impact on the equating results and the format representativeness started to become a contributing factor to the performance of the concurrent calibration when the two groups were not equivalent and the degree of multidimensionality due to item formats further increased. Interestingly, the content representativeness factor did not show statistically and practically significant effect on the equating results even when the degree of multidimensionality due to multiple content areas increased. This finding is inconsistent with several previous research papers (Cook & Peterson, 1987; Harris, 1991; Klein & Jarjoura, 1985) in which the content representativeness of common-item sets played an important role in the equating for MC-only tests under CINEG design. This inconsistency could be attributed to the specific design properties of this simulation study. As noted before, in this study, the test consisted of items classified into two content areas and the correlation between these two content areas was set at 0.9 or 0.75, which represented a realistically strong relationship between the two content areas. The degree of multidimensionality caused by highly correlated content areas might not be sufficient to invoke the significant effect of the content representativeness factor. Furthermore, the effect of the content representativeness factor in the previous study results was explained

in terms of the mean performance differences of the nonequivalent groups in various content areas (Harris, 1991; Kirkpatrick, 2005; Klein & Jarjoura, 1985). However, this study assumed that the mean performance on various content areas is completely parallel for two examinee groups. It might also be the reason to cause the inconsistency between the results from the current study and previous research.

Fifth, the performance of the concurrent calibration with the unidimensional IRT models declined significantly with the increase of the degree of multidimensionality caused by different content areas or multiple item formats, which indicated that the concurrent calibration was not quite robust to the violation of the unidimensionality assumption. This finding agrees with Beguin, Hanson, & Glas (2000)'s study in which they only examined MC tests but found that the performance of concurrent calibration was sensitive to the multidimensionality of the data. However, this finding differs from that from Kim & Kolen (2006), in which the concurrent calibration was found to be quite robust to the violation of the unidimensionality assumption compared to the separate calibration. But the evidence also indicated that this robustness of the concurrent calibration did not seem to be consistent across various test types.

Implications for Practice

With the profession's recognition of the importance of equating and with the steady increase in the use of mixed-format tests in large-scale assessments, mixed-format test equating attracts more and more attention and interest especially under CINEG design because in practice, the groups of examinees taking a test on different test dates are usually not considered simply equivalent from the same population. Under such circumstances, equating is conducted through a common-item set which is assumed to

provide accurate information about how the two examinee groups differ from one another with regard to performance on the total test. If two examinee groups performed equally well on the common-item set, but one group taking the new form scored higher than the other one taking the old form, it would appear that the new form was easier than the old form. Through equating, the scores of examinees taking the new and more difficult form would be adjusted accordingly.

Given the importance of the information obtained on the basis of the common-item set, it is a widely held belief that the composition of the common-item set is crucial to the CINEG equating practice and a typical common-item set should be a parallel miniature version of the total test. More specifically, it is recommended that a common-item set should be proportionally representative of the total test in terms of content and statistical specifications (Kolen & Brennan, 2004). In mixed-format test equating, it also should be format representative. However, the reality is far from ideal, and circumstances may arise in practice that cause the common-item set to become non-representative. For example, because of time and budget limitations, an insufficient number of CR items may be included in the common-item set. Or because of a recent change in the current knowledge system, an error in the printing of the test booklets, or an unauthorized rearranging of the item options, some items have to be removed from the common-item set. The removal of some common items might in turn result in the statistical, content, or format non-representativeness of the common-item set to the total test. Therefore, the research on the composition of common-item sets, especially about which requirements of statistical, content and/or format representativeness could be less restrictive, becomes helpful.

The major contribution of this study is to provide guidance on how to construct an "optimum" common-item set in terms of the statistical, content and format representativeness for the mixed-format test equating under CINEG design. Here, "optimum" means that under a certain practical condition, some requirement(s) in terms of the statistical, content and format representativeness should be taken more seriously and some other requirement(s) could be relaxed. Before making any decisions for constructing a common-item set, it should always be kept in mind that every real testing program has its own special properties, thus it is reasonable in practice to avoid blindly following the suggestions made below, although these are good initial guidelines.

Based on the discussion of the major findings of this study, the following two pieces of suggestions are recommended for constructing a common-item set in practice.

ONE: When the test forms are constructed to be truly unidimensional, the statistical representativeness will be the most crucial characteristic in the composition of common-item sets. In other words, when constructing a common-item set, the item statistical specification in the common-item set should be as similar as possible to that in the total test. The requirements of content and format representativeness could be loosened, such as using only MC items in the common-item set.

TWO: More realistically, it is likely that all tests will contain a certain degree of multidimensionality, and that this multidimensionality might be caused by many different factors. No matter what kind of multidimensional test structure it is, the statistical characteristics of common-item sets should remain proportionally representative of the total test. Furthermore, when the level of multidimensionality relating to the use of multiple item formats is moderate to severe, the common-item set should include both

MC and CR items and the proportion of MC and CR items should be roughly equal to that of the total test being equated. In addition, requiring the common-item set to mimic the content characteristics of the total test may be too restrictive and could be relaxed as long as various content areas are moderately to highly correlated with one another.

This study also provides information on the robustness of the concurrent calibration to various levels of multidimensionality. In general, the concurrent calibration with the unidimensional IRT models was not very robust to the violation of the unidimensionality assumption and the adequacy of its performance declined with increase in the degree of multidimensionality. This finding should ring an alarm bell for the equating practice in which the unidimensional IRT equating methods are sometimes used without examining the unidimensionality assumption. When multidimensionality exists, the use of concurrent calibration with the unidimensional IRT models might bias the results and thus the concurrent calibration using multidimensional IRT models should be considered. For example, a multidimensional compensatory or non-compensatory model could replace the 3PL model for MC items and a multidimensional version of the GRM could be used instead of the unidimensional version of the GRM.

This study not only has straightforward implications for the equating practice such as better selection of common items and equating methods, but also calls for great caution when choosing the evaluation criteria. Throughout this chapter, the word "optimum" has appeared in quotes. This is to emphasize that all the conclusions reached and the implications of this study are based on the specific evaluation criteria used (i.e., the BIAS, the RMSE, and the classification consistency). These criteria are among a number of criteria that could be used to evaluate the equating accuracy (e.g., Pearson

correlation coefficient used by Yang, 2000; moments for examinee ability distributions used by Cao, Yin, & Gao, 2007; unweighted and weighted root mean square differences, Harris & Crouse, 1993; weighted mean absolute error used by Beguin, Hanson, & Glas, 2000), with different criteria possibly resulting in different "optimum" common-item sets. Overall, the BIAS and RMSE, although having small discrepancies in estimating the equating accuracy, agreed with each other in their general pattern of results. The classification consistency measure, however, led to somewhat different conclusions about the equating accuracy. The differences in classification consistency proportions across all the simulation conditions were small. However, even small differences can have policy implications depending on how scores are reported, how large the examinee base is, and how serious the consequences are associated with decisions based on test scores. In the current K-12 state testing setting, many educational tests are used to classify examinees into levels of achievement, and stakes associated with the score reporting are usually at the highest level for the state policymaking body. Therefore, if the use of different evaluation criteria results in different conclusions and suggestions, it will be extremely important to select evaluation criteria that suit the state needs, are thoroughly examined and have well known features.

Overall, this study informs the equating practice in many aspects, such as better composition of common-item sets, better employment of equating methods and better selection of evaluation criteria, which ultimately will lead to a more precise, efficient, and fairer testing practice.

Limitations and Suggestions for Future Research

The above summary and discussion of the results must be kept in perspective, keeping in mind the following limitations of this study, which could be inspiration for a future research agenda.

As with any simulation study, the results are used in an attempt to understand the real world. However, real testing problems are far from simple and simulations cannot include all features of the real testing environment. In this study, only a small number of factors with limited levels were investigated. And many other factors that are often believed to affect the equating results were fixed. Among the fixed factors are the sample size, the proportion of MC items to CR items in the total test, the length of the common-item set, the IRT models and the calibration program. For example, the length of the common-item set in this study was fixed as one third of the total test to guarantee sufficient number for accurate equating. However, previous research found that the length of the common-item set interacted with the content representativeness. Meanwhile, in many practical settings, repeating large numbers of common items might not be possible because of the need to frequently update the item pool and the potential unfair advantage to failed candidates who take tests on successive dates. Therefore, future research could utilize shorter common-item sets and examine their interaction with other characteristics of common-item sets like the statistical, content and format representativeness. Also, future research could expand by investigating more levels of the existing five factors in this study. For example, this study did not find a significant effect of the content representativeness on the performance of the concurrent calibration. One plausible reason is that the correlation between content areas was set high, and thus

moderate to low correlations could be manipulated to see whether the effect of the content representatives will emerge. Moreover, this study only investigated one type of statistical non-representativeness, that is, the 0.3 mean difficulty difference between items in the common-item set and in the total test. Lower magnitude of the mean difficulty difference as well as the variability differences of the difficulty parameter could be explored in the future. It may be that if more factors were included in the simulation study, more generalizable findings could be obtained, although the complexity of the design would have made interpretation that much more difficult.

In this study, the multidimensionality was caused by different content areas and item formats and the two examinee groups were assumed to perform equally well on both content areas and on both item formats. However, in real test administration, the performance of two examinee groups is seldom identical. Therefore, future research might allow a differential ability in population means and variances on various content areas and different item formats.

This study only investigated the performance of the concurrent calibration under various simulation conditions. However, separate calibration is also widely used in real testing programs. One potential benefit of separate calibration is for diagnostic purposes, that is, having two sets of item parameter estimates can help to identify possible individual item problems (Hanson & Beguin, 2002; Kolen & Brennan, 2004). Moreover, concurrent calibration puts more of a burden on the program than separate estimation does, which may result in some performance problems. This could especially be true in the situations where more than two forms were being equated simultaneously. In separate calibration convergence problems seldom occur. All in all, there is insufficient evidence

to recommend completely avoiding separate calibration in favor of concurrent calibration. Therefore, this study could be expanded in the future to explore the performance of separate calibration (at least characteristic curve methods) under various conditions and compare it to that of concurrent calibration.

Next, the complex factorial model for multidimensional test structure assumed in this study, although specified to reflect one reasonable configuration for a large-scale assessment, should be tested in practice to investigate how closely it mimics the real tests. Even so, this factorial model is only one among many possible complex underlying factorial structures. In real testing situations, each test has its unique properties and thus its underlying factorial structure could vary largely and become even more complex. For example, various content and format factor combinations could impose different influences on examinee's correct response to each item, in contrast to the assumption of equivalent influences among factor combinations made in this study. Moreover, according to the robustness check conducted in this study, when multidimensionality exists, the use of concurrent calibration with unidimensional IRT models might bias the results and thus the concurrent calibration using applicable multidimensional IRT models should be considered. However, multidimensional IRT equating especially for mixed-format tests is a barely explored area. More systematic investigations should be conducted to develop new multidimensional IRT equating methods, and explore their characteristics and behavior with different simulation conditions or in real situations.

Last but not least,, this study only involved part of the whole equating process. In testing programs, raw-to-scale score conversions are often conducted as the final step of equating and the resulting scale scores instead of the expected total scores are reported.

Thus, test practitioners might be more interested in the impact of different factors on the final scale scores. Since weighing and score conversion processes in mixed-format tests involve many complicated issues, additional effort in this area should be invested in the future.

