This is to certify that the

dissertation entitled

THE EFFECTS OF CONTENT HOMOGENEITY AND EQUATING METHOD ON THE ACCURACY OF COMMON-ITEM TEST EQUATING

presented by

Wen-Ling Yang

has been accepted towards fulfillment
of the requirements for

Ph.D. degree in Counseling, Educational Psychology, and Special Education

Richard M. Jaeger
Major professor

Date Dec. 29, 1997

DATE DUE	DATE DUE	DATE DUE
MAR 0 3 2003		

THE EFFECTS OF CONTENT HOMOGENEITY AND EQUATING METHOD ON THE ACCURACY OF COMMON-ITEM TEST EQUATING

By

Wen-Ling Yang

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Counseling, Educational Psychology, and Special Education

1997

ABSTRACT

THE EFFECTS OF CONTENT HOMOGENEITY AND EQUATING METHOD ON THE ACCURACY OF COMMON-ITEM TEST EQUATING

By

Wen-Ling Yang

Often in educational testing and measurement, people use alternate test forms to achieve comparable test scores for measuring growth or ensuring test security. To obtain valid comparisons between groups and to enhance test fairness, they rely on various equating techniques to equate forms of the same test. It is important to evaluate the adequacy of these equating methods and the accuracy of their outcomes. In my dissertation, I studied the effects of test characteristics on the accuracy of equating outcomes when different methods were used to equate two test forms of a test. Specifically, I wanted to know whether equating accuracy improves with a test made of content-homogeneous items, whether it improves with an anchor test that is content-representative of its total test, and whether such content effects depend on the particular equating method used for equating. My major goal is to improve test results, which often lead to critical educational decisions.

The data I analyzed is the test results from a professional in-training examination. It has a negatively skewed score distribution because the test was written for a minimum-competency examination. In equating practice, such test outcome receives less attention

than it should have. The common-item equating design was used because the two groups of examinees taking different forms were not randomly formed or assigned. I used an item-sampling design to create four tests that differ in the content homogeneity of their items and the content representativeness of their anchor items. All the items in these tests are from one overall content domain, but fall into 23 different content areas. Each of the four tests has two forms, and a set of common anchor items is embedded in each form. I applied linear, equipercentile, and two IRT-based equating methods to equate the two forms of each test. By means of the item-sampling designs, I was able to establish two innovative criteria based on true score for evaluating the accuracy of equating outcomes from these methods. I also used two other criteria based on the outcomes of arbitrary equatings to examine how well equating accuracy is estimated with such criteria. I also elaborated on issues of construct validity and test dimensionality, which are relevant to test equating.

Overall, I found that all the equating methods yielded accurate results to a moderate degree. They all produced more accurate results when the anchor items were more representative of the total test, or when the items in a test had homogeneous content. Therefore, to improve equating accuracy, I recommend an inclusion of anchor items that fully reflect the overall test content. I also found that the IRT-based equating outcomes were more accurate than the outcomes from the other equating methods. However, the differences are small thus may not have practical significance. If the degree of equating accuracy is critical for decision-makings of a testing program, such as high-stake examinations, IRT-based equating methods are recommended.

To my loving parents, for their great expectations in their children's education.

To my many dear friends for their faithful friendships, which comforted me on my long way to reach this goal, far away from home.

ACKNOWLEDGMENTS

For my dissertation director, Dr. Richard T. Houang, I thank him for all his assistance in generating research ideas, acquiring test data, critiquing my study design and method, and reviewing my analysis results. Most of all, I would like to thank him for being a great mentor to me and making me believe in myself.

Dr. Irvin Lehmann is my long-time academic advisor, who always listens to my problems and questions. I enjoyed being his advisee because he is such an understanding professor, who always has good solutions or useful suggestions for me. I owe special thanks to him for his patience and superb academic advice and counseling.

I am grateful to Dr. Betsy J. Becker for her encouragement and her thorough review of an earlier draft of my dissertation. Her timely and challenging feedback on my studies enlightened my view of conducting in-depth and precise research.

I appreciate the efforts of Dr. Mary Ann Reinhart and Dr. Maria T. Tatto in making time from their busy schedules to attend the proposal and defense meetings for my dissertation. Their valuable comments on practical equating and related research issues are also appreciated.

Lastly, I would like to acknowledge the help from Dr. Michael J. Kolen. He not only provided an extended version of computer program to assist me with equipercentile equating, but also answered, via e-mail, many of my questions concerning practical equating issues.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PART I: COMMON-ITEM EQUATING ISSUES -- AN OVERVIEW

INTRODUCTION

In educational testing, to ensure test security, alternate test forms are often used so that all examinees do not need to take the same test at the same time. The need for interchangeable test forms is especially important for licensure examinations and other tests used to make critical decisions. Comparable test forms are also used to measure growth or trends. In theory, alternate test forms are created by careful test construction such that their items will have similar item difficulties. However, results of test construction are often not satisfactory because test forms are seldom parallel in the straight theoretical sense. A practical strategy to achieve comparable test scores is to establish equivalency between different forms via equating.

Equating procedures are based on the idea of making statistical adjustments in pursuit of four conditions: same construct, equity, symmetry, and population invariance (Hambleton & Swaminathan, 1990; Lord, 1980). By satisfying these conditions, in theory, one can obtain comparable test scores from equated test forms. Equating can be linear or non-linear depending on how test scores are transformed across forms. Various equating models vary substantially in their assumptions and equating functions. Selection of an equating model should take into account purpose of equating, theoretical plausibility and applicability of a model, and characteristics of the examinees and test forms being equated.

Conventional linear equating methods, such as Tucker linear method, are straightforward and convenient in computation. Therefore, they have been used widely in

practice for years. Nevertheless, their results do not always meet all criteria for equivalent tests. For example, equivalent scores from linear equating can vary across examinee groups and item samples. To overcome such drawbacks of linear equating, different equating models based on item response theory (IRT) have been developed. The use of the IRT-based methods has recently increased in its popularity. IRT-based equating methods are especially useful for equating based on the common-item design, where random assignment of examinees is not required. They are often used when the assumptions made by linear equating are not likely to hold (Cook & Eignor, 1991; Crocker & Algina, 1986).

Research has shown that IRT-based methods are more robust than linear equating, and they will lead to greater stability when tests to be equated differ somewhat in content and length (Petersen, Cook, & Stocking, 1983). Despite their theoretical appeal and empirical advantages, IRT-based equating methods are often under scrutiny because of the inconsistency in their equating outcomes. Another issue is the possible IRT-based equating method by test interaction (Hills, Subhiyah, & Hirsch, 1988; Peterson, Cook, & Stocking, 1983). In equating practice, there are also concerns about its cost and the practical significance of improvement in equating accuracy due to IRT-based methods.

Dimensionality of a test is an issue relevant to equating accuracy. It is particularly critical for IRT-based equating that assumes an unidimensional trait (Hambleton & Swaminathan, 1990). An IRT-based equating model assuming unidimensionality is not likely to work well on a test of multidimensionality. Test dimensionality may also affect the effectiveness of conventional equating methods. This is because the conventional approach also assumes unidimensionality but in an implicit way (Green, Yen, & Burket,

1989). Often, a broad test domain is defined to encompass a variety of knowledge or skills. It is less likely that only one trait or one single dominant trait influences the examinee performance on the test. To ensure equating accuracy, it is therefore crucial to check the IRT assumption of unidimensionality. It is also important to evaluate the robustness of IRT applications when the assumption is violated.

This study addresses practical common-item equating issues concerning equating methods and test characteristics. Four pairs of sampled test forms, varying in their content homogeneity, are equated by the Tucker linear method, frequency-estimation equipercentile method, and two IRT-based equating methods. Various equating results from these methods are evaluated using four types of criteria for evaluating equating accuracy. Resulting equating outcomes are compared and discussed, with considerations of restrictions on this study. Suggestions are made for equating practice and future research.

The major goal of this study is to inform testing practice, leading to improved measures of ability or achievement and more valid comparisons of different groups. The study on the effect of content homogeneity and content representativeness on equating accuracy should improve the precision of equivalent scores, test construction, and test efficiency in the context of common-item equating. The findings and conclusions reached in this study will provide sound groundwork for future studies. The unique part of the research design, such as the item-sampling design and the use of multiple innovative criteria for evaluating equating accuracy, should cast insights on improving the estimation of equating accuracy for future studies.

Chapter 1

RESEARCH PURPOSES AND QUESTIONS

To better understand the effectiveness of various equating methods, and the influences on equating accuracy from the characteristics of a test and its items, this study has these specific goals:

- To estimate and compare the equating accuracy of linear equating, equipercentile equating and IRT-based equating.
- To investigate the effects of content homogeneity of test items, and content representativeness of anchor items, on the estimation of equating accuracy yielded by various equating methods.
- To assess the effectiveness of various criteria for evaluating equating accuracy.
- To address dimensionality issues related to the test data, and to investigate their influences on the IRT-based equating results, such as the robustness of a unidimensional IRT model.
- To inform testing practice and future studies about useful ways for (1) improving anchor-item equating design, (2) selecting an equating method, and (3) evaluating equating accuracy.

Pursuing these goals, this study outlined a set of research questions to direct the design, method, and analysis of its equating research. These questions not only reflect specific research interests but also address important issues and concerns about equating practice.

- To what extent does the results of Tucker linear equating, equipercentile equating, and the IRT-based equating agree?
- Does equating result depend on content homogeneity of test items and content representativeness of anchor items? In other words, does the accuracy of equating improve when the items in a test are content homogeneous? Does it improve when the content of anchor items becomes more representative of the total test?
- How accurate are the results of various equating methods, based on these criteria for evaluating equating accuracy: (a) a raw-score-based true-score estimate, (b) an IRT-based true-score estimate, (c) the result of the equipercentile equating method on equating the two forms of a longer and, in theory, more reliable test, and (d) the result of the equipercentile equating method on equating the two forms of a shorter subtest, sampled from the longer original in-training test?
- Which criterion, among various criteria for evaluating equating accuracy, is relatively better than the other criteria, for this particular minimum competency examination?
- Does the IRT assumption of unidimensionality hold for the IRT-based equatings? Mathematically or conceptually speaking, how can we describe the dimensionality of the test?
- Will the resulting outcomes of the IRT-based equatings suggest that the three parameter logistic (3PL) IRT model is appropriate for a minimum competency test that has a negatively skewed score distribution? Is the IRT model sound in theory?

In the following chapters, literature for relevant equating issues are first reviewed and summarized. They include conditions of equivalency, equating guidelines, assumptions and procedures of various equating methods, features of the common anchor-item equating, and estimation of equating accuracy. Then, the data, item sampling schemes, common-item non-equivalent group design, and particular equating methods used in this dissertation are described. After the results of various equating methods on different tests are discussed, suggestions are made for the equating practice and future research.

PART II: REVIEW OF LITERATURE

Chapter 2

CONDITIONS AND GENERAL GUIDELINES FOR EQUATING

Important requirements of equating, including the conditions of equivalency, general guidelines for conducting equating studies, and the criteria for selecting equating methods, are reviewed in this chapter. The limitations of equating are also discussed.

Conditions of Equivalency

If test Y is to be equated to test X, no matter what equating procedure is chosen, the following conditions must be satisfied to conclude that scores on test X and test Y are made equivalent (Angoff, 1984; Dorans, 1990; Hambleton & Swaminathan, 1990; Kolen & Brennan, 1995; Lord, 1980; Petersen, Kolen, & Hoover, 1989):

- Both tests measure the same construct.
- The equating achieves equity. That is, for individuals of identical proficiency, the conditional frequency distributions of scores on the two tests are the same.
- The equating transformation is symmetric; that is, the equating of Y to X is the inverse of the equating of X to Y.
- The equating transformation is invariant across sub-groups of the population, from which it is derived.

In addition to the above conditions, equating using the IRT model also requires the assumption of unidimensionality. These conditions of equivalency are elaborated below.

Same Construct

The requirement of the same construct is a matter of test construction and can be achieved by carefully selecting items that measure the same construct. Formulas relating tests of different constructs to each other can be computed for the purpose of regression or prediction, but it is meaningless to compare tests measuring different constructs. Since equating is a matter of transforming scores for the sake of comparison, it makes no sense for the forms of a test to measure different constructs.

Equity

The condition of equity implies that individuals of the same proficiency obtain the same score, no matter which tests are taken. The proficiencies of individuals taking two different tests are usually estimated via their performance on the common items or an anchor test. At every ability level, the conditional frequency distributions on different forms should be the same. The corresponding percentile ranks in any given group should be equal for equivalent scores.

Symmetry

The score transformation should be invertible to achieve symmetry. Regardless of equating from X to Y or from Y to X, the same score on one test should correspond to one given score on another test.

The condition of symmetry requires that the function (e_X) used to equate a score (y) on Form Y to the scale of Form X be the inverse of the function (e_Y) used to equate a score (x) on Form X to the Form Y scale: $e_X(y) = e_Y^{-1}(y)$ and $e_Y(x) = e_X^{-1}(x)$.

Population Invariance

Equating results are desired to be independent of the unique characteristics of the examinee samples used in the equating process. No matter which groups of examinees are used, the equating results should not change with the characteristics of the particular examinee groups. The results should depend on only the underlying construct measured by the test. Among various equating models, the procedures based on regression inherently fail to achieve this condition, while IRT-based equating is expected to result in population invariance by assigning the same estimated ability score to all the examinees at the same ability level.

The condition of population invariance is one of the ultimate goals of test equating, and can be assessed by examining the equivalency of the test forms across sub-groups. If population invariance is not achieved, one possible reason is that the tests or test forms may not measure the same construct. In this case, the procedures of test construction and the resulting test items should remain under scrutiny.

Unidimensionality for IRT-Based Equating

Unidimensionality is an underlying assumption for the equating based on item response theory, although it is not explicitly recognized as a condition of equating. The IRT-based equating is more restrictive because it requires unidimensional test items.

Equating Guidelines

When test data from different forms of a test are very similar or very different, equating may not be desired. Other than reducing errors, inappropriate equating may introduce more error to test scores, and unnecessary equating will increase the cost of testing. Once it is determined that equating is preferred, factors such as feasibility, cost, and any unique testing context should all be considered to carry out the equating. However, there are no absolutely superior criteria for the selection of equating design or

method (Harris & Crouse, 1993). As a result, arbitrary judgments and decisions are necessary and should be based on equating expertise and experience.

Brennan and Kolen (1987, 1995) proposed a set of rules to guide test equating. They argued that the test content and statistical specifications for tests being equated ought to be defined precisely and be stable over time. In the process of test construction, item statistics should be obtained from pre-testing or a previous use of the test. Each test should be reasonably long, with at least 35 items, and the scoring keys should be consistent. The stems for common items, alternatives, and stimulus materials should be identical for the forms to be equated. The characteristics of examinee groups should be stable over time, too. The groups should be relatively large, larger than roughly 400 examinees. The curriculum, training materials, and field of study should also be stable. The test items should be administered and secured under standardized conditions.

Brennan and Kolen (1987) also have a set of ideal suggestions for test equating:

- Embed two sets of common items in the full-length test;
- Length of an anchor test should be at least 1/5 of the total-test length, and the anchor test should mirror the total test in content specification and statistical characteristics;
- Administer at least one link form no earlier than one year in the past, and administer at least one link form in the same month as the form to be equated; and
- Place each common item in approximately the same position in the two forms.

Criteria for Selecting Equating Methods

Usually, an equating method is selected or tailored to accommodate a particular testing situation. If guessing is explicitly encouraged during testing time and its effect cannot be overlooked, a fair equating should account for this factor. Suppose accurate equivalent scores are strongly desired by some testing programs, it is critical to select an equating method that yields the most accurate equating for that particular test.

Three major aspects to be considered for the selection of an equating method are: (1) Are the required underlying assumptions tenable? (2) Is the equating procedure practical? and, (3) How good is the equating result? (Crocker & Algina, 1986) Common equating methods are compared on each of these three aspects below.

Tenability of Model Assumptions

The premise of model application is that all the underlying assumptions of the selected model hold. Linear equating assumes that the score distributions of the tests being equated have identical shapes, and is appropriate for equating use when score distributions only differ in the means and/or standard deviations. Due to this assumption, the derived equivalent scores will have identical percentile ranks.

Equipercentile equating requires fewer assumptions than linear equating. It does not assume the same shapes for score distributions, but determines which scores on the different tests will have the same percentile rank (Crocker & Algina, 1986). However, in theory, the equipercentile method is associated with larger errors than linear equating (Lord, 1982a). Also, it is less practical to apply equipercentile equating because it is far more complicated.

Both linear and equipercentile equatings assume that the tests being equated measure the same trait and have equal reliability. Given two tests that have different average difficulty, however, the assumption of equal reliability usually does not hold. In such case, these two equating methods are likely to yield erroneous results. The results of the two methods also depend on the particular test items used, and fail to meet the condition of equity for equating. Furthermore, the methods do not meet the requirement

of population invariance (Hambleton & Swaminathan, 1990). Unlike these methods, IRT-based equating does not have the same drawbacks and could be a better alternative.

Applicability of Design and Method

Random group design, single group design with counter-balancing, and the common-item nonequivalent groups design are three common designs used to collect data before equating (Kolen & Brennan, 1995). Equating designs differ in the need for randomly formed examinee groups, single or multiple test administrations, test length, or the examinee sample size needed. Depending on various conditions in real testing situations, feasible equating designs are chosen and corresponding methods are used to equate the test results of different forms. For example, traditional equating may be adequate if examinees are randomly assigned to form groups and each group takes a different test form, or if different forms are assigned to examinees randomly, or if the groups take both test forms in randomly assigned orders. Otherwise, IRT-based methods are more appropriate.

Random group design is often desired because each examinee only has to take one form and several forms can be equated at the same time. Nevertheless, this approach requires the test forms to be available and administered at the same time, which is sometimes not practical. One solution to this problem is the use of the anchor design. Either multiple test forms with embedded anchor items (the internal anchor) can be given to different examinee groups, or a third test (the external anchor) can be given to two examinee groups that take different test forms. Without random assignment, the anchor-score distributions for different sub-populations may be markedly different and the

assumption of equity is unlikely to hold (Crocker & Algina, 1986). In this case, the linear and equipercentile methods are likely to yield inaccurate results, while IRT-based equating tends to be more adequate.

Equating Accuracy

One major justification for test equating is its effectiveness, that is, the extent to which the equating method used yields adequately equivalent scores. Nevertheless, because true scores can never be known in practice, perfect equivalency can never be determined in a strict sense. As a consequence, there is no best criterion for evaluating equating accuracy, and there is also no definite procedure for determining which equating methods should be used in a given context (Harris & Crouse, 1993). The interest for assessing equating accuracy thus is to find adequate equating methods that are appropriate for a given context. Issues regarding the assessment of equating accuracy are discussed further in Chapter 7.

Limitations of Equating

Test equating cannot solve problems originating through crude and improper test construction. It is meant to overcome the insufficiency of good test-construction practice that has failed to yield test forms of the same difficulty level.

Both traditional equating and IRT-based equating are primarily designed to remedy minor differences in difficulty between test forms. Cook and Eignor (1991) indicated that no equating method could satisfactorily equate tests that were markedly different in difficulty, reliability or test content. This perspective raises doubt over the practicality of

vertical equating, which transforms scores across levels of achievement onto a single scale. Theoretically and operationally, vertical equating is much more difficult to accomplish than horizontal equating. In addition, vertical equating often results in a lack of test invariance, which can be accounted for by multidimensionality (Skaggs & Lissitz, 1988).

Equal reliability is usually assumed by equating models, such as the linear equating and the equipercentile equating. Due to floor and ceiling effects, however, tests that differ in difficulty are not likely to be equally reliable for all sub-groups of examinees, and the relationship between the tests can be nonlinear (Skaggs & Lissitz, 1986). It is implied that observed scores on tests of different difficulty cannot be equated. Therefore, in such cases, equating is actually carried out in a loose sense.

From a pragmatic point of view, however, equating aims to arrive at a conversion equation that approximates an ideal equivalency. Therefore, despite its limitations by nature, test equating is of great use in comparing scores on test forms with minor differences.

Chapter 3

TUCKER LINEAR EQUATING

Linear equating has the appeal of simplicity in terms of score transformation and is used most often with the common-item design (Kolen & Brennan, 1987). Among the many linear methods, Tucker linear equating is one of the methods employed most frequently.

Synthetic Population

For the common-item design, the Tucker method involves the use of a synthetic population (Braun & Holland, 1982). A synthetic population is usually defined as a combination of the proportionally weighted (proportional to sample sizes) populations of examinees taking different test forms. Typically, an equating function is viewed as being defined for a single population, and the two examinee populations must be combined as one single population for defining an equating relationship (Kolen & Brennan, 1987).

Model Assumptions

In an anchor-item equating design, suppose examinees in Population 1 take Form X, those in Population 2 take Form Y, and V is the embedded set of anchor items in both

forms. To equate scores on Form X to the scale of Form Y, Tucker linear equating requires strong statistical assumptions, as follows (Kolen & Brennan, 1987; Kolen & Brennan, 1995):

1. The linear regression function (slope and intercept) for the regression of X on V is the same for Populations 1 and 2. The function for the regression of Y on V is also the same for the two populations.

2. The variance of X given V is the same for the two populations, and the variance of Y given V is also the same for the two populations.

Under the above assumptions, the linearly transformed scores on one form, yielded by Tucker's method, will have the same mean and standard deviation as the scores on another form. Because of the assumptions about the variances and regression functions in relation to the two populations, Tucker linear equating is more accurate when examinee groups are similar.

Equating Procedures

Using the proportional weights to form a synthetic population, Tucker linear equating basically involves the following concepts and procedures (Kolen & Brennan, 1987; Kolen & Brennan, 1995):

1. Find the weights for Populations 1 and 2 by using these formula: $w_1=n_1/(n_1+n_2)$ and $w_2=n_2/(n_1+n_2)$, where n_1 and n_2 are the sample sizes of examinees from populations 1 and 2, respectively.

2. Let α_1 and α_2 be the regression slopes for the populations. For Population 1,

$$\alpha_1(X|V)=\sigma_1(X,V)/\sigma_1^2(V) \text{ and } \alpha_1(Y|V)=\sigma_1(Y,V)/\sigma_1^2(V), \quad (3.1)$$

and for population 2,

$$\alpha_2(X|V)=\sigma_2(X,V)/\sigma_2^2(V) \text{ and } \alpha_2(Y|V)=\sigma_2(Y,V)/\sigma_2^2(V). \quad (3.2)$$

In addition, let β_1 and β_2 be the regression intercepts for the two populations, and μ_1 and μ_2 be the population means, then

$$\beta_1(X|V)=\mu_1(X)-\alpha_1(X|V)\mu_1(V) \text{ and } \beta_1(Y|V)=\mu_1(Y)-\alpha_1(Y|V)\mu_1(V), \quad (3.3)$$

and

$$\beta_2(X|V)=\mu_2(X)-\alpha_2(X|V)\mu_2(V) \text{ and } \beta_2(Y|V)=\mu_2(X)-\alpha_2(Y|V)\mu_2(V). \quad (3.4)$$

To compute the $\hat{\alpha}_1(X|V)$ and $\hat{\alpha}_2(Y|V)$, observed data could be plugged in to the above equations.

3. By assumptions about the slopes and intercepts for the two populations, $\alpha_1(X|V)=\alpha_2(X|V)$, $\alpha_1(Y|V)=\alpha_2(Y|V)$, $\beta_1(X|V)=\beta_2(X|V)$, and $\beta_1(Y|V)=\beta_2(Y|V)$. And, by assumptions about the same variances for the two populations,

$$\sigma_1^2(X)[1-\rho_1^2(X,V)]=\sigma_2^2(X)[1-\rho_2^2(X,V)], \quad (3.5)$$

and

$$\sigma_1^2(Y)[1-\rho_1^2(Y,V)]=\sigma_2^2(Y)[1-\rho_2^2(Y,V)]. \quad (3.6)$$

4. With the above assumptions, it can be demonstrated that

$$\mu_1(Y)=\mu_2(Y)+\alpha_2(Y|V)[\mu_1(V)-\mu_2(V)],\ \mu_2(X)=\mu_1(X)-\alpha_1(X|V)[\mu_1(V)-\mu_2(V)], \quad (3.7)$$

$$\sigma_1^2(Y)=\sigma_2^2(Y)+\alpha_2^2(Y|V)[\sigma_1^2(V)-\sigma_2^2(V)],\ \sigma_2^2(X)=\sigma_1^2(X)-\alpha_1^2(X|V)[\sigma_1^2(V)-\sigma_2^2(V)], \quad (3.8)$$

and

$$\sigma_1(Y,V)=\sigma_2(Y,V)[\sigma_1^2(V)/\sigma_2^2(V)],\ \sigma_2(X,V)=\sigma_1(X,V)[\sigma_2^2(V)/\sigma_1^2(V)]. \quad (3.9)$$

5. The weights and the parameters of Populations 1 and 2 can express the parameters for the synthetic population. The equations for the population means are (a) $\mu_s(X)=w_1\mu_1(X)+w_2\mu_2(X)$, (b) $\mu_s(Y)=w_1\mu_1(Y)+w_2\mu_2(Y)$, and (c) $\mu_s(V)=w_1\mu_1(V)+w_2\mu_2(V)$. And, the population variances are

$$\sigma_s^2(X)=w_1\sigma_1^2(X)+w_2\sigma_2^2(X)+w_1w_2[\mu_1(X)-\mu_2(X)]^2, \quad (3.10)$$

$$\sigma_s^2(Y)=w_1\sigma_1^2(Y)+w_2\sigma_2^2(Y)+w_1w_2[\mu_1(Y)-\mu_2(Y)]^2, \quad (3.11)$$

and

$$\sigma_s^2(V)=w_1\sigma_1^2(V)+w_2\sigma_2^2(V)+w_1w_2[\mu_1(V)-\mu_2(V)]^2, \quad (3.12)$$

where s denotes the synthetic population.

6. Substitute the equations in step 4 in the equations in step 5, the means and variances for the synthetic population on Form X and Form Y can be derived as follows:

$$\mu_s(X)=\mu_1(X)-w_2\alpha_1(X|V)[\mu_1(V)-\mu_2(V)], \quad (3.13)$$

$$\mu_s(Y)=\mu_2(Y)+w_1\alpha_2(Y|V)[\mu_1(V)-\mu_2(V)], \quad (3.14)$$

$$\sigma_s^2(X)=\sigma_1^2(X)-w_2\sigma_1^2(X|V)[\sigma_1^2(V)-\sigma_2^2(V)]+w_1w_2\alpha_1^2(X|V)[\mu_1(V)-\mu_2(V)]^2, \quad (3.15)$$

and

$$\sigma_s^2(Y)=\sigma_2^2(Y)+w_1\sigma_2^2(Y|V)[\sigma_1^2(V)-\sigma_2^2(V)]+w_1w_2\alpha_2^2(Y|V)[\mu_1(V)-\mu_2(V)]^2. \quad (3.16)$$

To obtain estimates for the means and variances for the synthetic population, plug in observed data to the above equations.

7. After taking the square roots of $\hat{\sigma}_s^2(X)$ and $\hat{\sigma}_s^2(Y)$, the equation for Tucker linear transformation, $\ell(x)=\hat{\sigma}_s(Y)/\hat{\sigma}_s(X)[x-\hat{\mu}_s(X)]+\hat{\mu}_s(Y)$, is obtained by replacing the parameters in the above equation with the estimated values obtained previously.

Practical Concerns

Equal reliability across test forms is required for Tucker linear equating. However, Kolen and Brennan (1987) argued that if test forms are designed to be as similar as possible in content and statistical characteristics, and the forms have the same length, small differences in reliability are not likely to have negative influences on equating outcomes.

Levine equally reliable method (Angoff, 1984; Livingston, Dorans, & Wright, 1990) is a frequently used linear equating method that assumes perfectly correlated true scores on the two forms. Compared to the Levine method, Tucker linear method is often considered more appropriate when examine groups are more similar and test forms less similar. The Levine method, however, is often said to be more appropriate when test forms are more similar and examinee groups less similar. Nevertheless, research findings have not yet provided clear evidence for the argument (Kolen & Brennan, 1987).

Chapter 4

EQUIPERCENTILE EQUATING

Equipercentile equating is an observed score equating that aims at finding a score on Form Y of a test that has the same percentile rank as a score on Form X of the same test. When used with the common-item design, like the linear equating methods, equipercentile equating requires a synthetic population composed of two weighted populations.

Equipercentile Function

Let x be a score on Form X, and K_X be the number of items on Form X. The equipercentile function for the synthetic population is

$$e_{Y_s}(x) = Q_S^{-1}[p_S(x)], -.5 \leq x \leq K_X + .5 ; \quad (4.1)$$

where the subscript "s" denotes the synthetic population, and (1) $e_{Y_s}(x)$ is the Form-Y equipercentile equivalent of score x on Form X, (2) $p_S(x)$ is the percentile rank function for Form X, and (3) Q_S^{-1} is the percentile function for Form Y (Kolen & Brennan, 1995).

Let F(x) be the discrete cumulative distribution function for Form X and x^* be the closest integer to x such that $x^*-.5 \leq x < x^*+.5$, then $p(x)$ can be expressed as follows:

$$\begin{aligned} p(x) &= 100\{F(x^*-1)+[x-(x^*-.5)][F(x^*)-F(x^*-1)]\}, \quad \text{if} \quad -.5 \leq x < K_x+.5, \\ &= 0, \quad \text{if} \quad x<-.5, \text{ and} \\ &= 100, \quad \text{if} \quad x \geq k_x+.5 \end{aligned} \tag{4.2}$$

The percentile function for Form Y, Q^{-1}, is the inverse of the percentile rank function, p^{-1}. It is used to find the Form-Y score corresponding to a particular percentile rank.

In equating practice, because of zero frequencies of some scores, the scores of a test are often discrete. As a consequence, equated Form-X score distribution will typically differ from the Form-Y score distribution. A variety of strategies have been developed and proposed to cope with such undesirable situations. Typically, smoothing methods are used to statistically adjust the ragged score distributions. Two commonly used equipercentile equating methods, the frequency estimation method and chained equipercentile equating, are briefly summarized below. The techniques of smoothing are also reviewed.

Frequency Estimation Method

The frequency estimation method assumes that, for both Form X and Form Y, the conditional distribution of total scores given each common-item score, is the same in both

populations. The more similar the two populations are, the more likely the above assumption will hold. Thus the frequency estimation method should be used only when the two populations are reasonably similar (Kolen & Brennan, 1995).

Assumptions

Let Population 1 take Form X and Population 2 take Form Y from the same test, and a set of common items (V) is embedded in both forms. The above assumption of the frequency estimation method is expressed as follows (Kolen & Brennan, 1995):

$$f_1(x|v) = f_2(x|v) \text{ for all } v, \text{ and } g_1(y|v) = g_2(y|v) \text{ for all } v, \qquad (4.3)$$

where $f_1(x|v)$ is the probability that total score X=x given that V=v in population 1, $f_2(x|v)$ is the probability that total score X=x given that V=v in population 2, $g_1(y|v)$ is the probability that total score Y=y given that V=v in population 1, and $g_2(y|v)$ is the probability that total score Y=y given that V=v in population 2.

Kolen and Brennan (1995) suggested that frequency-estimation equating should be conducted only when Populations 1 and 2 are reasonably similar to each other. They argued that the more similar the populations were to each other, the more likely the assumption of the method would hold. However, the decision to use the method depended on the context of the equating, as well as the empirical evidence of the degree of similarity.

Conditional distributions

The conditional distribution $f(x|v)$ is the probability of earning a score of x on Form X, given a score of v on common items V. It can be demonstrated that

$$f(x|v) = \mathrm{P}(\mathrm{X}=x \mid \mathrm{V}=v) = \mathrm{P}(\mathrm{X}=x, \mathrm{V}=v)/\mathrm{P}(\mathrm{V}=v) = f(x,v)/h_{\mathrm{V}}(v), \quad h_{\mathrm{V}}(v)>0 \qquad (4.4)$$

(Berry & Lindgren, 1990); where (1) $f(x,v)$ is the joint distribution of total score and common-item score, and it represents the probability of earning a score of x on Form X and a score of v on common items V, (2) for all x and v, $f(x,v) \geq 0$ and $\sum_{(x,v)} f(x,v) = 1$, and (3) $h_{\mathrm{V}}(v)$ represents the marginal distribution of scores on the common items, which is the probability that $\mathrm{V}=v$ and equals $\sum_{v} f(x,v)$ (Kolen & Brennan, 1995).

Procedures

Frequency-estimation equipercentile equating defines the distributions for the synthetic population on Forms X and Y as follows:

$$f_s(x) = w_1 f_1(x) + w_2 f_2(x) \text{ and } g_s(y) = w_1 g_1(y) + w_2 g_2(y), \qquad (4.5)$$

where $w_1 + w_2 = 1$.

To arrive at expressions composed of direct probability estimates of earning various scores, for the distributions for the synthetic population, the first main task is to express $f_2(x)$ and $g_1(y)$ in quantities for which direct estimates are available. This can be achieved as follows (Kolen & Brennan, 1995):

(1) By definition of density, $f(x|v) = f(x,v)/h(v)$, therefore

$$f_2(x,v) = f_2(x|v)h_2(v) \text{ and } g_1(y,v) = g_1(y|v)h_1(v). \quad (4.6)$$

(2) By assumption of identical conditional distributions,

$$f_2(x,v) = f_1(x|v)h_2(v) \text{ and } g_1(y,v) = g_2(y|v)h_1(v). \quad (4.7)$$

(3) By summing over common-item scores, there follow the marginal distributions:

$$f_2(x) = \Sigma_v f_2(x,v) = \Sigma_v f_1(x|v)h_2(v) \text{ and } g_1(y) = \Sigma_v g_1(y,v) = \Sigma_v g_2(y|v)h_1(v). \quad (4.8)$$

The distributions for the synthetic population therefore can be expressed in quantities that can be directly estimated from the data. The equations are

$$f_s(x) = w_1 f_1(x) + w_2 \Sigma_v f_1(x|v)h_2(v) \text{ and } g_s(y) = w_1 \Sigma_v g_2(y|v)h_1(v) + w_2 g_2(y), \quad (4.9)$$

where $w_1 + w_2 = 1$.

By summing $f_s(x)$ and $g_s(y)$ over values of x and y respectively, the cumulative distributions $F_s(x)$ and $G_s(y)$ can be derived. Define P_s as the percentile rank function for Form X and Q_s as the percentile rank function for Form Y, then P_s^{-1} and Q_s^{-1} are the percentile functions. For frequency estimation method, thus, the equipercentile function for the synthetic population is $e_{Y_s}(x) = Q_s^{-1}[p_s(x)]$.

Smoothing Techniques

Equipercentile equating is often not sufficiently precise due to sampling error. The lack of precision is typically indicated by irregular sample score distributions and equipercentile relationships (Kolen & Brennan, 1995). To obtain more accurate equating results, various smoothing methods have being used on an empirical base to produce smoothed estimates of the population score distributions that are supposed to have less estimation error than the sample score distributions (Kolen, 1991). Typical smoothing approaches include (1) presmoothing, such as the polynomial log-linear method (Holland & Thayer, 1987; Holland & Thayer, 1989; Rosenbaum & Thayer, 1987; Kolen, 1991) and the strong true score method (Lord, 1965; Hanson, 1991; Kolen & Brennan, 1995), and (2) postsmoothing, such as the cubic splines method (Kolen & Jarjoura, 1987; Kolen & Brennan, 1995). Rather than providing better descriptions for the score distributions, often the goal of smoothing is to improve the accuracy in estimating the population score distributions. In such case, it is important to select smoothing methods that improve the precision in estimation but do not introduce substantial bias into the smoothing process (Kolen, 1991).

It was found that both presmoothing and postsmoothing methods improve estimation of equipercentile equivalents to a similar degree. More specifically, smoothing in equipercentile equating can be expected to produce a modest decrease in mean-squared equating error when compared to unsmoothed equipercentile equating (Hanson, 1991; Hanson, Zeng, & Colton, 1994; Kolen and Brennan, 1995). Since postsmoothing directly smoothes the equipercentile relationship, it is more direct than presmoothing, which smoothes the score distributions. Because there is no statistical test for the fit of the postsmoothing method, Kolen and Brennan (1995) suggested applying and evaluating various degrees of smoothing to avoid adding equating error. Specifically, the graphs of the raw-to-raw equivalents for the various degrees of smoothing should be examined to find the relationship that is smooth but does not depart too much from the unsmoothed equivalents. Standard error bands could be constructed to facilitate the evaluation. In addition, the moments of the equated raw scores should be examined to study the similarity among the moments. Kolen and Brennan also offered recommendations for smoothing in scale-score equating, when raw scores are converted to scale scores for the sake of interpretation or presentation. Overall, the smoothing process requires judgments that are dependent on the sample sizes, distribution shapes, numbers of items, and other relevant characteristics of a testing program (Kolen & Brennan, 1995).

The cubic spline postsmoothing method fits a curve to the equipercentile relationship (Kolen & Jarjoura, 1987). It is designed to increase equating precision with frequency estimation method of equipercentile equating, for the common-item non-equivalent group design. For integer scores, x_i, the spline function is,

$$\hat{d}_Y(x) = v_{0i} + v_{1i}(x - x_i) + v_{2i}(x - x_i)^2 + v_{3i}(x - x_i)^3, \qquad x_i \le x < x_i + 1. \quad (4.10)$$

The v_{0i}, v_{1i}, v_{2i}, *and* v_{3i} are weights changing from one score point to the next such that a different cubic equation is defined for between each integer score. The spline is fit over the range of scores x_{low} to x_{high}, and $0 \le x_{low} \le x \le x_{high} \le K_x$. The spline function is minimized to achieve minimum curvature over score points and to satisfy the following constraint (Kolen & Brennan, 1995):

$$\frac{\sum_{i=low}^{high} \left[\frac{\hat{d}_Y(x_i) - \hat{e}_Y(x_i)}{\hat{s}e\ [\hat{e}_Y(x_i)]} \right]}{x_{high} - x_{low} + 1} \le S, \qquad (4.11)$$

where x_{low} is the lower integer score in the range and x_{high} is the upper integer score in the range. The $\hat{e}_y(x_i)$ is the estimate of the Form-Y equivalent of Form-X scores. The $\hat{s}e\ [\hat{e}_y(x_i)]$ is the estimated standard error of equipercentile equating, which standardizes the differences between the unsmoothed and smoothed relationships. The parameter S (≥ 0) is set to control the degree of smoothing. It has been found in practice that values of S between 0 and 1 produce adequate results.

To arrive at a more symmetric equating function, postsmoothing method averages two splines: the spline developed for converting Form X to the Form-Y scale and the

inverse of the spline developed for converting Form Y to the Form-X scale. The average is defined as follows (Wang & Kolen, 1994):

$$\hat{d}^*_Y(x) = \frac{\hat{d}_y(x) + \hat{d}_x^{-1}(x)}{2}, \quad -.5 \le x \le K_x + .5. \tag{4.12}$$

Chained Equipercentile Equating

Chained equipercentile equating involves a chain or sequence of two equipercentile equatings. To equate a Form-X score to a score on Form Y, the usual procedure is as follows (Angoff, 1971; Dorans, 1990; Kolen & Brennan, 1995; Marco et al., 1983):

(1) Find the equipercentile function, $e_{v_1}(x)$, that converts scores on Form X to the common items based on examinees from Population 1.

(2) Find the equipercentile function, $e_{y_2}(v)$, that converts scores on the common items to scores on Form Y based on examinees from Population 2.

(3) Convert the Form X score to a common-item score using $e_{v_1}(x)$, then equate the resulting common-item score to Form Y using $e_{y_2}(v)$.

Although the chained equipercentile method does not require the two populations to be very similar, Kolen and Brennan (1995) argued that it had the following drawbacks in theory: (1) Given the two test forms are essentially interchangeable, it is problematic to equate one full test form to the common items only; and (2) The population underlying the equating is not clearly defined, since it does not use the synthetic population.