Appendix A

	Group 1	Group 2	
		Equivalent	Nonequivalent
Unidimensionality	N (0, 1)	N (0, 1)	N (0.5, 1)
Multidimensionality			
$\rho_{content1,content2} = 0.9$	$\Sigma = \begin{bmatrix} 1 & & & \\ 0.9 & 1 & & \\ 0 & 0 & 1 & \\ 0 & 0 & 0 & 1 \end{bmatrix}$ $\mu = [0 \quad 0 \quad 0 \quad 0]$	$\Sigma = \begin{bmatrix} 1 & & & \\ 0.9 & 1 & & \\ 0 & 0 & 1 & \\ 0 & 0 & 0 & 1 \end{bmatrix}$ $\mu = [0 \quad 0 \quad 0 \quad 0]$	$\Sigma = \begin{bmatrix} 1 & & & \\ 0.9 & 1 & & \\ 0 & 0 & 1 & \\ 0 & 0 & 0 & 1 \end{bmatrix}$ $\mu = [0.5 \quad 0.5 \quad 0.5 \quad 0.5]$
$\rho_{content1,content2} = 0.75$	$\Sigma = \begin{bmatrix} 1 & & & \\ 0.75 & 1 & & \\ 0 & 0 & 1 & \\ 0 & 0 & 0 & 1 \end{bmatrix}$ $\mu = [0 \quad 0 \quad 0 \quad 0]$	$\Sigma = \begin{bmatrix} 1 & & & \\ 0.75 & 1 & & \\ 0 & 0 & 1 & \\ 0 & 0 & 0 & 1 \end{bmatrix}$ $\mu = [0 \quad 0 \quad 0 \quad 0]$	$\Sigma = \begin{bmatrix} 1 & & & \\ 0.75 & 1 & & \\ 0 & 0 & 1 & \\ 0 & 0 & 0 & 1 \end{bmatrix}$ $\mu = [0.5 \quad 0.5 \quad 0.5 \quad 0.5]$

Note: The order of factors in the multivariate normal distribution is Content 1, Content 2, MC_F, and CR_F.

Appendix B

Unidimensionality, Format Rep., Content Rep., Statistical Rep.

Item Set		Content Area 1	**Item#**	Content Area 2	**Item#**
Unique item set for Form 1	MC	a~LN(0, 0.5) b~N(0, 1) c~BETA(8, 32)	**16**	a~LN(0, 0.5) b~N(0, 1) c~BETA(8, 32)	**16**
	CR	a~LN(0, 0.5) b1~N(-1.5, 0.2) b2~N(-0.5, 0.2) b3~N(0.5, 0.2) b4~N(1.5, 0.2)	**2**	a~LN(0, 0.5) b1~N(-1.5, 0.2) b2~N(-0.5, 0.2) b3~N(0.5, 0.2) b4~N(1.5, 0.2)	**2**
Common-item set	MC	a~LN(0, 0.5) b~N(0, 1) c~BETA(8, 32)	**8**	a~LN(0, 0.5) b~N(0, 1) c~BETA(8, 32)	**8**
	CR	a~LN(0, 0.5) b1~N(-1.5, 0.2) b2~N(-0.5, 0.2) b3~N(0.5, 0.2) b4~N(1.5, 0.2)	**1**	a~LN(0, 0.5) b1~N(-1.5, 0.2) b2~N(-0.5, 0.2) b3~N(0.5, 0.2) b4~N(1.5, 0.2)	**1**
Unique item set for Form 2	MC	a~LN(0, 0.5) b~N(0.5, 1) c~BETA(8, 32)	**16**	a~LN(0, 0.5) b~N(0.5, 1) c~BETA(8, 32)	**16**
	CR	a~LN(0, 0.5) b1~N(-1.0, 0.2) b2~N(0, 0.2) b3~N(1.0, 0.2) b4~N(2.0, 0.2)	**2**	a~LN(0, 0.5) b1~N(-1.0, 0.2) b2~N(0, 0.2) b3~N(1.0, 0.2) b4~N(2.0, 0.2)	**2**

Unidimensionality, Format Rep., Content Rep., Statistical Non-Rep.

Item Set		Content Area 1	**Item#**	Content Area 2	**Item#**
Unique item set for Form 1	MC	a~LN(0, 0.5) b~N(0, 1) c~BETA(8, 32)	**16**	a~LN(0, 0.5) b~N(0, 1) c~BETA(8, 32)	**16**
	CR	a~LN(0, 0.5) b1~N(-1.5, 0.2) b2~N(-0.5, 0.2) b3~N(0.5, 0.2) b4~N(1.5, 0.2)	**2**	a~LN(0, 0.5) b1~N(-1.5, 0.2) b2~N(-0.5, 0.2) b3~N(0.5, 0.2) b4~N(1.5, 0.2)	**2**
Common-item set	MC	a~LN(0, 0.5) b~N(0.3, 1) c~BETA(8, 32)	**8**	a~LN(0, 0.5) b~N(0.3, 1) c~BETA(8, 32)	**8**
	CR	a~LN(0, 0.5) b1~N(-1.2, 0.2) b2~N(-0.2, 0.2) b3~N(0.8, 0.2) b4~N(1.8, 0.2)	**1**	a~LN(0, 0.5) b1~N(-1.2, 0.2) b2~N(-0.2, 0.2) b3~N(0.8, 0.2) b4~N(1.8, 0.2)	**1**
Unique item set for Form 2	MC	a~LN(0, 0.5) b~N(0.5, 1) c~BETA(8, 32)	**16**	a~LN(0, 0.5) b~N(0.5, 1) c~BETA(8, 32)	**16**
	CR	a~LN(0, 0.5) b1~N(-1.0, 0.2) b2~N(0, 0.2) b3~N(1.0, 0.2) b4~N(2.0, 0.2)	**2**	a~LN(0, 0.5) b1~N(-1.0, 0.2) b2~N(0, 0.2) b3~N(1.0, 0.2) b4~N(2.0, 0.2)	**2**

Unidimensionality, Format Rep., Content Partially Under-Rep., Statistical Rep.

Item Set		Content Area 1	**Item#**	Content Area 2	**Item#**
Unique item set for Form 1	MC	a~LN(0, 0.5) b~N(0, 1) c~BETA(8, 32)	**13**	a~LN(0, 0.5) b~N(0, 1) c~BETA(8, 32)	**19**
	CR	a~LN(0, 0.5) b1~N(-1.5, 0.2) b2~N(-0.5, 0.2) b3~N(0.5, 0.2) b4~N(1.5, 0.2)	**2**	a~LN(0, 0.5) b1~N(-1.5, 0.2) b2~N(-0.5, 0.2) b3~N(0.5, 0.2) b4~N(1.5, 0.2)	**2**
Common-item set	MC	a~LN(0, 0.5) b~N(0, 1) c~BETA(8, 32)	**11**	a~LN(0, 0.5) b~N(0, 1) c~BETA(8, 32)	**5**
	CR	a~LN(0, 0.5) b1~N(-1.5, 0.2) b2~N(-0.5, 0.2) b3~N(0.5, 0.2) b4~N(1.5, 0.2)	**1**	a~LN(0, 0.5) b1~N(-1.5, 0.2) b2~N(-0.5, 0.2) b3~N(0.5, 0.2) b4~N(1.5, 0.2)	**1**
Unique item set for Form 2	MC	a~LN(0, 0.5) b~N(0.5, 1) c~BETA(8, 32)	**13**	a~LN(0, 0.5) b~N(0.5, 1) c~BETA(8, 32)	**19**
	CR	a~LN(0, 0.5) b1~N(-1.0, 0.2) b2~N(0, 0.2) b3~N(1.0, 0.2) b4~N(2.0, 0.2)	**2**	a~LN(0, 0.5) b1~N(-1.0, 0.2) b2~N(0, 0.2) b3~N(1.0, 0.2) b4~N(2.0, 0.2)	**2**

Unidimensionality, Format Rep., Content Partially Under-Rep., Statistical Non-Rep.

Item Set		Content Area 1	**Item#**	Content Area 2	**Item#**
Unique item set for Form 1	MC	a~LN(0, 0.5) b~N(0, 1) c~BETA(8, 32)	**13**	a~LN(0, 0.5) b~N(0, 1) c~BETA(8, 32)	**19**
	CR	a~LN(0, 0.5) b1~N(-1.5, 0.2) b2~N(-0.5, 0.2) b3~N(0.5, 0.2) b4~N(1.5, 0.2)	**2**	a~LN(0, 0.5) b1~N(-1.5, 0.2) b2~N(-0.5, 0.2) b3~N(0.5, 0.2) b4~N(1.5, 0.2)	**2**
Common-item set	MC	a~LN(0, 0.5) b~N(0.3, 1) c~BETA(8, 32)	**11**	a~LN(0, 0.5) b~N(0.3, 1) c~BETA(8, 32)	**5**
	CR	a~LN(0, 0.5) b1~N(-1.2, 0.2) b2~N(-0.2, 0.2) b3~N(0.8, 0.2) b4~N(1.8, 0.2)	**1**	a~LN(0, 0.5) b1~N(-1.2, 0.2) b2~N(-0.2, 0.2) b3~N(0.8, 0.2) b4~N(1.8, 0.2)	**1**
Unique item set for Form 2	MC	a~LN(0, 0.5) b~N(0.5, 1) c~BETA(8, 32)	**13**	a~LN(0, 0.5) b~N(0.5, 1) c~BETA(8, 32)	**19**
	CR	a~LN(0, 0.5) b1~N(-1.0, 0.2) b2~N(0, 0.2) b3~N(1.0, 0.2) b4~N(2.0, 0.2)	**2**	a~LN(0, 0.5) b1~N(-1.0, 0.2) b2~N(0, 0.2) b3~N(1.0, 0.2) b4~N(2.0, 0.2)	**2**

Unidimensionality, Format Rep., Content Completely Under-Rep., Statistical Rep.

Item Set		Content Area 1	**Item#**	Content Area 2	**Item#**
Unique item set for Form 1	MC	a~LN(0, 0.5) b~N(0, 1) c~BETA(8, 32)	**8**	a~LN(0, 0.5) b~N(0, 1) c~BETA(8, 32)	**24**
	CR	a~LN(0, 0.5) b1~N(-1.5, 0.2) b2~N(-0.5, 0.2) b3~N(0.5, 0.2) b4~N(1.5, 0.2)	**1**	a~LN(0, 0.5) b1~N(-1.5, 0.2) b2~N(-0.5, 0.2) b3~N(0.5, 0.2) b4~N(1.5, 0.2)	**3**
Common-item set	MC	a~LN(0, 0.5) b~N(0, 1) c~BETA(8, 32)	**16**	a~LN(0, 0.5) b~N(0, 1) c~BETA(8, 32)	**0**
	CR	a~LN(0, 0.5) b1~N(-1.5, 0.2) b2~N(-0.5, 0.2) b3~N(0.5, 0.2) b4~N(1.5, 0.2)	**2**	a~LN(0, 0.5) b1~N(-1.5, 0.2) b2~N(-0.5, 0.2) b3~N(0.5, 0.2) b4~N(1.5, 0.2)	**0**
Unique item set for Form 2	MC	a~LN(0, 0.5) b~N(0.5, 1) c~BETA(8, 32)	**8**	a~LN(0, 0.5) b~N(0.5, 1) c~BETA(8, 32)	**24**
	CR	a~LN(0, 0.5) b1~N(-1.0, 0.2) b2~N(0, 0.2) b3~N(1.0, 0.2) b4~N(2.0, 0.2)	**1**	a~LN(0, 0.5) b1~N(-1.0, 0.2) b2~N(0, 0.2) b3~N(1.0, 0.2) b4~N(2.0, 0.2)	**3**

Unidimensionality, Format Rep., Content Completely Under-Rep., Statistical Non-Rep.

Item Set		Content Area 1	**Item#**	Content Area 2	**Item#**
Unique item set for Form 1	MC	a~LN(0, 0.5) b~N(0, 1) c~BETA(8, 32)	**8**	a~LN(0, 0.5) b~N(0, 1) c~BETA(8, 32)	**24**
	CR	a~LN(0, 0.5) b1~N(-1.5, 0.2) b2~N(-0.5, 0.2) b3~N(0.5, 0.2) b4~N(1.5, 0.2)	**1**	a~LN(0, 0.5) b1~N(-1.5, 0.2) b2~N(-0.5, 0.2) b3~N(0.5, 0.2) b4~N(1.5, 0.2)	**3**
Common-item set	MC	a~LN(0, 0.5) b~N(0.3, 1) c~BETA(8, 32)	**16**	a~LN(0, 0.5) b~N(0.3, 1) c~BETA(8, 32)	**0**
	CR	a~LN(0, 0.5) b1~N(-1.2, 0.2) b2~N(-0.2, 0.2) b3~N(0.8, 0.2) b4~N(1.8, 0.2)	**2**	a~LN(0, 0.5) b1~N(-1.2, 0.2) b2~N(-0.2, 0.2) b3~N(0.8, 0.2) b4~N(1.8, 0.2)	**0**
Unique item set for Form 2	MC	a~LN(0, 0.5) b~N(0.5, 1) c~BETA(8, 32)	**8**	a~LN(0, 0.5) b~N(0.5, 1) c~BETA(8, 32)	**24**
	CR	a~LN(0, 0.5) b1~N(-1.0, 0.2) b2~N(0, 0.2) b3~N(1.0, 0.2) b4~N(2.0, 0.2)	**1**	a~LN(0, 0.5) b1~N(-1.0, 0.2) b2~N(0, 0.2) b3~N(1.0, 0.2) b4~N(2.0, 0.2)	**3**

Unidimensionality, Format Non-Rep., Content Rep., Statistical Rep.

Item Set		Content Area 1	**Item#**	Content Area 2	**Item#**
Unique item set for Form 1	MC	a~LN(0, 0.5) b~N(0, 1) c~BETA(8, 32)	**15**	a~LN(0, 0.5) b~N(0, 1) c~BETA(8, 32)	**15**
	CR	a~LN(0, 0.5) b1~N(-1.5, 0.2) b2~N(-0.5, 0.2) b3~N(0.5, 0.2) b4~N(1.5, 0.2)	**3**	a~LN(0, 0.5) b1~N(-1.5, 0.2) b2~N(-0.5, 0.2) b3~N(0.5, 0.2) b4~N(1.5, 0.2)	**3**
Common-item set	MC	a~LN(0, 0.5) b~N(0, 1) c~BETA(8, 32)	**9**	a~LN(0, 0.5) b~N(0, 1) c~BETA(8, 32)	**9**
	CR	a~LN(0, 0.5) b1~N(-1.5, 0.2) b2~N(-0.5, 0.2) b3~N(0.5, 0.2) b4~N(1.5, 0.2)	**0**	a~LN(0, 0.5) b1~N(-1.5, 0.2) b2~N(-0.5, 0.2) b3~N(0.5, 0.2) b4~N(1.5, 0.2)	**0**
Unique item set for Form 2	MC	a~LN(0, 0.5) b~N(0.5, 1) c~BETA(8, 32)	**15**	a~LN(0, 0.5) b~N(0.5, 1) c~BETA(8, 32)	**15**
	CR	a~LN(0, 0.5) b1~N(-1.0, 0.2) b2~N(0, 0.2) b3~N(1.0, 0.2) b4~N(2.0, 0.2)	**3**	a~LN(0, 0.5) b1~N(-1.0, 0.2) b2~N(0, 0.2) b3~N(1.0, 0.2) b4~N(2.0, 0.2)	**3**

Unidimensionality, Format Non-Rep., Content Rep., Statistical Non-Rep.

Item Set		Content Area 1	**Item#**	Content Area 2	**Item#**
Unique item set for Form 1	MC	a~LN(0, 0.5) b~N(0, 1) c~BETA(8, 32)	**15**	a~LN(0, 0.5) b~N(0, 1) c~BETA(8, 32)	**15**
	CR	a~LN(0, 0.5) b1~N(-1.5, 0.2) b2~N(-0.5, 0.2) b3~N(0.5, 0.2) b4~N(1.5, 0.2)	**3**	a~LN(0, 0.5) b1~N(-1.5, 0.2) b2~N(-0.5, 0.2) b3~N(0.5, 0.2) b4~N(1.5, 0.2)	**3**
Common-item set	MC	a~LN(0, 0.5) b~N(0.3, 1) c~BETA(8, 32)	**9**	a~LN(0, 0.5) b~N(0.3, 1) c~BETA(8, 32)	**9**
	CR	a~LN(0, 0.5) b1~N(-1.2, 0.2) b2~N(-0.2, 0.2) b3~N(0.8, 0.2) b4~N(1.8, 0.2)	**0**	a~LN(0, 0.5) b1~N(-1.2, 0.2) b2~N(-0.2, 0.2) b3~N(0.8, 0.2) b4~N(1.8, 0.2)	**0**
Unique item set for Form 2	MC	a~LN(0, 0.5) b~N(0.5, 1) c~BETA(8, 32)	**15**	a~LN(0, 0.5) b~N(0.5, 1) c~BETA(8, 32)	**15**
	CR	a~LN(0, 0.5) b1~N(-1.0, 0.2) b2~N(0, 0.2) b3~N(1.0, 0.2) b4~N(2.0, 0.2)	**3**	a~LN(0, 0.5) b1~N(-1.0, 0.2) b2~N(0, 0.2) b3~N(1.0, 0.2) b4~N(2.0, 0.2)	**3**

Unidimensionality, Format Non-Rep., Content Partially Under-Rep., Statistical Rep.

Item Set		Content Area 1	**Item#**	Content Area 2	**Item#**
Unique item set for Form 1	MC	a~LN(0, 0.5) b~N(0, 1) c~BETA(8, 32)	**12**	a~LN(0, 0.5) b~N(0, 1) c~BETA(8, 32)	**18**
	CR	a~LN(0, 0.5) b1~N(-1.5, 0.2) b2~N(-0.5, 0.2) b3~N(0.5, 0.2) b4~N(1.5, 0.2)	**3**	a~LN(0, 0.5) b1~N(-1.5, 0.2) b2~N(-0.5, 0.2) b3~N(0.5, 0.2) b4~N(1.5, 0.2)	**3**
Common-item set	MC	a~LN(0, 0.5) b~N(0, 1) c~BETA(8, 32)	**12**	a~LN(0, 0.5) b~N(0, 1) c~BETA(8, 32)	**6**
	CR	a~LN(0, 0.5) b1~N(-1.5, 0.2) b2~N(-0.5, 0.2) b3~N(0.5, 0.2) b4~N(1.5, 0.2)	**0**	a~LN(0, 0.5) b1~N(-1.5, 0.2) b2~N(-0.5, 0.2) b3~N(0.5, 0.2) b4~N(1.5, 0.2)	**0**
Unique item set for Form 2	MC	a~LN(0, 0.5) b~N(0.5, 1) c~BETA(8, 32)	**12**	a~LN(0, 0.5) b~N(0.5, 1) c~BETA(8, 32)	**18**
	CR	a~LN(0, 0.5) b1~N(-1.0, 0.2) b2~N(0, 0.2) b3~N(1.0, 0.2) b4~N(2.0, 0.2)	**3**	a~LN(0, 0.5) b1~N(-1.0, 0.2) b2~N(0, 0.2) b3~N(1.0, 0.2) b4~N(2.0, 0.2)	**3**

Unidimensionality, Format Non-Rep., Content Partially Under-Rep., Statistical Non-Rep.

Item Set		Content Area 1	**Item#**	Content Area 2	**Item#**
Unique item set for Form 1	MC	a~LN(0, 0.5) b~N(0, 1) c~BETA(8, 32)	**12**	a~LN(0, 0.5) b~N(0, 1) c~BETA(8, 32)	**18**
	CR	a~LN(0, 0.5) b1~N(-1.5, 0.2) b2~N(-0.5, 0.2) b3~N(0.5, 0.2) b4~N(1.5, 0.2)	**3**	a~LN(0, 0.5) b1~N(-1.5, 0.2) b2~N(-0.5, 0.2) b3~N(0.5, 0.2) b4~N(1.5, 0.2)	**3**
Common-item set	MC	a~LN(0, 0.5) b~N(0.3, 1) c~BETA(8, 32)	**12**	a~LN(0, 0.5) b~N(0.3, 1) c~BETA(8, 32)	**6**
	CR	a~LN(0, 0.5) b1~N(-1.2, 0.2) b2~N(-0.2, 0.2) b3~N(0.8, 0.2) b4~N(1.8, 0.2)	**0**	a~LN(0, 0.5) b1~N(-1.2, 0.2) b2~N(-0.2, 0.2) b3~N(0.8, 0.2) b4~N(1.8, 0.2)	**0**
Unique item set for Form 2	MC	a~LN(0, 0.5) b~N(0.5, 1) c~BETA(8, 32)	**12**	a~LN(0, 0.5) b~N(0.5, 1) c~BETA(8, 32)	**18**
	CR	a~LN(0, 0.5) b1~N(-1.0, 0.2) b2~N(0, 0.2) b3~N(1.0, 0.2) b4~N(2.0, 0.2)	**3**	a~LN(0, 0.5) b1~N(-1.0, 0.2) b2~N(0, 0.2) b3~N(1.0, 0.2) b4~N(2.0, 0.2)	**3**

Unidimensionality, Format Non-Rep., Content Completely Under-Rep., Statistical Rep.

Item Set		Content Area 1	**Item#**	Content Area 2	**Item#**
Unique item set for Form 1	MC	a~LN(0, 0.5) b~N(0, 1) c~BETA(8, 32)	**6**	a~LN(0, 0.5) b~N(0, 1) c~BETA(8, 32)	**24**
	CR	a~LN(0, 0.5) b1~N(-1.5, 0.2) b2~N(-0.5, 0.2) b3~N(0.5, 0.2) b4~N(1.5, 0.2)	**3**	a~LN(0, 0.5) b1~N(-1.5, 0.2) b2~N(-0.5, 0.2) b3~N(0.5, 0.2) b4~N(1.5, 0.2)	**3**
Common-item set	MC	a~LN(0, 0.5) b~N(0, 1) c~BETA(8, 32)	**18**	a~LN(0, 0.5) b~N(0, 1) c~BETA(8, 32)	**0**
	CR	a~LN(0, 0.5) b1~N(-1.5, 0.2) b2~N(-0.5, 0.2) b3~N(0.5, 0.2) b4~N(1.5, 0.2)	**0**	a~LN(0, 0.5) b1~N(-1.5, 0.2) b2~N(-0.5, 0.2) b3~N(0.5, 0.2) b4~N(1.5, 0.2)	**0**
Unique item set for Form 2	MC	a~LN(0, 0.5) b~N(0.5, 1) c~BETA(8, 32)	**6**	a~LN(0, 0.5) b~N(0.5, 1) c~BETA(8, 32)	**24**
	CR	a~LN(0, 0.5) b1~N(-1.0, 0.2) b2~N(0, 0.2) b3~N(1.0, 0.2) b4~N(2.0, 0.2)	**3**	a~LN(0, 0.5) b1~N(-1.0, 0.2) b2~N(0, 0.2) b3~N(1.0, 0.2) b4~N(2.0, 0.2)	**3**

Unidimensionality, Format Non-Rep., Content Completely Under-Rep., Statistical Non-Rep.

Item Set		Content Area 1	**Item#**	Content Area 2	**Item#**
Unique item set for Form 1	MC	a~LN(0, 0.5) b~N(0, 1) c~BETA(8, 32)	**6**	a~LN(0, 0.5) b~N(0, 1) c~BETA(8, 32)	**24**
	CR	a~LN(0, 0.5) b1~N(-1.5, 0.2) b2~N(-0.5, 0.2) b3~N(0.5, 0.2) b4~N(1.5, 0.2)	**3**	a~LN(0, 0.5) b1~N(-1.5, 0.2) b2~N(-0.5, 0.2) b3~N(0.5, 0.2) b4~N(1.5, 0.2)	**3**
Common-item set	MC	a~LN(0, 0.5) b~N(0.3, 1) c~BETA(8, 32)	**18**	a~LN(0, 0.5) b~N(0.3, 1) c~BETA(8, 32)	**0**
	CR	a~LN(0, 0.5) b1~N(-1.2, 0.2) b2~N(-0.2, 0.2) b3~N(0.8, 0.2) b4~N(1.8, 0.2)	**0**	a~LN(0, 0.5) b1~N(-1.2, 0.2) b2~N(-0.2, 0.2) b3~N(0.8, 0.2) b4~N(1.8, 0.2)	**0**
Unique item set for Form 2	MC	a~LN(0, 0.5) b~N(0.5, 1) c~BETA(8, 32)	**6**	a~LN(0, 0.5) b~N(0.5, 1) c~BETA(8, 32)	**24**
	CR	a~LN(0, 0.5) b1~N(-1.0, 0.2) b2~N(0, 0.2) b3~N(1.0, 0.2) b4~N(2.0, 0.2)	**3**	a~LN(0, 0.5) b1~N(-1.0, 0.2) b2~N(0, 0.2) b3~N(1.0, 0.2) b4~N(2.0, 0.2)	**3**

Appendix C

Multidimensionality, Format Rep., Content Rep., Statistical Rep.