Chapter 5

IRT-BASED EQUATINGS

Classical methods of equating, developed for equating observed raw scores, are criticized for not meeting some of the conditions of equating: equity, symmetry, and invariance (Hambleton & Swaminathan, 1990). Equating based on item response theory, however, does not suffer from the same drawbacks, given the IRT model fits the data (Hambleton & Swaminathan, 1990; Kolen, 1981). The result of IRT-based equating, however, varies with the particular equating technique or procedure used. This chapter provides an overview of various IRT-based equating methods for the anchor-item design.

Conceptual Steps of IRT-Based Equating

Typically, IRT-based equating involves the following steps (Hambleton & Swaminathan, 1990):

- Choose an appropriate equating design that takes into account the nature of the test and the group of examinees.
- Determine an appropriate item response model to estimate IRT parameters for alternate forms, and assess model-data fit by gathering a variety of goodness-of-fit measures; including statistical tests of significance and the checks of model assumptions.

- Establish a common metric for ability and item parameters by determining the equating constants (the slope and intercept of a linear equation) that relate either ability parameters or item parameters.
- Make decisions on the scale of the test scores to be reported; the scores can be ability scores, estimated true scores, or observed scores.

Linear Transformation of IRT Scales

IRT parameter estimates obtained from alternate forms of a test can be converted to the same scale via linear transformation (Kolen & Brennan, 1995). Assuming item and person invariance, linear transformation is reasonable for the non-equivalent-group anchor-item design because the difficulty and discrimination parameters for the common items from the alternate forms are linearly related (Petersen, Cook, & Stocking, 1983; Hills, Subhiyah, & Hirsch, 1988).

In theory, given that a 3PL IRT model fits the data, transformation equations relating IRT parameters for alternate forms of a test (say, Form X and Form Y) are defined as follows (Hambleton & Swaminathan, 1990; Kolen & Brennan, 1995):

(1) For person i, the equation for the ability parameter is $\theta_{y_i} = A\theta_{x_i} + B$, where A and B are constants and θ_{y_i} and θ_{x_i} are the values of person i's ability on the scales of Forms Y and X.

(2) Let a_{y_j}, b_{y_j}, and c_{y_j} be the item parameters for item j on Form-Y scale, and a_{x_j}, b_{x_j}, and c_{x_j} be the parameters on Form-X scale, (a) the equation for item discrimination parameter is $a_{y_j} = a_{x_j} / A$, (b) the equation for item difficulty parameter is

$b_{y_j} = Ab_{x_j} + B$, and (c) the equation for lower asymptote (guessing) parameter is $c_{y_j} = c_{x_j}$.

For a group of persons or items, Kolen and Brennan (1995) showed that the transformation constants (A and B) could be expressed as follows:

$$A = \sigma(b_y)/\sigma(b_x) = \mu(a_x)/\mu(a_y) = \sigma(\theta_y)/\sigma(\theta_x), \tag{5.1}$$

and

$$B = \mu(b_y) - A\mu(b_x) = \mu(\theta_y) - A\mu(\theta_x). \tag{5.2}$$

In the above equations, the means $\mu(a_x)$, $\mu(a_y)$, $\mu(b_x)$, and $\mu(b_y)$, as well as the standard deviations $\sigma(b_x)$ and $\sigma(b_y)$, are defined over items. And, the means $\mu(\theta_x)$ and $\mu(\theta_y)$, as well as the standard deviations $\sigma(\theta_x)$ and $\sigma(\theta_y)$, are defined over persons.

In practice, IRT parameters are unknown and thus need to be estimated. In the anchor-item equating design, parameter estimates for anchor items can be obtained and used to replace the parameters in the above equations to find the scaling constants. Basically, linear transformation of IRT scales involves two stages: (a) first, alternate test forms are calibrated separately, (b) the information on anchor items obtained from the two IRT calibrations is then used to derive transformation equations for person and item parameters, which can be used to arrive at equivalent scaled scores for examinees taking different test forms.

In addition to the above scale-transformation procedure, various techniques for transforming IRT scales have been proposed. Regression techniques can be applied, but

the established relationship is not symmetric (Hambleton & Swaminathan, 1990). The mean/sigma method (Marco, 1977), the mean/mean method (Loyd & Hoover, 1980), and the method involving the use of the geometric means of the a-parameters (Mislevy & Bock, 1990) are all straightforward and similar to the procedure described above. Taking into account individual standard error of estimate, the robust mean and sigma method (Linn, Levine, Hastings, & Wardrop, 1981) and robust iterative weighted mean and sigma method (Stocking & Lord, 1983) use variance-weighted means and standard deviations to find the transformation constants. In short, poorly estimated parameters with larger variances receive less weights. The iterative method also weights outliers less.

The above methods suffer from one common flaw, however. They do not take into account all of the item parameters at the same time. Using different transformation methods, various combinations of a-, b-, and c-parameter estimates may result in very similar item characteristic curves (Kolen & Brennan, 1995). The characteristic curve methods have been proposed to overcome such potential problems. They will be briefly reviewed in this chapter, following the fixed-b method.

Fixed-b Method

The fixed-b IRT-based equating method sequentially calibrates test items following these steps:

(1) Estimate bs and other item parameters for Form-A items;

(2) Calibrate Form-B items by fixing bs for the anchor items at the values obtained from the previous step;

(3) Form B scale is then fixed onto the scale of Form A (Petersen, Cook, & Stocking, 1983; Hills, Subhiyah, & Hirsch, 1988).

Although the fixed-b method is usually used with LOGIST (Hills, Subhiyah, & Hirsch, 1988), previous research also found this method work well with BILOG (Yang & Houang, 1996; Yang, 1997). Specifically, the equating results yielded by the fixed-b method using BILOG were consistent with the results from IRT-based linear transformation method.

Characteristic Curve Transformation (Formula) Methods

Unlike the above methods, characteristic curve methods developed by Haebara (1980) and Stocking and Lord (1983) consider the parameter estimates simultaneously. The two methods estimate the difference between the item characteristic curves on the two scales, for a given θ and over items, differently. However, both methods rely on iterative algorithms that minimize the overall differences over examinees to find the transformation constants (A and B).

It has been found from comparison studies (Baker & Al-Karni, 1991) that the characteristic curve transformation methods yielded more accurate results than the other methods. Nevertheless, Baker and Al-Karni (1991) found that the results did not differ much sometimes. In addition to requiring computationally intensive iteration procedures, the characteristic curve methods also have the limitation of not explicitly accounting for the error in estimating item parameters (Kolen & Brennan, 1995).

When the characteristic curve methods are practically not applicable, Kolen and Brennan (1995) proposed a strategy to improve the equating results of the mean/sigma and mean/mean methods. Basically, scatter plots (Form Y vs. Form X) of the item-parameter estimates on the common items are used to identify potential outliers, and the transformation results with and without the outliers are compared.

IRT True-Score Equating

In theory, true scores on alternate test forms can be obtained and equated. To eliminate negative scores, values on the θ (ability) scale may be transformed to their corresponding true-score values (Hambleton, Swaminathan, & Rogers, 1991). Then the true scores can be equated via linear transformation.

IRT True Scores

Let θ be the parameter for ability and n be the number of items in a test, then the true score (ξ) can be defined as follows: $\xi = \sum_{1}^{n} p_i(\theta)$ (Crocker & Algina, 1986; Hambleton & Swaminathan, 1990; Lord, 1980). The true score of an examinee with ability θ on a test is the sum of the conditional probabilities of correct responses across the item characteristic curves (*ICC*s). The *ICC* is a monotonically increasing function that describes the relationship between examinees' item performance and the abilities underlying item performance. Graphically, the *ICC* shows that as the level of the trait increases, the probability of a correct response to an item increases (Hambleton, Swaminathan, & Rogers, 1991).

Alternatively, ξ can also be interpreted in terms of the test characteristic curve (*TCC*), which is the sum of item characteristic curves (*ICC*s) (Hambleton, Swaminathan, & Rogers, 1991). Though monotonically related, the true score ξ and ability θ are expressed on different scales of measurement. The scale for ξ depends on the number of items on the test, while the scale for θ is independent of the number of items on the test

(Lord, 1980). However, ξ is useful in reporting ability estimates, since it is on the same scale as the number-right score.

When comparing tests or test forms of different lengths, true proportion correct or domain scores (π) can be reported instead of ξ. Ranging between 0 and 1, π is computed by dividing ξ by the number of items (n) in test forms. That is, $\pi = \xi/n = 1/n\ [\sum_{1}^{n} p_i(\theta)]$ (Hambleton & Swaminathan, 1990; Hambleton, Swaminathan, & Rogers, 1991).

Taking into account the number of alternative options, which has substantial influence on guessing, the true score formula can be rewritten as:

$$\xi = \sum_{1}^{n} \{[(k_i + 1) / k_i] \times p_i(\theta) - 1 / k_i\}, \tag{5.3}$$

where n is the number of test items, and $k_i + 1$ is the number of alternate answers for item i (Petersen, Cook, & Stocking, 1983).

Equating True Scores

Suppose the ability level of an examinee on test form X is θ_x and ξ_x is the corresponding true score, and the ability level of the same examinee on alternate test form Y is θ_y and ξ_y is the corresponding true score. Then the equating equations for true scores are

$$\xi_x = \sum_{i=1}^{n} p_i(\theta_x) \text{ and } \xi_y = \sum_{j=1}^{m} p_j(\theta_y) \equiv \sum_{j=1}^{m} p_j(\alpha\theta_x + \beta), \tag{5.4}$$

where n is the number of items on test X and m is the number of items on Y, $p_i(\theta_x)$ is the probability of a correct answer to item i by an examinee, whose ability level on test X is θ_x, $p_j(\theta_y)$ is the probability of a correct answer to item j by an examinee, whose ability level on test Y is θ_y, and (4) $\theta_y = \alpha\theta_x + \beta$ expresses the linear relationship between θ_y and θ_x (Hambleton & Swaminathan, 1990). In theory, for a given value θ_x, the pair of true scores (ξ_x, ξ_y) on test forms X and Y can be determined. In practice, however, true scores can only be estimated.

Concurrent Calibration Method

Using LOGIST, an IRT calibration program on the mainframe computer, item and ability parameters can be estimated simultaneously in the following manner:

(1) Treat examinees taking Book A and Book B as one sample. Treat data as if all the examinees have taken a test composed of all the items in both of the forms.

(2) Code the scores on Book B items as "not reached" for the examinees taking Book A. Code similarly for the examinees taking Book B.

(3) Calibrate, in a single LOGIST run, the ability parameters for all the examinees and the item parameters for all the items. The ability estimates for the examinees taking either Book A or Book B are automatically placed on the same scale, and no further step is needed (Hambleton & Swaminathan, 1990).

Conceptually, the concurrent calibration method is expected to yield more stable equating results because it does not make any assumptions about the relationship between the item parameter scales for separate calibration runs (Petersen, Cook, & Stocking, 1983; Hills, Subhiyah, & Hirsch, 1988).

Advantages of IRT-Based Equating

Traditional equating methods can yield good results if the test forms are sufficiently parallel (Lord, 1980). However, when the tests to be equated differ in difficulties and nonrandom groups of examinees differ in ability, IRT-based methods are in theory considered better than the other methods (Cook & Eignor, 1991). Major advantages of IRT-based equating are summarized below.

Curvilinear Equating

IRT-based methods are useful for modeling either a linear or curvilinear relationship between the raw scores on the two editions of a test. The methods make no assumption of equal reliability or identical observed score distributions (Cook & Eignor, 1983; Kolen, 1981). The result of IRT-based equating often agrees with linear equating to a surprising degree. One possible explanation is that the tests being equated were constructed to be considerably similar (Berk, 1982).

Item-Free and Person-Independent Measures

The most distinguishing advantage of IRT-based equating is its "item-free" estimates for persons and "person-free" item characteristics (Lord, 1977). Ideally, examinees of the same ability will get the same ability score, no matter which items are taken.

In addition, IRT-based methods estimate errors of measurement at each ability level, while traditional equating methods only yield a single standard error of measurement for all examinees. Green, Yen, and Burket (1989) suggested that the IRT method would yield equivalent ability estimates for item sets differing in difficulty and/or discrimination,

despite the fact that the equivalent estimates might be associated with different standard errors of measurement.

Practical Appeal

In addition to its theoretical advantages, IRT-based equating is found to have the following advantages in practice (Cook & Eignor, 1983, 1991):

- It provides better equating than equipercentile equating at the upper ends of score scales, where important decisions are often made.
- It improves flexibility in equating a new test form to any of the previous editions of a test, given that their item parameter estimates are placed on the same scale.
- If re-equating is needed, which usually occurs when certain items are added or dropped, it is easier to reequate through IRT-based methods since it is more convenient to obtain IRT true score estimates.
- It enables item-level pre-equating, where the equating conversions between a new test edition and a previous edition or editions can be derived before the new edition is administered operationally, if item-level pretest data are available and item parameter estimates can be placed on a common scale.
- When equating test forms that differ somewhat in content and length, the 3PL-IRT-based equating may reduce equating bias or scale drift in equating chains (Petersen, Cook, & Stocking, 1983).
- The stability of the scales near the extreme values may also be increased (Hills, Subhiyah, & Hirsch, 1988).

The above arguments suggest that IRT-based equating is superior to conventional equating methods. Nevertheless, Kolen and Brennan (1995) indicated that IRT models gained their flexibility by making strong statistical assumptions, which were not likely to

hold precisely in real testing situations. As a result, the robustness of IRT-based applications to the violations of IRT model assumptions needs to be studied. In the mean time, the relative efficacy of IRT-based equating remains uncertain. Green, Yen, and Burket (1989) noted that it was not safe to say that IRT methods would yield equivalent ability estimates if the items in different forms were different in content coverage. Therefore, despite past findings that content variation has substantially smaller effects on ability estimates than on item parameters (Yen, 1980), test content should be accounted for in IRT-based equating.

IRT Assumption of Test Dimensionality

For achievement testing, the IRT assumption that examinee performance on a test is unidimensional is likely to be violated (Dorans & Kingston, 1985; Green, Yen, & Burket, 1989). This is mainly because most sets of test items measure a complex of abilities rather than one single trait (Reckase, Ackerman, & Carlson, 1988). To justify the applications of the unidimensional IRT models, the robustness of the unidimentional IRT models to the violations of the IRT unidimensionality assumption becomes a major point of interest. The impact of such violations on IRT-based equating thus deserves to be examined.

Green, Yen, and Burket (1989) indicated that the degree of test unidimensionality also limited the usefulness of classical equating approaches, such as equipercentile equating, but in a less visible way. They argued that unidimensionality was an implicit assumption in classical equating. Given the fact that the concept of unidimensionality is important in both classical and IRT-based test equatings, literature on test dimensionality is reviewed below.

Definition of Test Dimensionality

Test dimensionality can be defined as the number of latent variables (traits or abilities) accounting for the correlations among item responses in a particular data set. Camilli, Wang, and Fesq (1995) thus argued that dimensionality was not a property of a test per se but was context dependent. In addition to a particular set of items and a particular set of examinees, Camilli and his colleagues indicated that test dimensionality was also dependent on test use. Therefore, they suggested that arguments regarding dimensionality should take into account both test content and the evidence from statistical analyses. Camilli, Wang, and Fesq (1995) further differentiated between two kinds of dimensionality: while functional dimensionality depends on the testing situation and the use of test scores, statistical dimensionality is defined by local independence (Lord, 1982b).

Definition of Unidimensionality

Unidimensionality requires that there is a dominant factor or trait that influences test performance on a set of test data (Hambleton, Swaminathan, & Rogers, 1991). Test scores are most meaningful when all the items depend on a single (dominant) trait. If the IRT assumption of unidimensionality holds, local independence should be observed. Statistically, local independence requires that, for fixed ability level θ, the item characteristic functions for any pair of items i and j should be independent (Lord, 1982b). If the probabilities for the given responses to the given items i and j are not independent at fixed θ, the responses may depend on some trait other than the θ. Hence, the IRT assumption of unidimensionality is violated. In practice, local independence is examined

by checking the variance-covariance or correlation matrices for examinees within different intervals on the ability or test score scale (Hambleton, Swaminathan, & Rogers, 1991).

In addition to studying the assumption of local independence, there are several other ways to checking the unidimensionality assumption, including examining an eigenvalue plot of the inter-item correlation matrix, comparison of eigenvalue plots based on the test data and random data, fitting a nonlinear one-factor analysis model to the inter-item correlation matrix to study the residuals, and examining the estimated parameter values of items that are more likely to violate the unidimensionality assumption (Hambleton, Swaminathan, & Rogers, 1991).

Robustness of IRT Unidimensionality Assumption

It has been shown that violations of unidimensionality might have an impact on equating, but the effects might not be substantial (Camilli, Wang, & Fesq, 1995; Dorans & Kingston, 1985; Reckase, Ackerman, & Carlson, 1988). Nevertheless, if a test is influenced by several equally potent dimensions, IRT procedures are likely to yield inconsistent estimates of ability (Reckase, 1979).

Many researchers have studied the impact of violations of IRT unidimensionality assumption on IRT-based equating outcomes. Dorans and Kingston (1985) separated a test into two "item groups", according to the two dimensions suggested by factor analysis. They then assessed the comparability of IRT-based equatings based on (1) homogeneous IRT calibrations (calibrating the two item groups separately), and (2) heterogeneous IRT calibration (calibrating the entire test at one time). They argued that if the IRT model fit the data, the item-grouping should not have effect on IRT-based equating. That is, the homogeneous and the heterogeneous calibrations should result in the same item parameter

estimates. Although it was found that the homogeneous calibrations resulted in higher discrimination parameter estimates than the heterogeneous calibration, which caused an asymmetry of equating, other study results of Dorans and Kingston suggested that IRT-based equating might be sufficiently robust to the dimensionality violation. Therefore, Dorans and Kingston argued that it was reasonable to regard the assumed IRT single trait as a weighted composite of different traits underling the test.

Using both simulated and real test data, Reckase, Ackerman, and Carlson (1988) also proved theoretically and empirically that the IRT unidimensionality assumption was robust. The simulated data were generated to be two-dimensional by the multidimensional 2PL IRT model, and the real test data had two rather strong dimensions -- mathematics achievement and reading ability. The statistic used to detect the violation of the unidimensionality assumption was the Q_3 suggested by Yen (1984). Reckase, Ackerman, and Carlson concluded that even though more than one ability is required for successful performance on a test, a set of items measuring the same weighted composite of the multiple abilities will meet the assumptions of a unidimensional IRT model.

Yen (1984) hypothesized that when test data is generated from several correlated underlying traits, the unidimensional three-parameter IRT model assumes a unidimensional trait that is conceptually a combination of the multiple underlying traits. This hypothesis was supported by Yen's derivations and simulated data. Yen further hypothesized that when defining the unidimensional three-parameter trait, the unidimensional IRT model ignores other independent traits that influence only a few items of a test (Yen, 1984). Dorans (1990) further argued that although tests to be equated must measure the same construct and contain items of the same content mix, they do not have to be composed of unidimensional items.

Chapter 6

ISSUES IN COMMON-ITEM EQUATING

The common-item equating design was used in this study to equate the two forms of a test, because these forms were taken by two nonequivalent examinee groups and a set of common anchor items were embedded in each of the test forms. This chapter reviews issues relevant to common-item equating, including the effects on the accuracy of equating results of ability differences and sampling, characteristics of anchor items, and the forms of score distributions of the test forms being equated.

Effects of Ability Differences and Sampling on Equating

Sample invariance is a desirable outcome in test equating. Ideally, equating results should be independent of the sub-populations of the same ability. Lawrence and Dorans (1990) suggested that population independence should be investigated under circumstances where the examinee samples differ in ability because the equating results might rely on the examinee samples of approximately equal ability.

Effects of Ability Differences

Ability differences between examinee samples may have substantial impacts on equating results (Cook, Eignor, & Schmitt, 1988). Theoretically, the closer the groups in

ability, the more accurate the equating should be. However, Marco, Petersen, and Stewart (1983) found that if the anchor test mirrored the content and the difficulty level of the total test, the sample differences had relatively small and unsystematic effects on the quality of the equating results.

To cope with the threat of inaccuracy from the ability discrepancy, particular sampling strategies may be employed to draw samples with similar ability. The effectiveness of different sampling strategies have been studied with various equating methods, since sampling effects are likely to vary with equating methods. Current literature generally supports the use of representative sampling. "Matched sampling" that matches examinee groups on the anchor test score, however, is not favored (Eignor, Stocking, & Cook, 1990; Livingston, Dorans, & Wright, 1990). It is also suggested that some other ability measure be used to match the samples when populations differ in ability (Livingston, Dorans, & Wright, 1990).

Representative vs. Matched Sampling

Representative sampling typically requires random sampling of examinees so that the examinee sample is representative of the target population being studied. "Matched sampling", however, is to stratify examinee samples on some ability measure such as the anchor test score. As a result of "matched sampling", the two examinee groups will have the same score distribution on the particular ability measure (Dorans, 1990; Lawrence & Dorans, 1990; Livingston, Dorans, & Wright, 1990).

Lawrence and Dorans (1990) found it useful to match samples on the anchor test score in terms of abridging the disagreement among the equating outcomes yielded by various equating methods (Tucker linear equating, Levine equally reliable linear equating, frequency-estimation equipercentile equating, chained equipercentile equating, and IRT true-score equating). Such disagreement usually occurs with representative sampling. Schmitt, Cook, Dorans, and Eignor (1990) also had similar findings. However, Lawrence

and Dorans (1990) also found that IRT-based equating procedures were seriously affected by differences in group ability, despite the fact that matching improved the equating results to at least somewhat degree.

Livingston, Dorans, and Wright (1990) suggested that when populations differ in ability, "matched sampling" may yield little improvement in equating accuracy. Dorans (1990) further indicated that "matched sampling" can be problematic when the two samples differ widely in ability. Schmitt, Cook, Dorans, and Eignor (1990) cautioned that matching on a set of internal common items may introduce some unknown degree of bias for certain equating methods, and the magnitude and effect of this bias is not clear. Recognizing that matching samples on the anchor test score would introduce more errors than representative sampling in estimating item difficulty, Eignor, Stocking, and Cook (1990) did not recommend the use of such matching strategy with the 3PL IRT true-score equating, or Levine equally reliable equating, or chained equipercentile equating. It is speculated that stratifying on the anchor test score might lead to violations of statistical assumptions of sample invariance (Lawrence & Dorans, 1990).

Characteristics of Anchor items

The characteristics of anchor items, particularly their content representativeness and the length of the anchor test, influence equating results. It is thus crucial to select adequate anchor items when test equating is desired. The issues of the length and the content representativeness of the anchor are briefly reviewed in this section.

Anchor Length

For the anchor-item equating design, it has been shown that the efficiency of linear equating depends on the correlation between the anchor test and the test forms, and this correlation is a monotonically increasing function of the reliability of the total test and the

length of the anchor test (Budescu, 1985). The anchor length thus raises a reasonable concern about the efficiency of equating, since from time to time equating has to be based on short anchors.

Although there is no absolute standard for setting the length of an anchor, a rule of thumb (Angoff, 1984) is as follows: Include at least 20 items or 20% of the total number of items in a test, whichever is larger. Several studies have shown that as few as five to six carefully selected anchor items performed satisfactorily for the IRT-based equating. In such cases, the item parameters of different tests were estimated by the IRT concurrent method (Raju, Edwards, & Osberg, 1983; Wingersky & Lord, 1984; Raju, Bode, Larsen, & Steinhaus, 1988; Hills, Subhiyah, & Hirsch, 1988). Nevertheless, using the IRT concurrent method, Hills, Subhiyah, and Hirsch (1988) found that randomly selected anchor items (five items from a mathematics test) were not sufficient to produce satisfactory equating results, but an anchor of ten items was sufficient.

To determine the adequate length for an anchor, each testing program has to decide an acceptable level of equating efficiency to meet its particular needs, while taking into account the time, cost, and context constraints for equating (Budescu, 1985).

Content Representativeness

Whether the anchor items are representative of the entire tests, in terms of its content and statistical properties, is especially important when the examinee groups vary in ability (Cook & Petersen, 1987). Budescu (1985) demonstrated that the magnitude of the correlation between the anchor test and the unique component of each test form was the single most important determinant for the efficiency of linear equating using the anchor-item equating design. Thus, to achieve a high correlation and to enhance equating

efficiency, Budescu recommended the use of an anchor test that represented the same psychological task to both examinee groups and required the same psychological operations as the two test forms did.

In a study using chained equating, Klein and Jarjoura (1985) compared the equating accuracy of content-representative anchors versus non-representative but substantially longer anchors. The testing program in their study was a professional certification examination and the 250 multiple-choice items were classified into six content areas. Tucker linear equating was used to equate test forms and the results revealed that the content representation of the anchors was critical for test equating.

Equating Tests with Skewed Distributions

Large scale achievement tests that have approximately symmetrical and bell-shaped score distributions are often the focus of equating. It is necessary from time to time, though, to equate tests that have skewed score distributions, such as minimum-competency tests and licensure exams with high passing standards. For licensure or certification programs, test forms are often equated with special interest in a particular cut-off score, or a range of scores, to inform decision making. To maximize the precision of the decision, it is reasonable to direct more attention to equating in the cutting score region, even at the expense of poorer equating at other scores (Brennan & Kolen, 1987).

Hills, Subhiyah, and Hirsch (1988) equated one version of the Florida Statewide Student Assessment test, a minimum-competency test, to an early version administered two years before. The test items were from the same content domain, item difficulties were similar, and the examinees were essentially from the same population. They found that five equating methods (Angoff's Design IVA linear method, Rasch model, 3P-IRT

concurrent method, 3P-IRT fixed-parameter method, and 3P-IRT formula method) yielded similar results. Thus they concluded that IRT-based equating methods could be applied to equate minimum-competency tests with extremely skewed distributions.

Chapter 7

EVALUATION OF EQUATING ACCURACY

The task of test equating does not stop after an equating method is applied and equivalent scores are established. After the forms of a test are equated by a particular method, it is important to know whether they are really equated. There are many ways to study the effectiveness of test equating. In addition to computing statistical measures that directly estimate the overall accuracy of equating, the effectiveness of equating can also be determined by examining scale stability (Kolen, 1981) or checking the assumptions of equating, such as population invariance (Lawrence & Dorans, 1990).

Harris and Crouse (1993) provided a thorough and recent overview of various approaches used in equating research and practice for evaluating equating outcomes. In this chapter, the variety of approaches identified by them are first summarized. Important issues regarding the estimation of equating accuracy, including equating errors and problems due to the use of some arbitrary criterion for equating accuracy, are then elaborated. Two common indices of equating accuracy are also reviewed.

Approaches for Evaluating Equating Accuracy

Harris and Crouse (1993) reviewed the approaches and criteria proposed in the literature for assessing equating outcomes to explore this largely undeveloped area.

According to their classifications, the evaluation approaches and the particular criteria used in these approaches include:

- Using the equivalent-expected-scores criterion, which is based on the definition of "weak equity" (Yen, 1983). "Weak equity" is developed from Lord's definition of equity for equating (Lord, 1980). It only requires that the means of the conditional distributions of scores on each test form be equal after equating.
- Using indices that summarize overall accuracy of equating, such as the root-mean-squared deviation (RMSD) (Klein & Jarjoura, 1985; Livingston, Dorans, & Wright 1990).
- Computing standard errors of equating associated with sampling of examinees.
- Using simulated data that has known true equating relationship to enable absolute criterion for evaluating equating accuracy.
- Equating a test to itself, directly or through a chain of intervening test forms. This approach is mainly for the study of scale drift (Petersen, Cook, & Stocking, 1983) attributed to equating method. The typical criterion for equating accuracy is whether the resulting conversion is an identity of the "origin" test. Using chained equating, Petersen, Cook, and Stocking (1983) found that IRT score conversions were often associated with less discrepancy from the initial scale than the other equating methods. In addition, the equating methods based on the three-parameter logistic IRT model were found to be more stable than the other equating methods when tests differed somewhat in content and length.
- Treating an equating based on a very large sample of examinees as an estimate of a population equating. The results of equatings for smaller groups are then compared to

the "population" equating to evaluate their equating accuracy. For instance, Livingston (1993) studied the accuracy of equatings for some very small samples (25, 50, 100, and 200) by comparing their equating results to the equating outcome in the full population of 93,283 examinees. Livingston, Dorans, and Wright (1990) also used the equating result for two Scholastic Aptitude Test (SAT) forms on large equivalent populations (more than 115,000 students for each form) as a criterion for equating accuracy. They studied the effectiveness of five equating methods, used in combination with two sampling strategies respectively, by comparing their equating results to the large-sample criterion. Such large-sample criteria, however, is usually difficult to obtain in practice.

- Comparing the consistency or agreement of equating results yielded by various equating methods.
- Conducting replication studies, such as cross-validation that uses independent samples of examinees, to examine sample invariance. Cross-validation is often used to study the stability of equating. Typically, one wishes to replicate results found in one sample in another independent sample. Harris & Crouse (1993) indicated, however, such study of sample invariance does not provide evidence for equating accuracy.
- Practical yet sometimes subjective approaches, such as examining the frequencies of rounded scale scores, inspecting the conversion tables for gaps, or inspecting smoothing outcomes.
- Other atypical approaches, such as examining correlation between anchor and unique items (Budescu, 1985), studying the relationship between equivalent scores and criterion scores, and inspecting test information functions.

Estimating Equating Errors

Equating errors due to estimation can be random or systematic. Random equating error results from sampling of examinees. By using large samples of examinees and choosing appropriate equating designs, random errors can be reduced. Violations of the assumptions and conditions of equating, however, cause systematic equating errors. Systematic errors sometimes can be so large that the results of equating may be worse than no equating (Kolen & Brennan, 1995). Therefore, the conditions of equating and the assumptions of equating models should be examined carefully to control for the systematic errors.

In equating practice, standard errors of equating are useful in indicating the amount of random errors in equating such that the effectiveness of equating can be determined. For many equating designs and methods, approaches for estimating standard errors of equating have been developed and applied (Kolen & Brennan, 1995). For instance, the delta method (Kendall & Stuart, 1977), which is based on a Taylor series expansion, is used commonly for deriving standard error expressions. In addition, Jarjoura and Kolen (1985) derived the standard errors of equipercentile equating to be used in the design of common-item nonequivalent populations. Both of their simulated and real test data suggested that the derived standard errors were useful in estimating equating errors.

From time to time, only small samples of examinees are available for equating study. In such cases, the amount of random equating errors is a major concern. To determine the usefulness of common-item linear equating with small samples, Parshall, Houghton, and Kromrey (1995) used bootstrap standard errors of equating (Efron, 1982; Efron & Tibshirani, 1993) and statistical bias in equating to study the adequacy of equating. Their results showed that although the levels of equating bias with small

samples were trivial, bootstrap standard errors increased substantially as sample size decreased. Thus, Parshall and his colleagues argued that for small samples, bootstrap approach may provide more accurate estimates of standard errors.

Arbitrary Nature of Equating Criteria

In equating practice, because true equivalent relationship can never be known, equating results are often compared to some arbitrary but sound criteria for estimating the accuracy of equating (Dorans & Kingston, 1985). In such cases, the consistency or agreement between the equating outcomes and the criteria is typically measured to represent equating accuracy. The estimation of equating accuracy thus depends on the nature and quality of the arbitrary criteria used. Moreover, consistency measures do not address the issue of equating accuracy directly.

Usually, it is unreasonable to compare the results of all equating methods to one single criterion for evaluating equating accuracy, since equating designs and models vary with the particular context of testing. Equivalent scores derived from conventional equating methods that have been known to yield accurate results, or have been used in practice for some time, are typically used as the criteria for IRT-based equating (Hills, Subhiyah, & Hirsch, 1988). Generally, results from equipercentile equating are a good candidate for such a criterion. For example, Yen (1985) suggested using the results from equipercentile equating as a criterion because it is as accurate as the IRT-based equating results.

Some empirical findings (Petersen, Cook, & Stocking, 1983; Hills, Subhiyah, & Hirsch, 1988) have suggested that IRT-based equating is more accurate than conventional equating, including various linear equatings and equipercentile equating. IRT methods

were thought to be superior for their capacity to equate both parallel and non-parallel tests or test forms (Kolen, 1981). It was also found that IRT-based procedures were effective for both inter-level and inter-form equating (Green, Yen, & Burket, 1989). Nevertheless, conclusions should be made carefully to take into account the fact that the criteria used in equating practice for evaluating equating accuracy are usually arbitrary in practice.

Root-Mean-Squared Deviation (RMSD)

The root-mean-squared deviation (RMSD), also known as the root-mean-squared error of equating (RMSE), is a commonly used overall accuracy measure for equating (Klein & Jarjoura, 1985; Kolen & Harris, 1990; Livingston, Dorans, & Wright, 1990; Schmitt, Cook, Dorans, & Eignor, 1990). Suppose Form B of a test is equated to Form A of the same test using an equating method E, and another sound equating method E_C is used as a "criterion equating", whose results for the same test forms are used as a criterion for evaluating the accuracy of the equating results of the method E. Then

$$RMSD = \{[\Sigma n_y(\hat{x}_y - x_y)^2]/\Sigma n_y\}^{1/2} , \tag{7.1}$$

where (a) n_y is the number of examinees with raw score y on Form B, (b) $\hat{x}_y$ is the scaled score on Form A corresponding to y, determined by the equating method E, (c) x_y is the scaled score on Form A corresponding to y, determined by the criterion equating E_C, and (d) the summation is over the raw-score levels on Form-B.

Several other summary statistics, such as the "mean equating error" (Klein & Jarjoura, 1985), have also been used as the indices of equating accuracy. Some of them

are variations of the RMSD. In a study investigating the adequacy of several curvilinear equating models on the verbal portion of the SAT, Marco, Petersen, and Stewart (1983) used a "weighted mean square difference" to evaluate the effectiveness of the various equating models. Representing total error, the "weighted mean square difference" weighted those values occurring more often more heavily. Marco, Petersen, and Stewart further standardized this summary index of accuracy to enable comparisons of the total errors across equating situations or equating methods.

Using the same notations as defined in Formula 7.1, the overall bias of the equating method E can be estimated by computing the following statistic:

$$\Sigma n_y(\hat{x}_y - x_y) \; / \Sigma n_y \; . \tag{7.2}$$

This bias statistic measures the tendency for the equating method E to yield equated scores that are systematically too high or too low (Livingston, Dorans, & Wright, 1990). Such bias statistic, however, is not a good index of equating accuracy. It has the drawback of underestimating the overall bias when some of the resulting equated scores are estimated too high and some are estimated too low. This is because the negative bias at individual raw-score levels will cancel out the positive bias at individual raw-score levels.

PART III: METHODOLOGIES AND RESULTS

Chapter 8

DATA, DESIGN AND METHOD

Nature of the data analyzed in this study, including characteristics of the test and test takers (examinees) from the professional examination, are described in this chapter. This chapter also presents research designs developed for this study and the methods used to analyze the data.

Description of Data

The test data analyzed were the scores on the two forms, Book A and Book B, of an in-training examination taken by the candidates for doctors of a medical specialty. The candidates took the test while participating in various in-training programs located at different sites (usually in hospitals), to prepare for the formal board certification examination. The cutoff for a passing score was a minimum of 75% of the test items being correctly answered.

The test takers of the in-training examination were strongly motivated to become board-certified and to participate in the in-training programs in preparation for the certification exams. Since the in-training test provided candidates a valuable opportunity to get familiar with the formal certification exams, it was assumed that the candidates had taken the test seriously. After receiving the instructions and training from the in-training

programs, most of the test takers were expected to master the knowledge or skills being tested and pass the examination.

Test Content and Format

The test forms included multiple-choice items, and each test item had five alternatives. The overall content of all the items was related to a medical specialty. The core content for the specialty, representing the scope of the medicine practice or the universe of the specialty, was developed in part to outline the material for testing on the formal certification examination. Twenty-three core content areas (categories) were identified, and the proportional distribution of the test items across these core content areas was determined for the in-training, formal certification, and recertification examinations. Originally, Book A and Book B each had 225 items. For each test form, the proportional distribution of the test items across the 23 core content areas is as follows:

Table 1 - Proportional Distribution of Test Items Across the 23 Core Content Areas

Core Content Area	Number of Items	%	Core Content Area	Number of Items	%
1	13	5.8	13	13	5.8
2	23	10.2	14	7	3.1
3	3	1.3	15	5	2.2
4	14	6.2	16	15	6.7
5	5	2.2	17	13	5.8
6	19	8.4	18	25	11.1
7	5	2.2	19	8	3.6
8	3	1.3	20	5	2.2
9	6	2.7	21	4	1.8
10	9	4.0	22	9	4.0
11	8	3.6	23	9	4.0
12	4	1.8			
			Total	225	100.0

An analysis on item discrimination indicated that a few items in the original item pool were unsatisfactory, because they misled the examinees to give irrelevant answers. After these poorly-written items were excluded, a pool of 255 items was available for this study. The item responses of all the items were all scored as right or wrong (coded as 1 or 0). Book A had 203 items, of which 58 items were unique to Book A. The total number of items in Book B was 197, among them 52 were unique items. A total of 145 anchor items were identically worded and embedded in both test forms in the same location (same item numbers in both test forms). Characteristics of these anchor items, such as their difficulties and content representativeness, will be discussed in next chapter. Outcomes of item analyses for four subtests, sampled from the professional in-training test, will also be presented in next chapter. The negatively skewed distributions of test scores on various test forms will be examined.

Examinee Groups

A total of 2,242 candidates took the in-training test. One examinee who had an extremely low total score, compared to the total scores of the others, was determined to be an outlier. It is very likely that this outlying case did not take the test seriously or guessed throughout the entire test. To avoid possible contamination introduced by the outlier to the entire data set, the outlier was excluded from this study.

The two examinee groups taking different forms of the in-training test were not randomly formed, neither were the two test forms randomly assigned to the 2,241 subjects. A total of 1,092 subjects took Book A, while 1,149 others took Book B. A preliminary inspection on the test data showed that the examinee group taking Book B scored slightly higher on the 145 anchor items (mean=107.721, sd=13.113) than the group

taking Book A (mean=105.457, sd=13.767). Therefore, it was possible that the group taking Book B had higher ability. Nonetheless, as Lord (1981) noted, difference in ability level would not influence equating results, given that an anchor-test design was employed. The differences between the two examine groups will be further discussed in next chapter.

Appropriateness for Equating

The test data analyzed in this study generally met the requirements for equating, described earlier in the section for equating guidelines in Chapter 2. Specifically, the test forms had sufficient number of items, and the sampled test forms created from them for the equating study (described in next section) were reasonably long. Most of the test items were common anchor items, and all the items were from one single content domain. The test items were administered and secured under standardized conditions. Some items had been administered in previous years under the same standardized testing situations and found to be satisfactory.

In addition, the stems, alternatives, and stimulus materials for the common anchor items were identical for the two test forms. The scoring keys were clear and consistent for the two forms. The overall examinee group, exceeding 2,200 subjects, was reasonably large. Under the guidance of the core content specification, the curriculum and training materials received by different examinees were expected to be consistent. The curriculum and training also should be stable over time.

Research Design

Items included in Books A and B of the in-training test were from 23 core content areas. These items formed an item pool for one overall content domain, representing the

medical specialty. Four tests, shorter than the in-training test, were assembled with items sampled from the overall item pool for this study. Each of the four sampled tests had two test forms -- one subtest of Book A and one subtest of Book B. While various item-sampling schemes sampled items from the overall item pool, it also manipulated the content homogeneity of the test items and the content representativeness of the anchor items. The resulting four sampled tests, therefore, differed in content homogeneity. They also had anchor items differing in their content representativeness. However, all the sampled items in different sampled test forms were from the same overall content domain.

Item sampling results are presented in Table 2. All of the sampled test forms were much shorter in length than the original test forms. In testing practice, short tests are often used for practical reasons such as limited time for testing and concerns about the effect of fatigue. It is also of great interest to study the outcomes of equating when test forms to be equated are shorter in length. Therefore, equating results based on the shorter sampled test forms in this study were expected to provide useful insights for the common practice of testing and equating. All sampled test forms were also expected to have negatively skewed score distributions, as the original test forms did. In equating research, data with such distributions received less attention than they should have.

Each pair of the sampled test forms in this study were equated using internal anchor-item equating design with non-equivalent examinee groups. Four equating methods -- the Tucker linear method, the non-linear equipercentile method, and two IRT-based methods -- were used to equate each pair of test forms. Equating results of these methods were compared using four criteria for evaluating equating accuracy. Overall, the variables delineating the entire study included content homogeneity of test items, content representativeness of anchor items in relation to total test, equating method, and criterion

Table 2 - Number of Items Sampled Under Four Sampling Schemes

Item Sampling Scheme		Simple random sampling (SR)	Equal-weight domain random sampling (EW)	Proportional-weight domain random sampling (PW)	Purposeful sampling (PS)
Resulting Test Form	A	60 items (including 30 anchor items)	69 items (including 49 anchor items)	60 items (including 40 anchor items)	60 items (including 45 anchor items)
	B	60 items (including 30 anchor items)	69 items (including 49 anchor items)	60 items (including 40 anchor items)	57 items (including 45 anchor items)

for evaluating equating accuracy. The manipulation of the content representation of test items via item sampling enabled this study to investigate the effect of content homogeneity and content representativeness on the accuracy of common-item test equating. The use of multiple equating methods and the availability of multiple criteria for evaluating equating accuracy allowed this study to better estimate the effect of equating method on equating accuracy.

Two fundamental aspects of the study design -- the common-item equating design and manipulation of content representation (item-sampling schemes and their outcomes) are elaborated below. The equating methods and criteria for evaluating equating accuracy used are described in subsequent sections.

Common-Item Design for Equating

The two examinee groups taking the two alternate test forms were not formed by random selection or assignment. Therefore, equating was made possible by the common items embedded in the two test forms. The anchor item design for equating was also appropriate for this study because: (1) for the original test forms, the content of anchor items was made representative of the entire test, and (2) anchor items with the same wording were embedded in the alternate test forms at the same positions.

Manipulation of Content Representation

All items in the two original test forms were from one overall content domain for the medical specialty. The overall content domain, however, was mapped by 23 sub-content domains (represented by the 23 core content areas). Drawing on a hypothesis that items from these 23 sub-content domains differed somewhat in their content, despite the

fact that they were all from one overall content domain, this study employed various item-sampling schemes to manipulate the content representation of the sampled tests. Pooling together the items from the 23 sub-content domains, four subsets of items were drawn from the overall item pool. Each of the resulting four sampled tests had two alternate forms, and each pair of test forms shared a common set of anchor items. Although the sampled tests were made to have similar numbers of anchor items, the content representativeness of the anchor items varied across sampled test. Also, the tests were made to have items differed in their content homogeneity.

The four item-sampling schemes used in this study were simple random sampling (SR), equal-weight domain random sampling (EW), proportional-weight domain random sampling (PW), and purposeful sampling (PS). The various assumptions, the sampling procedures, and the sampling outcomes of these schemes are briefly described below. The measures taken to control the effects of test length and anchor length on equating outcome are also explained.

Simple random sampling. The SR scheme was based on the assumption that items from different core content areas did not differ substantially, since all of the items were from one single content domain relating to the medical specialty. This sampling scheme thus disregarded the existence of the 23 core content areas and randomly drew items from the overall item pool using a random number table. The overall item pool had three compartments: an anchor-item pool filled with the 145 anchor items that were embedded in both Book A and Book B, a unique-item pool with all the unique items that were included in Book A, and a unique-item pool with all the unique items from Book B.

To create the pair of sampled test forms, SR-A and SR-B (see Table 2), a set of 30 anchor items were first randomly sampled from the anchor-item pool. The sampled 30

anchor items were built into both SR-A and SR-B. Then, 30 unique items were randomly sampled from the Book A unique-item pool and built into SR-A, and another 30 unique items were randomly sampled from the Book B unique-item pool and built into SR-B. Consequently, the resulting sampled test forms had 60 items in each. Table 3 summarizes the sampling results of the SR scheme. As shown in Table 3, the sampled items in SR-A were from 19 core content areas, and the sampled items in SR-B were from 20 core content areas.

Equal-weight domain random sampling. The EW scheme assumed that the 23 core content areas represented equally significant parts of the overall medical content domain. Disregarding the number of items within an area, this sampling scheme randomly sampled three items from each of the 23 areas to form a sampled test form. For each area, ideally, two of the three sampled items should be anchor items. It was to account for the fact that there were more anchor items than unique items in the overall item pool, and to make sure that the anchor items in the sampled test forms were evenly drawn from the 23 core content areas. Such ideal sampling was achieved for most of the areas, except for the few areas where (1) there were less than three items, or (2) there was not any unique item, or (3) there was only one anchor item.

Technically, the EW scheme was random sampling stratified on core content areas. After the first sampled test form was created in accordance to the above sampling condition, its anchor items were used as the anchor set for the second sampled test form. Unique items were then randomly sampled from various areas to make up the second sampled test form. The resulting sampled test forms, EW-A and EW-B (see Table 2), had 69 items in each, and shared a total of 49 common anchor items.

Table 3 - Number of Items Sampled from the Original Test by Simple Random Sampling

Core Content Area	Sampled Test Form	
	SR-A	SR-B
1	3	3
2	5	3
3	2	2
4	3	4
5	2	2
6	8	7
7	2	3
8	0	0
9	1	2
10	4	3
11	2	2
12	0	0
13	2	2
14	4	3
15	0	0
16	1	1
17	6	4
18	8	8
19	2	4
20	1	2
21	2	2
22	2	2
23	0	1
Total =	60	60

Proportional-weight domain random sampling. The PW scheme assumed that the size of a core content area in the original item pool reflected its significance. Therefore, it randomly drew from each of the 23 areas a number of items proportional to the total number of items in that area. Using the proportional distribution presented in Table 1, the number of items to be sampled from an area for a 60-item test form was calculated. Table 4 summarizes the calculation results. Each of the resulting test forms, PW-A and PW-B (see Table 2), of the PW scheme had 60 items. In addition, PW-A and PW-B shared a total of 40 common anchor items.

Purposeful sampling. The PS scheme included only the items from the largest three core content areas, assuming that the number of items in a core content area reflected the importance of the particular content, but simplifying to focus only on the 3 most important areas. Fewer core content areas in a test lends more confidence to the homogeneity of its test items or the IRT assumption of unidimensionality of the test.

The PS scheme resulted in two sampled test forms that shared 45 common anchor items. The two forms, PS-A and PS-B (see Table 2), had 15 and 12 unique items respectively.

Controlling test length. As mentioned earlier, most of the eight sampled test forms were composed of 60 items such that the effect of test length on equating outcome was hold constant over various sampled tests. However, the PS scheme deliberately included all the items from the largest three core content areas only, which inevitably resulted in a slightly shorter test with 57 items for PS-B. In addition, because the EW scheme called for equal number of items from all the 23 areas, both EW-A and EW-B ended up with 69 items, a few more than 60. These small differences in test length, nevertheless, should not be serious.

Table 4 - Number of Items Sampled Using Proportional-Weight Domain Random Sampling

Core Content Area	Proportion of Test Items (%)	Number of Items Sampled
1	5.8	5.8 * 60 = 3.48 ≅ 4
2	10.2	10.2 * 60 = 6.12 ≅ 6
3	1.3	1.3 * 60 = 0.78 ≅ 1
4	6.2	6.2 * 60 = 3.72 ≅ 4
5	2.2	2.2 * 60 = 1.32 ≅ 1
6	8.4	8.4 * 60 = 5.04 ≅ 5
7	2.2	2.2 * 60 = 1.32 ≅ 1
8	1.3	1.3 * 60 = 0.78 ≅ 1
9	2.7	2.7 * 60 = 1.62 ≅ 2
10	4.0	4.0 * 60 = 2.40 ≅ 2
11	3.6	3.6 * 60 = 2.16 ≅ 2
12	1.8	1.8 * 60 = 1.08 ≅ 1
13	5.8	5.8 * 60 = 3.48 ≅ 4
14	3.1	3.1 * 60 = 1.86 ≅ 2
15	2.2	2.2 * 60 = 1.32 ≅ 1
16	6.7	6.7 * 60 = 4.02 ≅ 4
17	5.8	5.8 * 60 = 3.48 ≅ 4
18	11.1	11.1 * 60 = 6.66 ≅ 7
19	3.6	3.6 * 60 = 2.16 ≅ 2
20	2.2	2.2 * 60 = 1.32 ≅ 1
21	1.8	1.8 * 60 = 1.08 ≅ 1
22	4.0	4.0 * 60 = 2.40 ≅ 2
23	4.0	4.0 * 60 = 2.40 ≅ 2
Total	100.0	60

Controlling anchor length. A previous study using the same set of data found that equating accuracy depended on the number of anchor items in the test forms being equated. Specifically, equating results from test forms that had more anchor items tended to be more accurate (Yang & Houang, 1996). To hold the effect of anchor length on equating accuracy constant over the various sampled tests, the anchor tests in various sampled tests therefore were fixed to be sufficiently lengthy. For any of the eight sampled test forms, at least half of the items included were anchor items (see Table 2).

Although the anchor lengths of all the sampled test forms were made sufficiently long, it was difficult for this study to sample test forms that all had the same number of anchor items. Mainly, this was because the number of items available for item sampling in some core content areas was limited. The types of these items (anchor or unique) further set limits for the results of item sampling. For instance, several core content areas failed to adequately support the EW scheme, because there were not enough items or items of certain type to yield ideal item-sampling outcomes. The outcome of the PS scheme is another good example to illustrate such difficulty in controlling anchor length. The PS scheme required an inclusion of all the items from the largest three core content areas, and it turned out that 75% of these items were anchor items.

However, by incorporating sufficiently large number of anchor items in all the sampled test forms, the effect of differential anchor lengths in this study became less likely to confound with the effect of content homogeneity (or content representativeness of anchor items). Though it might be intriguing to study the interaction of the anchor-length effect on equating accuracy in conjunction with the content-representativeness effect, such sidetrack was avoid in this research to keep the current study focused on the intended issues. The study design and interpretations of this study also remained straight forward.

Equating Methods

To investigate to what extent the results of various equating approaches agree, this study estimated and compared the effects of linear, equipercentile, and IRT-based equating approaches on the accuracy of equating. As reviewed in previous chapters, linear equating methods are straightforward and convenient in computation (Kolen & Brennan, 1987), but their results do not always meet all criteria for equivalent tests. Equipercentile equating is a frequently used non-linear equating approach, which is still based on observed score and has been known to have accurate results (Hills, Subhiyah, & Hirsch, 1988; Yen, 1985). Gaining their popularity in recent years, equating models based on item response theory have been found useful for equating based on the common-item design (Cook & Eignor, 1991; Crocker & Algina, 1986).

Tucker Linear Equating Method

This study chose Tucker Linear equating method to represent the linear equating approach. The Tucker method is one of the conventional (non-IRT) equating methods used most frequently. Under the assumptions of the Tucker method, the linearly transformed scores on one form will have the same mean and standard deviation as the scores on another form (Kolen & Brennan, 1987; Kolen & Brennan, 1995). Although the method requires equal reliability across test forms being equated, the impact of unequal reliability in this study was not very critical because each pair of the sampled test forms being equated were very similar in their content and had the same number of items (Kolen & Brennan, 1987).

Frequency-Estimation Equipercentile Method

Based on observed score, equipercentile equating aims at finding a score on Form Y of a test that has the same percentile rank as a score on Form X of the same test (Kolen & Brennan, 1995). The method used in this study to conduct equipercentile equating was the commonly used frequency estimation method (Kolen & Brennan, 1995), incorporated with the cubic spline postsmoothing method (Kolen & Jarjoura, 1987; Kolen & Brennan, 1995). The frequency estimation method was used because the two examinee groups were not too different. The cubic spline smoothing method was used to increase equating precision with frequency estimation method, for the common-item non-equivalent group design (Kolen & Jarjoura, 1987).

IRT-Based Equating Methods

The literature review in Chapter 5 has made clear that IRT-based equating methods have both theoretical appeal (Green, Yen, & Burket, 1989) and practical advantages (Cook & Eignor, 1983, 1991). For instance, IRT-based equating methods were found to be more useful than linear equating methods when tests to be equated differed somewhat in content and length (Petersen, Cook, & Stocking, 1983). However, IRT-based equating results are often inconsistent, and the practical significance of their improvement in estimating equating accuracy remains unclear. Therefore, this study selected two IRT-based equating methods that differed in their assumptions and procedures to do the comparisons. The results of these two methods were not only compared to the results of linear equating and equipercentile equating methods, but also compared to each other to study the relationship between IRT-based equatings.

Namely, the two IRT-based equating methods used in this study were the linear transformation method (Hambleton & Swaminathan, 1990; Kolen & Brennan, 1995) and the IRT fixed-b method (Petersen, Cook, & Stocking, 1983; Hills, Subhiyah, & Hirsch, 1988). As reviewed earlier, both methods were based on a three parameter logistic IRT model, which accounts for guessing (Hambleton & Swaminathan, 1990; Hambleton, Swaminathan, & Rogers, 1991). Guessing is likely to occur with multiple-choice items. Since all the test items in this study had multiple-choice item format, it was likely that examinees guessed on some difficult items.

Under the IRT assumptions of item and person invariance, linear transformation is reasonable for the non-equivalent-group anchor-item design of this study. This is because the difficulty and discrimination parameters for the common anchor items from alternate test forms are linearly related (Petersen, Cook, & Stocking, 1983; Hills, Subhiyah, & Hirsch, 1988). The fixed-b method is often used with LOGIST program (Hills, Subhiyah, & Hirsch, 1988). However, one previous study found that the fixed-b method also worked nicely with BILOG program (Yang & Houang, 1996; Yang, 1997). It was found that equating results yielded by the fixed-b method using BILOG were consistent with the results from IRT-based linear transformation method.

Criteria for Evaluating Equating Accuracy

By applying a variety of plausible criteria for evaluating equating accuracy, applicability of accuracy criteria could be explored in compensation for the lack of an absolute evaluation criterion. In this study, four criteria were developed for evaluating the accuracy of various equating results. Two of the criteria for evaluating equating accuracy were based on estimated true scores and the other two were arbitrary criteria based on

different equipercentile equating outcomes. The natures and uses of these criteria are described below.

True-Score Based Criteria

In this study, items were sampled from one overall item pool to form four pairs of sampled test forms. As a result, examinee performances on the complete set of the 145 common anchor items in the item pool could be regarded as the "anchor universe", relative to the anchor items embedded in the sampled test forms. "Pseudo true scores", the estimated true scores based on this "anchor universe", could thus be computed and used as criteria for evaluating equating accuracy.

The "pseudo true score" was estimated in two ways: (a) using the total raw score on the 145 anchor items, and (b) using the IRT estimated true score on the 145 anchor items. Although the raw-score based criterion had the drawbacks of being person and item dependent, it was conceptually superior to the IRT-based criterion because it was not biased in favor of the IRT-based equating. Nonetheless, the IRT-based criterion had advantages of being item-free and person-free.

The accuracy of equating result was expressed by Pearson's product-moment correlation coefficient (r) between true score estimates and the "pseudo true scores". A larger positive r would indicate a more accurate equating result. To evaluate the accuracy of the IRT-based equating outcomes, first, total true scores were estimated based on the IRT-based equating outcomes. Then, the true score estimates were correlated to the "pseudo true scores". To evaluate the accuracy of the non-IRT equating outcomes, the resulting equivalent scores were correlated to the "pseudo true scores".

Arbitrary Criteria

In addition to the two true-score criteria based on the 145 anchor items, the result of equipercentile equating on the sampled test forms was used as an arbitrary criterion to evaluate relative accuracy of the results produced by the other three equating methods (the Tucker method and the two IRT-based methods). Both the Pearson's *r* and the RMSD (Klein & Jarjoura, 1985; Livingston, Dorans, & Wright 1990) were computed to measure the agreement between this criterion and the results of the other three equating methods. The reason that the result of equipercentile equating method was chosen to represent the perfect equating outcome is because this method usually produced satisfactory results in the literature (Yen, 1985). In equating practice, because true equivalent relationship can never be known, it is common to compare equating results to such arbitrary but sound criteria for estimating equating accuracy (Dorans & Kingston, 1985). However, by using such an arbitrary criterion, the estimation of equating accuracy was subjected to the nature and quality of this arbitrary criterion. The strong assumption that the outcome of the equipercentile equating method is the true equating outcome may not be real. Moreover, the agreement between such a criterion and the results of the other three equating methods did not directly address the issue of equating accuracy. Therefore, other than being used to evaluate equating accuracy, this criterion was used to investigate the potential bias due to the use of an arbitrary criterion for evaluating equating accuracy.

The two original test forms, Book A and Book B, were also equated by the equipercentile method in this study. This equating result for the original test forms was used as the fourth criterion to evaluate the relative accuracy of the equating results produced for the sampled test forms by all of the four methods (Tucker linear, equipercentile, and the two IRT-based methods). The Pearson's *r* was computed to

measure the agreement between this criterion and the results produced by all the four equating methods. This criterion for evaluating equating accuracy had the same arbitrary nature as the previous criterion. It did not directly address the issue of equating accuracy, either. However, this criterion was considered more reliable because it was based on the equating results for the longer original test, which was similar to the sampled tests in content but had proportionally more anchor items.

Test Dimensionality

Issues of test dimensionality are relevant to the IRT-based equating in this study, since the IRT model used for the IRT-based equating assumed unidimensionality. These issues are also relevant to classical equating, which assumes unidimensionality in an implicit way. In return, such unidimensionality assumption of classic equating limits its use (Green, Yen, & Burket, 1989).

From the perspective of the single overall content domain, the test analyzed in this study was likely to be unidimensional. One could argue that the test performance was dominated by an underlying trait relating to the medical specialty. However, the variations in the content of items from the 23 core content areas also rendered chances for multidimensionality and raised reasonable doubts about the claim of unidimensionality. It also seemed likely that items from different areas measured different underlying traits. Thus, it was intriguing to know whether the assumption of unidimensionality was realistic for this equating study, and how robust this assumption was, given it was violated.

Statistically, the IRT unidimensionality assumption could only be tested indirectly by assessing the local independence (Lord, 1982b). As a condition for IRT-based equating, however, local independence remains a matter of assumption itself. It requires

that for fixed ability θ, the item characteristic functions for any pair of items i and j should be independent. Therefore, focusing on content homogeneity, this study assessed the dimensionality of the test by conducting confirmatory and exploratory factor analyses. Four confirmatory-factor models were developed by assuming various plausible content structures for the test. In addition to the empirical evidence from statistical analyses, arguments made about dimensionality should also take into account test content (Camilli, Wang, & Fesq, 1995). Therefore, this study also elaborated on theoretical literature to investigate the relationship between the content of the items from the different core content areas and the overall content of the test.

Although this study intended to investigate the dimensionality for the original test and all the sampled tests created in this study, such comprehensive plan, however, were neither realistic nor efficient. For an equating study that had already been complex in its design and labor-consuming, it was more practical to limit this part of study to a more manageable scope such that the analysis tasks and outcomes would not be overwhelming. Therefore, this study chose to focus on only a set of sampled items to study test dimensionality. Nevertheless, the narrowly focused investigation would still reasonably address the issue of dimensionality by casting useful insights on the dimensionality of the overall and sampled tests.

Group Disparity

The study outcome of test dimensionality was likely to be affected by the disparity between the non-equivalent groups. If the two examinee groups indeed possessed different types or amounts of knowledge and skills, one single underlying trait was not likely to account for the test performance for the examinees from these two groups. To

facilitate the study of dimensionality, it became important to study the group disparity. In fact, group disparity might occur because it was possible that examinees from different in-training programs at different sites received somewhat differential instructions and practices. Despite the fact that the overall core content was logically the basis of the in-training curricula, the directors of individual in-training programs had the authority to decide which part of the core content their programs were to emphasize. As a consequence, the instructions or practices of different programs might vary slightly and thus cause the disparity among the examinee performances.

To determine the adequacy of equating methods, which usually required groups with same or similar ability, it was also important to examine the degree of disparity between the non-equivalent groups. The need was especially true in this study because the examinee groups taking the two test forms were not randomly sampled or assigned. Therefore, available demographic data, such as program participation and years of experience, were analyzed in this study to help determine the degree of disparity.

Construct Validity Issues

In this study, an examinee's professional ability was in part dependent on the examinee's professional experience. Logically, the more years of professional experience, the more likely an examinee would score higher on the test. If the test forms of the test were truly equated, the resulting equivalent scores of both examinee groups taking different test forms should demonstrate such effect of professional experience. Therefore, this study compared the average equivalent scores of the examinee groups after equating to study the construct validity of the test.

For the sake of completeness and convenience, this study used the equivalent scores produced by frequency-estimation equipercentile method for the original test to do the group comparisons after equating. This set of equivalent scores was complete because it involved all the items in the original test, and it was convenient because it already had been made available earlier in the study for equating accuracy. Specifically, the construct validity of the test was studied by investigating the effects of test form, years of experience, and their interaction on the examinees' performance.

Research Tools

The IRT calibration program chosen for the analyses of this study was the advanced version of PC BILOG, BILOG-MG. One advantage of using BILOG is that BILOG yields marginal maximum likelihood (MML) estimates and the number of parameters estimated does not increase with the increasing number of examinees. Compared to BILOG, LOGIST simultaneously maximizes the joint likelihood function for the estimates of item and examinee parameters. However, the joint maximum likelihood (JML) estimates are likely to become inconsistent when the numbers of examinees or items increase (Baker, 1990; Mislevy & Stocking, 1989). Consequently, BILOG should yield more consistent results than LOGIST in such a situation.

PC-BILOG uses the estimated posterior θ distribution to establish the location and metric for the θ scale (Baker, 1990). In an earlier simulation study, Yen (1987) found that BILOG largely yielded more precise estimates of individual item parameters than LOGIST. She also expected the improvement in accuracy of BILOG to increase if sample size decreased substantially. In terms of estimating item and test characteristic functions, Yen (1987) found that the relative effectiveness of the two programs depended on test

length. Specifically, BILOG yielded more precise estimates than LOGIST on equating shorter test forms with ten items; However, the two programs yielded very similar estimates when longer tests with 20 and 40 items respectively were equated. Mislevy and Stocking (1989) also found that BILOG would yield more reasonable results when tests are shorter or the samples are smaller. Based on a Bayesian framework, BILOG imposes prior distributions on all item parameters of the 3PL model. If the prior information is not appropriate for the data, item parameter estimates are biased (Baker, 1990).

In addition to BILOG, SAS for Unix and Excel spreadsheets were also used in this study to assist with various equatings, as well as data management and other statistical analyses. The equipercentile equatings were facilitated by an extended version of the Common Item Program for Equating (CIPE) (Hanson, Zeng, & Kolen, 1995), which uses the frequency estimation method described by Kolen and Brennan (1995). The extended CIPE program--CIPE300 Plus is written in FORTRAN and has the capacity to handle long test forms with more than 200 items. It uses the cubic spline method (Kolen & Jarjoura, 1987; Kolen & Brennan, 1995) to post-smooth the resulting equipercentile relationship. Up to eight user-specified smoothing parameters are allowed to manipulate the degree of smoothing (Hanson, Zeng, & Kolen, 1995). Essentially, the smoothing parameter controls the average squared standardized difference between the smoothed and the unsmoothed equating outcomes. After the frequency-estimation equipercentile equatings were applied, the CIPE output of the unsmoothed equivalents and their corresponding standard errors were graphed using Excel, along with the other sets of smoothed equivalents yielded by the various smoothing parameters. The various graphs for the smoothed equivalents were inspected for their smoothness and compared to the unsmoothed equivalents. In addition, to evaluate the smoothing requirement of "moment

preservation" (Kolen & Brennan, 1995), the four moments -- mean, standard deviation, skewness, and kurtosis -- for the entire examinee population of the unsmoothed and smoothed equivalents were also computed using SAS. The moments of the smoothed equivalents were compared to the unsmoothed equivalents to identify the best smoothing parameter that yielded a smooth function not departing too much from the unsmoothed equivalents.

Research Restrictions and Limitations

The rich context of the test data analyzed in this study rendered opportunities for item sampling and data manipulation. With such advantages, this study was able to address a variety of research questions in depth. In addition, the complexity of data enriched the study design and helped expand the scope of research. However, the secondary nature of the data also restricted the study design in some ways and limited the generalizability of the study results. Restrictions on this study and limitations of the study results, caused by the data and the design used to accommodate the data, are briefly described below.

Data Manipulation

The secondary data used in this study was limited in a sense that it was collected before the study design was conceived. As a consequence, any manipulations before or during the data-collection process were not possible. The design of this study was therefore restricted by the nature of this secondary data. Typical consequences of such restrictions, and the consequent study limitations due to these restrictions, are discussed below:

Common-item equating design. The test forms of the in-training test were linked by a set of common anchor items, embedded in the forms and given to non-equivalent examinee groups. Therefore, the anchor-item design for equating was the only option for equating the test forms in this study.

Long anchors. Most of the items in the test were anchor items. As a result, the item sampling in this study naturally resulted in sampled test forms (PS-A, PS-B, PW-A, PW-B, EW-A, and EW-B) containing more anchor items than unique items. Given such long anchor tests, it was difficult for this study to evaluate the accuracy of equating under a situation where there were only few anchor items. Such restriction of long anchors might have caused the study result that -- all the equating methods yielded similarly satisfactory equating outcomes. Given the long anchors, all of the equating methods were likely to yield accurate outcomes. As a consequence, true differences among the equating accuracy of the various methods could not be detected or differentiated.

Unequal test length and anchor length. The number of items available for item sampling in each of the 23 core content areas was limited. In addition, the various item-sampling schemes had different demands in the types and amounts of test items. As a result, it was difficult for this study to obtain sampled test forms that all had the same number of items or anchor items. Although this study intentionally created sampled test forms that had similar test lengths and ensured that each sampled test form had a sufficient number of anchor items, there was still a slight chance that the equating results were under influence of the differential anchor lengths.

If the effect of anchor length existed indeed, such effect was likely to confound with the effect of content homogeneity and the effect of content representativeness of anchor items. Therefore, the study findings about the effects of content homogeneity and

content representativeness should be interpreted with special cautions for the confounding effect due to differential anchor lengths.

Interpretation of Equating Accuracy

The criteria used for evaluating equating accuracy and the index incorporated in this study to measure the accuracy of equating had inherent limitations. These limitations are summarized below.

Arbitrary nature and limited use of criteria. The arbitrary nature of the two criteria for evaluating equating accuracy based on the results of equipercentile equating method was self-evident. The major drawback of using these criteria was that they did not address equating accuracy directly. Only the consistency between the criteria and the results of the other equating methods were measured. Therefore, the evaluation outcomes based on these two criteria should be interpreted with cautions.

The other two criteria used in this study were based on the "pseudo true scores". They were only appropriate when the examinee population and the testing occasion were considered fixed. It was because the "pseudo true scores" estimated the true scores, at a particular point of time and for the particular examinee population of this study, on the complete set of the 145 common anchor items in the overall item pool. As a result of assuming such "pseudo true score" was the true score, the two criteria were only valid in an equating context where the examinees were from a population same as or similar to the one in this study and tested under a circumstance same as or similar to the one in this study.

Issues of Auto-correlation. The Pearson's *r* used to estimate the accuracy of equating was inflated by the artifact of auto-correlation. The auto-correlation was caused

by the overlapping of items from the sampled test and the items from the criterion used for evaluating equating accuracy. By excluding the overlapped items from the sampled test and then correlating the remaining items with the criterion, this study attempted to control the influence of the auto-correlation. The magnitude of the resulting Pearson's *r*, after controlling the auto-correlation, was used to measure the impact of the auto-correlation and to determine whether the problem of over-estimating equating accuracy was substantial. This strategy for controlling auto-correlation, however, did not completely eliminate the influence of the auto-correlation. Despite the difficulty, nevertheless, the strategy was still a useful way for improving the study on the effectiveness of various equating methods and the effect of content representativeness on equating accuracy.

Generalization of Results

Various sources of limitations on generalizing the study results are discussed below. They include the characteristics of the test items, the test, the groups of examinees, and the particular equating models used.

Item format and scoring system. In this study, all the test items had multiple-choice format and were dichotomously (right or wrong) scored. Logically, the research findings based on these items should not be generalized to an equating context where test forms have non-multiple-choice items (e.g., short-answer items or extended-response items) or a non-dichotomous scoring scheme (e.g., a partial credit system) is used.

Test items of different format are likely to induce differential examinee responses, therefore, they usually require different scoring keys or rubrics to provide adequate interpretation for the examinee responses. Equating involving non-dichotomously scored items entails different equating models with different assumptions, other than the ones

used in this study. However, the featured design of this study, such as the manipulation of content representativeness and the use of the true-score based criteria for evaluating equating accuracy, are useful in improving the design of other research involving items of different format or scoring system. The results and findings of this study also cast useful insights for equating context with slight variations.

Skewed score distributions. One important feature of the test data analyzed in this study is that the test was written for a minimum competency examination. The test therefore had a negatively skewed score distribution, since the examinees were expected to answer most of the items right. The four sampled tests also had negatively skewed score distributions. The study results based on such skewed score distributions should be carefully generalized to other testing situations that have similar score distributions.

Characteristics of examinee populations. The particular examinee population studied in this research also limited the generalization of the study results. As this study focused on a group of professionals in a medical field from a number of in-training programs, the results of this study should not be generalized to the other subject populations that differ from the one in this study.

Assumptions of equating models. The IRT-based equatings in this study assumed unidimensionality for the test forms being equating. Therefore, the equating results should not be generalized to other testing contexts where multi-dimensionality prevail. In addition, because the equatings were based on a 3-PL IRT model, which accounted for the chance of guessing, generalization of the study findings should be limited to the contexts where the 3-PL model applies. Similarly, generalization of the result from any other equating method should also take into account the particular assumptions made by that method.

Chapter 9

RESULTS AND DISCUSSIONS

To facilitate an inspection on characteristics of the four sampled tests, this chapter first summarizes the outcomes of reliability studies, item analyses, and correlation studies among the total scores on anchor items, unique items, and total test. Then, an examination of the examinee group differences is presented. Before presenting the main equating outcomes, the score distributions of various test forms are discussed. Adequacy of the 3PL IRT model, on which the IRT-based equatings are based, is also discussed.

Intermediate and final equating outcomes are presented and discussed in the following order: (1) results from various equating methods, using a raw-score-based true-score criterion, (2) equating results yielded by an IRT-based true-score criterion, (3) equating results produced by a criterion based on the outcome of the equipercentile method on equating the two original test forms, and (4) results yielded by an arbitrary criterion that was based on the outcomes of the equipercentile method on equating sampled test forms. These results are compared to explore the effectiveness of various criteria for evaluating equating accuracy.

At the end, this chapter addresses important issues relevant to the adequacy of test equating and the assumption of IRT-based equating. These issues relate to the construct

validity and dimensionality of a test. Investigation outcomes for the construct validity of the original test, and the adequacy of the method used to equate the two original test forms, are summarized. Empirical results and theoretical elaboration on issues of test dimensionality are also presented and discussed.

Characteristics of Tests and Examinee Groups

Upon inspecting the characteristics of the original and sampled tests: (a) reliability of these tests were studied by measuring their internal consistency, (b) item difficulties and item-total correlations were computed and examined, and (c) content homogeneity of the items in the same test, and content representativeness of anchor items in a test, were addressed. Also evaluated was the possible ability difference between the two examinee groups and its influence on the test equating results.

Internal Consistency of Tests

The reliability of internal consistency, measured by Cronbach's α is .894 for both original test forms. This indicates adequacy of the in-training test. All of the sampled test forms created by item sampling also show internal consistency. As shown in Table 5, the values of Cronbach's α ranges from .658 to .774 across sampled test forms. Although the numbers seem small, these indices suggest moderate reliability for these achievement tests. Typically, an achievement test has less emphasis on item homogeneity than an attitude or personality questionnaire does. In addition, an ability test usually has items that fall within wider ranges of item difficulty and item discrimination than an attitude questionnaire does.

Table 5 - Reliability of Sampled Test Forms:
Indices of Internal Consistency

	Sampled Test Form	Cronbach's α
Simple Random Sampling	SR-A	0.658
	SR-B	0.713
Equal-Weight Domain Random Sampling	EW-A	0.684
	EW-B	0.690
Proportional-Weight Domain Random Sampling	PW-A	0.662
	PW-B	0.691
Purposeful Sampling	PS-A	0.774
	PS-B	0.768

Internal consistency of an achievement test is therefore often lower than that of an attitude questionnaire, because of the lower correlations among individual items.

The measures of internal consistency for the sampled test forms are smaller than those measures for the original test forms. This is partly because there are fewer items in the sampled test forms. Comparing across the eight sampled test forms (see Table 5), as expected, the two forms based on the purposeful sampling have the highest internal consistency. The Cronbach's α is .774 for PS-A and .768 for PS-B. It is because all the items in PS-A or PS-B are from only three core content areas. The other six sampled test forms have similar internal consistency. Moreover, the two test forms from the same test are similar in their internal consistency. This provides some justification for using the Tucker linear method. As discussed in the section for equating methods in Chapter 8, the Tucker method requires equal reliability of test forms being equated.

Analyses of Item Difficulty

The results of item analyses, including analyses of item difficulty and item-total correlation, provide useful information on the sampled tests and their test items. The classical item difficulty (p) of item j, in a test taken by a group of N examinees, can be technically defined as $p_j = \frac{Number\ of\ examinees\ with\ a\ correct\ response\ on\ item\ j}{N}$.

In words, difficulty of an item is the proportion of examinees that answer an item correctly (Crocker & Algina, 1986). Analyses of item difficulty revealed that the items in all four sampled tests had moderate difficulties on average. Across various test forms, the average item difficulty range from 0.688 to 0.759.

In Figure 1a, individual items in various sampled test forms are sorted by their item difficulties in ascending order. The patterns of the graphed item difficulties for the two forms of each sampled test are similar, and the patterns across various sampled tests are also alike. Different from reporting average item difficulty, the graphs in Figure 1a give a closer look at the variations in item difficulty within and across test forms. They better describe how these sampled test forms resembled or differed from one another. Overall, they suggest that the sampled test forms had similar difficulties. For each of the test forms, the difficulties of the items spread rather evenly between 0.4 to 1. Such range of item difficulty (from medium to high) is expected, because it is typical for a test written for a minimum competency examination.

Figure 1b presents the cumulative frequency distributions of item difficulties. Items within a test form were sorted into ten intervals by their difficulties for a summary of the item-difficulty distribution of that test. The resulting distributions in Figure 1b show that all the sampled test forms had easy, moderate, and difficult items. Across test forms, the distributions look similar. It suggests that the sampled tests were similar in their difficulties, as suggested by Figure 1a.

Figure 1b also presents the mean and standard deviation of item difficulties for each sampled test form. It also includes significance test results for the differences between average item difficulties of each pair of sampled test forms. The equal-variance two-tailed Student's *t*-test was used to examine such mean differences. For all of the four sampled tests, none of the mean differences was statistically significant at the .05 level. These small and statistically non-significant differences attest to the adequacy of the sampled test forms for equating study. Typically, equating is only used to adjust for minor

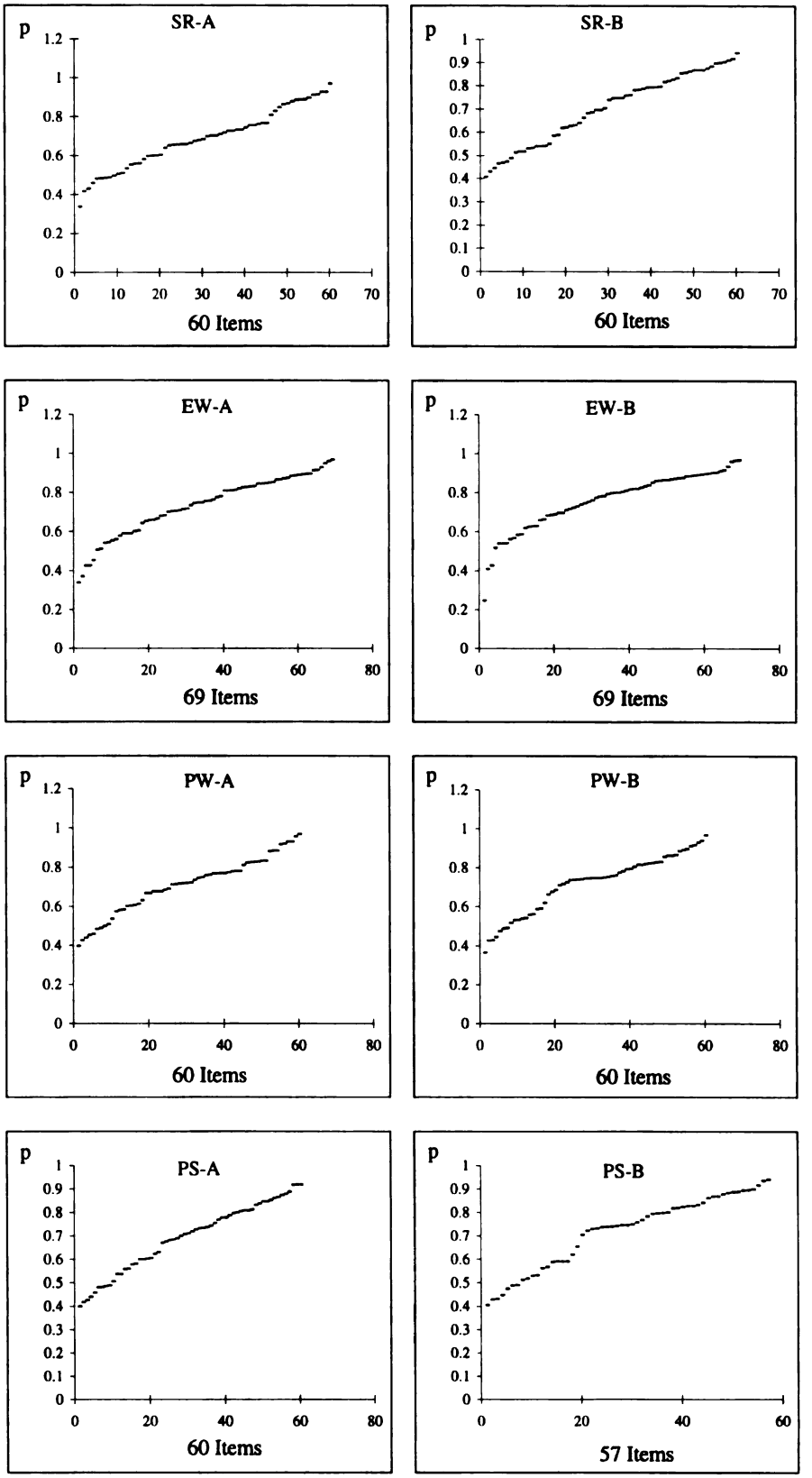


Figure 1a - Item Difficulty (p) for Items in Sampled Test Forms (in Ascending Order)

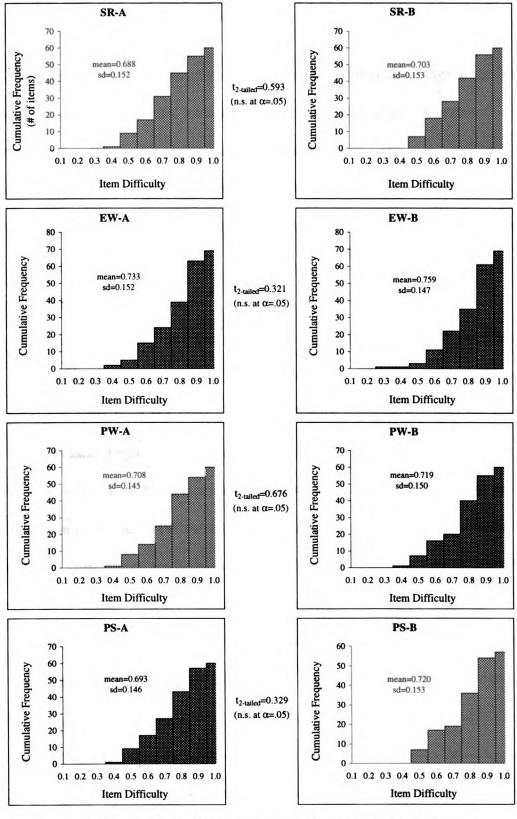


Figure 1b - Cumulative Frequency Distributions of Item Difficulty for Sampled Test Forms

differences in item difficulty between different forms of a sampled test (Cook & Eignor, 1991).