Item Set		Content Area 1	**Item#**	Content Area 2	**Item#**
Unique item set for Form 1	MC	MDISC~LN(0, 0.5) MDIFF~N(0, 1) c~BETA(8, 32)	**16**	MDISC~LN(0, 0.5) MDIFF~N(0, 1) c~BETA(8, 32)	**16**
	CR	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.5, 0.2) $MDIFF_2$~N(-0.5, 0.2) $MDIFF_3$~N(0.5, 0.2) $MDIFF_4$~N(1.5, 0.2)	**2**	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.5, 0.2) $MDIFF_2$~N(-0.5, 0.2) $MDIFF_3$~N(0.5, 0.2) $MDIFF_4$~N(1.5, 0.2)	**2**
Common-item set	MC	MDISC~LN(0, 0.5) MDIFF~N(0, 1) c~BETA(8, 32)	**8**	MDISC~LN(0, 0.5) MDIFF~N(0, 1) c~BETA(8, 32)	**8**
	CR	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.5, 0.2) $MDIFF_2$~N(-0.5, 0.2) $MDIFF_3$~N(0.5, 0.2) $MDIFF_4$~N(1.5, 0.2)	**1**	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.5, 0.2) $MDIFF_2$~N(-0.5, 0.2) $MDIFF_3$~N(0.5, 0.2) $MDIFF_4$~N(1.5, 0.2)	**1**
Unique item set for Form 2	MC	MDISC~LN(0, 0.5) MDIFF~N(0.5, 1) c~BETA(8, 32)	**16**	MDISC~LN(0, 0.5) MDIFF~N(0.5, 1) c~BETA(8, 32)	**16**
	CR	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.0, 0.2) $MDIFF_2$~N(0, 0.2) $MDIFF_3$~N(1.0, 0.2) $MDIFF_4$~N(2.0, 0.2)	**2**	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.0, 0.2) $MDIFF_2$~N(0, 0.2) $MDIFF_3$~N(1.0, 0.2) $MDIFF_4$~N(2.0, 0.2)	**2**

Multidimensionality, Format Rep., Content Rep., Statistical Non-Rep.

Item Set		Content Area 1	**Item#**	Content Area 2	**Item#**
Unique item set for Form 1	MC	MDISC~LN(0, 0.5) MDIFF~N(0, 1) c~BETA(8, 32)	**16**	MDISC~LN(0, 0.5) MDIFF~N(0, 1) c~BETA(8, 32)	**16**
	CR	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.5, 0.2) $MDIFF_2$~N(-0.5, 0.2) $MDIFF_3$~N(0.5, 0.2) $MDIFF_4$~N(1.5, 0.2)	**2**	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.5, 0.2) $MDIFF_2$~N(-0.5, 0.2) $MDIFF_3$~N(0.5, 0.2) $MDIFF_4$~N(1.5, 0.2)	**2**
Common-item set	MC	MDISC~LN(0, 0.5) MDIFF~N(0.3, 1) c~BETA(8, 32)	**8**	MDISC~LN(0, 0.5) MDIFF~N(0.3, 1) c~BETA(8, 32)	**8**
	CR	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.2, 0.2) $MDIFF_2$~N(-0.2, 0.2) $MDIFF_3$~N(0.8, 0.2) $MDIFF_4$~N(1.8, 0.2)	**1**	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.2, 0.2) $MDIFF_2$~N(-0.2, 0.2) $MDIFF_3$~N(0.8, 0.2) $MDIFF_4$~N(1.8, 0.2)	**1**
Unique item set for Form 2	MC	MDISC~LN(0, 0.5) MDIFF~N(0.5, 1) c~BETA(8, 32)	**16**	MDISC~LN(0, 0.5) MDIFF~N(0.5, 1) c~BETA(8, 32)	**16**
	CR	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.0, 0.2) $MDIFF_2$~N(0, 0.2) $MDIFF_3$~N(1.0, 0.2) $MDIFF_4$~N(2.0, 0.2)	**2**	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.0, 0.2) $MDIFF_2$~N(0, 0.2) $MDIFF_3$~N(1.0, 0.2) $MDIFF_4$~N(2.0, 0.2)	**2**

Multidimensionality, Format Rep., Content Partially Under-Rep., Statistical Rep.

Item Set		Content Area 1	**Item#**	Content Area 2	**Item#**
Unique item set for Form 1	MC	MDISC~LN(0, 0.5) MDIFF~N(0, 1) c~BETA(8, 32)	**13**	MDISC~LN(0, 0.5) MDIFF~N(0, 1) c~BETA(8, 32)	**19**
	CR	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.5, 0.2) $MDIFF_2$~N(-0.5, 0.2) $MDIFF_3$~N(0.5, 0.2) $MDIFF_4$~N(1.5, 0.2)	**2**	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.5, 0.2) $MDIFF_2$~N(-0.5, 0.2) $MDIFF_3$~N(0.5, 0.2) $MDIFF_4$~N(1.5, 0.2)	**2**
Common-item set	MC	MDISC~LN(0, 0.5) MDIFF~N(0, 1) c~BETA(8, 32)	**11**	MDISC~LN(0, 0.5) MDIFF~N(0, 1) c~BETA(8, 32)	**5**
	CR	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.5, 0.2) $MDIFF_2$~N(-0.5, 0.2) $MDIFF_3$~N(0.5, 0.2) $MDIFF_4$~N(1.5, 0.2)	**1**	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.5, 0.2) $MDIFF_2$~N(-0.5, 0.2) $MDIFF_3$~N(0.5, 0.2) $MDIFF_4$~N(1.5, 0.2)	**1**
Unique item set for Form 2	MC	MDISC~LN(0, 0.5) MDIFF~N(0.5, 1) c~BETA(8, 32)	**13**	MDISC~LN(0, 0.5) MDIFF~N(0.5, 1) c~BETA(8, 32)	**19**
	CR	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.0, 0.2) $MDIFF_2$~N(0, 0.2) $MDIFF_3$~N(1.0, 0.2) $MDIFF_4$~N(2.0, 0.2)	**2**	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.0, 0.2) $MDIFF_2$~N(0, 0.2) $MDIFF_3$~N(1.0, 0.2) $MDIFF_4$~N(2.0, 0.2)	**2**

Multidimensionality, Format Rep., Content Partially Under-Rep., Statistical Non-Rep.

Item Set		Content Area 1	**Item#**	Content Area 2	**Item#**
Unique item set for Form 1	MC	MDISC~LN(0, 0.5) MDIFF~N(0, 1) c~BETA(8, 32)	**13**	MDISC~LN(0, 0.5) MDIFF~N(0, 1) c~BETA(8, 32)	**19**
	CR	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.5, 0.2) $MDIFF_2$~N(-0.5, 0.2) $MDIFF_3$~N(0.5, 0.2) $MDIFF_4$~N(1.5, 0.2)	**2**	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.5, 0.2) $MDIFF_2$~N(-0.5, 0.2) $MDIFF_3$~N(0.5, 0.2) $MDIFF_4$~N(1.5, 0.2)	**2**
Common-item set	MC	MDISC~LN(0, 0.5) MDIFF~N(0.3, 1) c~BETA(8, 32)	**11**	MDISC~LN(0, 0.5) MDIFF~N(0.3, 1) c~BETA(8, 32)	**5**
	CR	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.2, 0.2) $MDIFF_2$~N(-0.2, 0.2) $MDIFF_3$~N(0.8, 0.2) $MDIFF_4$~N(1.8, 0.2)	**1**	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.2, 0.2) $MDIFF_2$~N(-0.2, 0.2) $MDIFF_3$~N(0.8, 0.2) $MDIFF_4$~N(1.8, 0.2)	**1**
Unique item set for Form 2	MC	MDISC~LN(0, 0.5) MDIFF~N(0.5, 1) c~BETA(8, 32)	**13**	MDISC~LN(0, 0.5) MDIFF~N(0.5, 1) c~BETA(8, 32)	**19**
	CR	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.0, 0.2) $MDIFF_2$~N(0, 0.2) $MDIFF_3$~N(1.0, 0.2) $MDIFF_4$~N(2.0, 0.2)	**2**	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.0, 0.2) $MDIFF_2$~N(0, 0.2) $MDIFF_3$~N(1.0, 0.2) $MDIFF_4$~N(2.0, 0.2)	**2**

Multidimensionality, Format Rep., Content Completely Under-Rep., Statistical Rep.

Item Set		Content Area 1	**Item#**	Content Area 2	**Item#**
Unique item set for Form 1	MC	MDISC~LN(0, 0.5) MDIFF~N(0, 1) c~BETA(8, 32)	**8**	MDISC~LN(0, 0.5) MDIFF~N(0, 1) c~BETA(8, 32)	**24**
	CR	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.5, 0.2) $MDIFF_2$~N(-0.5, 0.2) $MDIFF_3$~N(0.5, 0.2) $MDIFF_4$~N(1.5, 0.2)	**1**	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.5, 0.2) $MDIFF_2$~N(-0.5, 0.2) $MDIFF_3$~N(0.5, 0.2) $MDIFF_4$~N(1.5, 0.2)	**3**
Common-item set	MC	MDISC~LN(0, 0.5) MDIFF~N(0, 1) c~BETA(8, 32)	**16**	MDISC~LN(0, 0.5) MDIFF~N(0, 1) c~BETA(8, 32)	**0**
	CR	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.5, 0.2) $MDIFF_2$~N(-0.5, 0.2) $MDIFF_3$~N(0.5, 0.2) $MDIFF_4$~N(1.5, 0.2)	**2**	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.5, 0.2) $MDIFF_2$~N(-0.5, 0.2) $MDIFF_3$~N(0.5, 0.2) $MDIFF_4$~N(1.5, 0.2)	**0**
Unique item set for Form 2	MC	MDISC~LN(0, 0.5) MDIFF~N(0.5, 1) c~BETA(8, 32)	**8**	MDISC~LN(0, 0.5) MDIFF~N(0.5, 1) c~BETA(8, 32)	**24**
	CR	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.0, 0.2) $MDIFF_2$~N(0, 0.2) $MDIFF_3$~N(1.0, 0.2) $MDIFF_4$~N(2.0, 0.2)	**1**	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.0, 0.2) $MDIFF_2$~N(0, 0.2) $MDIFF_3$~N(1.0, 0.2) $MDIFF_4$~N(2.0, 0.2)	**3**

Multidimensionality, Format Rep., Content Completely Under-Rep., Statistical Non-Rep.