Specifically, the standard deviations of the item difficulties range from 0.145 to 0.153 over various test forms, indicating similar item difficulties. Combined with the patterns found in the distribution plots, these standard deviations suggest that the item difficulties for various sampled test forms spread in a similar way. The small standard deviations also suggest that the difficulties of the items within a form were not far apart.

Analyses of Item-Total Correlation

For each sampled test form, the items generally correlated positively and moderately to their total test. This provides evidence of homogeneity (in examinees' responses) for items from the same test form. Figure 2 presents graphs for item-total correlations in ascending order for various sampled test forms. For each form, most of the item-total correlation coefficients spread between .10 and .50. The average item-total correlations range between .199 to .293 across various forms. The standard deviations of the item-total correlations are between 0.103 and 0.126. These figures suggest that the overall patterns of item-total correlations were not too different across test forms.

The average item-total correlations for sampled test forms (SR-B, EW-B, PW-B, and PS-B) created from the original Book B were higher than those for their counterparts sampled from the original Book A. The equal-variance two-tailed Student's *t*-test was used to examine the mean differences for each pair of the sampled test forms. The significance test results show that none of the mean differences was statistically significant at the level of .05. For each sampled test, its two forms had similar item-total correlations.

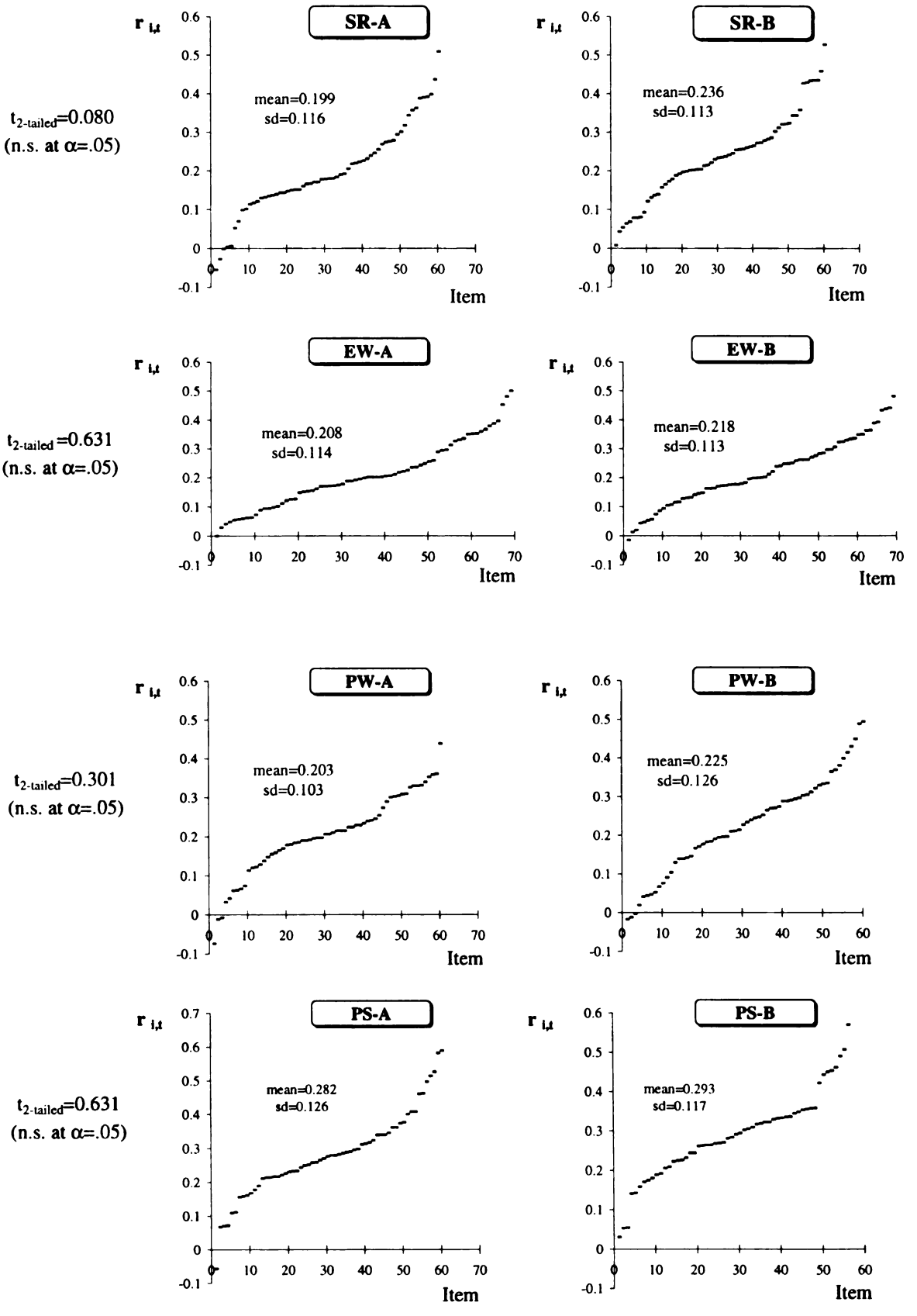


Figure 2 - Item-Total Correlation ($r_{i,t}$) for Sampled Test Forms (in Ascending Order)

Evidence of item sampling effect. Among the eight sampled test forms, PS-A and PS-B had the highest average item-total correlations (see Figure 2). This is because the two forms were created by purposeful sampling, which sampled items from only three out of the 23 core content areas. As a result of the sampling scheme, items in PS-A or PS-B were expected to correlate with one another to a higher degree than those items in the other six test forms. The largest average item-total correlations of PS-A and PS-B provide evidence for the effect of item sampling.

Test forms EW-A, EW-B, PW-A and PW-B all have items from each of the 23 core content areas. This fact could have contributed to the smaller average item-total correlation coefficients for these forms (than those coefficients for PS-A and PS-B). Also, it explains the similarities between the average item-total correlations of sampled tests EW and PW. It is plausible that the variations in average item-total correlation across various test forms are in part due to the four item sampling schemes, incorporated to the study design. These item-sampling schemes successfully manipulated the content homogeneity (or heterogeneity) of items in the sampled tests.

Inspecting anomaly cases. As shown in Figure 2, some sampled test forms had few items that correlate negatively to their corresponding total tests. These anomaly cases were identified and their item-total correlations (for the original and sampled test forms) were examined. Overall, the examination results show that, across various sampled test forms, the anomaly cases were not always the same items. Anomaly items in the original and sampled test forms are listed in Table 6.

Table 6 - Items with Negative Item-Total Correlations

Book A: (#27), (#52), #95, #206
SR-A: (#27), (#52), #78 EW-A: (#116) PW-A: (#27), (#52), #70 PS-A: #78
Book B: #66, #139
SR-B: None EW-B: #189 PW-B: #66, #139, #143 PS-B: None

Note. Items in "()" are anchor items.

As shown in Table 6, while some anchor items correlated negatively to one form of a test, they did not correlate negatively to the other form of the same test. For instance, anchor items #27 and #52 had negative correlation coefficients in SR-A, but they did not correlate negatively to SR-B. None of the items in SR-B had negative item-total correlations. Similarly, #27 and #52 had negative correlation coefficients in PW-A but not in PW-B. Item #116 had negative correlation coefficient in EW-A but not in EW-B. Therefore, to keep the set of 145 anchor items from the original test intact, these anomaly anchor items were kept for the subsequent equating studies. Fortunately, the magnitudes of these negative correlation coefficients were mostly less than .05.

Table 6 also shows that some of the anomaly items in sampled test forms were indeed anomaly items in the original test forms. Specifically, items #27 and #52 had negative item-total correlation coefficients in the original Book A, and they also had

negative coefficients in two of the sampled test forms (SR-A and PW-A). In addition, items #139 and #66 correlated to Book B negatively, and they also correlated to one sampled test form (PW-B) negatively. Based on the above findings, it may seem reasonable to exclude these anomaly items from the study. However, these anomaly items were not always anomalies across test forms. In addition, not all the anomaly items in sampled test forms correlated negatively to the original test forms. Items #70, #78 and #116 are examples for such case for sampled test forms PW-A, SR-A and EW-A respectively. Similarly, item #143 is an example for PW-B, and item #189 is an example for EW-B.

In part, the inconsistency in the anomaly cases across different test forms can be explained by the item sampling schemes used in this study. Such inconsistency can be part of the item-sampling outcomes, since the item sampling schemes manipulated content homogeneity of the sampled items in various test forms. Therefore, the second reason that the anomaly items were kept for further analyses is to maintain such effect of item sampling. It is determined that the impacts of the negative item-total correlations were not serious, since the negative correlations all had small magnitudes.

Characteristics of Anchor Items

The characteristics of anchor items in the original and sampled test forms were analyzed with different types of correlation coefficients. These analyses examined the relationship between the total score on anchor items and the total score on non-anchor (unique) items, and the relationship between the total score on anchor items and total test score. Their results are summarized in Table 7.

Table 7 - Coefficients of Correlations Between Total Scores on Anchor Items, Non-Anchor (Unique) Items, and Total Test

	Test Form	Correlation Coefficients		
		$r_{(anchor,unique)}$	$r_{(anchor,total)}$	$r_{(unique,total)}$
Original Test	Book A	.754**	.981**	.866**
	Book B	.736**	.979**	.859**
Sampled Test	SR-A	.503**	.861**	.873**
	SR-B	.537**	.863**	.889**
	EW-A	.439**	.939**	.721**
	EW-B	.488**	.939**	.758**
	PW-A	.443**	.924**	.752**
	PW-B	.482**	.925**	.778**
	PS-A	.486**	.968**	.690**
	PS-B	.451**	.968**	.660**

Note. **-Significance level less than 0.01 (two-tailed)

$r_{anchor,unique}$: Index of equating efficiency. Correlation between the total score on the anchor items and the total score on the unique items of a test ($r_{anchor,unique}$) provides a further check for test composition. In addition to providing empirical evidence of item homogeneity, this type of correlation can also be used to indicate efficiency for test equating. Budescu (1985) argued that the larger the $r_{anchor,unique}$ is, the more precisely the parameters will be estimated for the combined group in equating. Table 7 shows that all of the indices of equating efficiency for various test forms (see the first column of correlation coefficient) are statistically significant. Their values range from .44 to .54. This suggests that the anchor items and non-anchor items from the same test form were similar, and the common-item equating in this study would be efficient.

Moreover, the total scores on the unique items also correlate strongly and significantly with the scores on total test forms (see the third column of correlation coefficient in Table 7). This provides more evidence of adequate test composition.

$r_{anchor,total}$: Index of content representativeness for anchor items. Test scores on the anchor tests also correlate significantly with the total test scores to a considerable degree (see the second column of correlation coefficient in Table 7). The values of such correlations ($r_{anchor,total}$) range from .861 to .968 for various sampled tests. They show that the anchor items in the sampled tests were content representative of their total tests. However, these coefficients of $r_{anchor,total}$ were inflated by auto-correlation. The anchor test was correlated to itself when the $r_{anchor,total}$ was computed, since the anchor test was embedded in the total test. Despite the auto-correlation effect, from the perspective that these anchor items were an integral (inseparable) part of the total test, the coefficients of

$r_{anchor,total}$ still provides a sensible measure of equating efficiency (Budescu, 1985). Therefore, $r_{anchor,total}$ was used in this study, as an index of content representativeness for anchor test. Concerns about the influence of auto-correlation will be further addressed later in this section.

Evidence of item sampling effect. The various item sampling schemes used in this study intended to manipulate the content representativeness of the anchor items in the sampled tests. The differences in the $r_{anchor,total}$ across various sampled tests provide evidence of item-sampling effect for these schemes. Table 7 clearly shows that the various anchor tests were more or less representative of their corresponding total tests, despite the fact that all of the values of $r_{anchor,total}$ are large. This item-sampling effect improves the chance for the subsequent studies of the content-representativeness effect on equating accuracy to be valid.

As shown in Table 7, the $r_{anchor,total}$ decreases as the content specificity of the sampled test changes. The pattern of the changes for test form A is: from .968 (PS) to .939 (EW) to .924 (PW) to .861 (SR). The pattern for form B is: from .968 (PS) to .939 (EW) to .925 (PW) to .863 (SR). These two patterns show that test forms A and B had identical trends of decreases in $r_{anchor,total}$. Forms PS-A and PS-B had the most content-representative anchor items, because the purposeful item sampling scheme merely drew items from three core content areas. As a result, items in PS-A and PS-B were more homogeneous in their content, which logically led to the larger $r_{anchor,total}$.

Forms SR-A and SR-B had the least representative anchor items. This is because their randomly sampled items were from 19 and 20 core content areas respectively (see Table 3).

Not only the items in SR-A and SR-B were less content homogeneous, the content variations across items in SR-A and SR-B were also less predictable, due to the random sampling of items. The magnitudes of the $r_{anchor,total}$ for EW-A and PW-A were similar. For EW-B and PW-B, the magnitudes were also alike. They suggest that there were no substantial differences between the tests created by the equal-weight domain random sampling and the proportional-weight domain random samplings, in terms of the anchor-item content representativeness.

Cautions about auto-correlation and anchor-length effects. As mentioned earlier, $r_{anchor,total}$ was inflated by auto-correlation. Given same test length, a long anchor test would inflate the magnitude of $r_{anchor,total}$ more than a short anchor test.

As explained in the section of research limitations in Chapter 8, due to the limited availability of original test items in each of the 23 core content areas, and the different demands on test characteristics of the four item sampling schemes, the item-sampling design resulted in varying anchor lengths for different sampled tests. The percentage of anchor items was 50% for SR, 61% for EW, and 66% for PW. The two forms of PS had slightly different percentages: 75% for PS-A and 79% for PS-B. As a consequence of the differential anchor lengths, the impact of auto-correlation might be different on various sampled tests. Part of the stronger anchor-total correlations for PS-A and PS-B are attributed to their longer anchor lengths. Similarly, shorter anchor length partly accounts for the weaker correlations for SR-A and SR-B. Such anchor-length effect could affect the findings about the content representativeness of anchor items. Therefore, $r_{anchor,total}$ may not be sufficient in estimating how anchor items were representative of the total test.

Despite the limited empirical findings described above, the differential content representativeness for various anchor tests is in theory plausible. It is backed up by the particular content structure of the original test and the item-sampling design. In addition, although the anchor lengths were different across test forms, all of the anchor tests were controlled to be sufficiently long for test equating. The fact that all of the anchor tests are long in length is expected to lessen the impacts due to the differential lengths on equating accuracy. In each sampled test form, at least 50% of the items were anchor items (see Table 2). Such high percentage far exceeded the commonly recommended anchor lengths for adequate test equating (Angoff, 1984).

Inspecting Examinee Group Differences

The differences between the two examinee groups are studied with considerations of the ability, years of experience, and program participation of the examinees.

Ability differences. The examinee groups performed slightly differently on the anchor items. The group taking Book B (mean=107.721) scored slightly higher than the group taking Book A (mean=105.457) on the 145 anchor items. The difference between the group means was statistically significant at the .05 level (t=3.987, df=2,239, p=.0001). There were similar group differences across the four pairs of sampled test forms. Table 8 summarizes the statistical test results for the group mean differences on anchor items. Average item difficulties were computed for the anchor items and unique items separately to further examine the group differences. Table 9 summarizes these results. Across various sampled tests, there were slightly larger values of item difficulty on the anchor

Table 8 - Significance Test Results for Group Mean Differences on Anchor Tests

Test Form	n	Mean	s.d.	Mean Difference	t	df	p	Effect Size (d)
Book A	1092	105.457	13.767	2.264	3.987	2,239	0.0001	0.169
Book B	1149	107.721	13.113					
SR-A	1092	21.665	3.198	0.424	3.097	2,239	0.002	0.131
SR-B	1149	22.089	3.278					
EW-A	1092	37.595	4.585	0.658	3.493	2,239	0.0005	0.148
EW-B	1149	38.253	4.333					
PW-A	1092	28.810	4.112	0.536	3.106	2,239	0.002	0.132
PW-B	1149	29.346	4.041					
PS-A	1092	31.616	5.697	1.150	4.931	2,207.7	0.0001	0.209
PS-B	1149	32.766	5.321					

Note. All the t-tests are based on the assumption of homogeneity of variance, except for the t-test for the difference between forms PS-A and PS-B, which is based on unequal variances.

Table 9 - Average Item Difficulty of Anchor Items and Unique Items

Sampled Test Form	Anchor Items	Unique Items
SR-A	0.722	0.654
SR-B	0.736	0.670
EW-A	0.767	0.650
EW-B	0.781	0.705
PW-A	0.720	0.683
PW-B	0.734	0.690
PS-A	0.703	0.664
PS-B	0.728	0.691

Note. 1. The item difficulty, P_i, is defined to be the % of examinees getting item "i" correct, and the average item difficulty is:

$$\bar{p} = \frac{\sum_{i=1}^{n} p_i}{n}, \text{ where n is the total number of items}$$

2. In this study, 1,092 examinees took Book A, and 1,149 took Book B.

tests for the examinee group taking test form B than the values for the group taking test form A. It suggests small between-group differences in ability, as the *t*-tests in Table 8 have suggested.

However, the small differences between the two groups in ability across various tests might not have practical significance. As presented in Table 8, all the effect sizes for the group differences (ranging from 0.131 to 0.209) are relatively small, compared to their means and standard deviations. These group differences are attributed to the non-random selection or assignment of examinees in the examination.

Although the literature suggests that examinee-group disparity can be a threat to the accuracy of Tucker linear equating. In such case, Levine equally reliable method is often recommended (Kolen & Brennan, 1987). The Tucker equating method was used in this study because, in an earlier study of these data, Tucker method and Levine method yielded almost identical results.

Years of experience. The average years of experience for the group taking Book A was 1.981 (n=1091, s.d.=0.823), and the group taking Book B had 1.958 years of experience on average (n=1137, s.d.=0.822). Assuming equal variances (F=.093, p=0.76), a two-tailed *t*-test suggests that these two groups were not different in their experiences (*t*=.659, df=2,226, p=.510). The frequency distributions of years of experience by group, included in Table 10, further illustrate similarities between the two groups. At each level of "years of experience", the numbers of examinees from the two groups are similar. Also included in Table 10 is a 2×3 table (group by level of experience) containing sample counts and expected counts of examinees (N=2,228, after deleting cases with missing information on "years of experience"). The Chi-square test for this expectancy table

Table 10 - Group Comparisons on "Years of Experience"

Examinee Group	Years of Experience	n	%
I (taking Book A)	1	380	34.8%
	2	352	32.2%
	3	359	32.9%
	Unknown	1	0.1%
	Total	1092	
II (taking Book B)	1	409	35.6%
	2	367	31.9%
	3	361	31.4%
	Unknown	12	1.0%
	Total	1149	
I and II	1	789	35.2%
	2	719	32.1%
	3	720	32.1%
	Unknown	13	0.6%
	Grand Total	2241	

			Years of Experience			Total
			1	2	3	
Group	I	Count	380	352	359	1091
		Expected Count	386.4	352.1	352.6	1091.0
	II	Count	409	367	361	1137
		Expected Count	402.6	366.9	367.4	1137.0
Total		Count	789	719	720	2228
		Expected Count	789.0	719.0	720.0	2228.0

shows that the group membership of the examinees was independent of "years of experience" ($\chi^2(2) = 0.435$, p=.805). It suggests that the two examinee groups were similar in their experiences.

Program participation. The examinees participated in different in-training programs. Using information on examinee's program membership, Table 11 summarizes the program participation of the two examinee groups. Table 11 shows that the examinees taking Book A were from 64 in-training programs, and those taking Book B were from 62 programs. There were 109 programs in total. Most of the time, the examinees from the same program took the same test form. However, in 15 programs, all but one examinees took the same test form. In addition, in one program, all but 2 took the same test form, and in another program, all but 3 took the same test form. This is because when an examinee missed the scheduled testing time, that examinee was given a different test form at a different time.

There is a lack of Information on inter-program differences, such as different curricular designs and instructional methods. As a result, group differences originated in program variations can not be examined further. Nevertheless, the cumulative frequency distributions for the two groups, presented in Figure 3, provide some evidence of similarity between the two groups in program participation. The distribution plot shows that the examinees in both groups were from a variety of in-training programs, and the numbers of examinees varied across the many programs. Overall, the distributions for the two groups were similar.

Needs of demographic information. The above analyses on the demographic attributes of the two examinee groups indicate between-group similarities. However, if

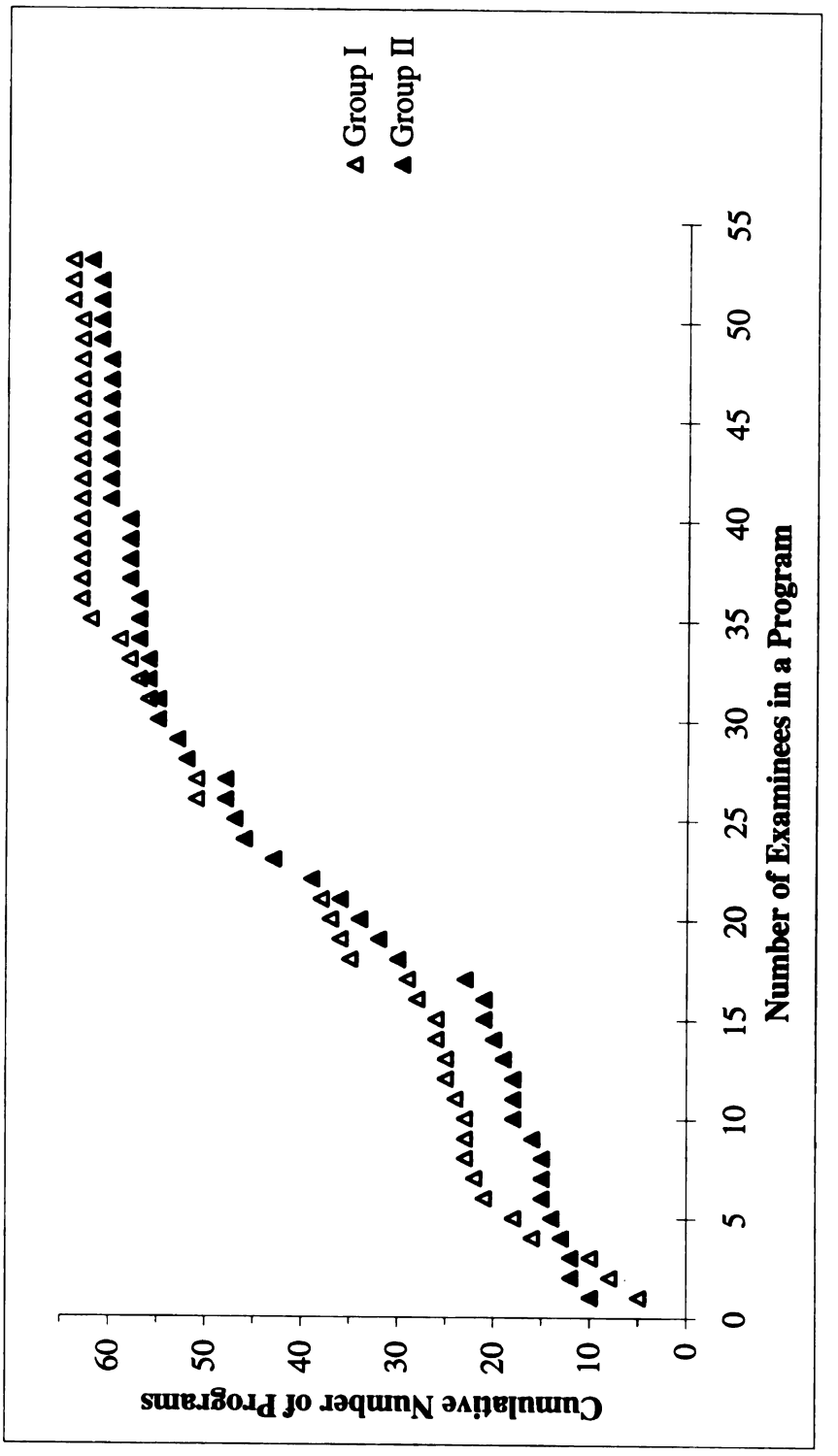
Table 11 - Program Participation: Numbers of Examinees from Different In-Training Programs by Group

Program Number	Group I	Group II	Total
1	19	1	20
2	30		30
3	18		18
4	32		32
5		22	22
6		41	41
7	26	1	27
8		49	49
10	18	1	19
11	1	21	22
12	1	34	35
14	21		21
15		28	28
16		22	22
17	1	23	24
18		28	28
19	35		35
21		28	28
22		25	25
23	1	20	21
24		53	53
25		24	24
26	3	9	12
27		41	41
28	34	1	35
29	24		24
30	26	1	27
31	29	1	30
32	24		24
33		13	13
34		19	19
35	18		18
36	35	1	36
37		22	22
38	1	37	38
39		30	30
40	30		30

Program Number	Group I	Group II	Total
41	20		20
43		26	26
44		23	23
45		18	18
46		21	21
47	35		35
48	33		33
49	22		22
50		23	23
51		20	20
52	16	1	17
53	23		23
54	18		18
56		15	15
63		28	28
74	4		4
79		17	17
81		30	30
82	26		26
83	23		23
84	16		16
89	23		23
90	18		18
91		17	17
93		19	19
94	28		28
95		29	29
96	25		25
97	17	1	18
98		18	18
99		24	24
103		18	18
105		18	18
106	5		5
107	4		4
108	4		4
109		2	2

Program Number	Group I	Group II	Total
110	11		11
111		24	24
114		18	18
115	18		18
116	36		36
121		32	32
122	51	2	53
123	23		23
124		14	14
125	26		26
126	14	1	15
127	24		24
128		18	18
129		18	18
130	31		31
132		23	23
133	12		12
134	2		2
136		6	6
137	6		6
138		10	10
148		10	10
150		5	5
152	8		8
154	6		6
801	4		4
802		4	4
803	7		7
804	6		6
805	5		5
806	3		3
807	4		4
901	2		2
902	2		2
903	4		4
Total	1092	1149	2241

Figure 3 - Cumulative Frequency Distribution of Number of Programs



more demographic information is available, in-depth group differences and their potential influences on dimensionality (defined empirically) and equating accuracy could be more thoroughly examined. The data analyzed in this study is secondary data, from which only two demographic variables (years of experience and program participation) were available. It restricted the investigation of this study on group differences.

Summary

In short, the sampled test forms were reliable in terms of their internal consistency. The items within the forms had moderate difficulties and the anchor items were representative of the total tests. The studies on the demographic attributes of the two examinee groups indicate between-group similarities. Although there was slight difference between the two examinee groups in their ability, the difference was not serious.

Score Distributions of Various Test Forms

The score distributions of the two original test forms, presented in APPENDIX A, were both negatively skewed. This is because the original test was written for a minimum competency examination. After the sampled test forms were created for this study, the score distributions of these test forms were also examined (included in APPENDIX A). Like the original Book A and Book B, all of the sampled test forms had more or less negatively skewed distributions.

Such properties of skewed score distributions for the four sampled tests are taken into account in the subsequent equating studies, for the discussions and interpretations of study results.

Adequacy of 3PL IRT Model for IRT-Based Equatings

The outcomes of the two IRT-based equating methods were based on a three parameter logistic IRT model. The IRT model incorporated a guessing parameter to account for the likely guessing factor in the minimum competency examination. Grounded in the 3 PL IRT model, the IRT-based equating outcomes therefore gained a logical and theoretical advantage of taking into account the chance of guessing.

In addition, the satisfactory equating outcomes of the two IRT-based equating methods (presented later in this chapter) provide empirical evidence of adequacy for the underlying 3PL IRT model. Combining the theoretical appeal with the empirical evidence of adequacy, this study concludes that the 3PL IRT model was appropriate for the IRT-based equatings conducted in this study, where test forms with negatively skewed score distributions were equated.

Equating Outcomes of IRT-Based Methods

The outcomes of IRT parameter estimation and the equating outcomes of the two IRT-based methods are summarized below. The equating results of the two IRT-based methods are found to be very similar.

Estimation of IRT Parameters

The results of fitting a 3-PL IRT model are summarized in Table 12. Over various test forms, for both IRT-based linear transformation method and fixed-b method, the intermediate outcomes of parameter estimation showed small variation in item

Table 12 - Results of IRT Parameter Estimation

Parameter Estimate		Sampled Test Form											
		Group I				Group II							
						IRT-Based Linear Transformation Method				IRT-Based Fixed-b Method			
		SR-A	EW-A	PW-A	PS-A	SR-B	EW-B	PW-B	PS-B	SR-B	EW-B	PW-B	PS-B
$\hat{a}$	mean	0.340	0.342	0.340	0.444	0.377	0.355	0.410	0.444	0.400	0.377	0.433	0.462
	s.d.	0.173	0.168	0.127	0.192	0.165	0.162	0.162	0.041	0.164	0.165	0.166	0.194
$\hat{b}$	mean	-0.884	-1.445	-1.090	-0.653	-1.008	-1.561	-0.380	-0.904	-0.591	-1.200	0.052	-0.650
	s.d.	2.239	2.231	2.043	1.848	1.891	2.240	2.361	1.705	1.951	2.374	2.301	1.774
$\hat{c}$	mean	0.252	0.260	0.256	0.247	0.241	0.270	0.328	0.231	0.311	0.347	0.384	0.277
	s.d.	0.046	0.033	0.029	0.052	0.034	0.030	0.052	0.041	0.053	0.048	0.056	0.049
$\hat{b}_{anchor}$	mean	-1.340	-1.750	-1.090	-0.750	-1.45	-1.88	-0.840	-0.180	N/A			
$\hat{\theta}$	mean	0.003	0.006	0.005	0.007	0.003	0.004	0.012	0.005	0.059	0.011	0.142	0.061
	s.d.	0.851	0.854	0.839	0.897	0.868	0.857	0.858	0.886	0.880	0.867	0.869	0.888

Note: a = item discrimination parameter
b = item difficulty parameter
c = guessing parameter
θ = person ability parameter

discrimination. Overall, the estimated values of the average item discrimination parameters were close and the standard deviations, relative to the means, were also small. In addition, the estimation for the guessing parameter yielded pretty similar results. However, the variation across sampled test forms in item difficulty seemed large.

When alternate forms of each sampled test were calibrated separately, as required by the IRT-based linear transformation method, the resulting average anchor-item difficulties differed across various sampled tests. Particularly, the anchor items of PS-A and PS-B had greater values in their average item difficulties than those of the other three pairs of test forms. These differences are attributed to the item sampling in this study, which sampled different numbers of items from different numbers of core content areas. For each of the sampled test, however, the average anchor-item difficulties for its two forms were not too far apart. This lends some evidence of similar ability for the two examinee groups.

Equated IRT Ability Estimates

The equivalent ability estimates yielded by the IRT-based linear transformation method and fixed-b equating method correlated strongly. It indicates similarities between the outcomes of the two methods. Across various sampled tests, all of the Pearson's *r*s were statistically significant and had values close to 1. Therefore, the two IRT-based methods did not differ much in determining individual examinee's standing in the entire examinee group.

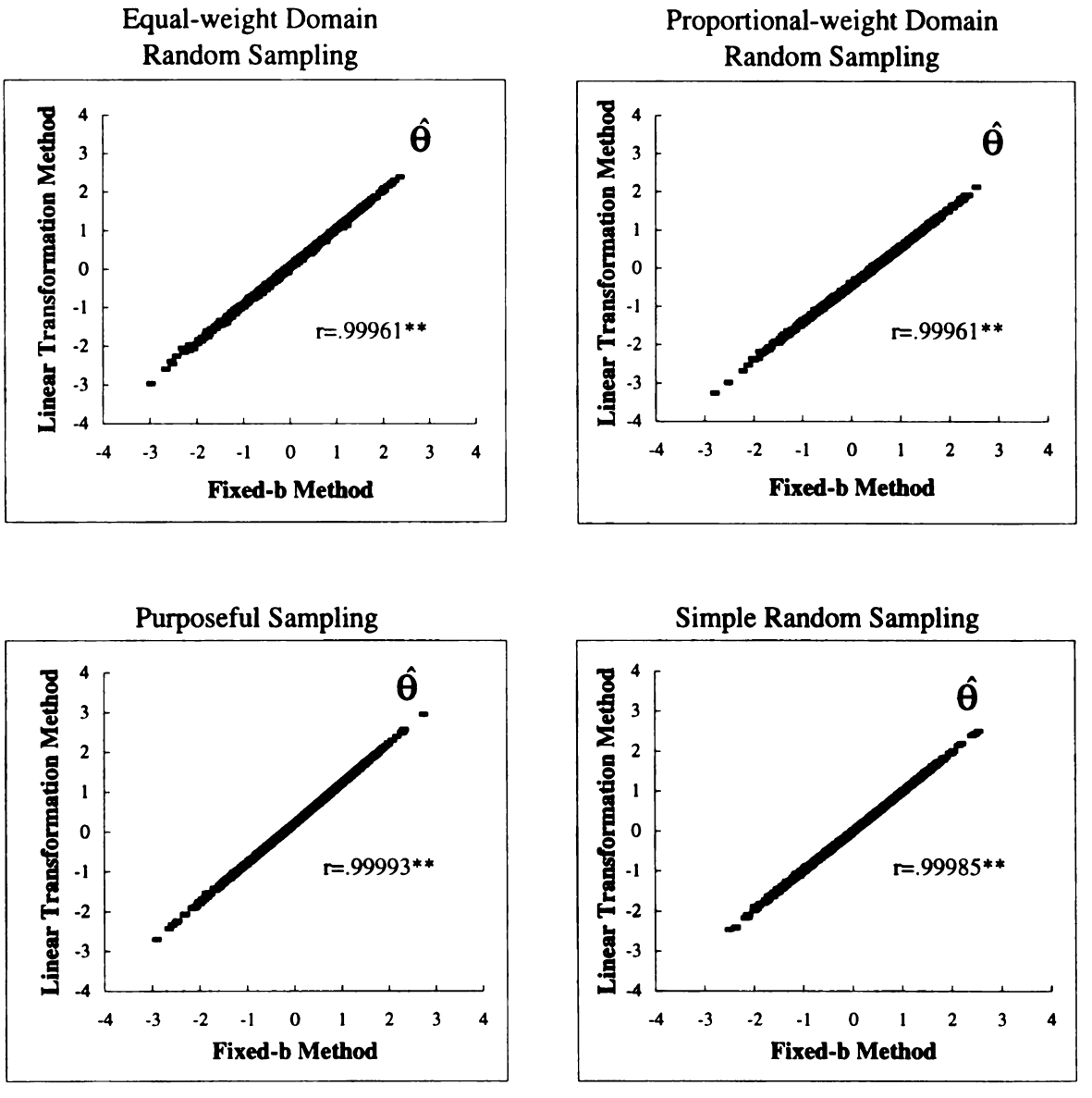
To further study how similar the results of the two IRT-based equating methods were, for each sampled test, the means and standard deviations of the two sets of resulting

ability estimates were compared. The mean difference between the ability estimates of the two IRT-based methods was also tested for its significance. As shown in Table 13, the average ability estimates of the two methods seem to differ only slightly and the standard deviations are very similar. However, the dependent-samples *t*-tests show that the outcomes of the two methods were significantly different ($p<.001$), no matter which pair of sampled test forms were equated. To control for the total error rate, which is likely to increase with the number of hypothesis tests, a more conservative significance level ($\alpha=0.01$) was chosen for the *t*-tests instead of the conventional $\alpha=0.05$. Overall, the test results suggest that the outcomes of the two IRT-based methods were not as close as represented by the Pearson's *r*s.

Although the statistical tests suggest significant differences, it should be noted that the large *t*-values in Table 13 are partly due to the small standard errors of mean difference and the large sample size, and hence more degrees of freedom. The effect sizes across the four tests are all very small (less than 0.5), implying practical insignificance. Graphing the resulting equivalent ability estimates of one IRT-based method against those of another method, the scatter plots in Figure 4 illustrate the positive relationships between the outcomes of the two methods across tests (differing in their item homogeneity). While the fairly solid straight lines in the plots suggest strong linear relationships, the slight thickness and coarseness of these lines indicate that the relationships were not as nearly perfect as indicated by the Pearson's *r*s. Overall, at the two ends of the ability scale, the outcomes of the two methods were more similar than the outcomes at the middle range of the ability scale.

Table 13 - Comparisons of the Resulting Ability Estimates of Two IRT-Based Equating Methods

Sampled Test	IRT-Based Equating Method		Paired Samples *t*-Test					
	Linear Transformation	Fixed-b	Difference	Std. Error of Mean Difference	*t*	df	Sig. (2-tailed)	Effect Size
PS	mean 0.335 s.d. 0.886	mean 0.061 s.d. 0.888	mean -0.275 s.d. 0.011	0.000	-863.344	1148	0.000	-0.310
EW	mean 0.134 s.d. 0.857	mean 0.011 s.d. 0.868	mean -0.123 s.d. 0.026	0.001	-159.430	1148	0.000	-0.143
PW	mean -0.238 s.d. 0.859	mean 0.142 s.d. 0.870	mean 0.379 s.d. 0.026	0.001	487.551	1148	0.000	0.439
SR	mean 0.113 s.d. 0.869	mean 0.059 s.d. 0.880	mean -0.055 s.d. 0.019	0.001	-97.545	1148	0.000	-0.062



Note. (1) θ is the examinee ability; (2) **- Significance level less than .01.

Figure 4 - Relationship Between the Resulting Ability Estimates of the Two IRT-Based Equating Methods

3PL IRT-Based Equivalent True Scores

Applying the following formula (Lord, 1980) to the outcomes of the 3PL IRT-based equatings, true score estimates were obtained:

$$\hat{T} = \sum_{i=1}^{n} p_i(\theta) = \sum_{i=1}^{n} \{c_i + (1 - c_i) / [1 + Exp^{-1.7a_i(\theta - b_i)}]\}, \qquad (9.1)$$

where $\hat{T}$ is the estimated true score, $p_i(\theta)$ is the probability of getting item i correct given examinee ability θ, n is the number of items, a_i is the item discrimination for item i, b_i is the item difficulty for items i, and c_i is the pseudo-chance level (guessing) for item i (Hambleton & Swaminathan, 1990). As expected, for each sampled test, correlation between the resulting true-score estimates from the two IRT-based equating methods was fairly strong and statistically significant. The Pearson's r ranged from .976 to .999 across the four sampled tests, indicating nearly perfect relationships between the outcomes of the two methods. These findings are similar to those based on the correlations for the ability estimates. Therefore, the two IRT-based methods used in this study were similar in determining individual examinee's standing in the entire examinee group.

Table 14 shows that the average true score estimates of the two IRT-based equating methods were very similar. So were their standard deviations. However, regardless of the pair of sampled test forms being equated, the dependent-samples t-test reveals significant difference between the outcomes of the two methods (α=0.01 and $p<.001$). According to these results of significance tests, across various sampled tests, the

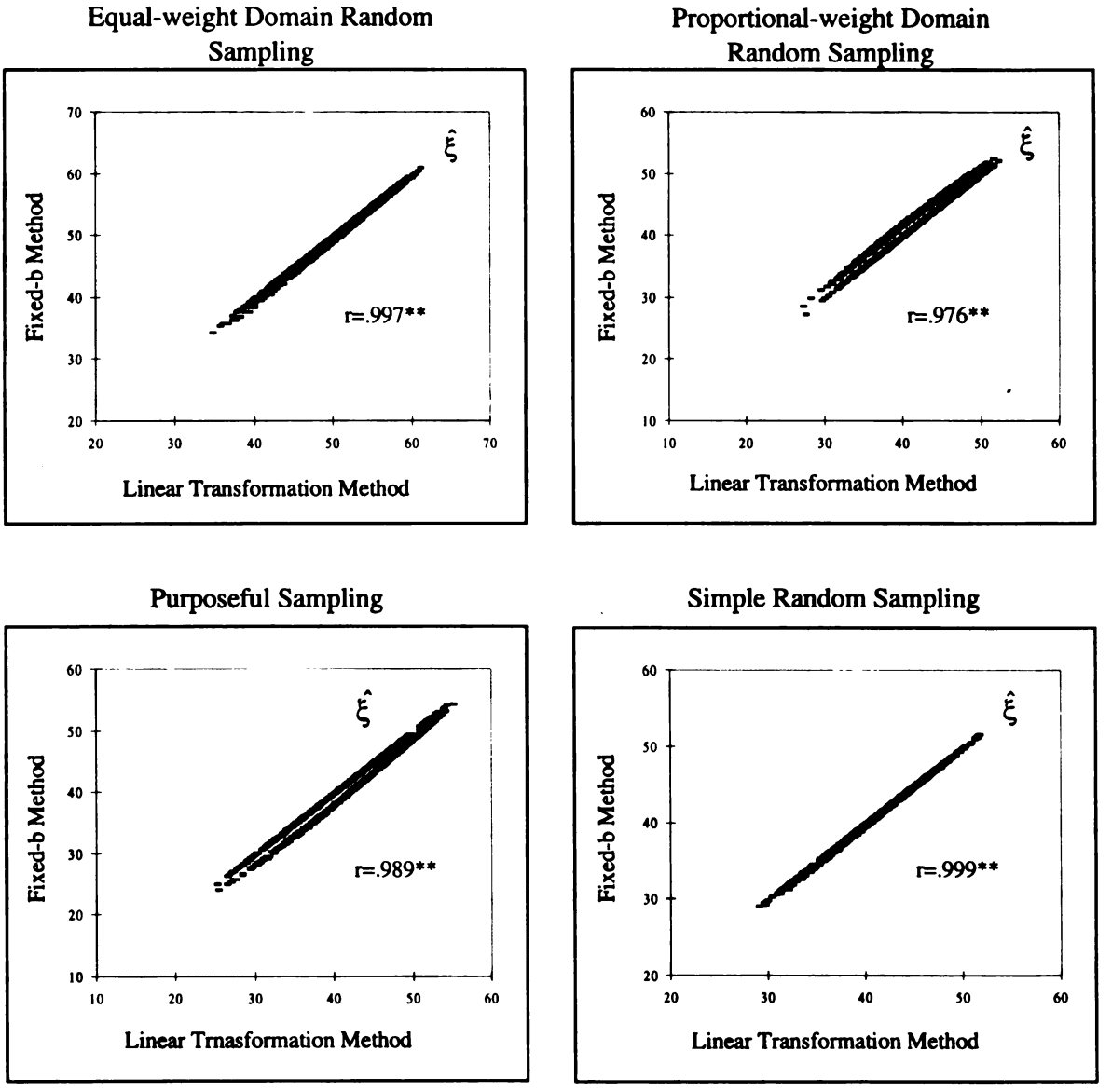
Table 14 - Comparisons of the Resulting True Score Estimates of Two IRT-Based Equating Methods

Sampled Test	IRT-Based Equating Method		Paired Samples *t*-Test					
	Linear Transformation	Fixed-b	Difference	Std. Error of Mean Difference	*t*	df	Sig. (2-tailed)	Effect Size
PS	mean 42.678 s.d. 5.570	mean 41.836 s.d. 5.595	mean 0.842 s.d. 0.843	0.018	47.250	2240	0.000	0.151
EW	mean 51.010 s.d. 4.389	mean 50.689 s.d. 4.445	mean 0.321 s.d. 0.340	0.007	44.720	2240	0.000	0.073
PW	mean 41.928 s.d. 4.157	mean 42.853 s.d. 4.081	mean -0.924 s.d. 0.911	0.019	-48.021	2240	0.000	-0.224
SR	mean 41.600 s.d. 4.179	mean 41.465 s.d. 4.210	mean 0.135 s.d. 0.151	0.003	42.513	2240	0.000	0.032

relationships between the two IRT-based methods were not as nearly perfect as suggested by the Pearson's *r*s. Nevertheless, the large *t*-values and significant test results in Table 14 can be attributed to the small standard errors of mean differences and the large sample size. In addition, the effect sizes for the differences across the four tests are all very small (less than 0.25). Thus, the differences between the two methods in estimating the true scores might not have practical significance.

The scatter plots in Figure 5 illustrate the delicate relationship between the two IRT-based equating methods, by graphing the resulting true score estimates of the fixed-b method against the resulting estimates of the linear transformation method. These plots are more revealing than those plots in Figure 4 (which are based on the resulting ability estimates) are, in showing the differences between the two equating methods. While the scattered data points form a pretty solid straight line for sampled tests EW and SR respectively, the data points for PW and PS clearly show more than one line. The two separate lines in the plots for PW and PS suggest that the resulting true score levels of one method did not correspond to the levels of another method on a one-to-one basis. On either one of the two sampled tests (PW or PS), when one IRT-based equating method was used, a group of examinees might receive the same scores, but the same group of examinees might receive different scores when another IRT-based method was used for equating. An inspection on the resulting estimates of equivalent true scores yielded by the two equating methods confirmed such speculation.

In addition, the formations of the data points for tests PW and PS in Figure 5 seem to be linear yet slightly elliptical. For each test, the two separate lines shown in the scatter plot not only look slightly curvilinear but also cave to the opposite directions. They



Note. (1) ξ is the equivalent true score estimate; (2) **- Significance level less than .01.

Figure 5 - Relationship Between the Resulting True Score Estimates of the Two IRT-Based Equating Methods

suggest that outcomes of the two IRT-based methods were more similar for cases receiving scores near or at the two ends of the true-score scale (than for the cases in the middle range of the scale). The slightly non-linear relationships between the two IRT-based equatings on PW and PS suggest that using Pearson's *r* for summarizing or comparing the outcomes of different equating methods could be misleading. Graphical displays contrasting the equivalent scores from different equating methods are recommended to improve the comparisons.

Smoothing Equipercentile Equating Outcomes

This study used the frequency estimation method (Kolen & Brennan, 1995) to conduct equipercentile equating. To increase equating precision, after obtaining the frequency-estimation equipercentile equivalent scores, this study applied the cubic spline postsmoothing method (Kolen & Jarjoura, 1987; Kolen & Brennan, 1995) to smooth the equivalent scores. A total of eight smoothing parameters were specified (s=.01, .05, .10, .20, .30, .50, .75, and 1) for postsmoothing. These parameters yielded smoothed equivalent scores differing in their degree of smoothing (Hanson, Zeng, & Kolen, 1995). That is, they controlled the amount of the average squared standardized difference between the smoothed and the unsmoothed equating outcomes.

Graphical Inspection on Smoothing Results

The resulting eight sets of smoothed equivalent scores were inspected graphically and statistically to determine which of the eight smoothing parameters resulted in the least amount of smoothing required for a smooth equipercentile equating function. For

graphical inspection, each of the eight sets of smoothed equivalent scores was graphed with the set of unsmoothed equivalent scores, and a standard error band was constructed around the unsmoothed equating outcomes to facilitate visual inspection. The adequacy of the various smoothed equating outcomes were in part judged by their smoothness and deviations from the unsmoothed equating outcomes, shown in such graphs. When there were more than one adequate smoothing parameters, judgment was made with considerations for the large sample size of this study and the numbers of items in the sampled test forms. Figure 6 presents a set of eight graphs for one sampled test (EW) to illustrate such graphical inspection techniques. Those graphs depict the changes (before and after a smoothing parameter is applied) in the relationship between the equivalent scores on sampled test forms EW-A and EW-B, when the degree of smoothing varies. The same type of graphs for the other sampled tests (PS, PW, and SR) are included in APPENDIX B.

In the case of smoothing the frequency estimation outcomes for sampled test EW, any values of smoothing parameter equal to or greater than .10 would result in smoothed equating outcomes that are too far from the unsmoothed equating outcomes (more than one standard error of the unsmoothed equivalent scores). This is illustrated in the graphs for smoothing parameters s=.10, s=.20, s=.30, s=.50, s=.75, and s=1.0 in Figure 6. These graphs show that using any one of these six parameters, some smoothed outcomes (between the scores 37 and 40 on the EW-B scale) would fall outside the standard error band around the unsmoothed outcomes. Figure 6 also shows that, using these parameters, there would be larger overall differences between the smoothed and unsmoothed outcomes (than when the parameters s=.01 and s=.05 were used). Therefore, both s=.01

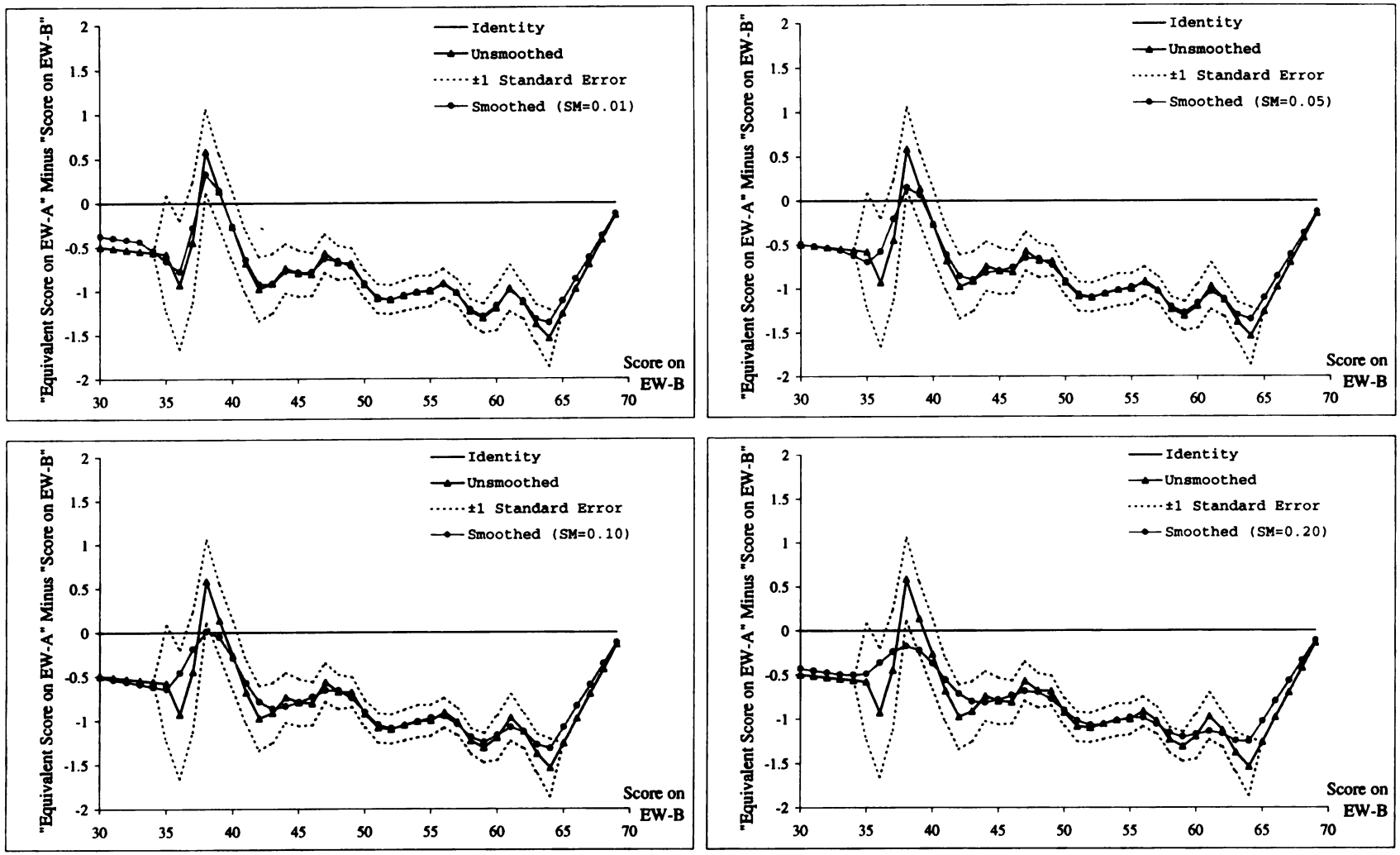


Figure 6 - Score to Score Equivalents by Various Degrees of Smoothing for Sampled Test EW

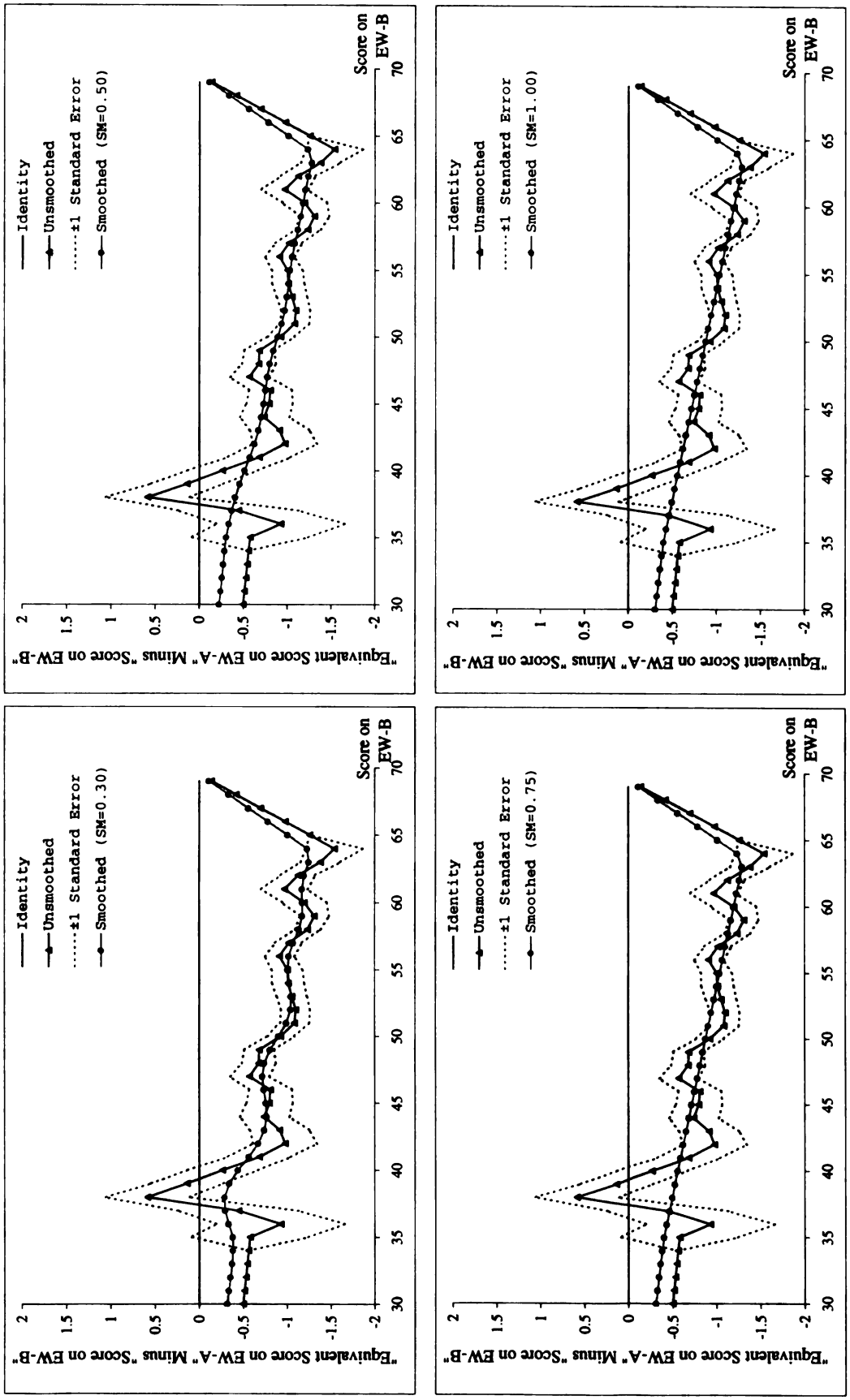


Figure 6 (cont'd)

and s=.05 were more appropriate for postsmoothing. Compared to the graph for s=.01, the graph for s=.05 suggests slightly smoother outcomes. Although s=.05 would result in slightly higher degree of smoothing, given the large sample size and number of items for EW, the amount of smoothing required by s=.05 should be scarcely more than the amount required by s=.01.

Evaluation of "Moment Preservation"

In addition to the graphical inspection, the four moments -- mean, standard deviation, skewness, and kurtosis -- of the resulting smoothed equivalent scores were estimated to evaluate the smoothing requirement of "moment preservation" (Kolen & Brennan, 1995). The estimation outcomes for the moments of the eight sets of equivalent scores are summarized in Table 15. The moments of the smoothed equivalent scores were compared to the moments of the unsmoothed equivalent scores such that the most appropriate smoothing parameter could be identified. An appropriate smoothing parameter will result in a smooth equipercentile equating function that does not depart too much from the unsmoothed equating outcomes.

Using the evaluation outcome for sampled test EW as an example, the assessment procedure of "moment preservation" is briefly illustrated. As shown in Table 15, for sampled test EW, the four moments of the smoothed equivalent scores resulted from s=.05 are more similar to the moments of the unsmoothed equivalent scores than those from s=.01. Combining this finding with the information from Figure 6, where smoothness and deviations from the unsmoothed outcomes of the smoothed outcomes were examined, s=.05 was therefore chosen to produce final smoothed equipercentile-

Table 15 - Moments for Postsmoothing Outcomes

Sampled Test	Smoothing Parameter	$\hat{\mu}$	$\hat{\sigma}$	Skewness	Kurtosis
PS	**Unsmoothed**	**42.2699**	**6.7032**	**-0.5084**	**0.0696**
	S=.01	42.2745	6.7038	-0.4997	0.0529
	S=.05	42.2739	6.7060	-0.5006	0.0583
	S=.10	42.2739	6.7078	-0.5017	0.0628
	S=.20*	42.2731	6.7096	-0.5030	0.0690
	S=.30	42.2728	6.7108	-0.5040	0.0733
	S=.50	42.2730	6.7140	-0.5049	0.0674
	S=.75	42.2769	6.7197	-0.5121	0.0489
	S=1.00	42.2796	6.7237	-0.5223	0.0518
EW	**Unsmoothed**	**51.0025**	**5.7939**	**-0.5517**	**0.6772**
	S=.01	51.0043	5.7973	-0.5465	0.6675
	S=.05 *	51.0031	5.7973	-0.5474	0.6724
	S=.10	51.0031	5.7974	-0.5478	0.6733
	S=.20	51.0029	5.7972	-0.5467	0.6678
	S=.30	51.0029	5.7972	-0.5467	0.6678
	S=.50	51.0027	5.7956	-0.5499	0.6679
	S=.75	51.0035	5.7963	-0.5550	0.6784
	S=1.00	51.0035	5.7963	-0.5550	0.6784
PW	**Unsmoothed**	**42.8099**	**5.5660**	**-0.3986**	**0.3033**
	S=.01	42.8117	5.5657	-0.3940	0.2846
	S=.05*	42.8116	5.5663	-0.3951	0.2932
	S=.10	42.8120	5.5649	-0.3934	0.2881
	S=.20	42.8130	5.5620	-0.3898	0.2739
	S=.30	42.8127	5.5601	-0.3864	0.2612
	S=.50	42.8148	5.5577	-0.3833	0.2467
	S=.75	42.8155	5.5613	-0.3917	0.2544
	S=1.00	42.8168	5.5686	-0.4183	0.2837
SR	**Unsmoothed**	**41.5984**	**5.7093**	**-0.3711**	**0.0754**
	S=.01*	41.6009	5.7101	-0.3687	0.0756
	S=.05	41.6020	5.7140	-0.3661	0.0824
	S=.10	41.6037	5.7166	-0.3621	0.0803
	S=.20	41.6052	5.7162	-0.3564	0.0584
	S=.30	41.6055	5.7128	-0.3513	0.0322
	S=.50	41.6111	5.7081	-0.3365	-0.0351
	S=.75	41.6091	5.7103	-0.3251	-0.0616
	S=1.00	41.6089	5.7106	-0.3227	-0.0651

Note. * Indicates the smoothing parameter selected for postsmoothing the frequency-estimation outcomes of equipercentile equating, after taking into account the information from this table and Figure 6.

equivalent scores for test form EW. Although some other smoothing parameters such as s=.10 and s=.50 seemed to yield moments more similar to those moments of the unsmoothed outcomes, they were not appropriate for smoothing the equated scores on EW because some of their smoothed outcomes would fall outside the standard error band.

It should be noted that the smoothing requirement of "moment preservation" also requires that the moments of the equated scores on one form of a test to be close to those on the other form of the same test (Kolen & Brennan, 1995). This property is desired for both random group equating design and common-item non-equivalent group design. However, for the non-equivalent group design used in this study, it is a lot more difficult to examine this property and the interpretation will not be as clear as for the random group design (M. J. Kolen, personal communication, May 6,1997). This study therefore did not directly assess the "moment preservation" on one form for the particular population taking the other form because of missing data. In addition, the moments in this study depended on the particular assumption made by the frequency-estimation method, used for the equipercentile equating. The frequency-estimation method assumes that, for both forms of a test, the conditional distribution of total score given each common-item score is the same in both populations.

Results of Selecting Smoothing Parameters

Despite the difficulty in assessing "moment preservation" across test forms, the graphical inspection on the smoothing results and the evaluation of "moment preservation" within test forms provide useful information for assessing the effectiveness of various smoothing parameters. Judgments were made about the relative estimation errors caused

by the eight smoothing parameters. Taking into account all the information, the following smoothing parameters were selected and used respectively for the four sampled tests, to improve the equated scores resulted from equipercentile equating: (a) S=.05 for sampled tests EW and PW, (b) S=.20 for sampled test PS, and (c) S=.01 for sampled test SR.

The final equipercentile equivalent scores yielded by the above smoothing parameters appeared to be smooth and were not too far apart from the unsmoothed results (see Figure 6 and APPENDIX B). Their four moments (see Table 15) were also close to those of the unsmoothed equivalent scores. Without introducing substantial bias into the smoothing process, the use of these smoothing parameters improved the precision of the equipercentile equating in estimating the equivalent scores (Kolen, 1991).

Results of Tucker Linear Method

For each of the four sampled tests, Tucker linear method found an equating equation that transformed scores on one form to a set of new scores comparable to the scores on the other form. Important intermediate outcomes of the Tucker method and the resulting equating equations are summarized in Table 16.

As reviewed in Chapter 3, the four Tucker equating equations presented in Table 16 were derived by defining a synthetic population, assuming equal conditional variances and same linear regression functions for the two populations, and estimating the means and variances for the synthetic population (Kolen & Brennan, 1987; Kolen & Brennan, 1995). Using these resulting Tucker equations, equivalent scores were established for the two forms of each sampled test.

Table 16 - Summary of the Results of Tucker Linear Equating Method

Sampled Test	Parameter Estimate						Tucker Equating Equation
	Form A			Form B			
	$\alpha_A(A\|V)$	$\mu_s(A)$	$\sigma_s^2(A)$	$\alpha_B(B\|V)$	$\mu_s(B)$	$\sigma_s^2(B)$	
SR	1.524	41.615	32.742	1.593	41.855	36.047	$\ell(b)$.953(b-41.855)+41.615
EW	1.218	51.003	33.830	1.257	51.941	35.559	$\ell(b)$.975(b-51.941)+51.003
PW	1.257	42.810	30.935	1.291	42.806	32.349	$\ell(b)$.978(b-42.807)+42.810
PS	1.169	42.271	44.882	1.150	40.408	43.036	$\ell(b)$1.021(b-40.408)+42.271

Note.

1. "*A*" represents the score on test form A, "*B*" represents the score on test form B, and "*V*" represents the score on common anchor items.
2. α_A is the regression coefficient for the population taking test from A, and α_B is the regression coefficient for the population taking form B.
3. "*s*" denotes the synthetic population, and "*b*" is the observed score on form B.
4. The weight for the population taking form A is .487, and the weight for the population taking form B is .513.

Similarities Among Outcomes of Various Equating Methods

The equating results of the Tucker method and the other equating methods are compared in this section. The positive and significant strong relationships among these results (see the underlined correlation coefficients in Table 17) indicate similarities among these various equating outcomes. Individual examinees were ordered in a similar way, regardless of the equating method used.

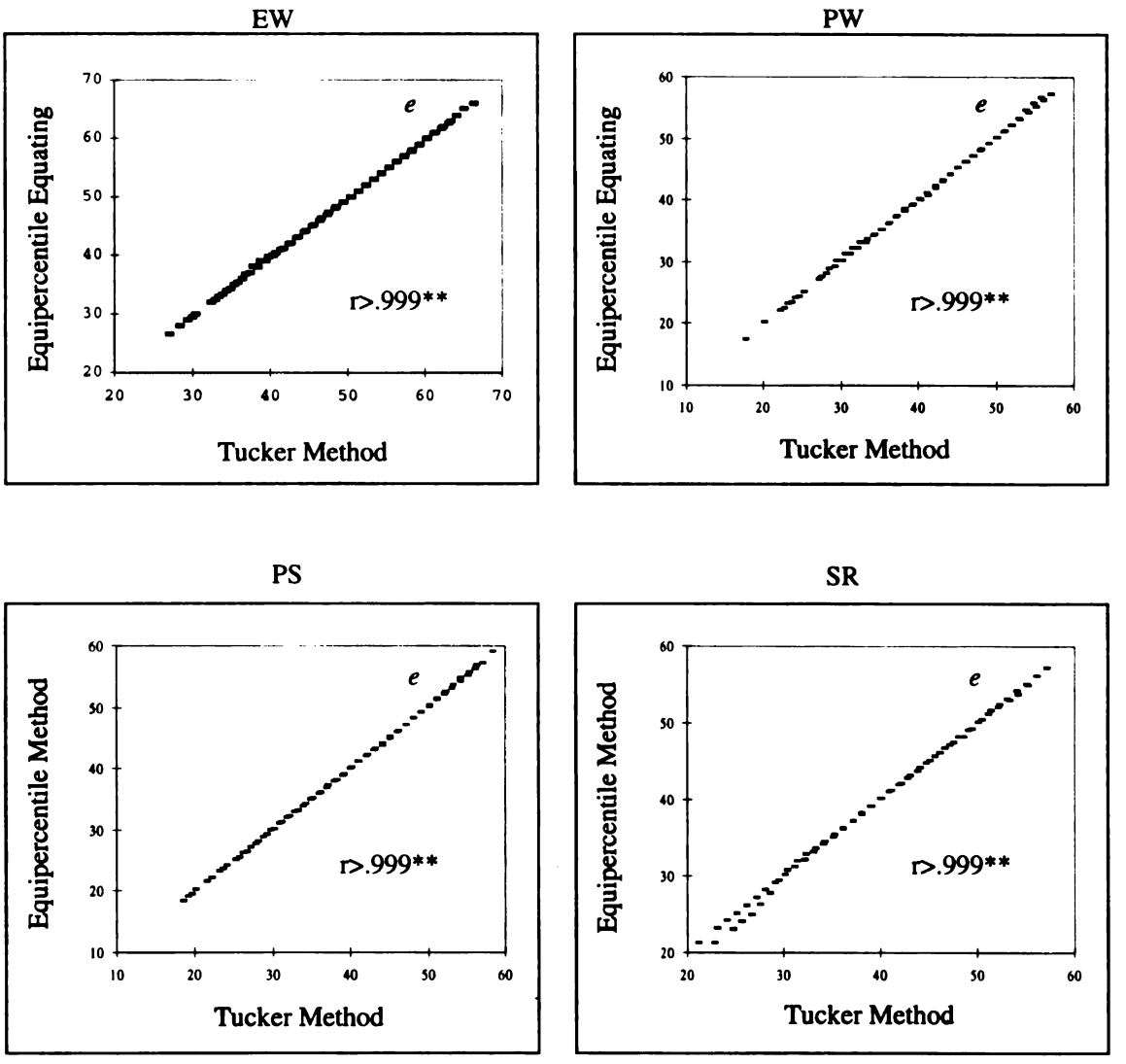
Comparisons between the outcomes of the two IRT-based methods have been discussed previously and considerable similarities are found. Using the same strategies -- Pearson's *r*, dependent-samples *t*-test, and scatter plot, results of the Tucker method and the frequency-estimation equipercentile method are compared. The large Pearson's *r*s between the outcomes of these two methods in Table 17 suggest that these methods yielded almost identical rank order for individual examinees. For each of the four sampled tests, the correlation is nearly perfect ($r > .999$). The scatter plots in Figure 7 further confirm the similarities.

Except for the plot for sampled test SR, each of the scatter plots in Figure 7 clearly shows one single narrow straight line. This indicates great resemblance of Tucker equating outcomes to the outcomes of the equipercentile method. The plot for sampled test SR shows that the outcomes of the Tucker and equipercentile methods were similar when their resulting equivalent scores were in the middle range or at the high end of the score scale. However, they differed slightly when their resulting scores were at the low end (between scores 20 and 30) of the score scale. This suggests that the equating outcomes of the two methods were very similar when examinees had medium or higher scores. The outcomes only differed slightly for the examinees with pretty low scores.

Table 17 - Relationships Among Various Equating Outcomes for Different Sampled Tests

Pearson's Correlation Coefficient (r)		Equating Method															
		Tucker Linear method				Equipercentile Method				IRT-Based Linear Transformation Method				IRT-Based Fixed-b Method			
		SR	EW	PW	PS	SR	EW	PW	PS	SR	EW	PW	PS	SR	EW	PW	PS
Tucker Linear Method	SR	1.000															
	EW	0.782	1.000														
	PW	0.795	0.782	1.000													
	PS	0.810	0.755	0.795	1.000												
Equipercentile Method	SR	1.000	0.783	0.795	0.810	1.000											
	EW	0.782	1.000	0.782	0.754	0.783	1.000										
	PW	0.795	0.781	0.999	0.794	0.795	0.781	1.000									
	PS	0.810	0.755	0.795	1.000	0.811	0.754	0.794	1.000								
IRT-Based Linear Transformation Method	SR	0.965	0.780	0.791	0.834	0.964	0.780	0.790	0.834	1.000							
	EW	0.786	0.964	0.787	0.776	0.786	0.964	0.786	0.776	0.811	1.000						
	PW	0.768	0.750	0.944	0.773	0.767	0.750	0.944	0.774	0.785	0.776	1.000					
	PS	0.795	0.739	0.771	0.973	0.794	0.739	0.770	0.973	0.842	0.771	0.748	1.000				
IRT-Based Fixed-b Method	SR	0.963	0.779	0.790	0.831	0.962	0.779	0.789	0.831	0.999	0.810	0.791	0.837	1.000			
	EW	0.783	0.961	0.785	0.770	0.784	0.960	0.784	0.770	0.808	0.997	0.789	0.759	0.810	1.000		
	PW	0.785	0.771	0.963	0.801	0.784	0.771	0.963	0.801	0.804	0.798	0.976	0.793	0.802	0.795	1.000	
	PS	0.799	0.740	0.774	0.972	0.799	0.739	0.773	0.972	0.846	0.772	0.782	0.989	0.846	0.771	0.795	1.000

Note. All of the Pearson's rs are significant at α=.01.



Note. (1) *e* is the resulting equivalent score; (2) **- Significance level less than .01.

Figure 7 - Relationship Between Equating Outcomes of the Tucker Method and the Frequency-Estimation Equipercentile Method

The relationships between the IRT-based equating outcomes and the non-IRT equating outcomes were not as strong as the relationship between the outcomes of the two IRT-based methods. They were also less strong than the relationship between the outcomes of the Tucker and equipercentile methods. The Pearson's *r*s between the IRT-based and the non-IRT equating outcomes ranged from .944 to .973 (see the underlined correlation coefficients in Table 17). This finding reflects the logical differences between the IRT-based equating approach and the conventional equating approach.

Evaluation of Equating Accuracy

The accuracy of equating outcomes was evaluated using four different criteria. An index of equating accuracy was computed by correlating resulting equivalent scores from different methods to each of these criterion scores: (a) total raw scores on the 145 anchor items (Raw-145), (b) IRT-estimated true scores on the 145 anchor items (IRT-145), (c) resulting equivalent scores of the frequency-estimation equipercentile method on equating the two original test forms (FE-long), and (d) resulting equivalent scores of the equipercentile method on equating the sampled test forms (FE-short). The last two criteria, FE-long and FE-short, were arbitrary criteria for evaluating equating accuracy. However, FE-long was expected to be more reliable. FE-short was used to facilitate an examination on evaluation bias caused by using an arbitrary criterion for evaluating equating accuracy. By correlating the outcomes of the other three equating methods to the outcomes of the equipercentile method on the sampled tests, this study examined estimation bias due to the arbitrary nature of FE-short.

Preview of Important Results

Before presenting the massive information, in details, regarding findings from studies of equating accuracy, this section first previews selected important results to highlight major findings.

In brief, using Raw-145 and IRT-145, this study found that the IRT-based equating outcomes were more accurate than those outcomes of the linear and equipercentile methods were. Although the differences between the estimated equating accuracy of the IRT-based methods and the non-IRT-based methods were small, statistical significance tests for the differences concluded that they were statistically significant at α=.05. However, the statistically significant but little improvement of the IRT-based methods in equating accuracy might not have practical significance. Among various sampled tests, equating results for sampled test PS were often the most accurate, regardless of the equating method used. The twofold implications of this finding, in improving equating accuracy for common-item equating practice, are:

- It is important to include anchor items that are representative of the total test in content.
- It is also useful to construct test forms with items that are more homogeneous in their content, or to limit the content coverage of test forms to a small number of topics.

In addition, the findings from EF-long and EF-short confirmed that the use of an arbitrary criterion would lead to erroneous assessment outcomes of equating accuracy, as concluded in the literature (Dorans & Kingston, 1985; Harris & Crouse, 1993).

Table 18 summarizes the estimation results of equating accuracy for outcomes from various equating methods on different sampled tests, using three different evaluation

Table 18 - Accuracy of Equating Outcomes from Various Equating Methods on Different Sampled Tests

Index of Equating Accuracy (Pearson's *r*)		Criterion for Evaluating Equating Accuracy		
		"Pseudo True Score"		FE-long
		Raw-145	IRT-145	
Tucker Linear Method	SR	0.832	0.819	0.884
	EW	0.859	0.829	0.880
	PW	0.860	0.839	0.883
	PS	0.892	0.898	0.903
Equipercentile Equating Method	SR	0.832	0.819	0.884
	EW	0.858	0.829	0.880
	PW	0.859	0.838	0.882
	PS	0.892	0.898	0.903
IRT-Based Linear Transformation Method	SR	0.856	0.860	0.897
	EW	0.877	0.870	0.893
	PW	0.845	0.839	0.864
	PS	0.894	0.917	0.897
IRT-Based Fixed-b Method	SR	0.854	0.858	0.896
	EW	0.873	0.865	0.889
	PW	0.870	0.867	0.888
	PS	0.895	0.916	0.898

Note. All of the indices of equating accuracy (the Pearson's *r* s between the criterion scores and the resulting equivalent scores of an equating method) are significant at α=.01.

criteria -- Raw-145, IRT-145, and FE-long. The evaluation results of equating accuracy using FE-short are included in Table 17 (see the bordered Pearson's correlation coefficients). Details of analysis outcomes on equating accuracy are presented below. First, the estimation of equating accuracy using Raw-145 is discussed. The results based on IRT-145 follows. Then, results from FE-long are examined, followed by an inspection on the results from FE-short.

Evaluation Using Raw-145 as a Criterion

The total raw scores on all of the 145 common anchor items (Raw-145) in the original item pool were treated as "pseudo true scores" of individual examinees. Therefore, it could be used as one type of criterion for evaluating equating accuracy. Specifically, to study equating accuracy of the IRT-based methods, the equivalent true scores estimated by the two IRT-based equating methods were correlated to Raw-145. Raw-145 was also correlated to the equivalent scores resulted from the Tucker method and the equipercentile method to estimate the accuracy of these equating outcomes. The resulting Pearson's *r*s indicated the degrees of accuracy for the equating outcomes of these four methods. These evaluation outcomes of equating accuracy are summarized in the first numeric column of Table 18.

Comparing equating accuracy of various methods. Using Raw-145 as a criterion, the indices of equating accuracy ranged from .832 to .895 for various sampled tests and equating methods. Overall, all four equating methods yielded accurate results to a moderate degree for the four sampled tests. For each sampled test, the accuracy of equating outcomes from the four equating methods differed slightly. The outcomes of the

IRT-based methods were consistently more accurate than those of the non-IRT methods, regardless of the sampled test forms being equated. The only exception occurred on sampled test PW, where the Pearson's *r*s of the Tucker method (*r*=.860) and the equipercentile method (*r*=.859) were slightly larger than the IRT-based linear transformation method (*r*=.845). Therefore, these differences were tested for their statistical significance.

Suppose equating accuracy of two equating methods, Y and Z, are compared. Let X be the criterion Raw-145. The significance test statistic appropriate for the dependent samples in this study is

$$t = \frac{(r_{xy} - r_{xz})\sqrt{(n-3)(1+r_{yz})}}{\sqrt{2(1-r_{xy}^2 - r_{xz}^2 - r_{yz}^2 + 2r_{xy}r_{xz}r_{yz})}}, \tag{9.2}$$

where x is the Raw-145 criterion-score (the total raw scores on all of the 145 common anchor items), y is the resulting equivalent score of method Y, z is the resulting equivalent score of method Z, and n is the sample size (Hinkle, Wiersma, & Jurs, 1979). In essence, the statistic r_{xy} represents the estimated equating accuracy of method Y, and r_{xz} represents the estimated equating accuracy of method Z. The statistic r_{yz} estimates the relationship between the equating outcomes of methods Y and Z. The underlying distribution of this test statistic is the Student's *t*-distribution with $n-3$ degrees of freedom. The critical values of the test statistic for this study are ±1.962, because all the tests will be non-directional, the level of significance is set at α=.05, and there will be

1,146 degrees of freedom.

For sampled tests SR and EW, where the IRT-based methods had slightly larger indices of equating accuracy than the non-IRT methods, the significance tests found all those differences significant ($|t|>1.962$). For sampled test PS, although the IRT-based methods also had slightly larger indices of equating accuracy than the non-IRT methods, the significance tests found no significant differences because the differences were so small. For sampled test PW, no matter the IRT-based methods had larger or smaller indices of equating accuracy than the non-IRT methods, none of the differences were significant.

Across sampled tests, the Tucker method and the equipercentile method had almost identical indices of equating accuracy. Statistically, none of the differences between the two methods was significant. It suggests that these two methods produced equally accurate outcomes, when Raw-145 was used as a criterion for evaluating equating accuracy. This finding coincides with the similarities previously found between the two methods (as shown in Figure 7).

In summary, there is a clear pattern across sampled tests in Table 18 showing that the IRT-based equating outcomes were more accurate than those outcomes of the Tucker linear and the equipercentile methods. Despite the fact that the improvements of the IRT-based equating methods were not much, they were statistically significant. While such small improvements may not have practical significance for equating in some occasions, they can be very valuable in some other occasions, especially when there is a strong demand for precise equated scores such as high-stake certification examination. Information on the degree of equating accuracy will help to make decisions about which

equating method to use for a particular testing program in a particular equating context.