Item Set		Content Area 1	**Item#**	Content Area 2	**Item#**
Unique item set for Form 1	MC	MDISC~LN(0, 0.5) MDIFF~N(0, 1) c~BETA(8, 32)	**8**	MDISC~LN(0, 0.5) MDIFF~N(0, 1) c~BETA(8, 32)	**24**
	CR	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.5, 0.2) $MDIFF_2$~N(-0.5, 0.2) $MDIFF_3$~N(0.5, 0.2) $MDIFF_4$~N(1.5, 0.2)	**1**	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.5, 0.2) $MDIFF_2$~N(-0.5, 0.2) $MDIFF_3$~N(0.5, 0.2) $MDIFF_4$~N(1.5, 0.2)	**3**
Common-item set	MC	MDISC~LN(0, 0.5) MDIFF~N(0.3, 1) c~BETA(8, 32)	**16**	MDISC~LN(0, 0.5) MDIFF~N(0.3, 1) c~BETA(8, 32)	**0**
	CR	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.2, 0.2) $MDIFF_2$~N(-0.2, 0.2) $MDIFF_3$~N(0.8, 0.2) $MDIFF_4$~N(1.8, 0.2)	**2**	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.2, 0.2) $MDIFF_2$~N(-0.2, 0.2) $MDIFF_3$~N(0.8, 0.2) $MDIFF_4$~N(1.8, 0.2)	**0**
Unique item set for Form 2	MC	MDISC~LN(0, 0.5) MDIFF~N(0.5, 1) c~BETA(8, 32)	**8**	MDISC~LN(0, 0.5) MDIFF~N(0.5, 1) c~BETA(8, 32)	**24**
	CR	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.0, 0.2) $MDIFF_2$~N(0, 0.2) $MDIFF_3$~N(1.0, 0.2) $MDIFF_4$~N(2.0, 0.2)	**1**	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.0, 0.2) $MDIFF_2$~N(0, 0.2) $MDIFF_3$~N(1.0, 0.2) $MDIFF_4$~N(2.0, 0.2)	**3**

Multidimensionality, Format Non-Rep., Content Rep., Statistical Rep.

Item Set		Content Area 1	**Item#**	Content Area 2	**Item#**
Unique item set for Form 1	MC	MDISC~LN(0, 0.5) MDIFF~N(0, 1) c~BETA(8, 32)	**15**	MDISC~LN(0, 0.5) MDIFF~N(0, 1) c~BETA(8, 32)	**15**
	CR	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.5, 0.2) $MDIFF_2$~N(-0.5, 0.2) $MDIFF_3$~N(0.5, 0.2) $MDIFF_4$~N(1.5, 0.2)	**3**	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.5, 0.2) $MDIFF_2$~N(-0.5, 0.2) $MDIFF_3$~N(0.5, 0.2) $MDIFF_4$~N(1.5, 0.2)	**3**
Common-item set	MC	MDISC~LN(0, 0.5) MDIFF~N(0, 1) c~BETA(8, 32)	**9**	MDISC~LN(0, 0.5) MDIFF~N(0, 1) c~BETA(8, 32)	**9**
	CR	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.5, 0.2) $MDIFF_2$~N(-0.5, 0.2) $MDIFF_3$~N(0.5, 0.2) $MDIFF_4$~N(1.5, 0.2)	**0**	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.5, 0.2) $MDIFF_2$~N(-0.5, 0.2) $MDIFF_3$~N(0.5, 0.2) $MDIFF_4$~N(1.5, 0.2)	**0**
Unique item set for Form 2	MC	MDISC~LN(0, 0.5) MDIFF~N(0.5, 1) c~BETA(8, 32)	**15**	MDISC~LN(0, 0.5) MDIFF~N(0.5, 1) c~BETA(8, 32)	**15**
	CR	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.0, 0.2) $MDIFF_2$~N(0, 0.2) $MDIFF_3$~N(1.0, 0.2) $MDIFF_4$~N(2.0, 0.2)	**3**	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.0, 0.2) $MDIFF_2$~N(0, 0.2) $MDIFF_3$~N(1.0, 0.2) $MDIFF_4$~N(2.0, 0.2)	**3**

Multidimensionality, Format Non-Rep., Content Rep., Statistical Non-Rep.

Item Set		Content Area 1	**Item#**	Content Area 2	**Item#**
Unique item set for Form 1	MC	MDISC~LN(0, 0.5) MDIFF~N(0, 1) c~BETA(8, 32)	**15**	MDISC~LN(0, 0.5) MDIFF~N(0, 1) c~BETA(8, 32)	**15**
	CR	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.5, 0.2) $MDIFF_2$~N(-0.5, 0.2) $MDIFF_3$~N(0.5, 0.2) $MDIFF_4$~N(1.5, 0.2)	**3**	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.5, 0.2) $MDIFF_2$~N(-0.5, 0.2) $MDIFF_3$~N(0.5, 0.2) $MDIFF_4$~N(1.5, 0.2)	**3**
Common-item set	MC	MDISC~LN(0, 0.5) MDIFF~N(0.3, 1) c~BETA(8, 32)	**9**	MDISC~LN(0, 0.5) MDIFF~N(0.3, 1) c~BETA(8, 32)	**9**
	CR	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.2, 0.2) $MDIFF_2$~N(-0.2, 0.2) $MDIFF_3$~N(0.8, 0.2) $MDIFF_4$~N(1.8, 0.2)	**0**	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.2, 0.2) $MDIFF_2$~N(-0.2, 0.2) $MDIFF_3$~N(0.8, 0.2) $MDIFF_4$~N(1.8, 0.2)	**0**
Unique item set for Form 2	MC	MDISC~LN(0, 0.5) MDIFF~N(0.5, 1) c~BETA(8, 32)	**15**	MDISC~LN(0, 0.5) MDIFF~N(0.5, 1) c~BETA(8, 32)	**15**
	CR	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.0, 0.2) $MDIFF_2$~N(0, 0.2) $MDIFF_3$~N(1.0, 0.2) $MDIFF_4$~N(2.0, 0.2)	**3**	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.0, 0.2) $MDIFF_2$~N(0, 0.2) $MDIFF_3$~N(1.0, 0.2) $MDIFF_4$~N(2.0, 0.2)	**3**

Multidimensionality, Format Non-Rep., Content Partially Under-Rep., Statistical Rep.

Item Set		Content Area 1	**Item#**	Content Area 2	**Item#**
Unique item set for Form 1	MC	MDISC~LN(0, 0.5) MDIFF~N(0, 1) c~BETA(8, 32)	**12**	MDISC~LN(0, 0.5) MDIFF~N(0, 1) c~BETA(8, 32)	**18**
	CR	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.5, 0.2) $MDIFF_2$~N(-0.5, 0.2) $MDIFF_3$~N(0.5, 0.2) $MDIFF_4$~N(1.5, 0.2)	**3**	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.5, 0.2) $MDIFF_2$~N(-0.5, 0.2) $MDIFF_3$~N(0.5, 0.2) $MDIFF_4$~N(1.5, 0.2)	**3**
Common-item set	MC	MDISC~LN(0, 0.5) MDIFF~N(0, 1) c~BETA(8, 32)	**12**	MDISC~LN(0, 0.5) MDIFF~N(0, 1) c~BETA(8, 32)	**6**
	CR	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.5, 0.2) $MDIFF_2$~N(-0.5, 0.2) $MDIFF_3$~N(0.5, 0.2) $MDIFF_4$~N(1.5, 0.2)	**0**	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.5, 0.2) $MDIFF_2$~N(-0.5, 0.2) $MDIFF_3$~N(0.5, 0.2) $MDIFF_4$~N(1.5, 0.2)	**0**
Unique item set for Form 2	MC	MDISC~LN(0, 0.5) MDIFF~N(0.5, 1) c~BETA(8, 32)	**12**	MDISC~LN(0, 0.5) MDIFF~N(0.5, 1) c~BETA(8, 32)	**18**
	CR	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.0, 0.2) $MDIFF_2$~N(0, 0.2) $MDIFF_3$~N(1.0, 0.2) $MDIFF_4$~N(2.0, 0.2)	**3**	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.0, 0.2) $MDIFF_2$~N(0, 0.2) $MDIFF_3$~N(1.0, 0.2) $MDIFF_4$~N(2.0, 0.2)	**3**

Multidimensionality, Format Non-Rep., Content Partially Under-Rep., Statistical Non-Rep.

Item Set		Content Area 1	**Item#**	Content Area 2	**Item#**
Unique item set for Form 1	MC	MDISC~LN(0, 0.5) MDIFF~N(0, 1) c~BETA(8, 32)	**12**	MDISC~LN(0, 0.5) MDIFF~N(0, 1) c~BETA(8, 32)	**18**
	CR	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.5, 0.2) $MDIFF_2$~N(-0.5, 0.2) $MDIFF_3$~N(0.5, 0.2) $MDIFF_4$~N(1.5, 0.2)	**3**	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.5, 0.2) $MDIFF_2$~N(-0.5, 0.2) $MDIFF_3$~N(0.5, 0.2) $MDIFF_4$~N(1.5, 0.2)	**3**
Common-item set	MC	MDISC~LN(0, 0.5) MDIFF~N(0.3, 1) c~BETA(8, 32)	**12**	MDISC~LN(0, 0.5) MDIFF~N(0.3, 1) c~BETA(8, 32)	**6**
	CR	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.2, 0.2) $MDIFF_2$~N(-0.2, 0.2) $MDIFF_3$~N(0.8, 0.2) $MDIFF_4$~N(1.8, 0.2)	**0**	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.2, 0.2) $MDIFF_2$~N(-0.2, 0.2) $MDIFF_3$~N(0.8, 0.2) $MDIFF_4$~N(1.8, 0.2)	**0**
Unique item set for Form 2	MC	MDISC~LN(0, 0.5) MDIFF~N(0.5, 1) c~BETA(8, 32)	**12**	MDISC~LN(0, 0.5) MDIFF~N(0.5, 1) c~BETA(8, 32)	**18**
	CR	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.0, 0.2) $MDIFF_2$~N(0, 0.2) $MDIFF_3$~N(1.0, 0.2) $MDIFF_4$~N(2.0, 0.2)	**3**	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.0, 0.2) $MDIFF_2$~N(0, 0.2) $MDIFF_3$~N(1.0, 0.2) $MDIFF_4$~N(2.0, 0.2)	**3**

Multidimensionality, Format Non-Rep., Content Completely Under-Rep., Statistical Rep.

Item Set		Content Area 1	**Item#**	Content Area 2	**Item#**
Unique item set for Form 1	MC	MDISC~LN(0, 0.5) MDIFF~N(0, 1) c~BETA(8, 32)	**6**	MDISC~LN(0, 0.5) MDIFF~N(0, 1) c~BETA(8, 32)	**24**
	CR	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.5, 0.2) $MDIFF_2$~N(-0.5, 0.2) $MDIFF_3$~N(0.5, 0.2) $MDIFF_4$~N(1.5, 0.2)	**3**	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.5, 0.2) $MDIFF_2$~N(-0.5, 0.2) $MDIFF_3$~N(0.5, 0.2) $MDIFF_4$~N(1.5, 0.2)	**3**
Common-item set	MC	MDISC~LN(0, 0.5) MDIFF~N(0, 1) c~BETA(8, 32)	**18**	MDISC~LN(0, 0.5) MDIFF~N(0, 1) c~BETA(8, 32)	**0**
	CR	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.5, 0.2) $MDIFF_2$~N(-0.5, 0.2) $MDIFF_3$~N(0.5, 0.2) $MDIFF_4$~N(1.5, 0.2)	**0**	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.5, 0.2) $MDIFF_2$~N(-0.5, 0.2) $MDIFF_3$~N(0.5, 0.2) $MDIFF_4$~N(1.5, 0.2)	**0**
Unique item set for Form 2	MC	MDISC~LN(0, 0.5) MDIFF~N(0.5, 1) c~BETA(8, 32)	**6**	MDISC~LN(0, 0.5) MDIFF~N(0.5, 1) c~BETA(8, 32)	**24**
	CR	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.0, 0.2) $MDIFF_2$~N(0, 0.2) $MDIFF_3$~N(1.0, 0.2) $MDIFF_4$~N(2.0, 0.2)	**3**	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.0, 0.2) $MDIFF_2$~N(0, 0.2) $MDIFF_3$~N(1.0, 0.2) $MDIFF_4$~N(2.0, 0.2)	**3**

Multidimensionality, Format Non-Rep., Content Completely Under-Rep., Statistical Non-Rep.