Comparing equating accuracy for various sampled tests. Using the criterion Raw-145, this study also found that both conventional and IRT-based methods worked best on equating PS-A and PS-B. The average index of equating accuracy was .893. All methods but the IRT-based linear transformation method yielded the least satisfactory results on equating SR-A and SR-B, and the average accuracy was .839. The results of statistical significance tests concluded that, regardless of the equating method used, the equating outcomes for sampled test PS were significantly more accurate than the outcomes for SR. In addition, although the IRT-based linear transformation method had the least accurate result on equating PW-A and PW-B (r=.845), this outcome was not significantly different from the outcome of the same method on equating SR-A and SR-B. Overall, all of the four equating methods produced the least accurate outcomes for SR.

For sampled test PW, the average equating accuracy of the four methods was .858, and it was .867 for EW. The accuracy of equating outcome from any of the equating methods, except for the IRT-based linear transformation method, on equating PW-A and PW-B was not significantly different from the outcome of the same method on equating EW-A and EW-B. The linear transformation method yielded slightly more accurate outcome for EW than it did for PW.

In summary, regardless of the equating method used, the overall equating outcomes for EW and PW were equally accurate. These outcomes were less accurate than those outcomes for PS but more accurate than those outcomes for SR.

Equating method-test interaction. Given the above findings and conclusions about the equating accuracy for various sampled tests and of various equating methods, it is plausible that

overall there was no method-test interaction on estimating the accuracy of equating. That is, the relative equating accuracy of different equating methods did not depend on the particular test forms being equated, and the equating accuracy for different sampled tests were independent of the particular equating method used.

Effects of content homogeneity and content representativeness. As mentioned earlier, regardless of the equating method used, the equating results for sampled test PS were always the most accurate and the results for SR were always the least accurate. The twofold interpretations for these findings are:

- By the means of item-sampling designs, both PS-A and PS-B had items that were the most homogeneous in content, and both SR-A and SR-B had items that were the least homogeneous in content. The above findings therefore suggest that equating outcome based on a set of more content homogeneous items is likely to be more accurate than equating outcome based on a set of less content homogeneous items.
- Also, because of item sampling, PS-A and PS-B had anchor items that were the most representative to their total tests in content, and SR-A and SR-B had anchor items that were the least representative to their total tests. The above findings suggest that equating outcomes on a test containing content-representative anchor items is likely to be more accurate than the equating outcomes on a test containing less content-representative anchor items.

In short, equating accuracy depends on content homogeneity of test items in the test forms being equated. Equating accuracy also varies with the content representativeness of anchor items embedded in the test forms being equated. To improve equating accuracy, a testing program can use test forms composed with content-homogeneous items, given it is

realistic. For common-item equating, it is also important to include anchor items that adequately mirror a total test in content.

Evaluating effect of auto-correlation. The Pearson's *r*s used to represent the degrees of equating accuracy were contaminated by an artifact of auto-correlation. The auto-correlation was caused by the fact that the sampled tests overlapped with the criteria for evaluating equating accuracy on some items. In the cases where Raw-145 or IRT-145 was used as a criterion, all of the anchor items embedded in various sampled tests overlapped with part of the criterion, because these sampled anchor items were subsets of the "anchor universe" containing all the 145 anchor items. Due to such overlap of items, the resulting Pearson's *r*s were inflated. The resulting indices of equating accuracy were overestimated.

To study the impact of auto-correlation on the estimation of equating accuracy, the degrees of equating accuracy were estimated by excluding the anchor items from the sampled tests and then correlating the resulting IRT-estimated true scores on these reduced sampled tests to the scores of Raw-145. The correlation outcomes are summarized in APPENDIX C. This strategy for controlling artifact of auto-correlation, however, was not applicable in the cases where the Tucker linear method or the frequency-estimation equipercentile method was used for equating.

In the cases of the IRT-based equating outcomes, it is feasible and convenient to drop the anchor items and obtain a set of new total scores. This is because the IRT item parameters were estimated on an individual basis. Such advantage of IRT item calibration renders ease and convenience for test revision (to add or drop items). Based on observed test scores, the two non-IRT equating methods do not have the same flexibility. Their

resulting equivalent test scores can only be interpreted as a whole, that is, when the test forms being equated are kept intact. In addition, the necessary re-equating after dropping the anchor items is time-consuming and laborious. Although it is technically possible to obtain the difference scores between the resulting equivalent total scores and the subtotal scores based on only the common anchor items, such difference scores are not practical or logically sound. Therefore, the strategy for controlling artifact of auto-correlation was not used with the Tucker linear method and the frequency-estimation equipercentile method.

The magnitudes of the previously inflated indices of equating accuracy only attenuated slightly (less than .01), after controlling the auto-correlation (see the bordered and bolded Pearson's *r*s, before and after the adjustment for auto-correlation, in APPENDIX C). Also, the slight attenuation was not statistically significant. In addition, for various equating methods and sampled tests, the rank-order patterns of these indices remained the same as before the adjustment for auto-correlation. These similarities between the indices computed before and after the adjustment for auto correlation suggest that the impact of the auto-correlation was not serious or substantial on the estimation of equating accuracy. The unadjusted indices of equating accuracy remained valid. The findings, discussions, and conclusions based on these unadjusted indices were retained.

It is noted that the strategy used for controlling the artifact of auto-correlation did not completely eliminate the influence of the auto-correlation. Part of the artifact originated in the IRT-parameter estimation process and was therefore not easy to be controlled. Nevertheless, despite such difficulty, the strategy provides a useful alternative for improving the studies on the effectiveness of various equating methods and the effect of content representativeness.

Reliability and validity evidence on anchor tests. Validity and reliability evidence for anchor tests, embedded in the sampled tests, provide sound basis for validating findings from equating studies and reaching for plausible conclusions. By excluding the non-anchor items from the sampled tests, and then correlating the resulting IRT-estimated true scores on these reduced sampled tests (containing sampled anchor items only) to the scores of Raw-145, this study examined reliability and validity of anchor items, embedded in the various sampled tests. The investigation outcomes are summarized in APPENDIX D.

Raw-145, a type of the "pseudo true score", was regarded as a similar but more reliable measure to the score on the anchor test. This is because the anchor items included in the four sampled tests were all subsets of the "anchor universe", based on which the "pseudo true score" was computed. The "anchor universe" contained all the 145 anchor items available for this study, which was substantially more than the number of items in the sampled anchor tests. Therefore, the correlation coefficient between Raw-145 and the score on the anchor test could provide concurrent validity evidence for the anchor test. Moreover, from the perspective that a correlation coefficient was computed between an observed score (the score on the anchor test) and its true score (the "pseudo true score"), the Pearson's *r* also represented reliability of the anchor test.

The statistically significant Pearson's *r*s in APPENDIX D provide strong evidence of validity and reliability for the anchor tests in the four sampled tests. The average validity (reliability) was .895 for the anchor test embedded in PS. It was .875, .859 and .856 for the anchor tests of EW, PW, and SR respectively.

Evaluation Using IRT-145 as a Criterion

The second criterion for evaluating equating accuracy, IRT-145, was also based on the 145 common anchor items as the first criterion. It was the IRT-estimated true score on the 145 anchor items. Different from Raw-145, IRT-145 was not susceptible to the drawback of being person-dependent and item-dependent. Using IRT-145 as a criterion, Pearson's *r*s were computed to measure the degree of accuracy as before.

A summary of the evaluation outcomes of equating accuracy using IRT-145, for various equating methods and sampled tests, is included in Table 18 (see the second numeric column). As shown in the table, these outcomes were very similar to the evaluation outcomes resulted from Raw-145. More details of the evaluation outcomes from IRT-145 are presented below. Similarities and differences between the outcomes from IRT-145 and Raw-145 are discussed.

Accuracy of equating outcomes from various methods. Using the IRT-based criterion, the estimated equating accuracy ranged from .819 to .917 for various equating methods and sampled tests. As Raw-145, IRT-145 also found that all of the four methods yielded accurate results to a moderate degree, and the two IRT-based equating methods yielded significantly more accurate results than the Tucker linear method and the equipercentile method. The only exception is that the accuracy of the outcomes of the IRT-based linear transformation method on equating PW-A and PW-B was no different from those of the Tucker and the equipercentile methods. This exception, however, did not affect the overall conclusion that the IRT-based methods had more accurate outcomes than the non-IRT methods had.

Inspecting estimation bias due to IRT-145. As mentioned earlier, one of the concerns about using an IRT-based criterion is that such criterion may overestimate the equating accuracy of the IRT-based equating methods. In this study, however, IRT-145 did not seem to favor the IRT-based equating methods. Overall, IRT-145 did not systematically produce larger indices of equating accuracy for the IRT-based methods than for the non-IRT methods.

Contrasting the indices of equating accuracy yielded by Raw-145 and IRT-145 in Table 18, this study found that for only half of the time, IRT-145 yielded slightly larger indices than Raw-145 did for the IRT-based equating outcomes. For the other half of the time, Raw-145 yielded slightly larger indices than IRT-145 did for the outcomes from the IRT-based methods. In addition, most of these small differences between IRT-145 and Raw-145 were not statistically significant. Thus, IRT-145 was not biased in overestimating the equating accuracy of the IRT-based methods.

Comparing equating accuracy for various sampled tests. Using IRT-145, this study found that all of the equating methods worked best on equating sampled test forms PS-A and PS-B. For sampled test PS, the average index of equating accuracy across various equating methods was .907 (see Table 18). On equating SR-A and SR-B, however, all but the IRT-based linear transformation method produced the least accurate results. The average index of equating accuracy over various methods was .839 for SR.

Combining the results of statistical significance tests for the differences between the equating accuracy for sampled tests PS and SR, it is found that regardless of the equating method used, the equating outcomes for PS were significantly more accurate than the outcomes for SR. Although the IRT-based linear transformation method yielded the least

accurate result when equating PW-A and PW-B, the index of equating accuracy obtained for SR (.860) was not significantly larger than the index for PW (r=.839).

For sampled test PW, the average equating accuracy of the four methods was .846, and it was .848 for EW. The accuracy of equating outcome from any of the equating methods, except for the IRT-based linear transformation method, on equating PW-A and PW-B was not significantly different from the outcome of the same method on equating EW-A and EW-B. The linear transformation method yielded slightly more accurate outcome for EW than it did for PW.

In summary, regardless of the equating method used, the equating outcomes for EW and PW were equally accurate most of the time. These outcomes were less accurate than those outcomes for PS but more accurate than those outcomes for SR. Combined with all the other findings, this finding led to the same conclusion of Raw-145 -- there was no method-test interaction on estimating the accuracy of equating. Overall, using IRT-145 as a criterion for evaluating equating accuracy led to findings that are very similar to those yielded by Raw-145, and the conclusions reached by these two true-score-based criteria are exactly consistent.

Content homogeneity and content representativeness. Given the above findings and conclusions about the equating accuracy of various methods and for different sampled tests, IRT-145 also found that equating accuracy depended on content homogeneity of a set of test items and the content representativeness of anchor items. The finding suggests that when an anchor test was more content representative of its total test, regardless of the equating method used, the equating result for this test would be more accurate.

Evaluating effect of auto-correlation. As Raw-145, the estimates of equating accuracy yielded by IRT-145 are also susceptible to the artifact of auto-correlation, due to

overlap of items from the sampled tests and criterion tests. Therefore, the same strategy used to control the impact of the auto-correlation for the equating results based on Raw-145 was applied to improve the auto-correlation problem.

The resulting Pearson's *r*s, adjusted for the artifact of auto-correlation, are included in APPENDIX C. Compared to their corresponding unadjusted Pearson's *r*s, they show only a trivial amount of attenuation in their magnitudes (less than .01). In addition, the rank-order patterns of these indices of equating accuracy, before and after the adjustment, are pretty much the same. These findings suggest that the impact of the auto-correlation was not serious for the estimation of equating accuracy. The conclusions based on the unadjusted indices of equating accuracy should remain valid.

Reliability and validity evidences on anchor tests. Regarding IRT-145 as a "pseudo true score", the concurrent validity and reliability of the anchor tests of various sampled tests were estimated. The estimation was conducted in a way similar to the validity and reliability studies described earlier, where Raw-145 was regarded as the "pseudo true score". Using IRT-145, the evidences of validity and reliability for the anchor tests were found to be satisfactory. These assessment outcomes are recorded in APPENDIX D, with the outcomes from Raw-145.

The large positive Pearson's *r*s between the "pseudo true scores" and the scores on the anchor tests, shown in APPENDIX D, provide evidences of reliability and validity for each of the four sampled tests. These measures of validity (reliability) range from .839 to .917 and are all statistically significant at $\alpha = .01$. On average, the validity (reliability) of the anchor test embedded in sampled test PS is .917. The average validity (reliability) measures are .868, .854, and .860 respectively for the anchor tests of EW, PW, and SR.

These evidences of reliability and validity improve the chance for the research outcomes in this study to be valid.

Evaluation Using FE-long as a Criterion

The two original test forms -- Book A and Book B -- had many more items than the sampled test forms. These two forms were equated by the frequency-estimation equipercentile method, and the equating outcome was regarded as a more reliable criterion (FE-long) for evaluating equating accuracy because it was based on a test similar to but longer than the sampled tests. Using FE-long, the accuracy of various equating methods on equating the two forms of a sampled test was evaluated. As before, Pearson's *r* between the criterion score of FE-long and the resulting equivalent score on the sampled test was computed to represent the degree of equating accuracy of a particular equating method on a particular sampled test. The evaluation outcomes of equating accuracy are summarized in Table 18 (see the third numeric column).

Equating accuracy of various methods and auto-correlation. Using FE-long, all the outcomes of the four equating methods were estimated to be moderately accurate as before. The equating accuracy of various methods for different sampled tests ranged from .864 to .903. Across various sampled tests, the average index of equating accuracy for the IRT fixed-b method was .893 (see Table 18), very close to the index for the IRT-based linear transformation method (r=.888). The Tucker linear method and the equipercentile equating method had the same degree of accuracy on average (r=.887). However, compared to Raw-145 and IRT-145, FE-long often resulted in larger indices of equating accuracy. The larger Pearson's *r*s are in part attributed to the worsening artifact of auto-

correlation. In such cases, the impacts of the auto-correlation were more serious (than they were when the other criteria were used), because there was much more overlap between FE-long and the sampled tests. The sampled tests were all subsets of FE-long.

Moreover, using FE-long, the outcomes of the IRT-based equating methods were not always estimated to be more accurate than those outcomes from the non-IRT methods. This is somewhat different from the findings based on Raw-145 and IRT-145. Even when the outcomes of the IRT-based equating methods appeared to be more accurate, their improvements over the other methods were often not significant, or not as large as when the other criteria were used to estimate equating accuracy. The artifact of auto-correlation may explain such differences between the results from FE-long and the results from the previous two criteria. The worsening auto-correlation, associated with FE-long, could result in similar indices of equating accuracy. In such case, the dependent samples *t*-test could not detect the real differences. This explanation is supported by the facts that the indices of equating accuracy resulted from FE-long (shown in Table 18) had the narrowest range (.039) and the smallest standard deviation (.01). For Raw-145, the range was .063 and the standard deviation was .021. Evaluation outcomes of IRT-145 had the widest range (.098) and the largest standard deviation (.033).

Ideally, FE-long should be a more reliable criterion and thus provide an alternate way to study equating accuracy. However, such advantage of FE-long was smeared by its inherent problem of auto-correlation and its nature of being an arbitrarily selected criterion.

Equating accuracy for various sampled tests. The evaluation results from FE-long show that, regardless of the equating method used, the equating results for sampled test

PS were the most accurate among the results for all of the four sampled tests. On equating PS-A and PS-B, there were no statistical significant differences among the equating accuracy of the four methods. The average equating accuracy across the methods was r=.900. Across the methods, the average accuracy for SR was r=.890. However, on equating SR-A and SR-B, the outcomes of the IRT-based methods were slightly but significantly more accurate than the outcomes of the non-IRT methods.

Across the four sampled tests, the two IRT-based methods had the least accurate outcomes on equating PW-A and PW-B. For PW, while the outcome of the IRT-based linear transformation method was slightly (but significantly) less accurate than the outcomes of the non-IRT methods, significance test results also show that the outcome of the IRT-based fixed-b method was no different from the outcomes of the Tucker and the equipercentile methods.

Across the sampled tests, the non-IRT methods had the least accurate outcomes on equating EW-A and EW-B. For EW, the outcomes of the non-IRT methods were slightly (but significantly) less accurate than the IRT-based linear transformation method. However, significance test results also show that the outcomes of the Tucker and the equipercentile methods were no different from the outcome of the IRT-based fixed-b method.

In summary, many of the findings based on the criterion FE-long are not consistent with the findings based on the first two true-score based criteria. The improvement of the IRT-based equating methods over the non-IRT methods in equating accuracy is not clearly confirmed. In addition, the effect of content homogeneity of test items and the effect of content representativeness of anchor test, on estimating equating accuracy, are not

clear. Although the equating outcomes for PS seemed always more accurate than the outcomes for the other three sampled tests, regardless of the method used for equating its two forms, such advantage of PS was not always significant statistically. The patterns of the accuracy indices across the sampled tests were not clear, because the resulting indices of equating accuracy from FE-long had similar values. The inconsistencies between these findings and the previous findings are in part attributed to the problem of more serious auto-correlation underlying the criterion FE-long. They can also be partly accounted for by the vulnerability of FE-long, due to the fact that it was an arbitrarily selected criterion for evaluating equating accuracy. An implication from these conclusions is that accuracy of an arbitrary criterion itself is important, and should receive special attention, in evaluating effectiveness of the other equating outcomes.

Estimation Bias Due to an Arbitrary Criterion -- FE-short

It is common practice to use some arbitrary criterion for evaluating equating accuracy. However, the estimation of equating accuracy based on an arbitrary criterion is often biased because of the subjectivity of the particular criterion used. Therefore, it is one of the particular interests of this study to investigate the potential bias due to the arbitrary nature of a criterion for evaluating equating accuracy.

Measures of equating accuracy for EF-short. The arbitrary criterion being studied was the outcome of the equipercentile method on equating the two forms of a sampled test (EF-short). Such an arbitrary criterion was established for each of the four sampled tests so that the relative equating accuracy of the other three equating methods (the Tucker method and the two IRT-based methods) could be estimated. In addition to the

index of equating accuracy (Pearson's *r*), the root-mean-squared deviation (RMSD) statistic was also used as a second measure for equating accuracy. RMSD was appropriate for the estimation of equating accuracy in this study, because the outcomes of the equating methods being evaluated and the criterion FE-short were all based on the same sampled test. As a result, the resulting equivalent scores from the three methods and the criterion score were on the same scale.

Bias from the arbitrary nature of EF-short. Using FE-short, the indices of equating accuracy were computed for the Tucker method and the two IRT-based methods. The resulting indices are presented in Table 17 (see the bordered Pearson's *r*s). In summary, these indices differed somewhat from the indices resulted from the previous three criteria. While the previous indices suggest moderate equating accuracy of various methods on different sampled tests, the indices produced by FE-short suggest much higher degree of equating accuracy. The indices based on FE-short ranged from .944 to 1, while the indices based on FE-long ranged from .864 to .903, those based on Raw-145 ranged from .832 to .895, and those based on IRT-145 ranged from .819 to .917. These findings suggest potential bias due to the use of an arbitrary criterion.

Averaged across various sampled tests, the mean accuracy of the outcomes from the IRT-based linear transformation method and the fixed-b method were .961 and .964 respectively. The similarity between the two IRT-based methods is consistent with previous findings produced by the other criteria. Based on FE-short, a finding dramatically different from the previous findings is that the outcome of the Tucker linear method was significantly more accurate than the outcomes of the IRT-based methods. The indices of equating accuracy for the Tucker method over various sampled tests were

all close to 1, suggesting nearly perfect accuracy. This finding also suggests that FE-short was biased in evaluating equating accuracy. Arbitrarily selected to be a criterion, FE-short overestimated the accuracy of the outcome from the Tucker method. Relatively speaking, this non-IRT-based criterion might have underestimated the outcomes from the IRT-based equating methods. This conclusion is compatible to those reached in the literature about the bias against the IRT-based equating outcomes.

Despite its drawbacks discussed above, FE-short still found the equating results on PS the most accurate among the equating results on all four sampled tests, regardless of the equating method used. This finding provides some evidence for the effect of content representativeness of anchor test.

Bias due to index of equating accuracy. The evaluation outcomes of equating accuracy measured by the RMSD statistic are summarized in Table 19. Overall, these outcomes agreed with the previous findings about equating accuracy measured by the Pearson's *r*. However, there were still small discrepancies between the outcomes yielded by the RMSD and the Pearson's *r*.

The RMSDs for various equating methods on different sampled tests in Table 19 suggest that the equating outcomes of the Tucker method were more accurate than the outcomes of the two IRT-based methods all the time. This is consistent with the conclusion reached by the Pearson's *r*s about the differences between the equating methods. Across the four sampled tests, the RMSD for the Tucker method ranged from .102 to .177, while the RMSDs for the IRT-based methods ranged from 1.871 to 2.314. They indicate that the IRT-based equating outcomes deviated more from the criterion equating (FE-short) than the outcome of the Tucker Method.

Table 19 - Root-Mean-Squared-Differences for Evaluating Equating Accuracy

Equating Method	**Sampled Test**			
	PS	EW	PW	SR
Tucker Linear Method	0.109	0.102	0.177	0.170
IRT-Based Linear Transformation Method	1.871	1.955	2.314	2.019
IRT-Based Fixed-b Method	1.875	1.994	1.972	2.018

Different from the finding based on the Pearson's *r*s, the RMSDs suggest that only when IRT-based equating methods were used, the equating results for sampled test PS were more accurate than those for EW, PW, and SR. The Tucker method yielded more accurate result for EW than for PW, SR, and PS. In addition, when the IRT-based linear transformation method and the Tucker method were used, the equating results for PW were the least accurate among the results for all sampled tests. The IRT-based fixed-b method yielded the least accurate result for SR.

Comparing the estimation outcomes resulted from the Pearson's *r* and the RMSD, clearly, using different statistics to represent equating accuracy may lead to somewhat different estimations. Therefore, when assessing accuracy of equating outcomes, it is important to know how well a particular index of equating accuracy serves its purpose. The natures of the statistics used to represent the degree of equating accuracy should be taken into account when interpreting the estimation results of equating accuracy.

This study did not compute the RMSD statistic for all the equating outcomes and mainly relied on using Pearson's *r* to represent equating accuracy. The reason is that the resulting equivalent scores from different methods and the various criterion scores were not on the same scale most of the time. In such cases, Pearson's *r* provides an efficient and direct way to study the accuracy of equating outcomes. Although it is possible to transform the resulting equivalent scores and put these scores and the criterion scores onto the same scale, this study decided not to apply such transformation because: (a) score transformation may introduce more errors, and (b) score transformation may complicate the interpretations and implications of the analysis outcomes. In addition, from a practical perspective, transformed scores usually require additional explanations and justifications.

Therefore, this study chose to use Pearson's *r* to keep the resulting equivalent scores precise and the estimation outcomes straight, and to make the interpretations of analysis results direct. Taking into account the limitations of the Pearson's *r*, such as the issue of auto-correlation, this study presents and discusses the outcomes of equating accuracy with cautions.

Advantages of Using Multiple Criteria

The use of multiple criteria for evaluating equating accuracy in this study proves to be very informative. The comparisons among the resulting evaluation outcomes of various criteria render an opportunity to thoroughly study the effectiveness of various equating methods and the effect of content homogeneity on equating accuracy.

Uses of Raw-145 and IRT-145. The criteria Raw-145 and IRT-145 were both computed using the 145 common anchor items. These anchor items show adequate internal consistency. The Cronbach's α was .866 for the raw scores on the 145 items, and the Cronbach's α was .869 when the item scores were standardized to have unit variances (n=2,241). The evidence of internal consistency suggests that the criteria Raw-145 and IRT-145 were reliable. Raw-145 and IRT-145 also correlated positively and strongly to each other (r=.982 at α=.01). They were appropriate for evaluating the equating accuracy of the four methods in this study, because they were conceptually the "pseudo true scores".

Raw-145 and IRT-145 also complemented each other in improving the estimation of equating accuracy. On one hand, Raw-145 did not over-estimate the equating accuracy of the outcomes from the IRT-based methods. Instead, it provided conservative estimates

of equating accuracy. On the other hand, IRT-145 was not susceptible to the problem of person-dependent and item-dependent. By incorporating both Raw-145 and IRT-145, the assessment of equating accuracy in this study was less prone to biases. Overall, these two criteria yielded very similar estimates of equating accuracy.

Uses of FE-long and FE-short. Both the other two criteria -- FE-long and FE-short -- produced evaluation outcomes that were short of interpretable patterns and largely inconsistent with the outcomes from Raw-145 and IRT-145. This finding reflects the drawback of FE-long and FE-short for being subjective and arbitrary. FE-long was expected to be a more reliable criterion for evaluating the equating accuracy of the four methods on the sampled tests, but its assessment outcomes were influenced by serious auto-correlation and thus deviated considerably from those of Raw-145 and IRT-145. No better than FE-long, FE-short led to conclusions that were dramatically different from those of Raw-145 and IRT-145.

Despite the inability of FE-long and FE-short in producing precise estimates of equating accuracy, one implication from the findings about the flawed criteria is that it is critical to take into account the estimation errors accompanying an arbitrary criterion.

In summary, the use of multiple criteria and comparing their resulting assessment outcomes guarded the estimation of equating accuracy from being biased by a single arbitrary criterion. The results from using the strategy also cast valuable insights for equating practice and future research, on selecting appropriate criteria for evaluating equating accuracy.

Construct Validity Issues

The test of the professional examination analyzed in this study was written to measure an examinee's professional ability (knowledge or skills). The professional ability of an examinee partly depends on the examinee's professional experience. In theory, the more years of professional experience an examinee had accumulated, the more likely that the examinee would score higher on the test. Such effect of professional experience should exist for the two examinee groups taking different forms of a test, after the test scores from the different forms are equated. Therefore, after the test forms were equated, the construct validity of the test could be investigated by comparing the resulting average equivalent scores of the two examinee groups. Based on this scenario, this study conducted an investigation on the construct validity of the professional in-training test.

Specifically, using a set of equivalent scores on the original test, the effect of test form, the effect of years of experience, and the interaction between these effects on the examinee's performance were studied. Since the equivalent scores for the original test had been obtained by the equipercentile equating method in previous analyses for equating accuracy, for the sake of completeness and convenience, this set of equivalent scores were used for the group comparisons. The group means of the equivalent scores by test form and by years of experience are summarized in Table 20. These group means were graphed in Figure 8 to facilitate the inspection on the interaction effect of test form and years of experience. In summary, there are evidences of construct validity for the equated original test forms, and the equating outcomes were determined to be adequate.

If there were test form by experience interaction, the test would be lacking construct validity, or the statistical adjustment made via the equipercentile equating was

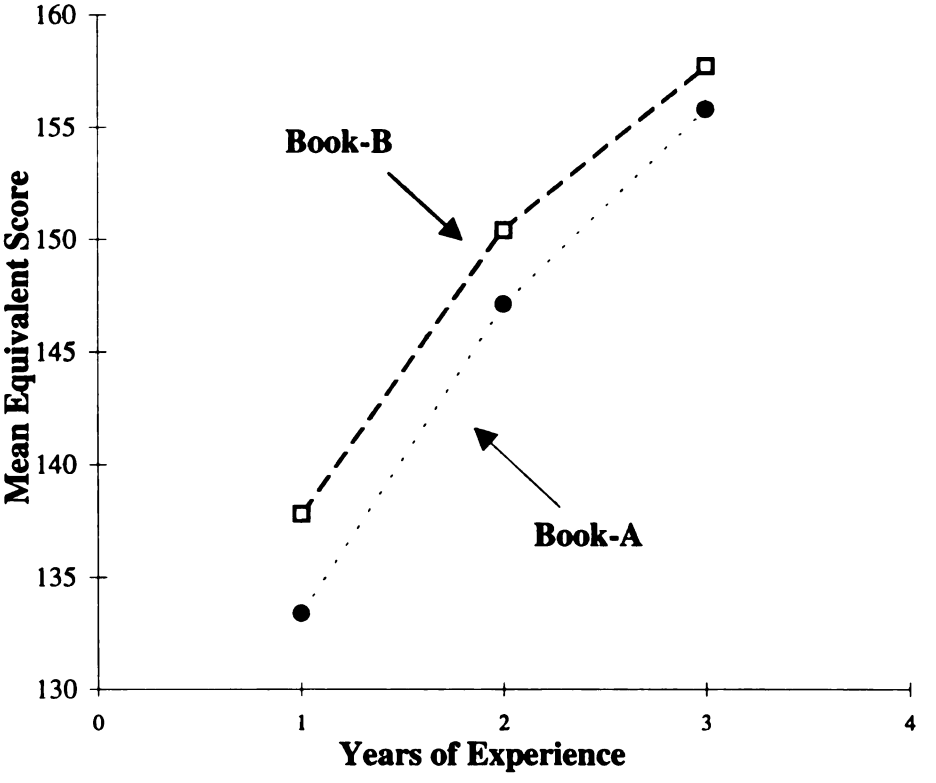
Table 20 - Average Equivalent Scores of Examinee Groups on the Original Test by Test Form and Years of Experience

Test Form	Years of Experience	Mean	Std. Dev.	n
Book A	1	133.374	17.682	380
	2	147.116	15.087	352
	3	155.816	13.185	359
Book B	1	137.826	16.366	409
	2	150.435	15.262	367
	3	157.761	13.148	361
Total		146.740	17.682	2228
2×3 ANOVA Result -- For "form×years" interaction: $F=1.27$, $p=.281$, $df_{form \times year}=2$, $df_{error}=2,222$, $\alpha=.05$				

Years of Experience	Mean	Std. Dev.	n
1	135.682	17.147	789
2	148.810	15.256	719
3	156.792	13.193	720

Test Form	Mean	Std. Dev.	n
Book A	145.192	18.060	1091
Book B	148.226	17.188	1137

Note. (1) The equivalent scores were obtained by the equipercentile equating method.
(2) The few cases missing information about years of experiences were excluded from the analysis.



Note. (1) The equivalent scores were obtained by the equipercentile equating method, using all the items from the original test.

(2) The few cases missing information about years of experiences were excluded from the analysis.

Figure 8 - "Test Form" by "Years of Experience" Interaction Effect

inadequate. As shown in Figure 8, there is no crossed interaction between test form and years of experience. The result of a significance test for the interaction effect further indicates that there was no statistically significant interaction at α=.05 (F=1.27, p=.281). Moreover, the means plot in Figure 8 and the group means presented in Table 20 show that the more experienced group always had higher average scores than the other group(s), no matter which test form was taken. The multiple comparisons using Tukey and Scheffe's tests further indicated significant differences among the groups differing in years of experience (p=.000 for all of the possible comparisons). All of these findings suggest that the equated test forms had construct validity.

As shown in Table 20, regardless of their professional experiences, the group taking Book B always scored higher on average than the group taking Book A. The significance test for the group difference further concludes that the two groups taking different test forms differed significantly (p=.000) in their test performances at a significance level of α=.05. This finding is not surprising, since it has been found in previous chapters that the two examinee groups were slightly different in their abilities.

Issues of Test Dimensionality

Assuming unidimensionality, the equating outcomes based on the 3PL IRT model were conducted and their outcomes were found satisfactory. However, because there were 23 core content areas nested within the single overall content domain underlying the sampled tests, whether the IRT assumption of unidimensionality held for the equatings in this study is not clear. If there were indeed more than one traits underlying the test being studied, regardless of the violation of the unidimensionality assumption, the satisfactory

IRT-based equating outcomes of this study would indicate robustness of the 3PL IRT model. Nevertheless, in such case, equatings based on multidimensional IRT models may be good alternatives. Therefore, to better understand the nature of the test being studied and to probe its impact on the equating practice, this study explored the dimensionality issues with a subset of manageable data.

Specifically, confirmatory and exploratory factor analyses were used to investigate the dimensionality of a 45-item sub-test (all the items were common anchor items). To avoid complicating the investigation with too many items or too many underlying factors (there were potentially 23 factors corresponding to the 23 core content areas), this study focused on the small sub-test. To include as many examinees as possible, the sub-test only contained anchor items. Responses of the entire examinee population (n=2,241) on these 45 anchor items were analyzed. In theory, there were three distinct factors underlying the 45-item sub-test, because all the items were drawn from the sampled test PS and PS only covered three of the 23 core content areas. The outcomes of the factor analyses are summarized and discussed below.

Confirmatory Factor Analyses

Considering the content structure of the sub-test, these models were appropriate for confirmatory factor analyses: (a) a model with three underlying factors, (b) a model with three first-order factors and one second-order factor, (c) a model with one overall factor only, and (d) a model with three single-factor sub-models, each dealt with items from the same core content area (14 of the 45 items were from the same area, another 12 were from a second area, and the remaining items were from the third area).

For dichotomously scored items, the Pearson's product-moment correlation based on normal scores is biased (inconsistent). The standard errors of the parameter estimates yielded by the generalized least square (GLS) method are not correct because of the wrong formula used (Jöreskog & Sörbom, 1993). Jöreskog & Sörbom (1993) recommended that tetrachoric correlation be estimated for each pair of the dichotomous items and the resulting correlation matrix be analyzed by the generally weighted least squares (WLS) method, using LISREL. Therefore, following these recommendations and using LISREL, tetrachoric correlations were estimated and used for the factor analyses. The inverse of the estimated asymptotic covariance matrix of these tetrachoric correlation coefficients was used as the weight matrix for the WLS method.

Results from the chi-square tests for overall model-data fit suggest that none of the theory-driven confirmatory factor analysis models fit the data. That is, the content structures specified in the various factor models for the sub-test were significantly different from the content structure of the sub-test implied by the actual data. However, it should be noted that the chi-square test was very sensitive to large sample size. Given the large sample size in this study, the test statistic was very likely to have large value. As a result, the test was more likely to show that there was significant difference between the theoretical model and the observed model.

Exploratory Factor Analyses

Exploratory factor analyses were conducted to further explore the dimensionality of the sub-test. The tetrachoric correlation matrix estimated previously by PRELIS was used as input data for various exploratory factor analysis models. Using statistical package

SAS on Unix, these factor analyses were conducted: principal component analysis, principal factor analysis, maximum likelihood factor analysis, and α factor analysis. In summary, these factor analyses suggest that there was more than one factor underlying the sub-test and the dimensionality issues were complex.

Principal component analysis. The results of the principal component analysis suggest one dominant component underlying the 45 items, although sixteen components were retained using the eigenvalues-greater-than-one criterion. The first component had an eigenvalue of 6.902, which accounted for 15.3% of the standardized variance of the correlation matrix. Each of the other components, however, had eigenvalue less than 1.55 and explained for less than 3% of the variance.

The first principal component appeared to be much more important than any other components, despite the fact that multiple components were required to provide an adequate summary of the data. In addition, all of the 45 items had positive loadings (ranging from .174 to .681) on the first principal component. The scree plot of eigenvalues in Figure 9 provides visual evidence of the single dominant factor.

Principal factor analysis. The principal factor analysis used the squared multiple correlations for the prior communality estimates. As a result, the total eigenvalue of the correlation matrix reduced to 9.714 and the average eigenvalue was 0.216. By the default "proportion" criterion (SAS Institute Inc., 1989), eight factors were retained. The first factor (eigenvalue equals 6.173) explained 63.6% of the variance, while each of the other factors accounted for no more than 8%. The resulting pattern of principle factors was similar to the pattern of principal components, and factor loadings of various items on the first factor were all positive. The resulting scree plot is presented in Figure 10.

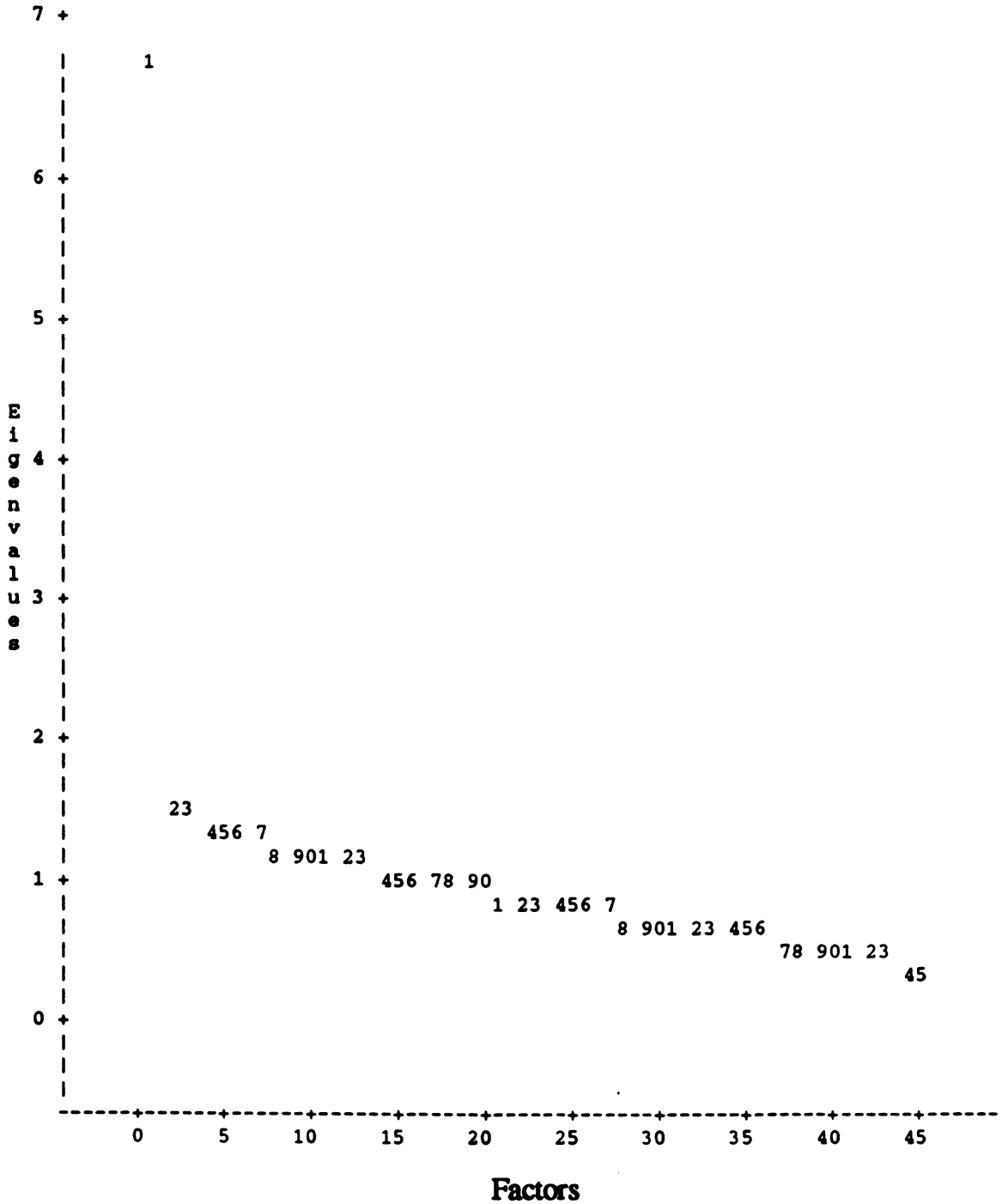


Figure 9 - Scree Plot of Eigenvalues for Principal Component Analysis

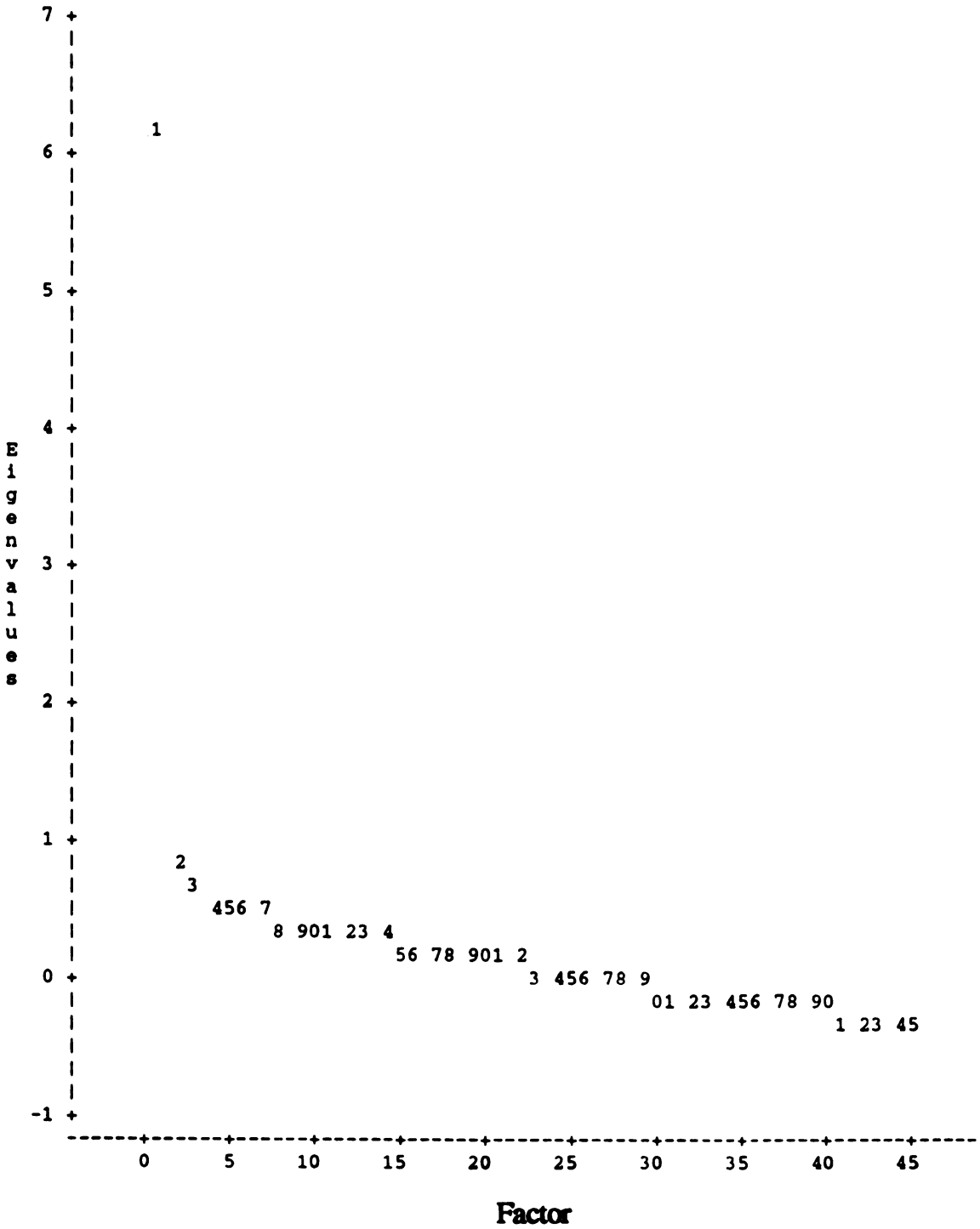


Figure 10 - Scree Plot of Eigenvalues for Principal Factor Analysis

Similar to the conclusion reached by the principal component analysis, the results of the principal factor analysis suggest one dominant factor underlying the 45 items, despite the fact that a total of eight factors were needed (according to the default "proportion" criterion) for an adequate summary of the data.