Item Set		Content Area 1	**Item#**	Content Area 2	**Item#**
Unique item set for Form 1	MC	MDISC~LN(0, 0.5) MDIFF~N(0, 1) c~BETA(8, 32)	**6**	MDISC~LN(0, 0.5) MDIFF~N(0, 1) c~BETA(8, 32)	**24**
	CR	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.5, 0.2) $MDIFF_2$~N(-0.5, 0.2) $MDIFF_3$~N(0.5, 0.2) $MDIFF_4$~N(1.5, 0.2)	**3**	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.5, 0.2) $MDIFF_2$~N(-0.5, 0.2) $MDIFF_3$~N(0.5, 0.2) $MDIFF_4$~N(1.5, 0.2)	**3**
Common-item set	MC	MDISC~LN(0, 0.5) MDIFF~N(0.3, 1) c~BETA(8, 32)	**18**	MDISC~LN(0, 0.5) MDIFF~N(0.3, 1) c~BETA(8, 32)	**0**
	CR	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.2, 0.2) $MDIFF_2$~N(-0.2, 0.2) $MDIFF_3$~N(0.8, 0.2) $MDIFF_4$~N(1.8, 0.2)	**0**	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.2, 0.2) $MDIFF_2$~N(-0.2, 0.2) $MDIFF_3$~N(0.8, 0.2) $MDIFF_4$~N(1.8, 0.2)	**0**
Unique item set for Form 2	MC	MDISC~LN(0, 0.5) MDIFF~N(0.5, 1) c~BETA(8, 32)	**6**	MDISC~LN(0, 0.5) MDIFF~N(0.5, 1) c~BETA(8, 32)	**24**
	CR	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.0, 0.2) $MDIFF_2$~N(0, 0.2) $MDIFF_3$~N(1.0, 0.2) $MDIFF_4$~N(2.0, 0.2)	**3**	MDISC~LN(0, 0.5) $MDIFF_1$~N(-1.0, 0.2) $MDIFF_2$~N(0, 0.2) $MDIFF_3$~N(1.0, 0.2) $MDIFF_4$~N(2.0, 0.2)	**3**

Appendix D

```
>PROBLEM RANDOM, INDIVIDUAL, DATA = 'D:\Simulation\response.dat',
      NITEMS = 90, NGROUPS = 2, NEXAMINEES = 6000, NCHARS = 4;
>TEST ITEMS = (1(1)32), L3;
>TEST ITEMS = (33(1)36), GRADED, NC = (5(0)4);
>TEST ITEMS = (37(1)52), L3;
>TEST ITEMS = (53, 54), GRADED, NC = (5, 5);
>TEST ITEMS = (55(1)86), L3;
>TEST ITEMS = (87(1)90), GRADED, NC = (5(0)4);
>EST NC=3000, IT=10;
>SAVE;
>END ;
5
01234
11111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111
1111111111
22222222222222222222222222222222222222222222222222222222222222222222222222222222222222222222222222222222
2222222222
00000000000000000000000000000003333000000000000000330000000000000000000000000000000000003333
00000000000000000000000000000004444000000000000000440000000000000000000000000000000000004444
00000000000000000000000000000005555000000000000000550000000000000000000000000000000000005555
(1X,4A1,T1,I1,4X,90A1)
```

```
>PROBLEM SCORE, INDIVIDUAL, DATA = 'D:\Simulation\response.dat',
      NITEMS = 90, NGROUPS = 2, NEXAMINEES = 6000, NCHARS = 4;
>TEST ITEMS = (1(1)32), L3;
>TEST ITEMS = (33(1)36), GRADED, NC = (5(0)4);
>TEST ITEMS = (37(1)52), L3;
>TEST ITEMS = (53, 54), GRADED, NC = (5, 5);
>TEST ITEMS = (55(1)86), L3;
>TEST ITEMS = (87(1)90), GRADED, NC = (5(0)4);
>EST NC=3000, IT=10;
>SAVE;
>END ;
5
01234
11111111111111111111111111111111111111111111111111111111111111111111111111111111
1111111111
22222222222222222222222222222222222222222222222222222222222222222222222222222222
2222222222
00000000000000000000000000000033330000000000000003300000000000000000000000000000
0000003333
00000000000000000000000000000004444000000000000000440000000000000000000000000000
0000004444
00000000000000000000000000000005555000000000000000550000000000000000000000000000
0000005555
(1X,4A1,T1,I1,4X,90A1)
```

Appendix E

Three-Way ANOVA (UNI_EQ: BIAS)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Effect Size ω^2
format	1.245	1	1.245	.830	.363	
content	.489	2	.245	.163	.850	
stats	199.325	1	199.325	132.843	.000	.099
format * content	1.087	2	.543	.362	.696	
format * stats	1.260	1	1.260	.840	.360	
content * stats	1.337	2	.669	.446	.641	
format * content * stats	.369	2	.184	.123	.884	
Error	1782.543	1188	1.500			
	1987.656	1199				

Note: Effect Size $\omega^2 = (SS_{effect} - (df_{effect})(MS_{error}))/(MS_{error} + SS_{total})$

Three-Way ANOVA (UNI_EQ: RMSE)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Effect Size ω^2
format	.799	1	.799	3.586	.059	
content	.182	2	.091	.409	.664	
stats	14.081	1	14.081	63.218	.000	.049
format * content	.233	2	.116	.522	.593	
format * stats	.147	1	.147	.659	.417	
content * stats	.021	2	.011	.048	.953	
format * content * stats	.018	2	.009	.040	.961	
Error	264.606	1188	.223			
Corrected Total	280.086	1199				

Three-Way ANOVA (UNI_EQ: CLASSIFICATION CONSISTENCY)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Effect Size ω^2
format	.007	1	.007	16.780	.000	.014
content	.000	2	.000	.509	.601	
stats	.001	1	.001	2.362	.125	
format * content	.000	2	5.53E-005	.137	.872	
format * stats	.000	1	.000	.679	.410	
content * stats	.001	2	.001	1.775	.170	
format * content * stats	.001	2	.000	.802	.449	
Error	.478	1188	.000			
Corrected Total	.488	1199				

Three-Way ANOVA (UNI_NEQ: BIAS)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Effect Size ω^2
format	2.033	1	2.033	1.167	.280	
content	.178	2	.089	.051	.950	
stats	148.198	1	148.198	85.090	.000	.066
format * content	.207	2	.103	.059	.942	
format * stats	3.994	1	3.994	2.293	.130	
content * stats	.058	2	.029	.017	.983	
format * content * stats	.767	2	.384	.220	.802	
Error	2069.099	1188	1.742			
Corrected Total	2224.534	1199				

Three-Way ANOVA (UNI_NEQ: RMSE)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Effect Size ω^2
format	.521	1	.521	.655	.418	
content	.051	2	.025	.032	.969	
stats	55.142	1	55.142	69.332	.000	.054
format * content	.036	2	.018	.022	.978	
format * stats	3.211	1	3.211	4.037	.045	.002
content * stats	.086	2	.043	.054	.947	
format * content * stats	.275	2	.138	.173	.841	
Error	944.865	1188	.795			
Corrected Total	1004.188	1199				

Three-Way ANOVA (UNI_NEQ: CLASSIFICATION CONSISTENCY)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Effect Size ω^2
format	.000	1	.000	.362	.547	
content	.001	2	.000	.277	.758	
stats	.046	1	.046	42.063	.000	.033
format * content	.000	2	8.40E-005	.077	.926	
format * stats	.005	1	.005	4.405	.036	.003
content * stats	.001	2	.000	.397	.672	
format * content * stats	.000	2	.000	.135	.874	
Error	1.301	1188	.001			
Corrected Total	1.354	1199				

Three-Way ANOVA (MUL ρ=0.9/α=10°_EQ: BIAS)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Effect Size ω^2
format	1.735	1	1.735	1.218	.270	
content	.242	2	.121	.085	.919	
stats	181.745	1	181.745	127.597	.000	.096
format * content	.922	2	.461	.324	.723	
format * stats	.108	1	.108	.076	.783	
content * stats	.093	2	.046	.032	.968	
format * content * stats	.545	2	.273	.191	.826	
Error	1692.146	1188	1.424			
Corrected Total	1877.537	1199				

Three-Way ANOVA (MUL ρ=0.9/α=10°_EQ: RMSE)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Effect Size ω^2
format	.314		.314	1.599	.206	
content	.351	2	.176	.895	.409	
stats	12.197	1	12.197	62.139	.000	.049
format * content	.070	2	.035	.179	.836	
format * stats	.314	1	.314	1.598	.206	
content * stats	.299	2	.149	.761	.467	
format * content * stats	.171	2	.086	.436	.647	
Error	233.183	1188	.196			
Corrected Total	246.899	1199				

Three-Way ANOVA (MUL ρ=0.9/α=10°_EQ: CLASSIFICATION CONSISTENCY)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Effect Size ω^2
format	.004	1	.004	9.891	.002	.009
content	.000	2	7.03E-005	.188	.829	
stats	.009	1	.009	23.462	.000	.020
format * content	3.58E-005	2	1.79E-005	.048	.953	
format * stats	.001	1	.001	1.359	.244	
content * stats	.003	2	.002	4.219	.015	.007
format * content * stats	.000	2	.000	.349	.705	
Error	.445	1188	.000			
Corrected Total	.461	1199				

Three-Way ANOVA (MUL ρ=0.9/α=10°_NEQ: BIAS)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Effect Size ω^2
format	3.074	1	3.074	1.966	.161	
content	.649	2	.324	.207	.813	
stats	229.739	1	229.739	146.902	.000	.109
format * content	.052	2	.026	.017	.984	
format * stats	.016	1	.016	.010	.919	
content * stats	.594	2	.297	.190	.827	
format * content * stats	.431	2	.215	.138	.871	
Error	1857.907	1188	1.564			
Corrected Total	2092.461	1199				

Three-Way ANOVA (MUL ρ=0.9/α=10°_NEQ: RMSE)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Effect Size ω^2
format	3.014	1	3.014	3.862	.050	.002
content	.080	2	.040	.051	.950	
stats	108.106	1	108.106	138.536	.000	.103
format * content	.138	2	.069	.088	.915	
format * stats	.032	1	.032	.041	.840	
content * stats	.566	2	.283	.362	.696	
format * content * stats	.136	2	.068	.087	.917	
Error	927.052	1188	.780			
Corrected Total	1039.123	1199				

Three-Way ANOVA (MUL ρ=0.9/α=10°_NEQ: CLASSIFICATION CONSISTENCY)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Effect Size ω^2
format	.006	1	.006	5.013	.025	.003
content	.002	2	.001	.696	.499	
stats	.135	1	.135	110.506	.000	.084
format * content	.001	2	.000	.394	.675	
format * stats	.000	1	.000	.321	.571	
content * stats	.003	2	.002	1.238	.290	
format * content * stats	.001	2	.000	.389	.678	
Error	1.453	1188	.001			
Corrected Total	1.601	1199				

Three-Way ANOVA (MUL ρ=0.9/α=35°_EQ: BIAS)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Effect Size ω^2
format	6.826	1	6.826	5.602	.018	.004
content	.130	2	.065	.053	.948	
stats	141.494	1	141.494	116.122	.000	.088
format * content	.499	2	.249	.205	.815	
format * stats	1.817	1	1.817	1.491	.222	
content * stats	1.231	2	.615	.505	.604	
format * content * stats	1.234	2	.617	.506	.603	
Error	1447.567	1188	1.218			
Corrected Total	1600.798	1199				

Three-Way ANOVA (MUL ρ=0.9/α=35°_EQ: RMSE)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Effect Size ω^2
format	2.54E-005	1	2.54E-005	.000	.990	
content	.038	2	.019	.130	.878	
stats	6.998	1	6.998	47.732	.000	.037
format * content	.431	2	.216	1.470	.230	
format * stats	.400	1	.400	2.727	.099	
content * stats	.929	2	.464	3.168	.042	.003
format * content * stats	.051	2	.026	.175	.840	
Error	174.169	1188	.147			
Corrected Total	183.016	1199				

Three-Way ANOVA (MUL ρ=0.9/α=35°_EQ: CLASSIFICATION CONSISTENCY)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Effect Size ω^2
format	.000	1	.000	.812	.368	
content	.000	2	8.99E-005	.350	.705	
stats	.001	1	.001	3.960	.047	.003
format * content	.000	2	7.28E-005	.283	.753	
format * stats	.000	1	.000	.393	.531	
content * stats	.002	2	.001	3.311	.037	.006
format * content * stats	9.07E-005	2	4.54E-005	.177	.838	
Error	.305	1188	.000			
Corrected Total	.309	1199				

Three-Way ANOVA (MUL ρ=0.9/α=35°_NEQ: BIAS)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Effect Size ω^2
format	100.563	1	100.563	66.057	.000	.047
content	1.186	2	.593	.389	.678	
stats	186.030	1	186.030	122.197	.000	.088
format * content	.098	2	.049	.032	.968	
format * stats	.237	1	.237	.155	.693	
content * stats	1.373	2	.687	.451	.637	
format * content * stats	.139	2	.069	.045	.956	
Error	1808.578	1188	1.522			
Corrected Total	2098.203	1199				

Three-Way ANOVA (MUL ρ=0.9/α=35°_NEQ: RMSE)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Effect Size ω^2
format	38.128	1	38.128	47.749	.000	.035
content	.742	2	.371	.465	.628	
stats	86.819	1	86.819	108.725	.000	.080
format * content	.068	2	.034	.043	.958	
format * stats	.808	1	.808	1.012	.315	
content * stats	1.046	2	.523	.655	.520	
format * content * stats	.028	2	.014	.017	.983	
Error	948.635	1188	.799			
Corrected Total	1076.276	1199				

Three-Way ANOVA (MUL ρ=0.9/α=35°_NEQ: CLASSIFICATION CONSISTENCY)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Effect Size ω^2
format	.029	1	.029	16.800	.000	.012
content	.002	2	.001	.588	.556	
stats	.136	1	.136	78.038	.000	.060
format * content	.001	2	.000	.243	.784	
format * stats	.000	1	.000	.090	.764	
content * stats	.001	2	.001	.333	.717	
format * content * stats	.001	2	.000	.239	.788	
Error	2.078	1188	.002			
Corrected Total	2.249	1199				

Three-Way ANOVA (MUL ρ=0.75/α=10°_EQ: BIAS)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Effect Size ω^2
format	1.464	1	1.464	1.172	.279	
content	2.803	2	1.401	1.122	.326	
stats	170.877	1	170.877	136.875	.000	.102
format * content	.148	2	.074	.059	.942	
format * stats	.070	1	.070	.056	.812	
content * stats	.193	2	.096	.077	.926	
format * content * stats	2.522	2	1.261	1.010	.364	
Error	1483.120	1188	1.248			
Total	3836.652	1200				
Corrected Total	1661.197	1199				

Three-Way ANOVA (MUL ρ=0.75/α=10°_EQ: RMSE)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Effect Size ω^2
format	.005	1	.005	.034	.854	
content	.067	2	.034	.224	.799	
stats	9.689	1	9.689	64.543	.000	.051
format * content	.045	2	.022	.149	.861	
format * stats	.117	1	.117	.782	.377	
content * stats	.008	2	.004	.026	.975	
format * content * stats	.227	2	.114	.757	.469	
Error	178.344	1188	.150			
Corrected Total	188.503	1199				

Three-Way ANOVA (MUL ρ=0.75/α=10°_EQ: CLASSIFICATION CONSISTENCY)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Effect Size ω^2
format	.001	1	.001	2.152	.143	
content	.001	2	.000	1.012	.364	
stats	.003	1	.003	10.038	.002	.007
format * content	.001	2	.000	1.084	.338	
format * stats	.000	1	.000	.826	.364	
content * stats	.002	2	.001	2.766	.063	
format * content * stats	.001	2	.000	.900	.407	
Error	.398	1188	.000			
Corrected Total	.407	1199				

Three-Way ANOVA (MUL ρ=0.75/α=10°_NEQ: BIAS)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Effect Size ω^2
format	.213	1	.213	.140	.708	
content	13.401	2	6.701	4.424	.012	.005
stats	147.571	1	147.571	97.427	.000	.074
format * content	1.220	2	.610	.403	.669	
format * stats	.177	1	.177	.117	.732	
content * stats	.045	2	.022	.015	.985	
format * content * stats	.442	2	.221	.146	.864	
Error	1799.437	1188	1.515			
Corrected Total	1962.506	1199				

Three-Way ANOVA (MUL ρ=0.75/α=10°_NEQ: RMSE)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Effect Size ω^2
format	.527	1	.527	.654	.419	
content	7.479	2	3.740	4.638	.010	.006
stats	70.719	1	70.719	87.702	.000	.067
format * content	1.094	2	.547	.678	.508	
format * stats	.160	1	.160	.199	.656	
content * stats	.012	2	.006	.007	.993	
format * content * stats	.003	2	.001	.002	.998	
Error	957.948	1188	.806			
Corrected Total	1037.943	1199				

Three-Way ANOVA (MUL ρ=0.75/α=10°_NEQ: CLASSIFICATION CONSISTENCY)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Effect Size ω^2
format	.001	1	.001	.932	.335	
content	.006	2	.003	2.416	.090	
stats	.055	1	.055	46.151	.000	.030
format * content	.001	2	.001	.573	.564	
format * stats	8.28E-006	1	8.28E-006	.007	.933	
content * stats	.002	2	.001	.837	.433	
format * content * stats	2.02E-005	2	1.01E-005	.009	.992	
Error	1.405	1188	.001			
Corrected Total	1.470	1199				

Three-Way ANOVA (MUL ρ=0.75/α=35°_EQ: BIAS)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Effect Size ω^2
format	1.801	1	1.801	1.575	.210	
content	2.504	2	1.252	1.095	.335	
stats	156.873	1	156.873	137.201	.000	.102
format * content	.104	2	.052	.046	.955	
format * stats	.000	1	.000	.000	.990	
content * stats	.660	2	.330	.289	.749	
format * content * stats	.040	2	.020	.018	.982	
Error	1358.336	1188	1.143			
Corrected Total	1520.318	1199				

Three-Way ANOVA (MUL ρ=0.75/α=35°_EQ: RMSE)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Effect Size ω^2
format	.012	1	.012	.098	.755	
content	.739	2	.370	3.021	.049	.003
stats	8.241	1	8.241	67.336	.000	.052
format * content	.008	2	.004	.032	.969	
format * stats	.234	1	.234	1.913	.167	
content * stats	.001	2	.000	.004	.996	
format * content * stats	.278	2	.139	1.134	.322	
Error	145.393	1188	.122			
Corrected Total	154.905	1199				

Three-Way ANOVA (MUL ρ=0.75/α=35°_EQ: CLASSIFICATION CONSISTENCY)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Effect Size ω^2
format	.000	1	.000	.510	.475	
content	.000	2	6.00E-005	.259	.772	
stats	.002	1	.002	7.041	.008	.007
format * content	.000	2	.000	.479	.619	
format * stats	.001	1	.001	6.220	.013	.004
content * stats	.000	2	8.66E-005	.375	.687	
format * content * stats	.000	2	5.18E-005	.224	.799	
Error	.275	1188	.000			
Corrected Total	.278	1199				

Three-Way ANOVA (MUL ρ=0.75/α=35°_NEQ: BIAS)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Effect Size ω^2
format	60.764	1	60.764	43.142	.000	.031
content	5.905	2	2.952	2.096	.123	
stats	145.672	1	145.672	103.427	.000	.076
format * content	.345	2	.172	.122	.885	
format * stats	.052	1	.052	.037	.847	
content * stats	.226	2	.113	.080	.923	
format * content * stats	.062	2	.031	.022	.978	
Error	1673.252	1188	1.408			
Corrected Total	1886.278	1199				

Three-Way ANOVA (MUL ρ=0.75/α=35°_NEQ: RMSE)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Effect Size ω^2
format	22.249	1	22.249	29.020	.000	.021
content	3.129	2	1.565	2.041	.130	
stats	73.859	1	73.859	96.335	.000	.072
format * content	.225	2	.113	.147	.863	
format * stats	.016	1	.016	.021	.885	
content * stats	.098	2	.049	.064	.938	
format * content * stats	.012	2	.006	.008	.992	
Error	910.828	1188	.767			
Corrected Total	1010.416	1199				

Three-Way ANOVA (MUL ρ=0.75/α=35°_NEQ: CLASSIFICATION CONSISTENCY)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Effect Size ω^2
format	.010	1	.010	5.298	.022	.003
content	.008	2	.004	2.285	.102	
stats	.146	1	.146	78.730	.000	.061
format * content	.004	2	.002	.998	.369	
format * stats	.001	1	.001	.724	.395	
content * stats	.000	2	.000	.124	.883	
format * content * stats	.001	2	.000	.177	.838	
Error	2.199	1188	.002			
Corrected Total	2.370	1199				

Appendix F

Multiple Comparisons (Tukey HSD)

Dependent Variable	(I) group	(J) group	Mean Difference (I-J)	Std. Error	Sig.
bias_eq	1.00	2.00	.00756	.04903	1.000
		3.00	-.12272	.04903	.090
		4.00	-.08763	.04903	.381
		5.00	-.11501	.04903	.131
	2.00	3.00	-.13027	.04903	.061
		4.00	-.09518	.04903	.296
		5.00	-.12257	.04903	.091
	3.00	4.00	.03509	.04903	.953
		5.00	.00771	.04903	1.000
	4.00	5.00	-.02738	.04903	.981
bias_neq	1.00	2.00	.46970(*)	.05342	.000
		3.00	.79801(*)	.05342	.000
		4.00	.77338(*)	.05342	.000
		5.00	.86006(*)	.05342	.000
	2.00	3.00	.32832(*)	.05342	.000
		4.00	.30368(*)	.05342	.000
		5.00	.39037(*)	.05342	.000
	3.00	4.00	-.02463	.05342	.991
		5.00	.06205	.05342	.773
	4.00	5.00	.08668	.05342	.483
rmse_eq	1.00	2.00	-.10665(*)	.01711	.000
		3.00	-.16401(*)	.01711	.000
		4.00	-.16690(*)	.01711	.000
		5.00	-.20961(*)	.01711	.000
	2.00	3.00	-.05736(*)	.01711	.007
		4.00	-.06025(*)	.01711	.004
		5.00	-.10295(*)	.01711	.000
	3.00	4.00	-.00289	.01711	1.000
		5.00	-.04559	.01711	.060
	4.00	5.00	-.04271	.01711	.092

rmse_neq	1.00	2.00	-.28519(*)	.03790	.000
		3.00	-.53326(*)	.03790	.000
		4.00	-.57246(*)	.03790	.000
		5.00	-.58832(*)	.03790	.000
	2.00	3.00	-.24807(*)	.03790	.000
		4.00	-.28728(*)	.03790	.000
		5.00	-.30314(*)	.03790	.000
	3.00	4.00	-.03920	.03790	.840
		5.00	-.05506	.03790	.593
	4.00	5.00	-.01586	.03790	.994
cons_eq	1.00	2.00	.00939(*)	.00074	.000
		3.00	.01303(*)	.00074	.000
		4.00	.01728(*)	.00074	.000
		5.00	.01518(*)	.00074	.000
	2.00	3.00	.00364(*)	.00074	.000
		4.00	.00788(*)	.00074	.000
		5.00	.00579(*)	.00074	.000
	3.00	4.00	.00424(*)	.00074	.000
		5.00	.00215(*)	.00074	.029
	4.00	5.00	-.00210(*)	.00074	.035
cons_neq	1.00	2.00	.01477(*)	.00159	.000
		3.00	.02395(*)	.00159	.000
		4.00	.03433(*)	.00159	.000
		5.00	.03529(*)	.00159	.000
	2.00	3.00	.00918(*)	.00159	.000
		4.00	.01956(*)	.00159	.000
		5.00	.02051(*)	.00159	.000
	3.00	4.00	.01038(*)	.00159	.000
		5.00	.01134(*)	.00159	.000
	4.00	5.00	.00096	.00159	.974

* The mean difference is significant at the .05 level.