Maximum likelihood and α factor analyses. Maximum likelihood method was used to study the factor analysis models with one, two, and three factors respectively. The factor analysis results indicated that more than three factors were needed to adequately summarize the data. This finding is consistent with the results of the principal component analysis and the principal factor analysis, which yielded 16 components and eight factors respectively.

Different from the outcomes of the three factor analyses described above, an α factor analysis extracted three factors out of the 45 items using rotation technique that results in maximum variance. However, there were no clear patterns to link these three factors to the three core content areas, to which the 45 items belonged. The result of the α factor analysis therefore reflects complexity of the dimensionality issues.

Unidimensionality Assumption for the Test

Overall, all the outcomes of the various factor analyses implied multidimensionality for the 45-item sub-test, despite the fact that these empirical outcomes were somewhat inconsistent. However, based on these outcomes, it is also likely that there was one dominant underlying factor. Given the single dominant factor underlying the test data, regardless of the other less influential factors, the unidimensional IRT model would be robust in fitting the test data.

It is believed and broadly accepted that, other than the strict requirement of one underlying trait for the IRT unidimensionality assumption, unidimensionality only requires a dominant factor or trait that influences test performance (Hambleton, Swaminathan, & Rogers, 1991). This is in part because most of the achievement tests, such as the one analyzed in this study, are not designed to be unidimensional and are unlikely to be unidimensional. Therefore, in our case, the empirical findings from various factor analyses suggest that the IRT assumption of unidimensionality was likely to hold, because there was a single dominant factor underlying the sub-test data.

The unidimensionality issue should also be addressed from a theoretical point of view. Given the theory that there were three correlated traits (each corresponds to one of the three core content areas relating to the single overall content domain) underlying the sub-test data. It can be hypothesized that there was a unidimensional trait that was conceptually a combination of these three underlying traits (Yen, 1984; Reckase, Ackerman, & Carlson, 1988). As a result, the unidimensional three-parameter IRT model was appropriate for the data, and the equatings based on the IRT model should be theoretically sound. The satisfactory outcomes of the IRT-based equating methods, discussed earlier in this chapter, provide substantial empirical evidence in supporting of this hypothesis. From the same theoretical perspective, the ability measured by the original professional test could also be considered as a composite unidimensional trait that had 23 component traits (corresponding to the 23 core content areas). However, more studies are needed to further research into the issues of test dimensionality.

Chapter 10

SUGGESTIONS

Grounded in the findings and conclusions reached in previous chapters, suggestions can be made to improve the equating practice and future research in common-item equating. With considerations of study limitations, my suggestions address these issues: (a) selection of equating method, (b) construction of test forms with embedded common anchor items, (c) controlling effect of anchor length, (d) use of multiple criteria for evaluating equating accuracy, (e) selection of index representing equating accuracy, (f) investigation of construct validity, (g) issues of test dimensionality, and (h) alternative approaches for common-item equating.

Selection of Equating Method

When equating test forms that have negatively skewed score distributions, using the common-item equating design, IRT-based equating methods are recommended to achieve highly precise estimation of equivalent scores if the degree of equating accuracy is a top priority. Otherwise, either the IRT-based methods, or the Tucker method, or the equipercentile method, will be adequate. This is because all these methods are likely to have moderately accurate results and these results will not be too different.

If it is allowed in time and cost, various equating methods should be applied to equate scores from different test forms of high-stake examinations. The resulting outcomes can be compared to one another to identify the set of equivalent scores that are the most precise. The choice of a particular equating method should take into account factors such as the theoretical appeal and empirical advantages of the method, the policy of an equating program regarding the acceptable level of accuracy, the cost and the amount of work required, and the possibility of finishing the equating work within a restricted timeline.

Construction of Test with Anchor Items

A testing program can limit the content coverage of a test to a small number of topics to increase the possibility of getting precise equating outcomes. This is because, according to the findings of this study, equating outcomes for a test with items that are more content homogeneous is likely to be more accurate than the equating outcomes of a set of less content homogeneous items.

In addition to manipulating content homogeneity of the items in a sampled tests, the item sampling design of this study also created various sets of anchor items that were more or less representative of their corresponding total tests in content. As a result, the same research findings leading to the effect of content homogeneity also suggest effect of content representativeness of anchor items on equating accuracy. This study found that the equating outcomes of a test with more content representative anchor items were likely to be more accurate than those outcomes of a test containing less content representative anchor items, regardless of the equating method used. Therefore, when alternate forms of

a test are constructed and a set of common anchor items are embedded in each of the form, it is important to select anchor items that are representative of the total test in content.

Controlling Effect of Anchor Length

If this study is to be replicated with more flexibility in data collection and manipulation, I would like to create various sampled tests with equal numbers of anchor items. By doing so, the interpretations of the effects of content homogeneity and content representativeness will be more definite and direct.

Equal Anchor Length

The sampled tests created in this study varied slightly in their lengths and had different numbers of anchor items. Such variations in test length and anchor length are due to the nature of the in-training test being analyzed. Because the original test did not have equal and sufficient numbers of items from all of the 23 core content areas, it was not easy to create ideal sampled tests with equal length and equal numbers of anchor items. For the control over a better condition to study the effects of content homogeneity and content representativeness, this study traded in its controls over test length and anchor length. As a result, there is a slight chance that the effect of content representativeness of anchor items was confounded with the effect of anchor length. From time to time, educational researchers have to choose between working with ideal but somewhat unrealistic research conditions and dealing with practical issues that lack perfect solutions. This study restriction on test length and anchor length represents one of such dilemmas.

The strategy used in this study to accommodate the imperfect situation caused by unequal anchor lengths was to include a sufficiently large number of common anchor items in each of the sampled test forms. It is hoped that the impact of the differential anchor lengths could be minimized. However, if the sampled tests had the same lengths and the same number of anchor items, the results and conclusions of this study about equating accuracy would be more convincing. Therefore, given better research conditions in the future, I would like to exercise better controls over test length and anchor length.

Fewer Anchor Items

In this study, all of the four equating methods produced adequate equating outcomes and these outcomes were similar to a great degree. For example, the Tucker method and the equipercentile equating method yielded almost identical equivalent scores. However, this result is somewhat bothersome, because one method is based on a linear approach and another method is based on a non-linear approach. There are three possible reasons for this interesting finding: (a) prior to various equating procedures, the two forms of the same sampled test were already very similar to each other due to careful test construction, (b) all of the four equating methods were indeed effective to a similar degree, and (c) because there were so many anchor items embedded in each test form (more than half of the items in a sampled test form were anchor items), the two forms of the same test were so similar such that any equating method (in our case, the Tucker or the equipercentile method) would produce good equating result.

In this study, due to the nature of the test data, each of the sampled test forms had many anchor items. Partly due to this reason, all of the equating outcomes tend to be

accurate. Therefore, all the indices of equating accuracy from different equating methods over different sampled tests looked quite similar to each other with large values.

If this study is to be replicated, I will only allow a small number of anchor items to be embedded in each of the sampled test forms. By allowing a small number of anchor items in a replication study, I should be able to detect the indeed differences among the equating outcomes from various methods, and the variations in the accuracy of these equating outcomes. I expect that these differences and variations will be larger than the cases when a longer anchor is used. Given a small number of anchor items, the effects of content homogeneity and content representativeness can also be better studied. I also expect these content-related effects to be more obvious.

Multiple Criteria for Evaluating Equating Accuracy

Equating accuracy can be better estimated, if the criterion used for evaluating equating accuracy is less prone to errors such as overestimation and underestimation. In equating practice, however, arbitrary equating outcomes are often used as convenient criteria. This is because no one single equating outcome can be determined as the best equating of all. The major drawback of an arbitrarily selected criterion is that it does not address equating accuracy directly. Thus, it can lead to erroneous estimation.

Coping with the disadvantageous situation due to the lack of an absolutely better criterion, this study employed multiple criteria for evaluating equating accuracy to gain extra information on the accuracy of various equating outcomes. By carefully comparing information from different criteria, with considerations of the strengths and limitations of various criteria in estimating equating accuracy, equating accuracy can be better evaluated.

In short, using multiple criteria, plausible estimation of equating accuracy is achievable and the risk of introducing bias by using one single criterion can be avoid or reduced. Therefore, I recommend this strategy for future studies of estimating equating accuracy.

Two of the four criteria used in this study were theoretically sound and yielded consistent estimation results for equating accuracy. The way that these two criteria were established and used to evaluate equating accuracy can be followed, or tailored and used, in future studies.

Selection of Index Representing Equating Accuracy

The use of Pearson's *r* and RMSD as alternate indices of equating accuracy in this study revealed that different indices representing the degree of accuracy did not always yield the same estimation outcomes. Small discrepancies were found between the estimations produced by the Pearson's *r* and RMSD. Although, the overall outcomes of these two statistics agreed most of the time. Future studies can be designed to explore the effectiveness of various statistics serving as indices of equating accuracy.

Construct Validity Issues

Due to limited availability of demographic information, only the effect of years of experience on examinee's test performance was investigated to provide evidence of construct validity for the in-training test. In the future, if more information regarding the characteristics of examinee groups is available, validity studies should be more thoroughly studied using this information.

Test Dimensionality Issues

In addition to gathering empirical evidence of test dimensionality, content specifications of test items and the overall content domain(s) underlying a test should also be considered to make realistic judgment about the dimensionality of the test. The characteristics of examinees and situational factors that are likely to influence examinees' test performances should also be considered. For instance, suppose the examinees in this study, who participated in different in-training programs, received differential amount of drills on previous-year tests and different type of coaching on test-taking skills. Then, the examinees' test performances would not only be influenced by their ability but also would be influenced by the variations in the in-training programs. Such program variations would affect the result of an empirical study on test dimensionality, and the result would be less likely to show unidimensionality.

Applications of Study Design and Techniques

The equating designs and techniques used in this study can be applied to the other contexts for equating research dealing with different types of scores, score distributions, or item formats. For instance, the item-sampling schemes and the strategy of multiple-criteria for evaluating equating accuracy may provide useful insights for designing an adequate equating program for testing programs such as performance assessment.

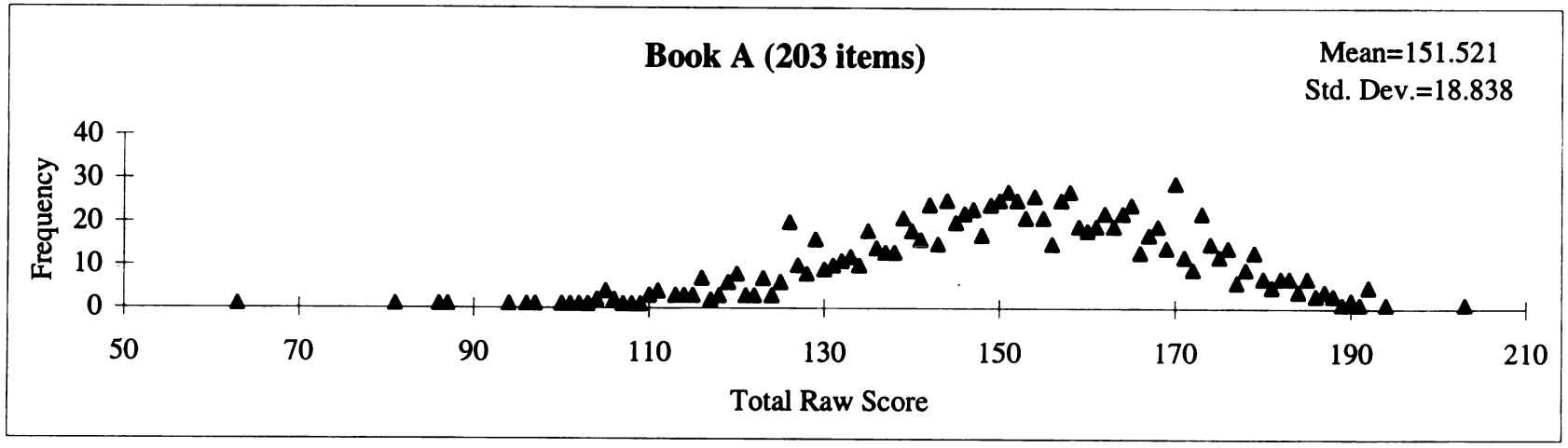
The applications may also be tailored to fit the needs of new equating contexts. For example, if a cut-off score in a particular score range is desired, because the criteria developed in this study did not care about any particular score range, a different criterion for accuracy can be developed to accommodate the need of higher precision at the cut-off.

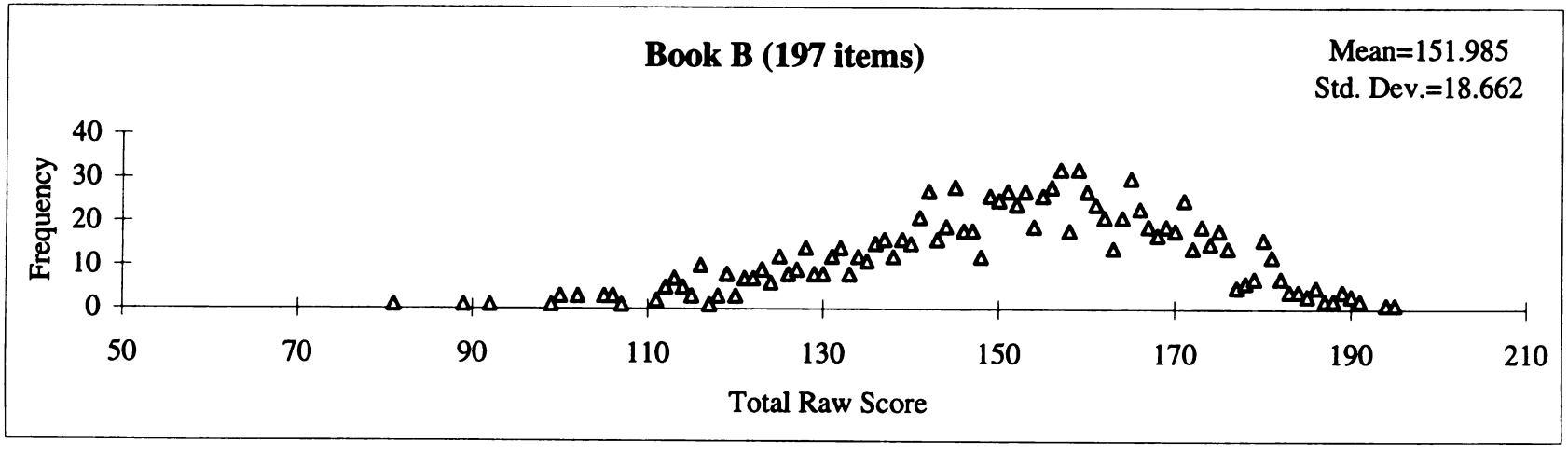
APPENDICES

APPENDIX A

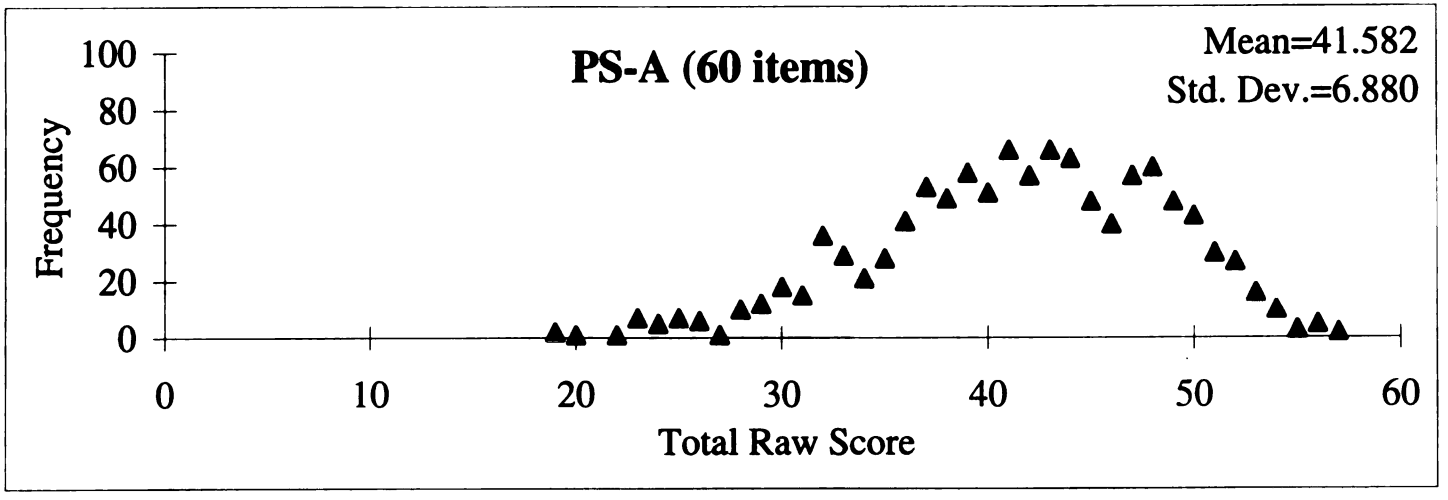
APPENDIX A

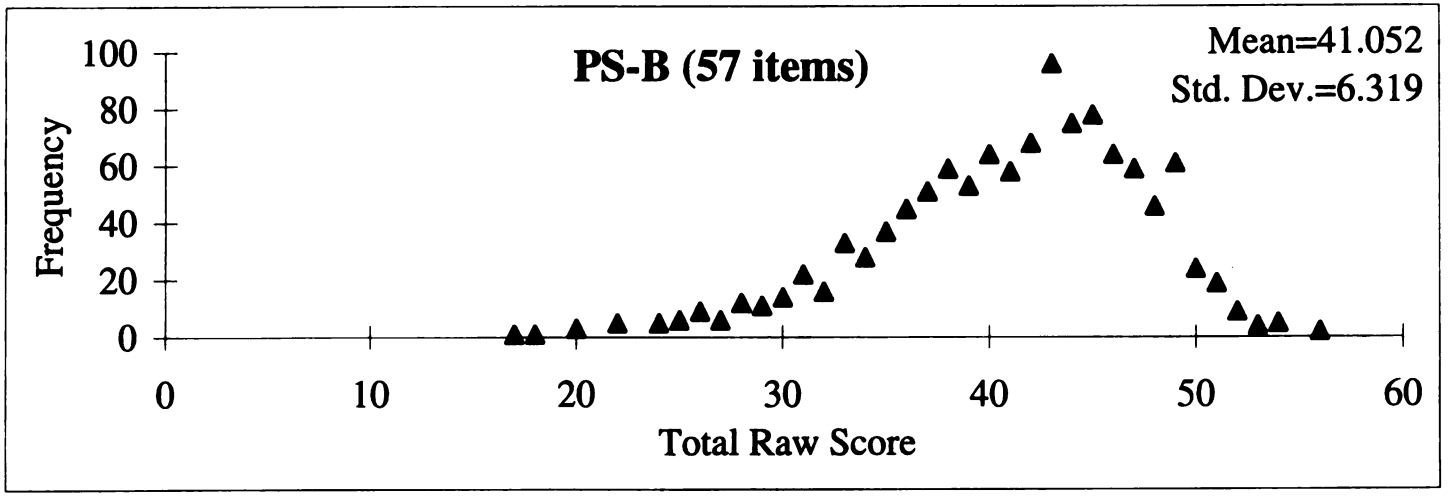
Distributions of Total Raw Scores for Various Test Forms

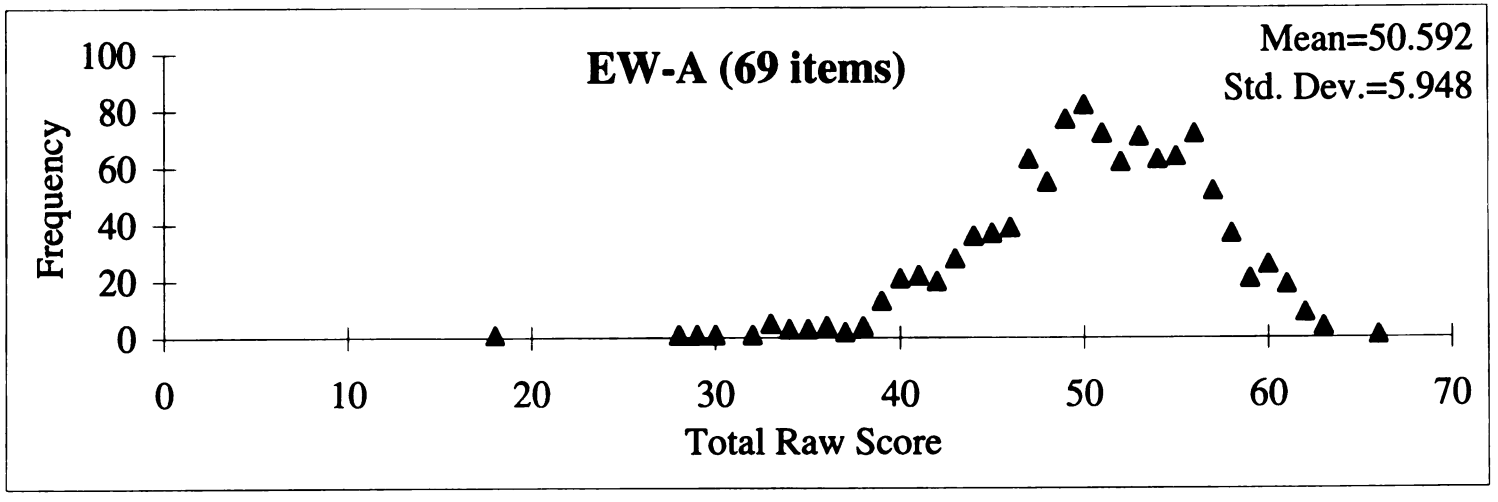


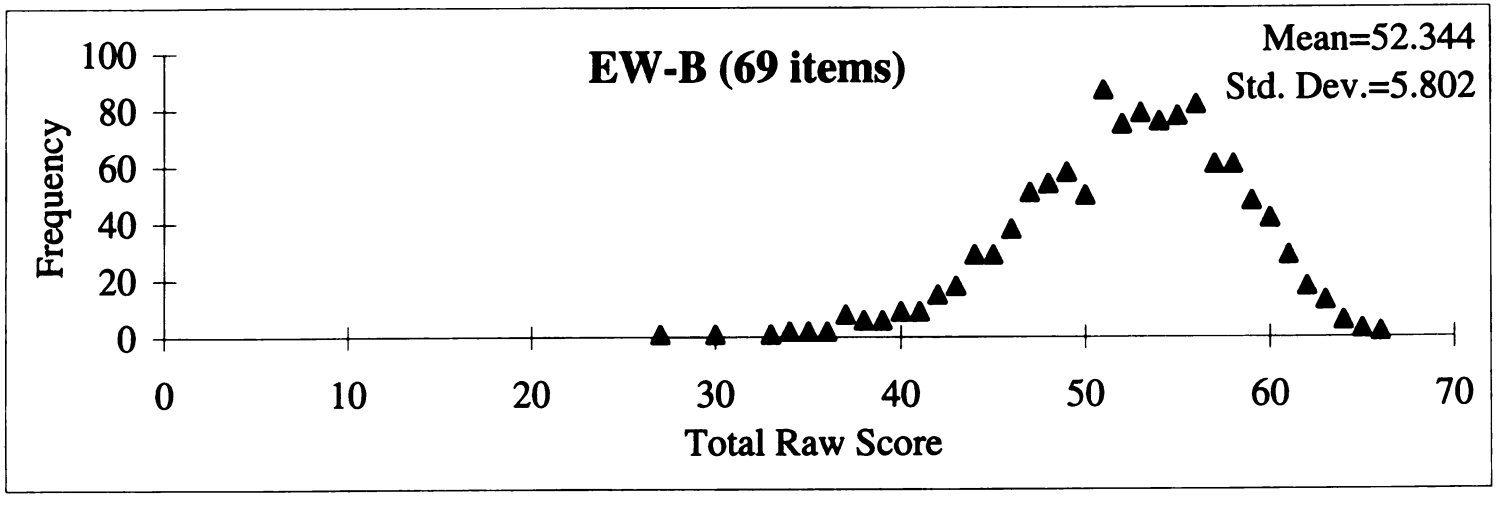


APPENDIX A (cont'd)

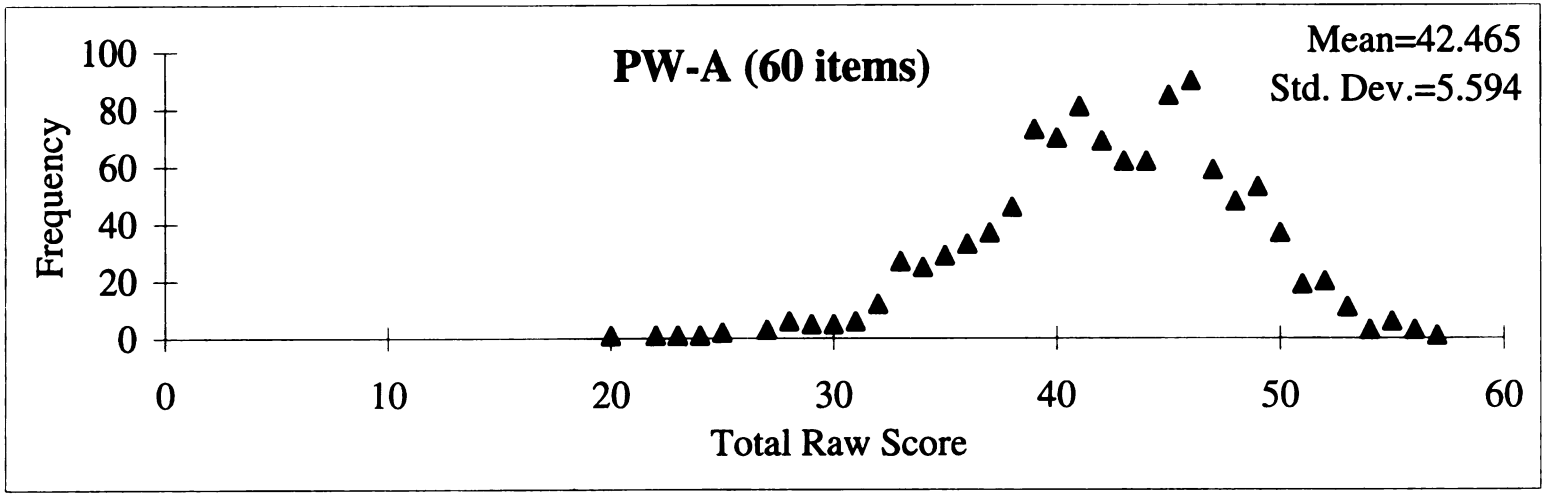


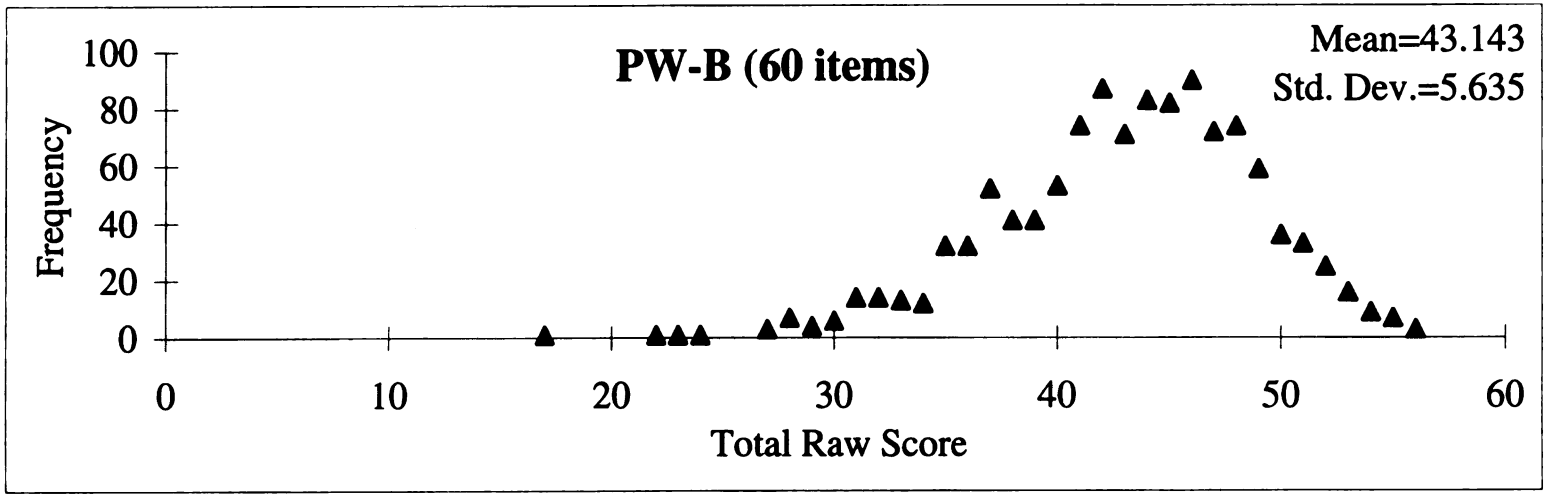



EW-A (69 items)
Mean=50.592
Std. Dev.=5.948
Frequency
100
80
60
40
20
0
0
10
20
30
40
50
60
70
Total Raw Score

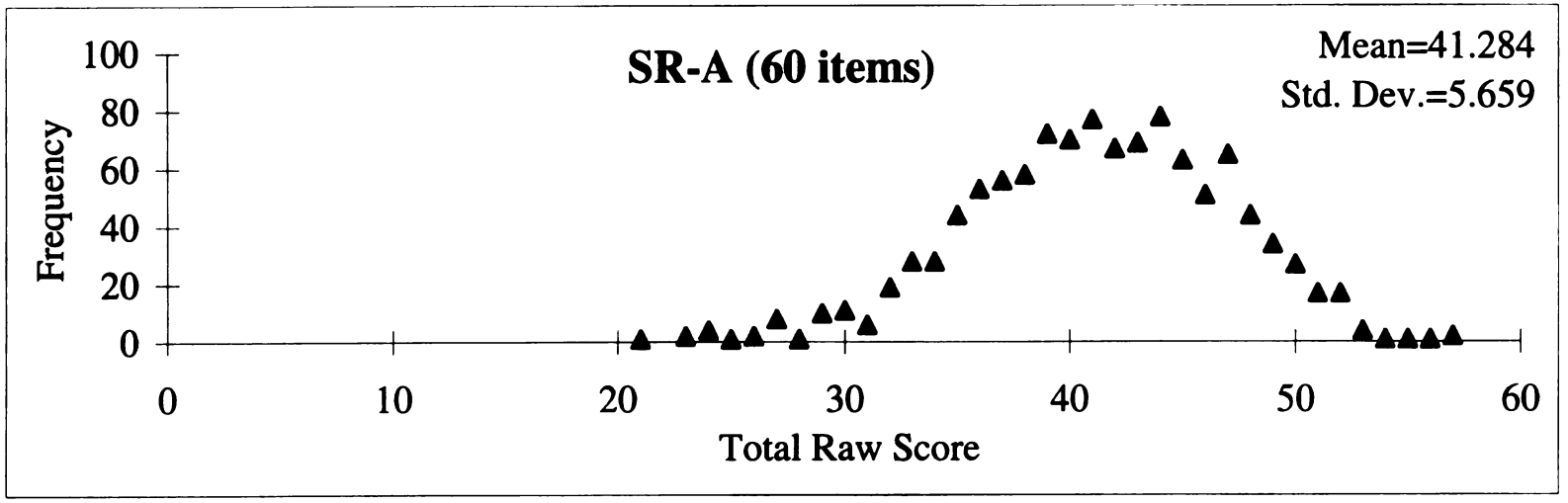

EW-B (69 items)
Mean=52.344
Std. Dev.=5.802
Frequency
100
80
60
40
20
0
0
10
20
30
40
50
60
70
Total Raw Score

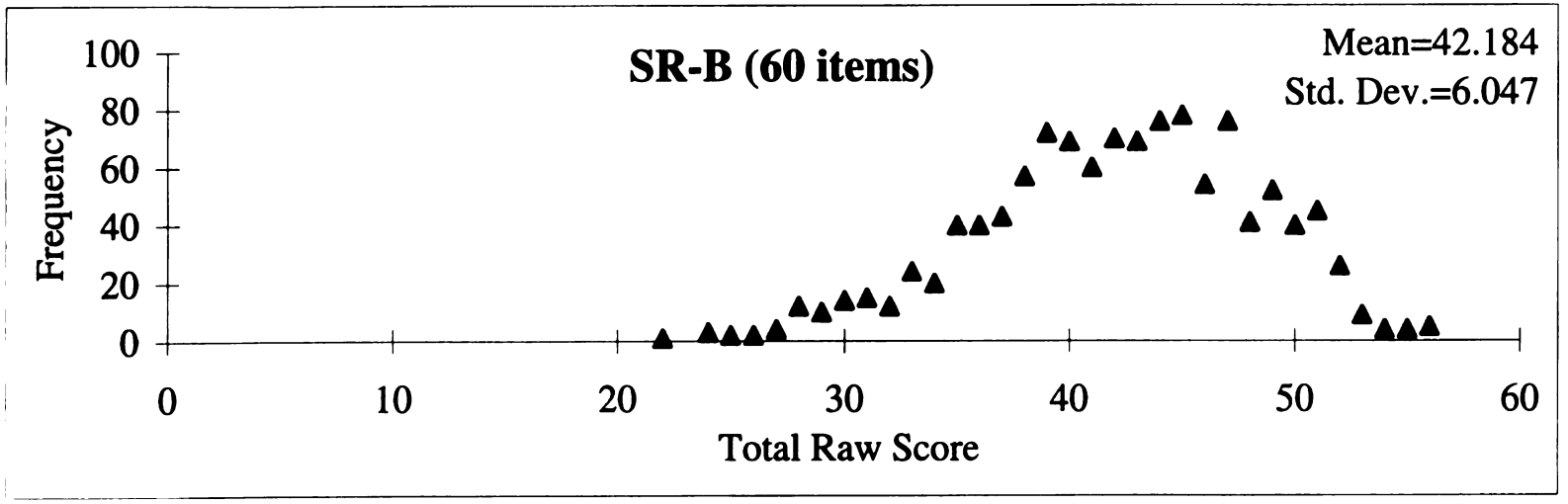
APPENDIX A (cont'd)





APPENDIX A (cont'd)