References

Ackerman, T. (1988). *An explanation of differential item functioning from a multidimensional perspective.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Baghi, H., Bent, P., DeLain, M., & Hennings, S. (1995). *A comparison of the results from two equating designs for performance-based student assessments*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

Baker, F. B. (1992). Equating tests under the graded response model. *Applied Psychological Measurement, 16,* 87-96.

Baker, F. B. (1993). Equating tests under the nominal response model. *Applied Psychological Measurement, 17,* 239-251.

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York: Marcel Dekker.

Bastari, B. (2000). *Linking multiple-choice and constructed-response items to a common proficiency scale.* Unpublished doctoral dissertation, University of Massachusetts.

Bennett, R. E. (1993). On the meanings of construed response. In R. E. Bennett, & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement* (pp. 1-27). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bennett, R. E., Rock, D. A., Braun, H. I., Frye, D., Spohrer, J. C., & Soloway, E. (1990). The relationship of expert-system scored constrained free-response items to multiple-choice and open-ended items. *Applied Psychological Measurement, 14,* 151-162.

Bennett, R. E., Rock, D. A., Wang, M-W. (1991). Equivalence of free-response and multiple-choice items. *Journal of Educational Measurement, 28,* 77-92.

Beguin, A. A., Hanson, B. A., & Glas, C. A. W. (2000). Effect of multidimesnioanlity on separate and concurrent estimation in IRT equating. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Beguin, A. A., & Hanson, B. A. (2001). *Effect of noncompensatory multidimensionality on separate and concurrent estimation in IRT observed score equating* (Measurement and Research Department Reports 2001-2). Citogroep, Arnhem, Maart.

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

Burton, R. (2001). Quantifying the effects of chance in multiple-choice and true/false tests: Question selection and guessing of answers. *Assessment and Evaluation in Higher Education, 26,* 41-50.

Cao, Y., Yin, P., & Gao, X. (2007). *Comparison of IRT and Classical Equating Methods for Tests Consisting of Polytomously-Scored Items*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

Childs, R. A., & Oppler, S. H. (2000). Implications of test dimensionality for unidimensional IRT scoring: An investigation of a high-stakes testing program. *Educational and Psychological Measurement, 60,* 939-955.

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

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (Revised ed.). Hillsdale, NJ: Erlbaum.

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.

Dodd, B. G. (1984). Attitude scaling: A comparison of the graded response and partial credit latent trait models. *Dissertation Abstracts International, 45*, 2074A.

Doody-Bogan, E., & Yen, W. (1983). *Detecting multidimensionality and examining its effect on vertical equating with the three parameter logistic model.* Paper presented at the annual meeting of the American educational Association, Montreal, Canada.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Ercikan, K., Schwarz, R. D., Julian, M. W., Burket, G. R., Weber, M. M., & Link, V. (1998). Calibration and scoring of tests with multiple-choice and constructed-response item types. *Journal of Educational Measurement, 35*, 137-154.

Finch, H. (2006). Comparison of the performance of varimax and promax rotations: factor structure recovery for dichotomous items. *Journal of Educational Measurement, 43*, 39-52.

Fitzpatrick, A. R. (2008). *The impact of anchor test configuration on students' proficiency classifications*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York, NY.

Fitzpatrick, A. R., Ercikan, K., Yen, W. M., & Ferrara, S. (1998). The consistency between raters scoring in different test years. *Applied Measurement in Education, 11*, 195-208.

Fitzpatrick, A. R., & Yen, W. M. (2001). The effects of test length and sample size on the reliability and equating of tests composed of constructed-response items. *Applied Measurement in Education, 14*, 31-57.

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

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

Haladyna, T. M. (1994). *Developing and validating multiple-choice items*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Haladyna, T. M. (1997). *Writing test items to evaluate higher order thinking*. Boston: Allyn & Bacon.

Hambleton, R. K. (1994). Item response theory: A broad psychometric framework for measurement advances. *Psicothema, 3*, 535-556.

Hambleton, R. K., & Cook, L. L. (1983). Robustness of item response models and effects of test length and sample size on the precision of ability estimates. In D. J. Weiss (Ed.), *New horizons in testing*. New York: Academic.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Norwell, MA: Kluwer Academic Publishers.

Hamilton, L., Nussbaum, E., & Snow, R. (1997). Interview procedures for validating science assessments. *Applied Measurement in Education, 10*, 181-200.

Hanick, P. L., & Huang, C-Y. (2002). *Effects of decreasing the number of common items in equating link item sets*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

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.

Harris, D. J. (1991). Equating with non-representative common item sets and non-equivalent groups. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

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

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 187-220). Westport, CT: Praeger Publishers.

Kamata, A., & Tate, R. (2005). The performance of a method for the long-term equating of mixed-format assessment. *Journal of Educational Measurement, 42,* 193-213.

Kim, S. H., & Cohen, A. S. (1995). A minimum χ^2 method for equating tests under the graded response model. *Applied Psychological Measurement, 19,* 167-176.

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

Kim, S. H., & Cohen, A. S. (2002). A comparison of linking and concurrent calibration under the graded response model. *Applied Psychological Measurement, 26,* 25-41.

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

Kim, S.-H., & Lee, W.-C. (2004). *IRT scale linking methods for mixed-format tests* (ACT Research Report Series 2004-5). Iowa City: ACT.

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

Kirkpatrick, R. K. (2005). *The effects of item format in common item equating.* Unpublished doctoral dissertation, University of Iowa.

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

Kolen, M. J. (1999-2000). Threats to score comparability with applications to performance assessments and computerized adaptive tests. *Educational Assessment, 6*, 73-96.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating: Methods and practices* (2nd ed.). New York, NY: Springer-Verlag.

Lane, S. (2005). *Status and future directions for performance assessments in education*. Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Canada.

Li, Y.-H., & Lissitz, R. W. (2000). An evaluation of the accuracy of multidimensional IRT linking. *Applied Psychological Measurement, 24*, 115-138.

Li, Y.-H., Lissitz, R. W., & Yang, Y.-N. (1999). *Estimating IRT equating coefficients for tests with polytomously and dichotomously scored items.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Montreal, Canada.

Linacre, J. M. (1988). *Many-faceted Rash measurement*. Chicago, IL: MESA Press.

Linn, R. L. (1995). High-stakes uses of performance-based assessments. Rationale, examples, and problems of comparability. In T. Oakland & R. K. Hambleton (Eds.), *International perspectives on academic assessment* (pp. 49-73). Norwell, MA: Kluwer Academic Publishers..

Livingston, S. A. (1994). *Equating constructed-response tests through a multiple-choice anchor: A small-scale empirical study.* Unpublished statistical report.

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

Lukhele, R., Thissen, D., & Wainer, H. (1994). On the relative value of multiple-choice, constructed response, and examinee-selected items on two achievement tests. *Journal of Educational Measurement, 31*, 234-250.

Manhart, J. J. (1996). *Factor analytic methods for determining whether multiple-choice and constructed-response tests measure the same construct.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York, NY.

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

Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist, 34*, 207-218.

Maydeu-Olivares, A., Drasgow, F., & Mead, A. D. (1994). Distinguishing among parametric item response models for polytomous ordered data. *Applied Psychological Measurement, 18*, 245-256.

Messick, S. (1995). Validity of psychological assessment. *American Psychologist, 50,* 741-749.

Min, K.-S. (2003). *The impact of scale dilation on the quality of the linking of multidimensional item response theory calibrations.* Unpublished doctoral dissertation, Michigan State University.

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

Oosterhof, A. (1996). *Developing and using classroom assessments*. New Jersey: Prentice Hall.

Perkhounkova, Y., & Dunbar, S. B. (1999). *Influences of item content and format on the dimensionality of tests combining multiple-choice and open-response items: An application of the Poly-DIMTEST procedure.* Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Quebec, Canada.

Petersen, N. S., Marco, G. L., & Stewart, E. E. (1982). A test of the adequacy of linear score equating models. In P. W. Holland and D. B. Rubin (Eds.) *Test equating*. New York: Academic Press, 71-135.

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

Rosa, K., Swygert, K., Nelson, L., & Thissen, D. (2001). Item response theory applied to combinations of multiple-choice and construed-response items – Scale scores for patterns of summed scores. In D. Thissen & H. Wainer (Eds.). *Test scoring* (pp.253-292). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Roussos, L., Stout, W., & Marden, J. (1998). Using new proximity measures with hierarchical cluster analysis to detect multidimensionality. *Journal of Educational Measurement, 35,* 1-30.

Snow, R. E. (1993). Construct validity and construed response tests. In R. E. Bennett, & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement* (pp. 45-60). Hillsdale, NJ: Lawrence Erlbaum Associates.

Spence, P. (1996). *The effect of multidimensionality on unidimensional equating with item response theory*. Unpublished doctoral dissertation, University of Florida, FL.

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

Sykes, R. C., Hou, L-L., Hanson, B., & Wang, Z. (2002). *Multidimensionality and the equating of a mixed-format math examination.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Sykes, R. C., & Yen, W. M. (2000). The scaling of mixed-item-format tests with the one-parameter and two-parameter partial credit models. *Journal of Educational Measurement, 37,* 221-244.

Tang, K. L., & Eignor, D. R. (1997). *Concurrent calibration of dichotomously and polytomously scored TOEFL items using IRT models* (TOEFL Technical Report 13). Princeton, NJ: Education Testing Service.

Tate, R. (1999). A cautionary note on IRT-based linking of tests with polytomous items. *Journal of Educational Measurement, 36,* 336-346.

Tate, R. (2000). Performance of a proposed method for the linking of mixed format tests with constructed response and multiple choice items. *Journal of Educational Measurement, 37*, 329-346.

Tate, R. (2003). Equating for long-term scale maintenance of mixed format tests containing multiple choice and constructed response items. *Educational and Psychological Measurement, 63,* 893-914.

Thissen, D. (1991). *MULTILOG: Item analysis and scoring with multiple category response models*. Chicago: International Educational Services.

Thissen, D., Wainer, H., & Wang, X-B. (1994). Are tests comprising both multiple-choice and free-response items necessarily less unidimensional than multiple-choice tests? An analysis of two tests. *Journal of Educational Measurement, 31,* 113-123.

Traub, R. E. (1993). On the equivalence of the traits assessed by multiple-choice and constructed-response tests. In R. E. Bennett, & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement* (pp. 29-44). Hillsdale, NJ: Lawrence Erlbaum Associates.

US Department of Education (2004). *Standards and assessments peer review guidance: Information and examples for meeting the requirements of the No Child Left Behind Act of 2001*. Washington, DC: Office of Elementary and Secondary Edcuation.

van der Linden, W. J., & Hambleton, R. K. (1996). *Handbook of modern item response theory*. New York: Springer.

Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: toward a marxist theory of test construction. *Applied Measurement in Education, 6,* 103-118.

Wainer, H., Wang, X-B., & Thissen, D. (1994). How well can we compare scores on test forms that are constructed by examinees' choice? *Journal of Educational Measurement, 31,* 183-199.

Way, W. D., Ansley, T. N., & Forsyth, R. A. (1988). The comparative effects of compensatory and noncompensatory two-dimensional data on unidimensional IRT estimates. *Applied Psychological Measurement, 12,* 239-252.

Wu, M., & Adams, R. (2006). Modelling mathematics problem solving item responses using a multidimensional IRT model. *Mathematics Education Research Journal, 18,* 93-113.

Yang, W-L. (2000). *The effects of content homogeneity and equating method on the accuracy of common-item test equating*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Yon, H. (2006). *Multidimensional item response theory (MIRT) approaches to vertical scaling*. Unpublished doctoral dissertation, Michigan State University.