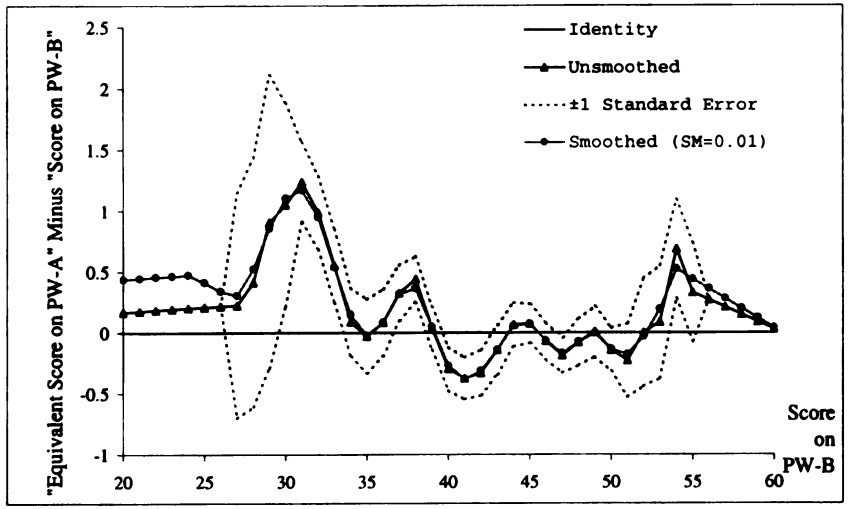


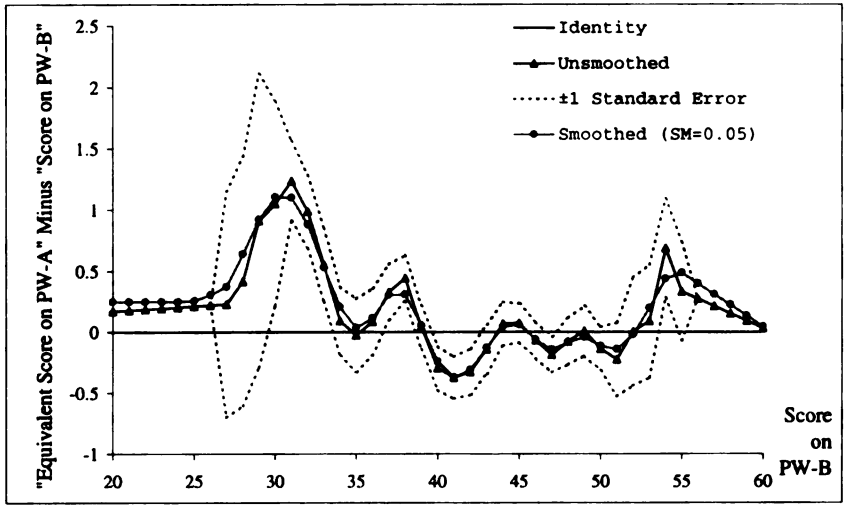


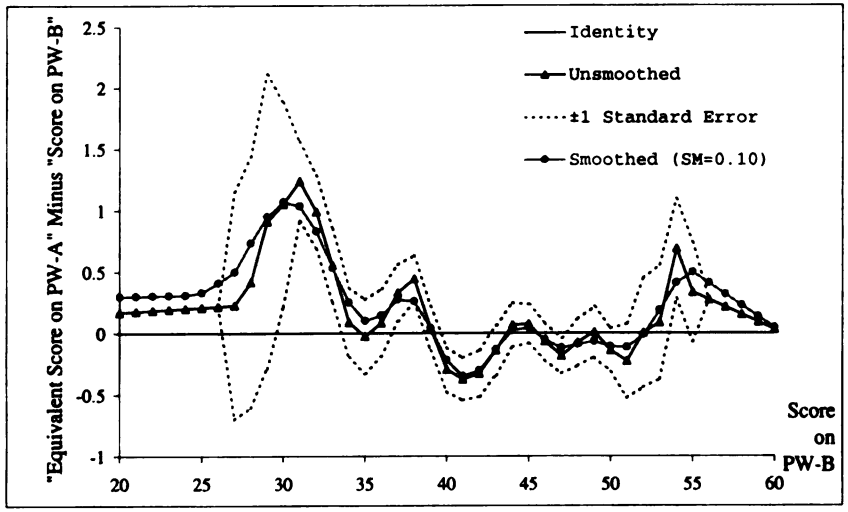
APPENDIX B

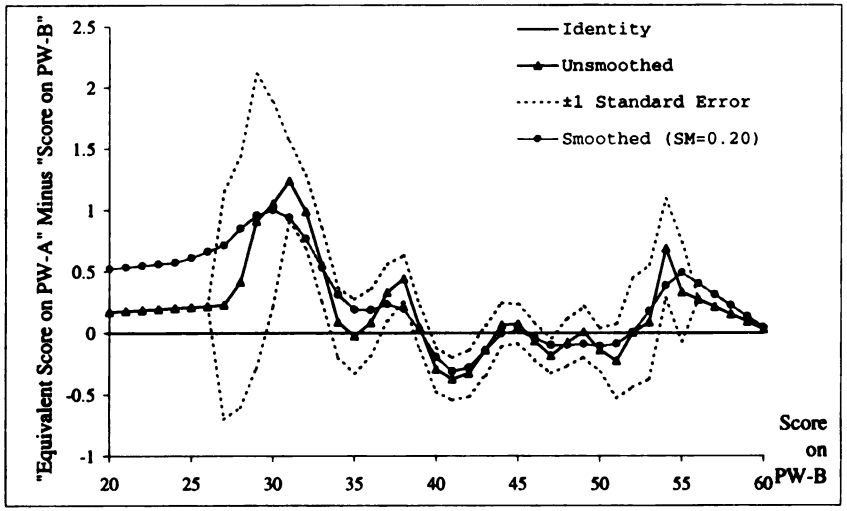
APPENDIX B-1

Score to Score Equivalents by Various Degrees of Smoothing for Sampled Test PW

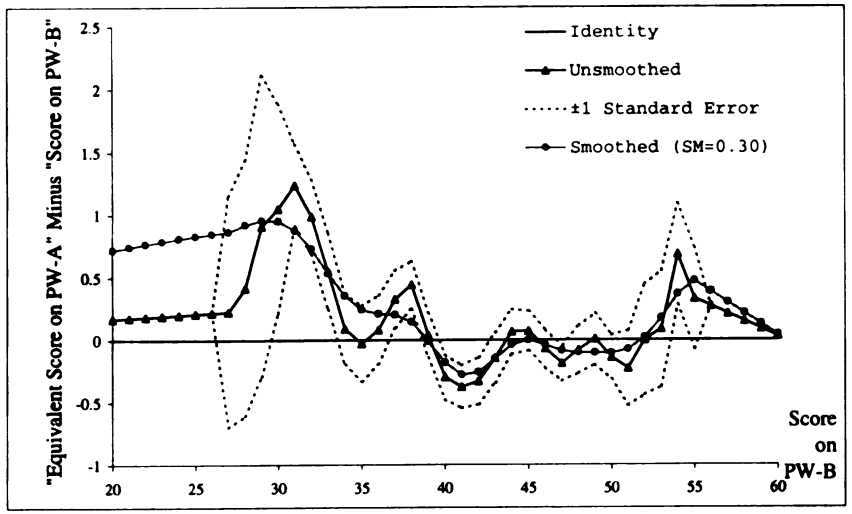


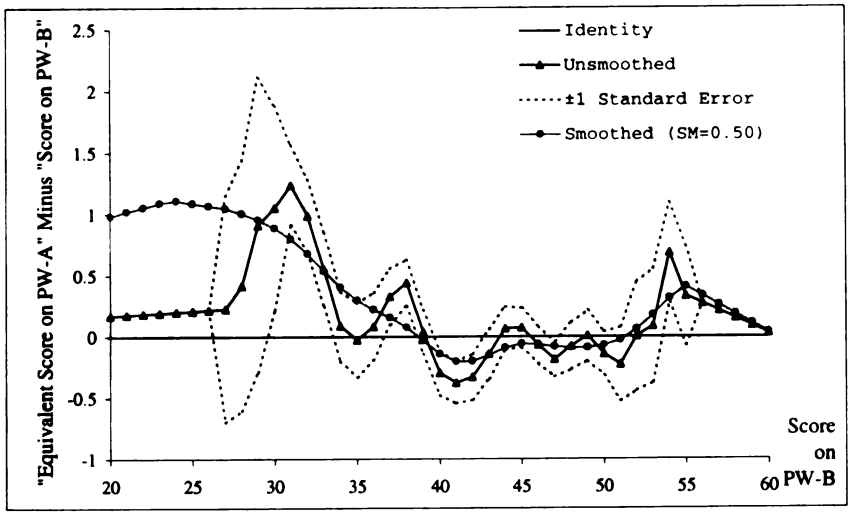


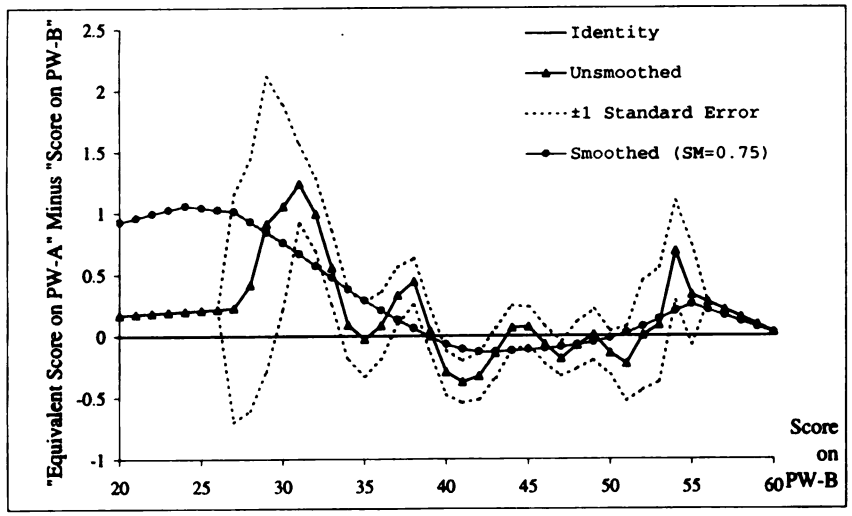


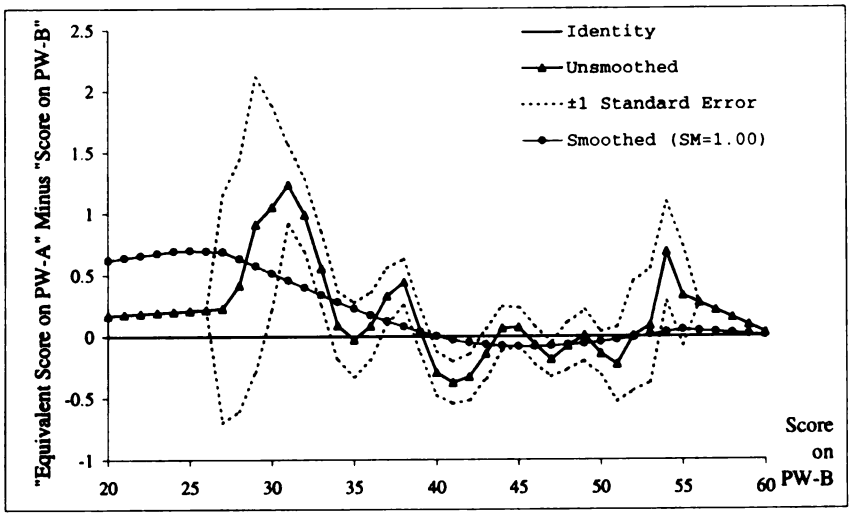


APPENDIX B-1 (cont'd)



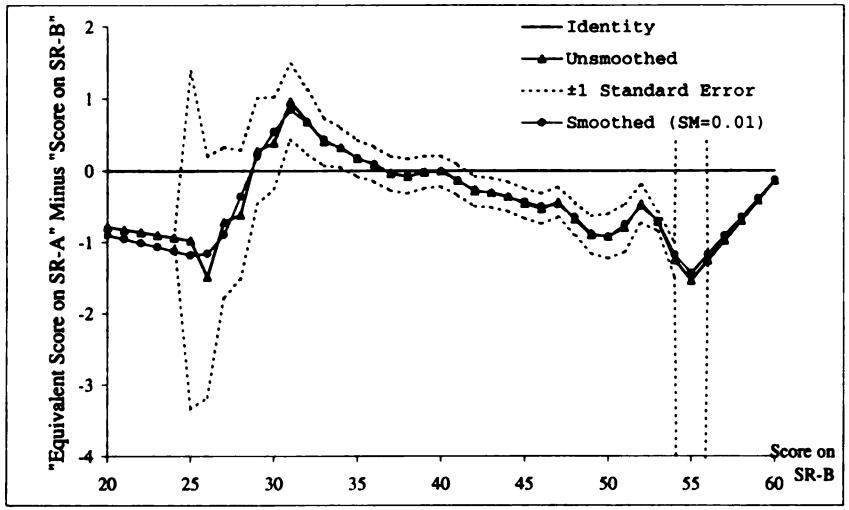


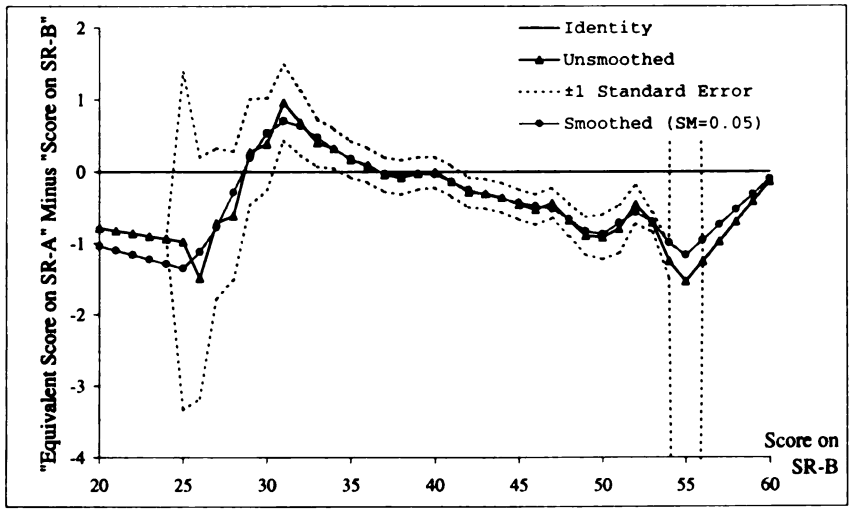


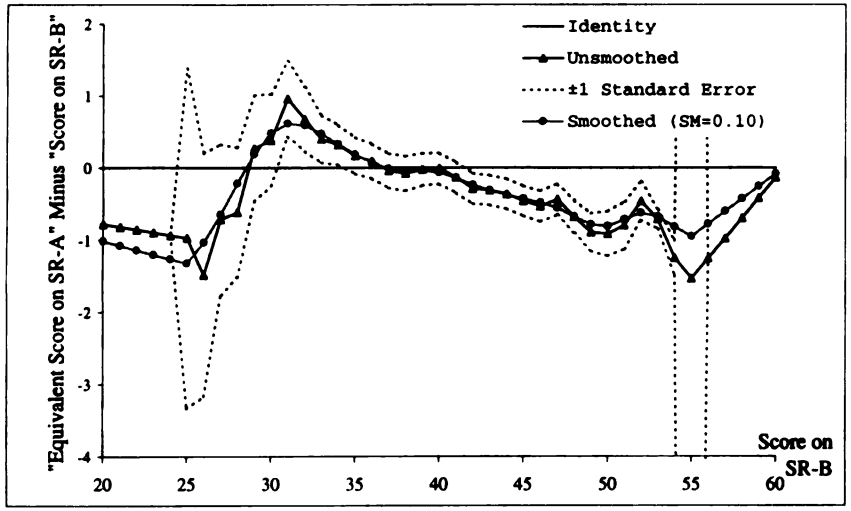


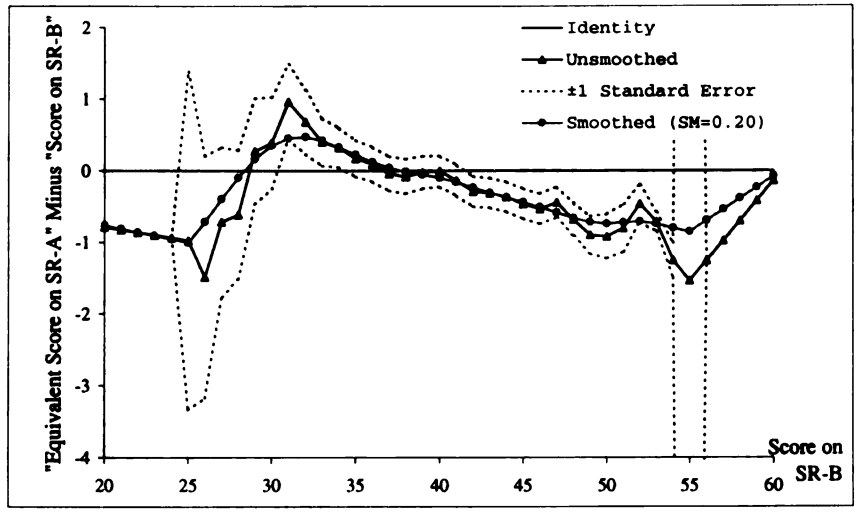
APPENDIX B-2

Score to Score Equivalents by Various Degrees of Smoothing for Sampled Test SR

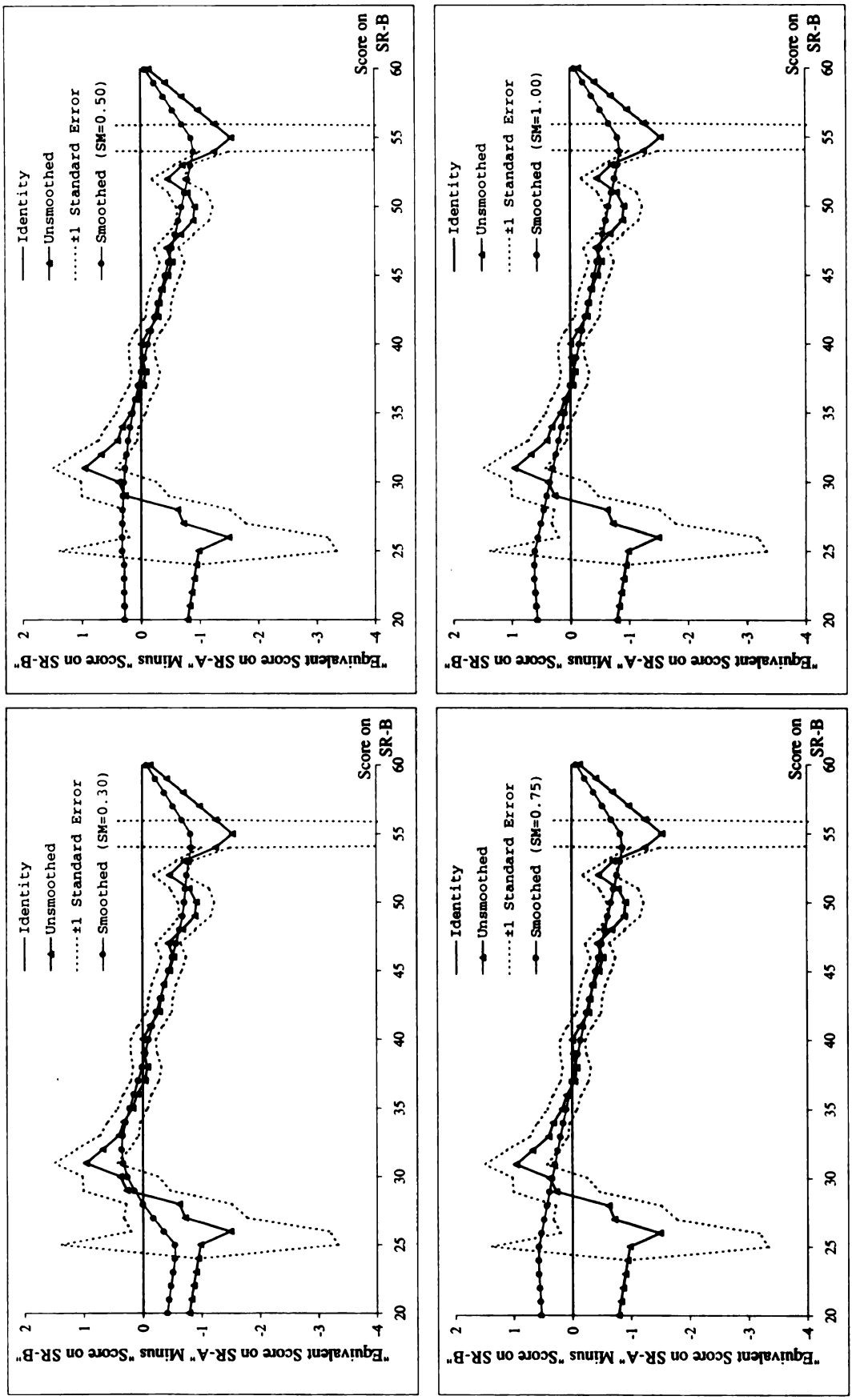






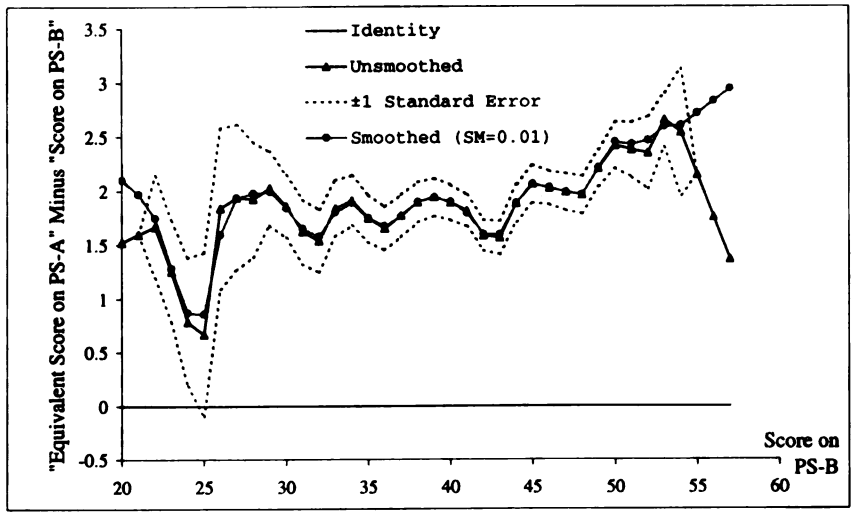


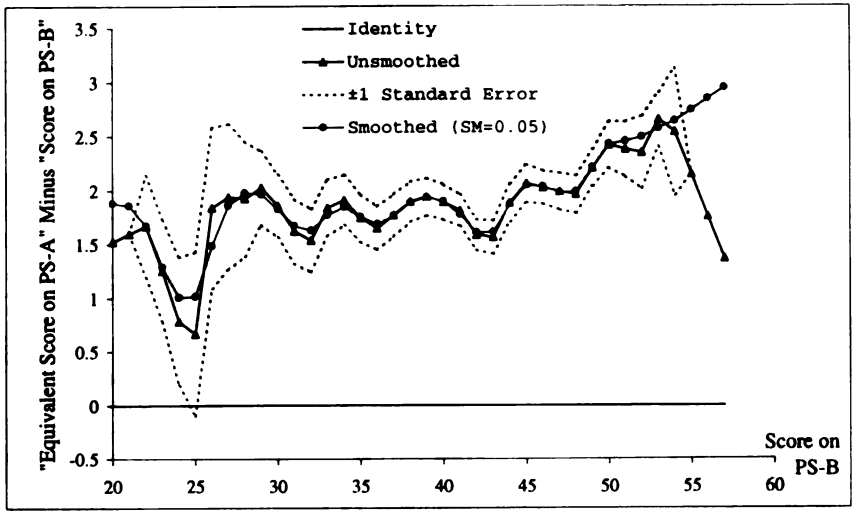
APPENDIX B-2 (cont'd)

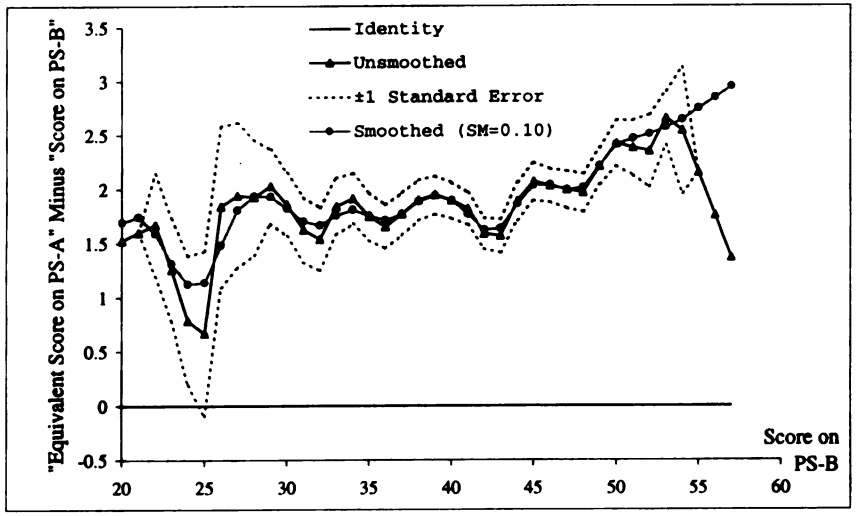


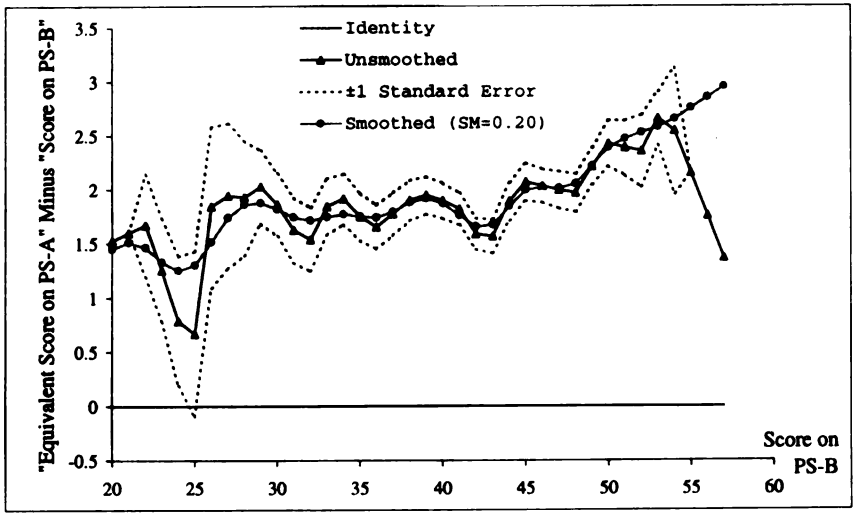
APPENDIX B-3

Score to Score Equivalents by Various Degrees of Smoothing for Sampled Test PS

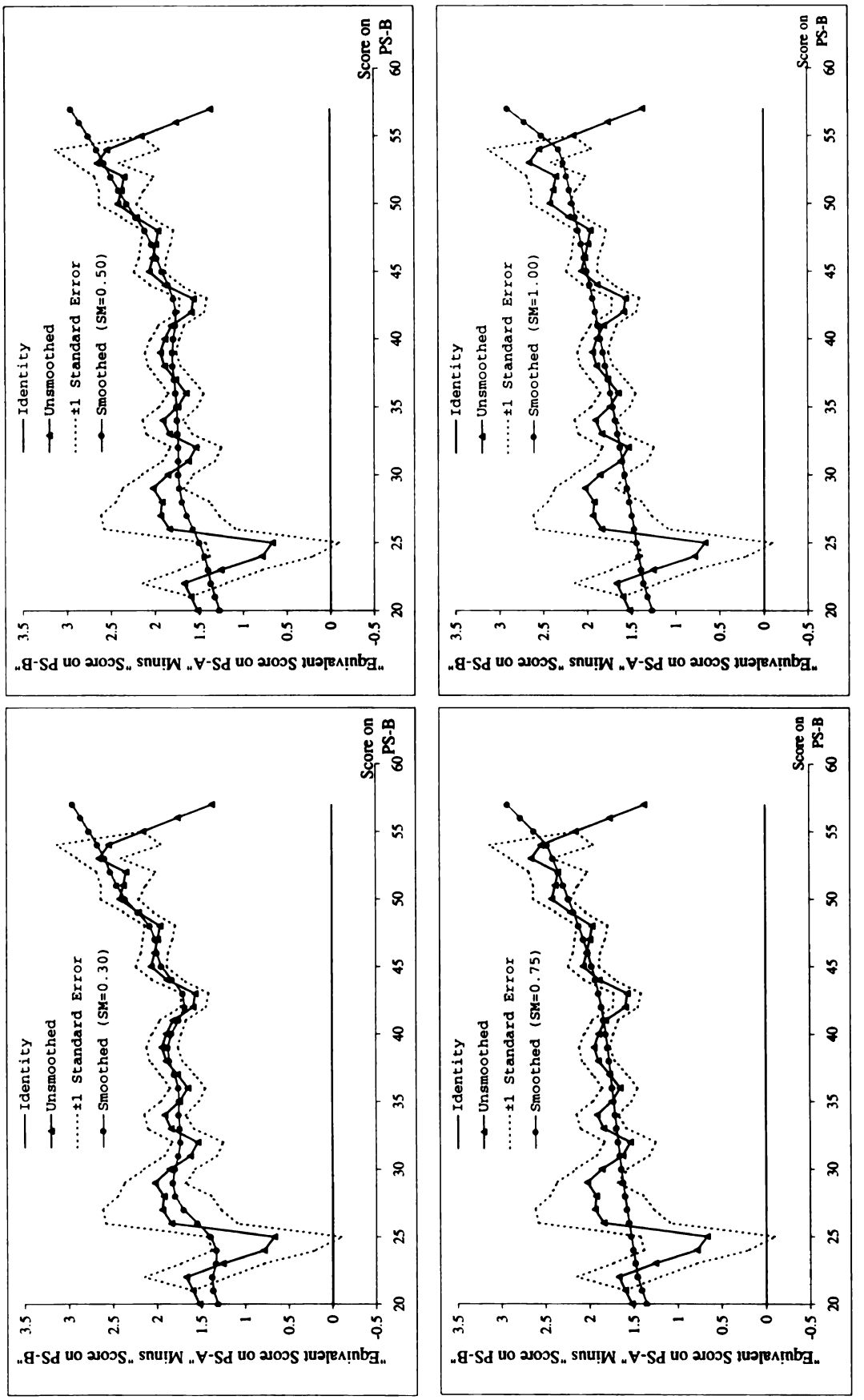








APPENDIX B-3 (cont'd)



APPENDIX C

APPENDIX C

Adjusted Correlation Matrix for Evaluating Equating Accuracy--Indices of Equating Accuracy after Controlling for Auto-Correlation

Adjusted Index of Equating Accuracy (Pearson's *r*)		Criterion		IRT-Based Linear Transformation Method				IRT-Based Fixed-b Method				**Pearson's *r* Before Adjustment**	
		Raw-145	IRT-145	EW	PS	PW	SR	EW	PS	PW	SR	Raw-145	IRT-145
Criterion	Raw-145	1.000											
	IRT-145	0.982	1.000										
IRT-Based Linear Transformation Method	EW	**0.868**	**0.861**	1.000								**0.877**	**0.870**
	PS	**0.891**	**0.914**	0.764	1.000							**0.894**	**0.917**
	PW	**0.842**	**0.836**	0.771	0.749	1.000						**0.845**	**0.839**
	SR	**0.854**	**0.858**	0.809	0.843	0.784	1.000					**0.856**	**0.860**
IRT-Based Fixed-b Method	EW	**0.865**	**0.856**	0.997	0.752	0.784	0.806	1.000				**0.873**	**0.865**
	PS	**0.892**	**0.913**	0.765	0.989	0.783	0.847	0.764	1.000			**0.895**	**0.916**
	PW	**0.867**	**0.864**	0.793	0.794	0.975	0.803	0.790	0.795	1.000		**0.870**	**0.867**
	SR	**0.852**	**0.856**	0.808	0.838	0.790	0.999	0.807	0.847	0.802	1.000	**0.854**	**0.858**

Note. (1) The correlation coefficients in gray background color were indices of equating accuracy.

They were the Pearson rs between the criterion and the IRT-estimated true scores computed using only the non-anchor items from the sample test.

(2) The underlined correlation coefficients indicated the relationship between the two IRT-based equatings.

(3) All of the Pearson correlation coefficients were significant at α=.01.

APPENDIX D

APPENDIX D

Reliability and Validity Evidence for the Anchor Tests of Four Sampled Tests

Validity/Relibility (Pearson's *r*)		"Pseudo True Score"	
		Raw-145	IRT-145
IRT-Based Linear Transformation Method	EW	0.877	0.870
	PS	0.894	0.917
	PW	0.846	0.839
	SR	0.857	0.861
IRT-Based Fixed-b Method	EW	0.873	0.865
	PS	0.895	0.916
	PW	0.871	0.868
	SR	0.855	0.859

Note. (1) The validity/reliability measure is the Pearson's *r* between the "pseudo true score" and the resulting IRT true score estimates on a sampled test containing anchor items only.

(2) All of the Pearson correlation coefficients were significant at α=.01.

LIST OF REFEERENCES

LIST OF REFEERENCES

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

Baker, F. B. (1990). Some observations on the metric of PC-BILOG results. Applied Psychological Measurement, 14, 139-150.

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

Berk, R. H. (1982). Discussion of item response theory. In P. Holland & D. B. Rubin (Eds.), Test equating. New York: Academic Press.

Berry, D. A., & Lindgren, B. W. (1990). Statistics: theory and methods. Belmont, CA: Brooks/Cole.

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

Brennan, R. L., & Kolen, M. J. (1987). Some practical issues in equating. Applied Psychological Measurement, 11, 279-290.

Budescu, D. (1985). Efficiency of linear equating as a function of the length of the anchor test. Journal of Educational Measurement, 22, 13-20.

Camilli, G., Wang, M., & Fesq, J. (1995). The effects of dimensionality on equating the law school admission test. Journal of Educational Measurement, 32, 79-96.

Cook, L. L., & Eignor, D. R. (1983). Practical considerations regarding the use of item response theory to equate tests. In R. K. Hambleton (Ed.), Applications of item response theory (pp.175-195). Vancouver, British Columbia: Educational Research Institute of British Columbia.

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

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

Cook, L. L., Eignor, D. R., & Schmitt, A. P. (1988). The effects on IRT and conventional achievement test equating results of using equating samples matched on ability (Research Rep. No. RR-88-52). Princeton, NJ: Educational Testing Service.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. Chicago: Holt, Rinehart and Winston, Inc.

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

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

Efron, B. (1982). The jackknife, the bootstrap, and other resampling plans. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Efron, B., & Tibshirani, R. J. (1993). An introduction to the bootstrap (Monographs on Statistics and Applied Probability 57). New York: Chapman & Hall.

Eignor, D. R., Stocking, M. L., & Cook, L. L. (1990). Simulation results of effects on linear and curvilinear observed- and true-score equating procedures of matching on a fallible criterion. Applied Measurement in Education, 3, 37-52.

Green, D. R., Yen, W. M., & Burket, G. R. (1989). Experiences in the application of item response theory in test construction. Applied Measurement in Education, 2, 297-312.

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

Hambleton, R. K., & Cook, L. L. (1977). Latent trait models and their use in the analysis of educational test data. Journal of Educational Measurement, 14, 75-96.

Hambleton, R. K., & Swaminathan, H. (1990). Item response theory: Principles and applications. Boston: Kluwer-Nijhoff Publishing.

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

Hanson, B. A. (1991). A comparison of bivariate smoothing methods in common-item equipercentile equating. Applied Psychological Measurement, 15, 391- 408.

Hanson, B. A., Zeng, L., & Colton, D. (1994). A comparison of presmoothing and postsmoothing methods in equipercentile equating (ACT Research Report 94-4). Iowa City, IA: American College Testing.

Hanson, B. A., Zeng, L., & Kolen, M. J. (1995, October). Equating Computer Programs. (Available from Michael Kolen, ACT, 2255 N. Dubuque Street, Iowa City, IA 52243).

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

Hills, J. R., Subhiyah, R. G., & Hirsch, T. M. (1988). Equating minimum-competency tests: Comparisons of methods. Journal of Educational Measurement, 25, 221-231.

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (1979). Applied statistics for the behavioral sciences. Chicago: Rand McNally.

Holland, P. W., & Thayer, D. T. (1987). Notes on the use of log-linear models for fitting discrete probability distributions (Technical Report No. 87-79). Princeton, NJ: Edcuational Testing Service.

Holland, P. W., & Thayer, D. T. (1989). The kernel method of equating score distributions (Technical Report No. 89-84). Princeton, NJ: Educational Testing Service.

Jarjoura, D., & Kolen, M. J. (1985). Standard errors of equipercentile equating for the common item nonequivalent populations design. Journal of Educational Statistics, 10, 143-160.

Jöreskog, K. G., & Sörbom, D. (1989). LISREL 7: A guide to the program and applications. Chicago: Scientific Software International, Inc.

Jöreskog, K. G., & Sörbom, D. (1993). LISREL 8: User's reference guide. Chicago: Scientific Software International, Inc.

Jöreskog, K. G., & Sörbom, D. (1995). PRELIS: A program for multivariate data screening and data summarization. Chicago: Scientific Software International, Inc.

Kendall,M., & Stuart, A. (1977). The advanced theory of statistics (4th ed., Vol.1). New York: Macmillan.

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. (1981). Comparison of traditional and item response theory methods for equating tests. Journal of Educational Measurement, 18, 1-10.

Kolen, M. J. (1991). Smoothing methods for estimating test score distributions. Journal of Educational Measurement, 28, 257-282.

Kolen, M. J., & Brennan, R. L. (1987). Linear equating models for the common-item nonequivalent-populations design. Applied Psychological Measurement, 11, 263-277.

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

Kolen, M. J., & Harris, D. J. (1990). Comparison of item preequating and random groups equating using IRT and equipercentile methods. Journal of Educational Measurement, 27, 27-39.

Kolen, M. J., & Jarjoura, D. (1987). Analytical smoothing for equipercentile equating under the common item nonequivalent populations design. Psychometrika, 52, 43-59.

Lawrence, I. M., & Dorans, N. J. (1990). Effect on equating results of matching samples on an anchor test. Applied Measurement in Education, 3, 19-36.

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

Livingston, S. A. (1993). Small-sample equating with log-linear smoothing. Journal of Educational Measurement, 30, 23-39.

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

Lord, F. M. (1965). A strong true score theory with applications. Psychometrika, 30, 239-270.

Lord, F. M. (1977). Practical applications of item characteristic curve theory. Journal of Educational Measurement, 14, 117-138.

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

Lord, F. M. (1982a). The standard error of equipercentile equating. Journal of Educational Statistics, 1, 165-192.

Lord, F. M. (1982b). Item response theory and equating- A technical summary. In P. Holland & D. B. Rubin (Eds.), Test equating. New York: Academic Press.

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

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

Marco, G. L., Petersen, N. C., & Stewart, E. E. (1983). A test of the adequacy of curvilinear score equating models. In D. J. Weiss (Ed.), New horizons in testing: Latent trait theory and computerized adaptive testing. New York: Academic Press.

Mislevy, R. J., & Bock, R. D. (1990). BILOG 3: Item analysis and test scoring with binary logistic models. Mooresville, IN: Scientific Software.

Mislevy, R. J., & Stocking, M. L. (1989). A consumer's guide to LOGIST and BILOG. Applied Psychological Measurement, 13, 57-75.

Parshall, C. G., Houghton, P. D., & Kromrey, J. D. (1995). Equating error and statistical bias in small sample linear equating. Journal of Educational Measurement, 32, 37-54.

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

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), Educational Measurement. New York: ACE/Macmillan.

Raju, N. S., Bode, R. K., Larsen, V. S., & Steinhaus, S. (1986, April). Anchor-test size and horizontal equating with the Rasch and three-parameter models. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Raju, N. S., Edwards, J. E., & Osberg, D. W. (1983, April). The effect of anchor test size in vertical equating with the Rasch and three-parameter models. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal.

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

Reckase, M. D., Ackerman, T. A., & Carlson, J. E. (1988). Building a unidimensional test using multidimensional items. Journal of Educational Measurement, 25, 193-203.

Rosenbaum, P. R., & Thayer, D. T. (1987). Smoothing the joint and marginal distributions of scored two-way contingency tables in test equating. British Journal of Mathematical and Statistical Psychology, 40, 43-49.

SAS Institute Inc. (1989). SAS/STAT® user's guide, version 6, fourth edition, volume 1. Cary, NC: SAS Institute Inc.

Schmitt, A. P., Cook, L. L., Dorans, N. J., & Eignor, D. R. (1990). Sensitivity of equating results to different sampling strategies. Applied Measurement in Education, 3, 53-71.

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

Skaggs, G., & Lissitz, R. W. (1988). Effect of examinee ability on test equating invariance. Applied Psychological Measurement, 12, 69-82.

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

Wang, T., & Kolen, M. J. (1994). A quadratic curve equating method to equate the first three moments in equipercentile equating (ACT Research Report 94-2). Iowa City, IA: American College Testing.

Wingersky, M. S., & Lord, F. M. (1984). An investigation of methods for reducing sampling error in certain IRT procedures. Applied Psychological Measurement, 8, 347-364.

Wingersky, M. S., & Barton, M. A. (1982). Logist user's guide: Logist 5, Version 1.0. Princeton, NJ: Educational Testing Service.

Yen, W. M. (1980). The extent, causes and importance of context effects on item parameters for two latent trait models. Journal of Educational Measurement, 17, 297-311.

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

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. Applied Psychological Measurement, 8, 125-145.

Yen, W. M. (1985). Tau equivalence of vertical equating using three-parameter item response theory and Thurstonian procedures. Paper presented at the meeting of the American Educational Research Association, Chicago.

Yen, W. M. (1987). A comparison of the efficiency and accuracy of BILOG and LOGIST. Psychometrika, 52, 275-291.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items. Chicago: Scientific Software International.